COLORED SLOPE GRAPH: VISUALIZATION OF TIME SERIES DATA

by

Apurva Nagarajan

Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer Graphics Technology

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Yingjie Victor Chen, Chair

Department of Computer Graphics Technology

Dr. Paul Parsons

Department of Computer Graphics Technology

Dr. Xavier Tricoche

Department of Computer Science

Approved by:

Dr. Nicoletta Adamo-Villani

I dedicate my work to Momo, my parents, and my sister. I don't think I could have completed this work without their support. I wish you all were here with me in West Lafayette for the successful completion of my work.

ACKNOWLEDGMENTS

I would like to thank all my committee members, Dr. Yingjie Victor Chen, Dr. Paul Parsons and Dr. Xavier Tricoche who have provided valuable feedback throughout my research study, and directed me to valuable resources that have helped me to conduct this research and write this thesis document. I would also like to thank Junhan Zhao, Xiaolei Guo, Lu Ding for their tremendous input and help in this research study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVG	Scalable Vector Graphics
HTML	Hyper Text Markup Language
CSS	Cascading Style Sheets

ABSTRACT

Time-series data draw extensive attention from many research domains, such as financial and biomedical engineering. Researchers often face the difficultly of visualizing multiple time-series data simultaneously. The predominant techniques suffer from visual clutter either due to over-plotting or an overwhelming number of small graphs that carries a high cognitive load. This research study proposes a Colored slope, a combination of Tufte's slope graph and time-series heatmap, to visualize multiple time-series data at once, balancing scalability and accuracy. Colored slope inherits the complementary advantages from each method, regarding visualizing temporal changes within a period and identifying precise values. The efficacy, effectiveness, and graphical perception of the Colored slope on visualizing multiple time-series data with publicly accessible stock data were evaluated and compared it to popular time-series visualizations, including line graphs, time-series heatmap, and small multiple spark area graphs. Analyzing the experimental data, this study concludes that Colored slope contributes to (1) allowing users to identify the extreme values (maximum and minimum), co-variability, the general trend of the data, and rate of change effectively for an optimally large number of visual marks(time-series data); (2) capable of display more instances of time-series data with a less visual clutter problem. Finally, several possible applications and limitations with the Colored slope are demonstrated.

CHAPTER 1. INTRODUCTION

In the era of big data, due to the advancement in computing and data warehousing, information is generated at a rapid rate. One such data that is widely produced is time-series data. The data stream is collected at a time interval such as hour, day, month, quarter, or year that are evenly spaced (Time Series Analysis and Forecasting, 2013), which is analyzed to understand causes and effects. As per Few (2009), six basic patterns are useful when a time-series data is analyzed. The patterns are a trend, variability, rate of change, co-variability, cycles, exceptions (Few, 2009). For users to analyze time-series data without the aid of visual analytics is difficult. This has intrigued many visualization researchers to explore various visualization techniques. The predominant visualization techniques for time-series data are a line graph, bar graph, radial graph, sparkline graph (Few, 2009). These techniques assist the users in examining time-series data to make prompt and accurate decisions.

For example, the financial industry is one of the highest generators of time-series data (Ko, Cho, Afzal, Yau, Chae, Malik, Beck, Ribarsky, & Ebert, 2016). Every second, thousands of stocks get traded in each sector (Ko, et al., 2016). Each stock has multi-variant attributes like prices, end price, closing price, volume. Financial analysts use visual analytics tools to generate various reports. These reports aid in the forecast and predict economic growth, risk assessment, etc. (Ko, et al., 2016). Some of the visualizations predominantly used in the financial industry include line graphs, sparkline graphs, and candlestick charts (Edwards, Magee, & Bassetti, 2018). These visualizations are popular among financial analysts for decision-making (Sorenson & Brath, 2013).

However, using these traditional techniques, which were designed decades before the intervention of the digital big data age, contributes to bottlenecks such as overplotting, which results in a visual clutter shown in Figure 1.1. In this research, a visualization design will be proposed to overcome such a cluttering problem when multiple time-series data are visualized.

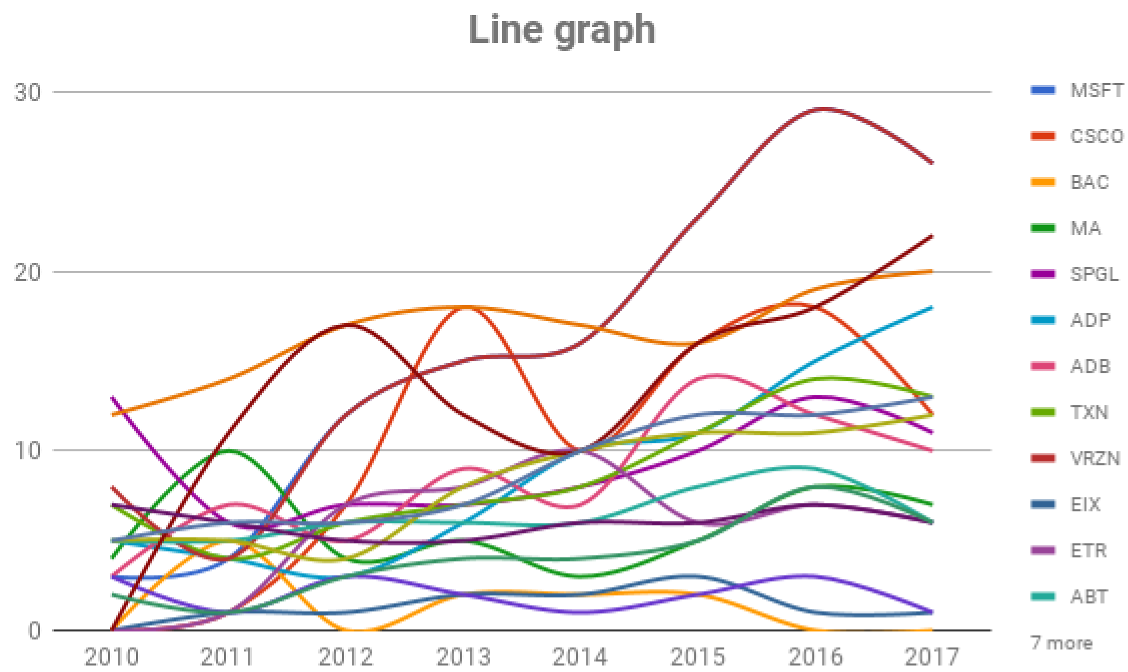


Figure 1.1: Sample Line Graph visualization with too many lines cluttered.

1.1 Problem Statement

"A random sample of 4000 graphics from 15 of the world's newspapers published found that more than 75% of them featured time-series" (Keogh, Hochheise, & Shneiderman, 2002). Time-series data visualization makes it easier for users to get a sense of the data and make informed decisions (Ko, et al., 2016). A single inaccurate decision or an accurate but delayed decision can affect an organization's growth and risk management strategies drastically.

The threshold after which the number of visual marks after which a visualization gets illegible is 50 (Saket, Endert, & Demiralp, 2018). The optimal number of visual marks in visualization is in the range of 5-30. A line graph is an ideal visualization for time-series data when the number of visual marks is less than ten. However, once the number of visual marks increases in the display, it is tough for the users to carry out basic tasks using line graph visualization. Also, a line graph is good at identifying the general trend of the data. However, a line graph is not the right choice when users want to determine a specific value (Saket, Endert, & Demiralp, 2018). precisely.

Research about time-series data visualization focuses on improving the effectiveness of these predominantly used graphs by using techniques such as “increased data density” (Keogh, Hochheise, & Shneiderman, 2002), statistical transformations of the data and “polar-coordinates” (Keogh, Hochheise, & Shneiderman, 2002). Several visualization techniques have tried addressing the visual clutter caused due to the occlusion and high density of the graphic in the visualization. However, the proposed techniques could not solve the problem of visual clutter when multiple time-series data are visualized simultaneously (Ko, et al., 2016). It is challenging to analyze and interpret various patterns like trends, variability, etc. from heavily cluttered temporal line graphs.

How can one design a visualization that reduces the visual clutter but still preserve the crux of the data? The question leads to the investigation of resolving the issues by: (1) utilizing a simple and intuitive visual form to enable the extension of visualizing multiple data sessions simultaneously; (2) allowing the users to easily obtain the critical attributes of time-series data, like trends, variability, or changes. Reviewing the conventional visualizations works, this study realized that Tufte's slope graph (Tufte, Edward R., 1983) provided a basic visual form, which demonstrates the trend for the entire time domain and the precise values at the beginning and ending points. Due to its simple form of straight lines, a slope graph is capable of putting many lines together at once. But due to is a simple form, it lacks the capability of providing details of the temporal change in the middle of the period. In the meanwhile, the heatmap can encode the values by colors in a given display space, but the nature of human's perception lack to capability to associate color/shade accurately with value and ratio (Ware, 2012). Therefore, this study seeks for a solution to combine those two conventional methods while maintaining the advantages of both.

1.2 Purpose Statement

The purpose of this proposed research was to design a visualization technique for time-series data and evaluate the visualization design based on the patterns stated by Few (2009). According to Few (2009), time-series data has six important characteristics for analysis. The patterns are trends, variability, rate of change, co-variation, cycles, exceptions. Few (2009) states that trend is the overall pattern to either increase, decrease, or remain constant in each period. Variability can be described as the average rate of change in a period (Few, 2009). As stated in Few (2009), Rate of Change is the percentage difference between any two values in each period, and Co-variation is the relationship between any two time-series data. According to Few (2009), cycles are the repetition of the pattern trend in each period, and Exception is defined as an outlier in each time-series data.

1.3 Research Questions

In this research, a new way of visualizing time-series data was presented along with the design and algorithm specifications. This visualization is referred to as “Colored Slope Graph” in the following sections. The colored slope graph design is inspired by slope graph visualization. In this technique, the first and the last temporal attributes are plotted using a slope line, and the intermediate points are encoded based on the color transform function. The efficacy, effectiveness and graphical perception of the Colored slope graph were studied by choosing a line graph, heatmap and spark area as the benchmark time-series visualization method. The performance of the colored slope graph was evaluated along with the predominant time-series visualizations based on several criteria and patterns listed in Few. From the evaluation method, this research aims to answer the following research questions:

1) Can users perceive the patterns, such as trends, coverability, variability, rate of change in the colored slope graph visualization?

    a) Can users identify the extreme (maximum value) in the colored slope graph?

    b) Can users identify the extreme (minimum value) in the colored slope graph?

    c) Can users identify high variability in the colored slope graph?

    d) Can users identify Co-Variability between two or more curves in the colored slope graph?

    e) Can users identify the maximum increase in the value between two temporal attributes?

    f) Can users identify the maximum decrease in the value between two temporal attributes?

    g) Can users identify the general trend of the time-series curves, i.e., can users identify the number of stocks increasing and the number of stocks decreasing in a single glance.

    A quantitative approach was adopted to address the above research questions, which are discussed in the Methodology section.

2) To what extent can users perceive the information presented in a slope line with color encoded on it?
    What is the correctness of the sketch drawn by the participants?

A quantitative approach was adopted to address the above research questions which are discussed in the Methodology section

3) Can the colored slope graph visualization reduce the cluttering problem, when multiple curves are visualized?

   A user evaluation was adopted to address the above research question which is discussed in the Methodology section

1.4 Definitions

Slope graph: A slope graph is very similar to a line graph. Time and non-temporal attributes are represented by the x-axis and the y-axis, respectively. Only the change in the attributes between the two temporal data points is represented in the slope graph without any relation to the intermediate temporal points.

Colored slope graph: The proposed visualization design in this research study which has been inspired by slope graph visualization.

Trends: Trend of a time-series data can be categorized in three ways, i.e., to increasing, decreasing or remaining constant. (Few, 2009).

Variability: In a concise form, a time-series curve is said to be highly variable when there is a high fluctuation of data which means the line or signal curve goes up and down (Few, 2009).

Rate of Change: "The percentage difference between any two points" (Few, 2009).

Co-variance: "When one time-series curves affected by the change in another time-series curve" (Few, 2009).

Cycle: "The patterns that repeat at regular intervals" (Few, 2009).

Exceptions: "The values that are well above or below the norm" (Few, 2009).

1.5 Assumptions

Below are the assumptions associated with the research

1. The sampling of the participants will be random, and the sample will represent the population.

2. The participants will answer the survey honestly.

3. The study results can be applied to a broader population

4. The data this research chooses poses the typical characteristics of other time-series data.

5. This study can be generalized to all time-series data.

1.6 Limitations

Below are the limitations associated with the research

1. The study will be limited to Purdue University students, who may not represent the entire population.

2. This study may use one or two types of time-series data as it is not feasible to cover all different unique types of time-series data.

3. This visualization is a design solution that will not consider any statistical transformation of the data.

4. This visualization study tackles the issue of visual clutter when 20-40 time-series data are visualized in each screen space.

5. The results of the study will be different if the number of time-series data is less than 20.

1.7 Delimitations

Below are the delimitations associated with the research

1. All the visualization will have no interaction.

2. The background-color, fonts will be kept constant in all the visualizations.

3. The primary dataset that has been chosen is a stock dataset because it has all the characteristics of a typical time-series data, and the stock market data is easily accessible.

1.8 Summary

This chapter summarizes the scope, problem significance, research purpose, research questions and other background information for the research project. The next chapter presents a review of the literature relevant to the visualization techniques in the financial industry.

CHAPTER 2. REVIEW OF LITERATURE

This chapter summarizes the literature review visualization about the time-series data and graphical perception of the visualization.

2.1 Time-series visualization

The very first time-series data visualization in history was presented in the18th century by William Playfair applied the line graph as their representation of their data (Tufte, Edward R., 1983). For better observing those essences, many visualization inventions happened in the following years. Muller and Schumann reviewed the time-dependent data techniques previously (MiilIer & Schumann, 2003). Aigner, Miksch, Müller, & Schumann (2007) systematically reviewed the variety of techniques for visualizing time-series data. In their introduction, Aigner pointed out the fact that most of the approaches were customized because of the complexity of considering time-series data from all aspects. They also suggested that visualizing time-series data had to consider about both representational and perceptional difficulties. Many time-series visualizations methods have been developed based on the implementations of the conventional line plot or bar chart. Tufte's sparklines (Tufte, Edward R., 1983) first integrated the line chart with small multiples for reviewing multiple time-series. For dealing with this issue, Heer, Kong, & Agrawala (2009) compared horizon graphs with line charts and found the trade-off curve of speed-accuracy observed from the subjects 'estimations between different charts. They later proposed optimal graphical perceptional approaches (Heer, Kong, & Agrawala, 2009). Considering the research limitation that Heer's work only used two times series, an investigation was conducted using their braided graphs to compare graphical perception with line graphs and horizon graphs (Javed, McDonnel, & Elmqvist, 2010). In M, E, and DA (2011), the authors proposed a visualization named Cloud Lines, which allowed the users to detect visual clusters in a compressed view within limited space. In this visualization, events data is encoded linearly on a timeline as circles and mapped the importance of the events by adjusting the sizes and opacity of the circles. These methods aim to entangle the visual mess by managing the space. Researchers also implement pixel visualization to encode the large time-series data visually. Kumar, Lolla, Keogh, Lonardi, and Ratanamahatana (2005) introduced a simple parameter-light method that allows the users to navigate through large time-series datasets.

Due to the lack of inputs from users, traditional techniques often have scalability issues due to overplotting (Keim, Kohlhammer, Ellis, & Mansmann, 2010). According to Sorenson and Brath, (2013), the line chart is still a popular technique for visualizing time-series data. To overcome the issue of overplotting, various statistical aggregators are used (Sorenson & Brath, 2013). Designing a new visualization often needs researchers in the field of data visualization to collaborate with users to evaluate their needs and difficulties and to design suitable alternatives for visual analysis (Ko, et al., 2016).

(Cleveland & McGill, 1985)graph, pie charts and similar techniques which are inspired by either line or pie charts (Murphy, 1999). These conventional graphs use 2D information, the fundamental concept of these techniques is to illustrate values of attributes (non-temporal) along the y-axis and time attribute along the x-axis (Ko, et al., 2016). In addition to the existing two axes', an essential dimension is added which aids the 3D visualization techniques to assist effective visualization (Ko, et al., 2016). While these techniques assist users to find patterns, trends, and outliers, these techniques suffer from problems such as over-plotting, occlusion, and inadequate support to compare various datasets (Sorenson & Brath, 2013).

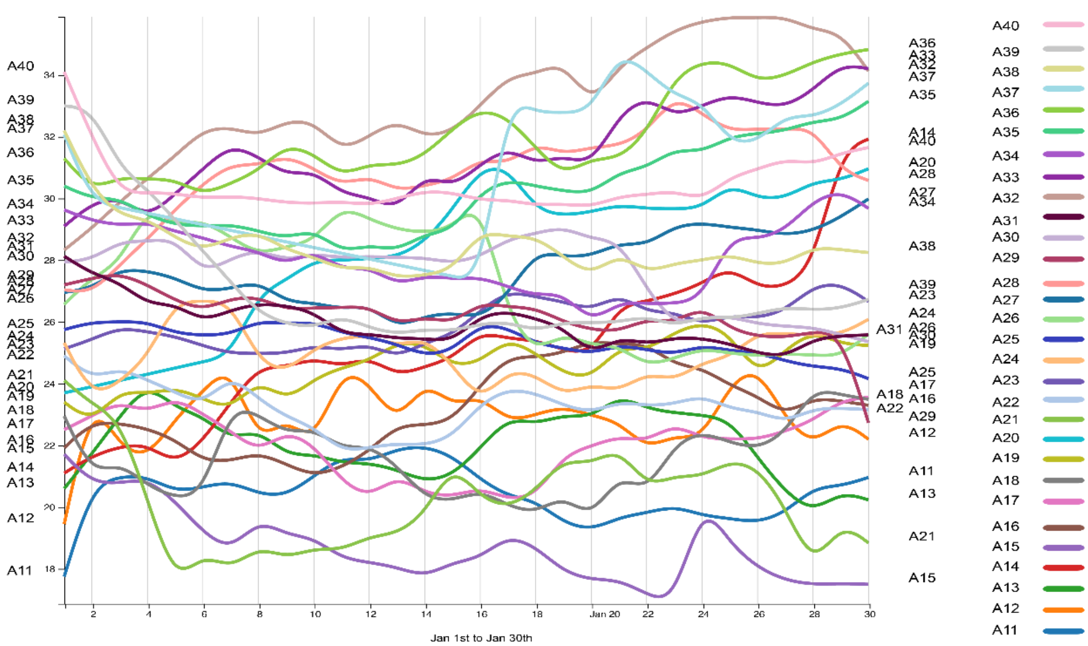


Figure 2.1: Sample Line Graph visualization

The 2D techniques mainly contain Line graphs that suffer from the issue of over-plotting which leads to visual clutter (Ko, et al., 2016). In Lin, Cao, & Zhang, (2005) visualization technique, the fish-eye technique was used (Furnas, 1986) to mitigate the overplotting issue. According to (Ko, et al., 2016) several visualizations that are not the variations of line graphs are also utilized. A wedge graph (Dao, Bazinet , Berthier , & Shneiderman , 2008) was utilized as an alternative to pie charts (Ko, et al., 2016). “Each wedge corresponds to stock in the wedge graphs, and the size and color of a wedge are mapped respectively to attributes” like opening price, closing price or volume (Ko, et al., 2016). The problem of over-plotting has intrigued researchers to propose and adopt other techniques of visualization of the time-series data for analysis (Ko, et al., 2016).

The concept of Sparklines was first introduced by (Tufte, Edward R., 1983). The initial version of it was called "Small Multiples" by Tufte, (1983). In this visualization, small, inch-sized versions of several displays are shown all in a single screen space, which helps the reader to instantly and parallelly compare the differences between frames. (Tufte, Edward R., 1983). The concept of Sparklines which was again coined by Tufte (2006), is closely based on small multiples. According to Tufte (2006), the sparklines consists of multiple tiny graphics in each display. Sparklines are tiny graphics often that have heavy data density yet with a design that is simple (Tufte, Beautiful Evidence, 2006). As stated by Tufte (1983), sparklines are well suited and adopted in comparing stocks.

Research has shown that the adoption of 3D visualizations for the analysis of time-series data like financial data, weather forecast, ranking data, etc., significantly improves the perception of the information compared to 2D visualizations (Ko, et al., 2016). In addition to the standard techniques, “3D visualizations also incorporate advanced techniques like 3D scatter plots, splat visualizations, and animating time-added 3D wedge charts for analyzing time-series data” (Ko, et al., 2016). One such 3D visualization technique is the use of spheres as a framework for visualization. This technique is a common tactic, as it utilizes Gestalt psychology (Koffka, 2014) which facilitates user navigation and creates aesthetically appealing visualizations (Ko, et al., 2016). Although the 3D visualization technique has power in certain aspects, it cannot solve visual clutter, as the volume of economic data becomes big (Ko, et al., 2016).

Geometrically transformed displays help users identify interesting aspects of the data (Keim D. A., 2002). “Scatterplot matrices, parallel coordinates, and visualizations of clusters” (Ko, et al., 2016) are included in this category (Ko, et al., 2016). In this technique, the multidimensional data undergoes an automated statistical algorithmic transformation before the visualization methods (Ko, et al., 2016). Usually, in a 2D space, a scatter plot displays time-series data as points. In a scatterplot, “change in the color, size, shape, and orientation of points with time-series data is the most common approach” (Ko, et al., 2016). According to Ko et al (2016), stocks are plotted based on monthly trading records i.e. on the x-axis and yearly trading records i.e. on the y-axis, and stocks are differentiated by assigning a specific color. The scatterplot is often coupled with streamlined systems by transforming the data with statistical algorithms (Ko, et al., 2016). Icon and glyph-based visualizations allocate an attribute of the data to the icon's function in the form of color and structure. The visualization is then constructed by arranging these icons in a spiral layout based on distinct time-period frames (Ko, et al., 2016). However, this technique is complex to perceive as it requires analysts to recognize the trends in the data (Ko, et al., 2016).

Another technique which was proposed by Alsakran, Zhao, & Zhao, (2010) but could not attract the researchers in the financial domain is the Parallel coordinate plots. This technique is hardly used in the financial domain (Ko, et al., 2016). This technique embraces the idea of a parallel coordinate plot by combining the tile-based technique (Alsakran, Zhao, & Zhao, 2010). The system conducts three procedures to display stocks (Alsakran, Zhao, & Zhao, 2010). First, the underlying background space is divided into a rectangular grid (Alsakran, Zhao, & Zhao, 2010). The density of an individual rectangular tile is calculated based on the number of lines passing through that tile (Alsakran, Zhao, & Zhao, 2010). Each tile is assigned a color and opacity based on the number of stocks passing through the tile (Alsakran, Zhao, & Zhao, 2010).

As stated in Ko et al. (2016), another graph that has recently acquired an application is a self-organizing map (SOM) which is produced from an unsupervised learning algorithm. According to Ko et al, (2016), the technique affects visualization's geometry. This method is used mainly to assess business financial efficiency. As the visualization is constructed based on the inputs from an unsupervised learning algorithm, the result is usually unpleasant (Ko, et al., 2016).

Users use visual analytics tools and techniques to produce various quantitative analyses, which aid in forecasting and projecting economic growth (Ko et al., 2016). According to Edwards, Magee, & Bassetti, (2018), the visualizations primarily used in the financial sector include line graphs and candlestick charts.

2.3 Graphical Perception

Information visualization is the science of representing data in the form of small graphics where data values are represented as graphical representations (Cleveland & McGill, 1985). Graphical representation of data is the fundamental goal of information visualization. Researchers have examined the impacts of visual encodes on their capacity to interpret and evaluate data portrayed in visualizations through human-subject studies. Graphical perception is described as the ability of the users to understand the visual encoding and with which the users can decipher the information presented in the visualization (Javed, McDonnel, & Elmqvist, 2010). Early research in graphical perception which will be discussed below primarily focused on examining the usefulness and value of various forms of graphic representations when performing visual tasks. In the works of Croxton and Stryker (1927), the authors compared the performance of a bar chart and circle diagram on effecting judgment accuracy. Peterson and Schramm (1954) compared the accuracy of reading eight different kinds of conventional graphs. Simkin and Hastie (1987) compared a simple bar chart, divided bar chart, and pie chart, and discussed speed and accuracy based on graph type and judgment type for information extraction. In the work of Spence and Lewandowsky (1991), work was presented which investigated the display of proportions and percentages in bar charts and pie charts. Later, many studies on graphical perception have focused on how visual coding affects the comprehension of the data set. The work of Bertin (1967) first provided the ranking of visual encodings according to their effectiveness for different tasks. Cleveland and McGill (1984) conducted a set of elementary perceptual experiments to determine the ranking of the visuals, which gives a scientific approach to ranking of visual coding. From the initial research which resulted in a detailed understanding of visual encoding, several other researchers began to study the graphic perception of different types of graphical representations. Some of the main studies measured the effect of visual coding variables (i.e., size, position, color, etc.) on the accuracy and/or response time of data estimates. Also, there were research studies that focused on the time-series visualization perception. For positional encoding, Heer, Kong, and Agrawala (2009) investigated the effects of chart size and layering of two types of visualization in value comparison tasks (Heer, Kong, & Agrawala, 2009). Javed, McDonnel, and Elmqvist

(2010) compared four types of multiple time-series visualizations that scattered or share the space under maximum, slope and discrimination tasks. In the work of Heer, Kong, and Agrawala (2009) it is demonstrated that share-space techniques reduce overall and visual clutter by comparing over limited visual space. For color encoding, Correll, Albers, Franconeri, and Gleicher, (2012) studied the efficiency of using position or color representations in the time-series. The experiment confirms that viewers were better at estimating averages when using color encoding. Hernandez, Paredes, Roseway, and Czerwinski (2014) conducted a set of comparison tasks to compare eight different time-series visualizations. Their results suggest that color encoding is more effective for summary comparisons. Chart size and aspect ratio on graphical perception are of concern when there are many data sets and need comparisons across them. Cleveland, Diaconis, and McGill (1982) studied a variety of axis ranges effect on correlation perception in scatterplots and proposed Banking to 45°, Cleveland (1993), which provides a design guideline that center slope around 45°. This helps in providing a good estimate in visual judgments of slope ratios. Based on Cleveland's works, Talbot, Gerth, and Hanrahan (2012) developed a model that fits more generally and can be used to aspect ratio selection. They found that selects flatter, wider aspect ratios can minimize the error in slope ratio estimation

2.4 Summary

This chapter summarizes the literature in the field of visualization of time-series data. The literature review also focuses on some of the drawbacks of the visualizations for time-series data. Also, literature on graphical perception of visual encoding is presented.

CHAPTER 3. FRAMEWORK AND METHODOLOGY

This chapter provides a high-level overview of the framework and the methodology used in this research study to evaluate the color slope visualization based on the benchmark of other predominant visualizations for time-series data such as line graph, spark area graph, and heatmap. An in-detail explanation of the experiments and the data analysis is described thoroughly in Chapter 5 and Chapter 6.

3.1 Framework

A quantitative study was adopted for the evaluation of the colored slope graph. This study consisted of a training process, two human subject experiments and a participant's survey. The motive of the experiments was to measure the efficacy and effectiveness of the colored slope visualization. In the first experiment, the colored slope graph was evaluated based on the benchmark of the line graph, spark area and heat map, which are the most widely used visualization techniques for time-series visualization which is illustrated in Figure 3.2. After the first experiment, the participants' demographic information was recorded. Later, the participants had to fill out surveys in which they were asked to rank the four visualizations based on several criteria. In the second experiment, the graphical perception of the color slope graph was studied, and the correctness of the participant's sketch was calculated. All the experiments were conducted in Heavilon hall, room 102. The participants finished the experiments in a single session. The three phases of the research study and the timeline of study that each participant goes through is listed in Figure 3.1.

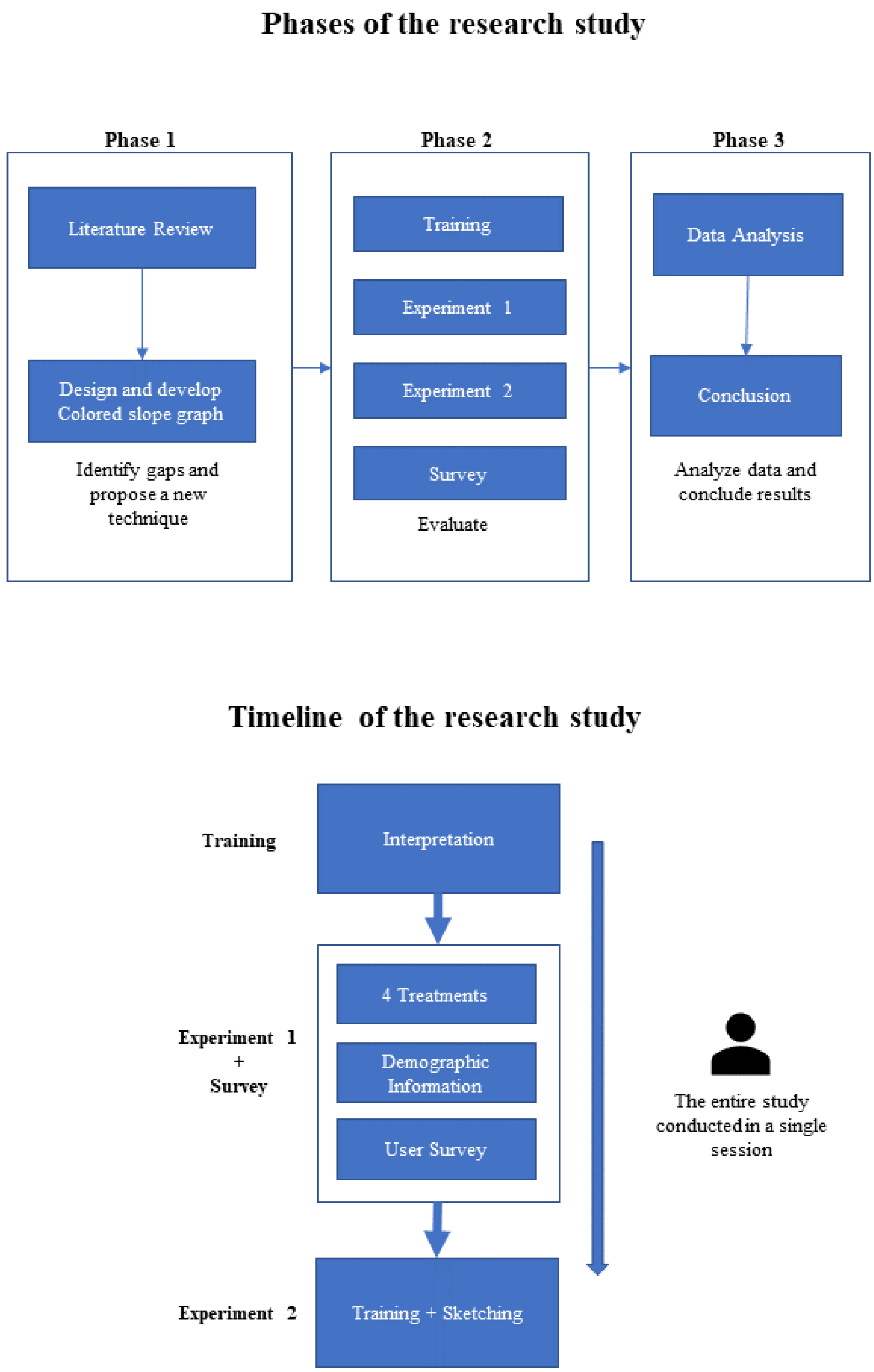


Figure 3.1: Research study framework describing the three phases of the study

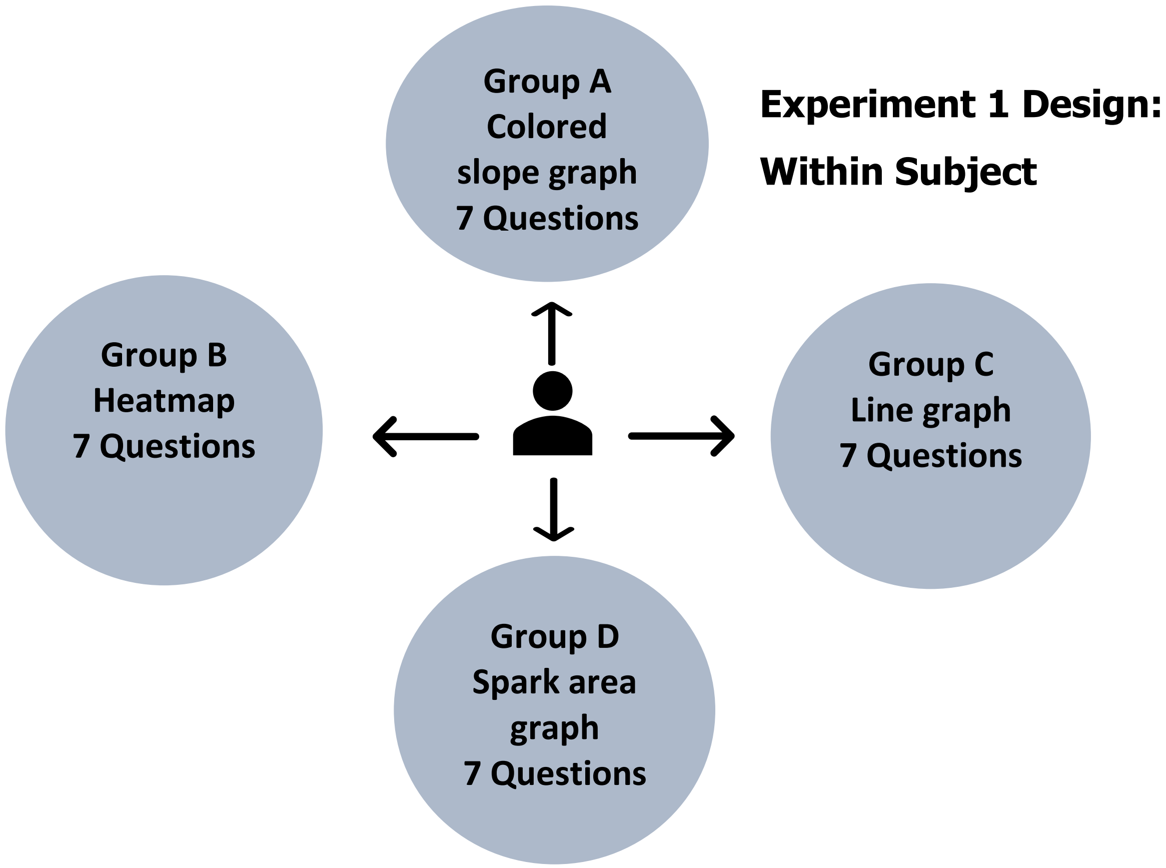


Figure 3.2: Research study Experiment 1 procedure set up.

3.2 Primary Dataset

The primary dataset was used from Kaggle (Kaggle, 2019), an online platform owned by Google where users can publish/access various data sets for either scientific, business or personal use (Kaggle Wikipedia, 2019). The dataset contains daily stock values of 7000 US companies from the year 1971 to 2017. Each year has 214 data points. Each Stock has 5 attributes i.e. Opening Price, Closing Price, Volume, Highest price, lowest price. For the visualization, the year 2017 is selected as it is the latest year. In the year 2017, the first 30 data points were selected for the period. Each visualization consisted of 30 stocks and each stock had 30 data points. For the visualization technique, the closing price of a stock will be used it determines.

3.3 Population and Sample

As this visualization is designed for everyday users with/without any visualization or analytical background, and not exclusively designed for experts in the field, college students were recruited for the study. A voluntary response sample was adopted for the study due to limitations and scale of the study. Emails were sent to students from CGT 575, CGT 370 and other departments for voluntary participation in the research study. 35 College students from various disciplines which include STEM, Arts, Management and at various levels which include Freshmen, Sophomore, Graduate voluntarily participated in the study. As stated by (Glassner, 1994) different individuals perceive color differently. Participants with any form of color blindness were asked to self-report and were eliminated from the experiments as they may distort the outcomes. Participants were rewarded with a $5 Amazon gift card for completing the whole study.

3.4 Hypotheses

Based on the patterns listed in (Few, 2009) the below hypothesis will be tested.

1. Which treatment performs better to identify the extreme(maximum) value in the time-series curve?

   HO1: There is no significant difference among the four groups in identifying extreme(maximum) value in the time-series curve.

   HA1: There is a significant difference in identifying among the four groups in identifying extreme(maximum) value in the time-series curve.

2. Which treatment performs better to identify the stock which covaries with a certain stock?

   HO2: There is no significant difference among the four groups in identifying the stock which covaries with a certain stock.

   HA2: There is a significant difference among the four groups in identifying the stock which covaries with a certain stock.

3. Which treatment performs better to identify the stock which has the highest variability?

   HO3: There is no significant difference among the four groups in identifying the stocks with the highest variability.

   HA3: There is a significant difference among the four groups in identifying the stocks with the highest variability.

4. Which treatment performs better to identify the stock which increases the most from 1^{st} Jan to 30^{th} Jan?

   HO4: There is no significant difference among the four groups identify the stock which increases the most from 1^{st} Jan to 30^{th} Jan.

   HA4: There is a significant difference among the four groups identify the stock which increases the most from 1^{st} Jan to 30^{th} Jan.

5. Which treatment performs better to identify the percentage of stocks increasing to the percentage of stocks decreasing?

   HO5: There is no significant difference among the four groups in identifying the percentage of stocks increasing to the percentage of stocks decreasing.

   HA5: There is a significant difference between the four groups in identifying the percentage of stocks increasing to the percentage of stocks decreasing.

6. Which treatment performs better to identify the extreme(minimum) value in the time-series curve?

   HO6: There is no significant difference among the four groups in identifying the extreme(minimum) value in the time-series curve.

   HA6: There is a significant difference between the four groups in identifying the extreme(minimum) value in the time-series curve.

7. Which treatment performs better to identify the stock which decreases the most from 1^{st} Jan to 30^{th} Jan?

   HO7: There is no significant difference among the groups in identifying the stock which decreases the most from 1^{st} Jan to 30^{th} Jan.

   HA7: There is a significant difference among the groups in identifying the stock which decreases the most from 1^{st} Jan to 30^{th} Jan.

3.4 Procedure

The first phase of this research study was to design and develop the algorithm of the colored slope graph. The details of the design and the algorithm are described thoroughly in Chapter4. After the design phase, to evaluate the visualization, the framework listed in section 3.1 was adopted. As this research carries out voluntary sample design, 35 participants from Purdue University participated across different programs and levels of education. Data collected from the experiments were analyzed and results were concluded. The procedure adopted by this research study is thoroughly described in Chapter 5.

3.4 Instruments

For the first experiment listed in section 3.1, a set of 7 questions were prepared for each visualization. For the second experiment, 10 colored slope lines along with their ground truth i.e. the corresponding line graph were created. For the participants' survey, a prevalidated survey from (Zhao, Liu, Guo, Qian, & Chen, 2019) and NASA TLX was incorporated. All the instruments are mentioned in detail in Chapter 5.

3.5 Data Analysis

In the first experiment, in each treatment, for each question, the participants score. The scoring criteria are thoroughly mentioned in Chapter 6. Statistical inference such as mean, median, the standard deviation was calculated based on the scores. The significance of the four treatments was calculated based on Mann – Whitney one-sided u test. Along with the scores, for each treatment, the time taken to complete the task was recorded and statistical inferences such as mean, median, the standard deviation were calculated. In the second experiment, from the participants' sketches, the correctness of the sketch is measured based on the ground truth of the actual line graph. This is calculated by extracting the pixel values from the sketches at 30 temporal points. From the extracted pixel values statistical inference such as mean, and the five quartiles of the data were calculated at 30 temporal points. From the surveys, where the participants ranked the four visualizations, the percentage of participants' choice and the average rank of colored slope graph was calculated. A thorough detail of the data processing, data analysis, interpretation of the analysis is listed in Chapter 5 and Chapter 6.

3.6 Summary

This chapter outlines the methodology and the framework for the research study. A brief introduction of the primary dataset, the hypothesis, data analysis, methods used, experimental collection and demographic sample. This gives the reader the chance to replicate and validate the work if one wishes to.

CHAPTER 4: DESIGN

In this chapter, a novel design, the inspiration of the design, the underlying algorithm, the technology used for the development is mentioned in detail.

4.1 Proposed Design

The proposed visualization technique illustrated in Figure 4.2 was designed by taking inspiration from a slope graph. A slope graph is very similar to a line graph where the change in the non-temporal attributes is visualized over a period illustrated in Figure 4.1. The temporal attributes are represented by the x-axis and the non-temporal attributes are represented by the y-axis respectively. The first y-axis is the starting value of the stock. The second y-axis is the last data point in a period. Unlike a line graph, a slope graph considers the change only between the two temporal points. Usually, a slope graph visualizes the trend between the temporal points which helps in handling the overplotting in a visualization. The data points which are in between the selected temporal points are not considered. However, this leads to the loss of information in the visualization.

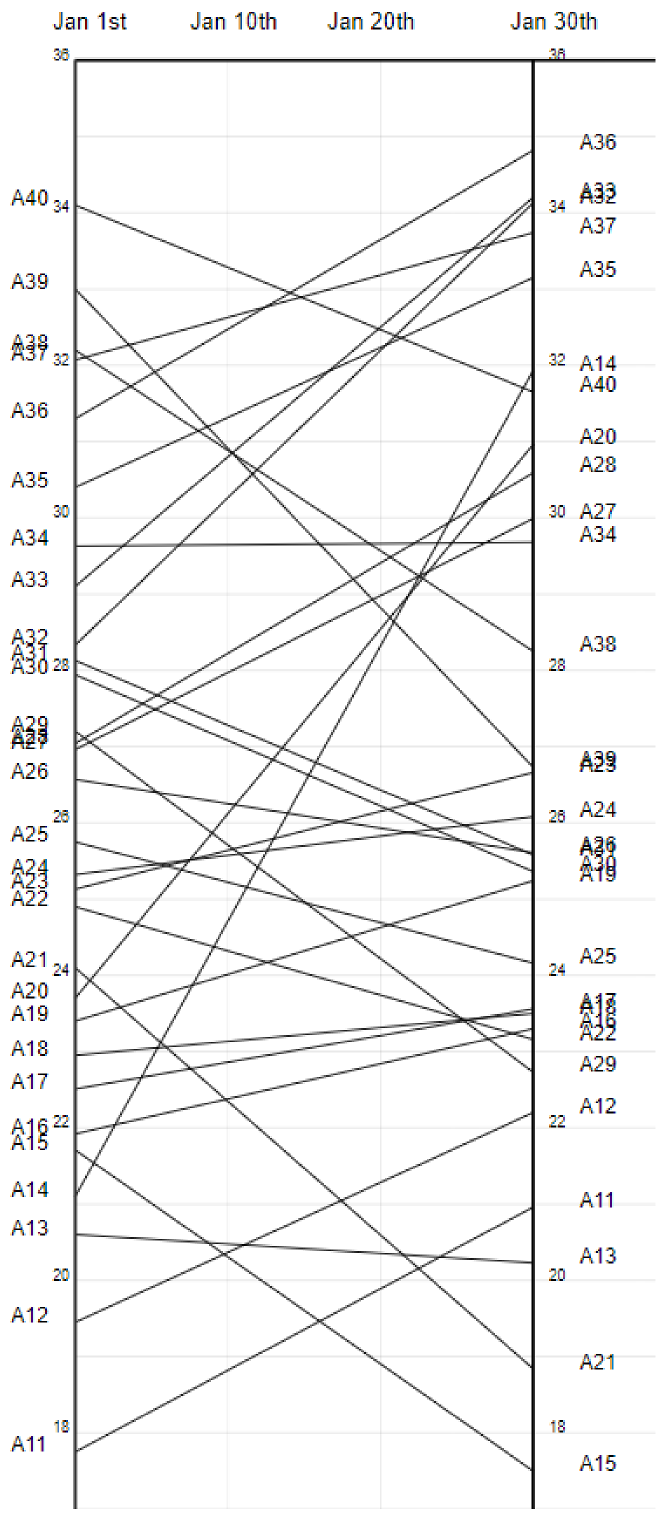


Figure 4.1: Slope Graph Visualization.

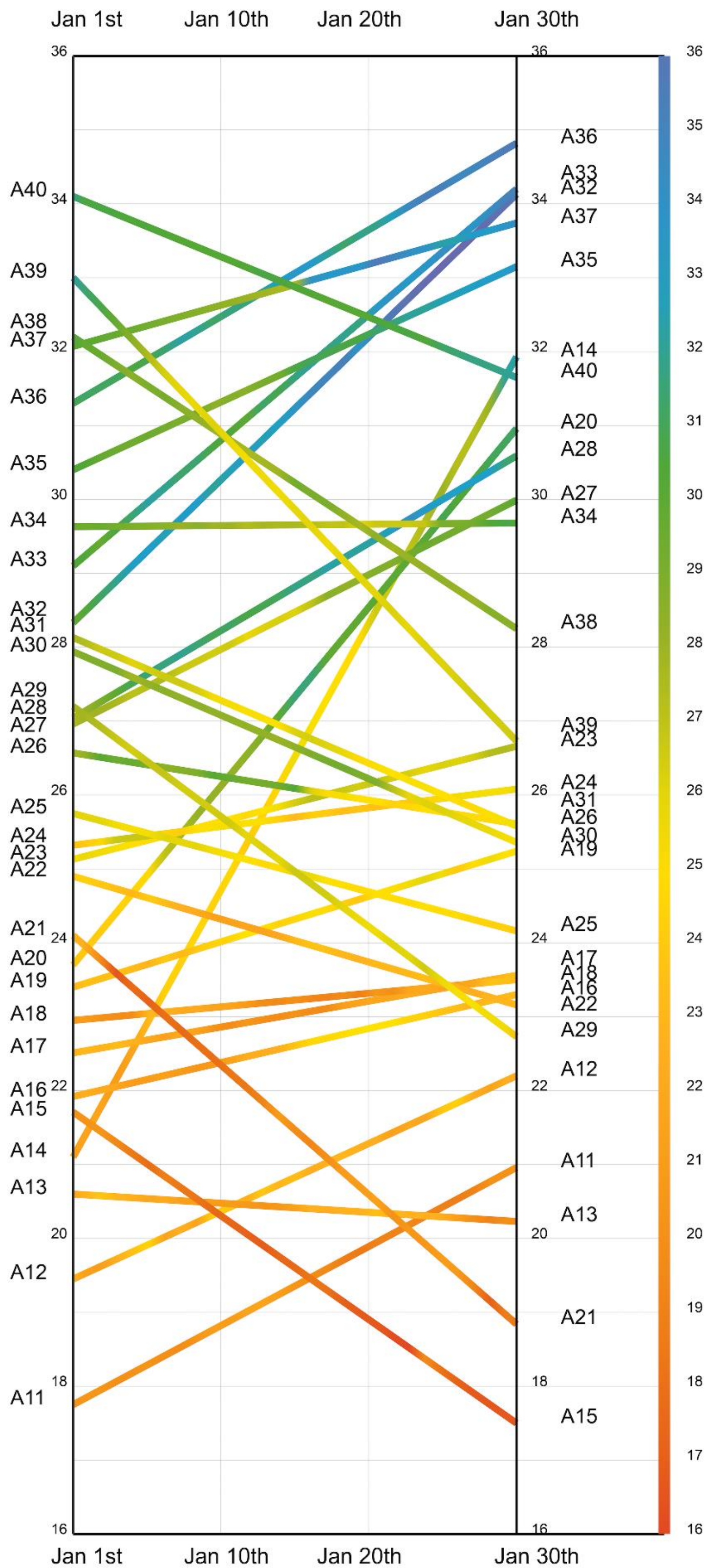


Figure 4.2: Colored slope graph.

The loss of information can be handled by incorporating the color scale in the visualization. In addition to visualizing the absolute values of starting and ending data points, the intermediate data points are represented by color-coding. Color-coding is specified using a color scale that reflects the absolute values of the data. The interpretation of colors in straight lines is given in Figure 4.5 and how a colored curved line is changed to a colored straight line is given in Figure 4.5.

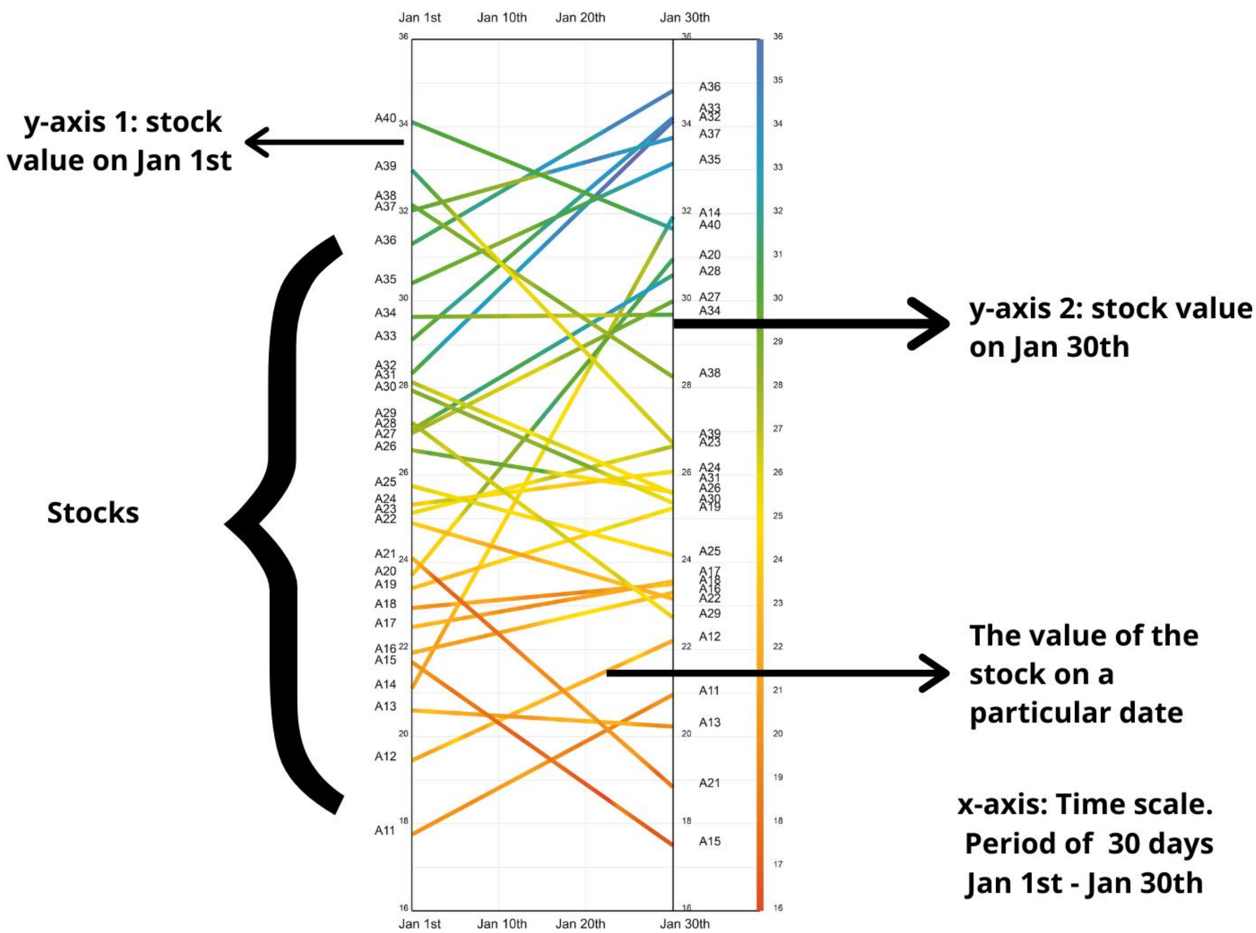


Figure 4.3: How to interpret the information visualized using the Colored slope graph.

4.2 Color Scale

An essential visual encoding method for the graphical representation of data in information visualization is using color and shade. Color-encoding aids the human eye for the identification of the right estimates of the values. Single hue with varying opacity and saturation or diverging scales is useful when the range of values visualized is limited, in terms of small number of investigating objects. However, when multiple time-series data are visualized simultaneously, differentiating an extensive range of values using a single hue might not be able to provide enough visual difference for small changes. As the primary purpose of this research study is to visualize multiple time-series data simultaneously, using a single hue will make readers hard to differentiate variations in smaller scales. Moreover, the deployment of Colored slope involves overlapping of the slopes. In that case, single-hue color scale may cause serious color perception issues because of the similarity of the colors. Additionally, gray-shade (saturation) is not ideal mainly because humans cannot perceive the comparison of a large range of values based on brightness and saturation alone. It is necessary to use a color scale which has large visual difference such as the diverging color scale (e.g., ranges from blue to red). Therefore, this research used a color scale from Color Brewer 2.0 (Color Brewer 2.0, n.d.) that has a broad-spectrum range which ranges from blue, green, yellow, orange, to red. The color scale used in this visualization is not color-blind friendly due to a wide spectrum. It is possible to use any other color scales, including color-blind safe options. However, due to the purpose of research, which is to find the maximum potential of the purposed solution, the color scale that could best differentiate values was chosen. The top and the bottom colors in the scale are mapped to the maximum value and minimum value in the dataset. Colors of other values in the middle are mapped accordingly. Colored slope can use any meaningful color scale as far as the designer provides a clear legend indicating the values of colors. In this study, stock data was used with a warmer tone to represent a lower value, and a cooler tone to represent a higher value, which is vice-versa in the traditional heat-map. The rationale behind this design decision is that, in a stock market data, cooler color (green) signifies growth, and warmer color (red) signifies a decline in the stock value.

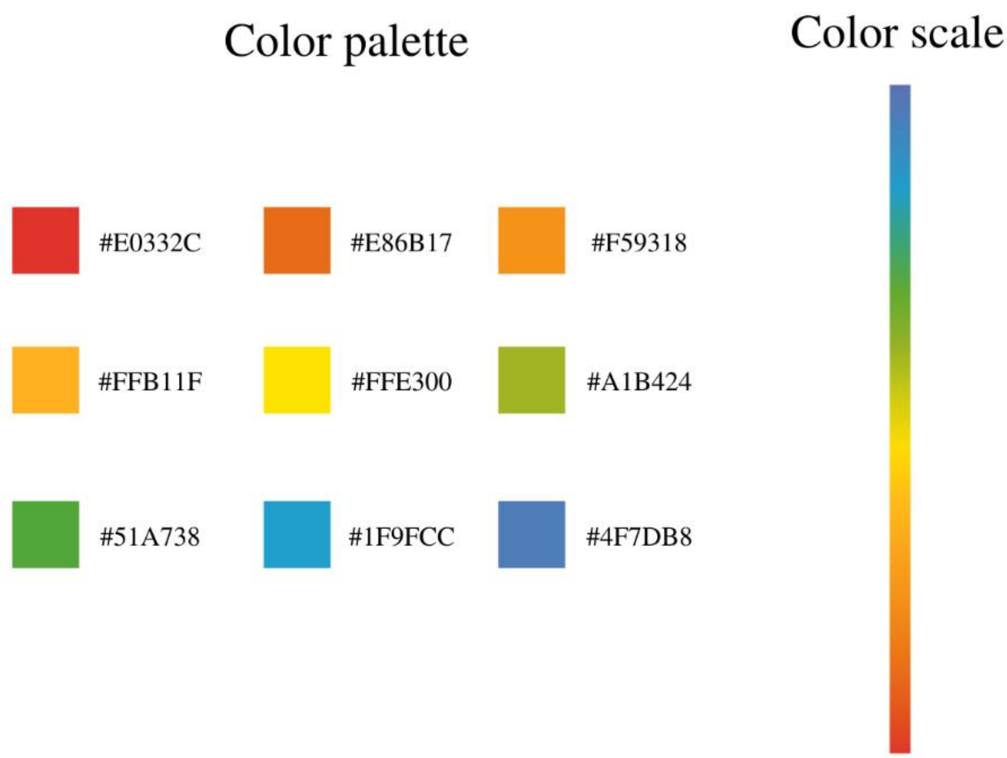


Figure 4.4: Color Scale and color palette used for Colored slope graph.

4.3 Color Transform Function

From the nine color codes, a color transform function is constructed with a linear gradience of the nine diverging colors. The total number of color stops is nine. The percentage at which the color stop occurs in the linear scale is calculated by (Max Val – Min Val)/ (Number of Colors - 1). For each time-series curve, the constructed linear scale is applied with a gradience in the direction of the slope line.

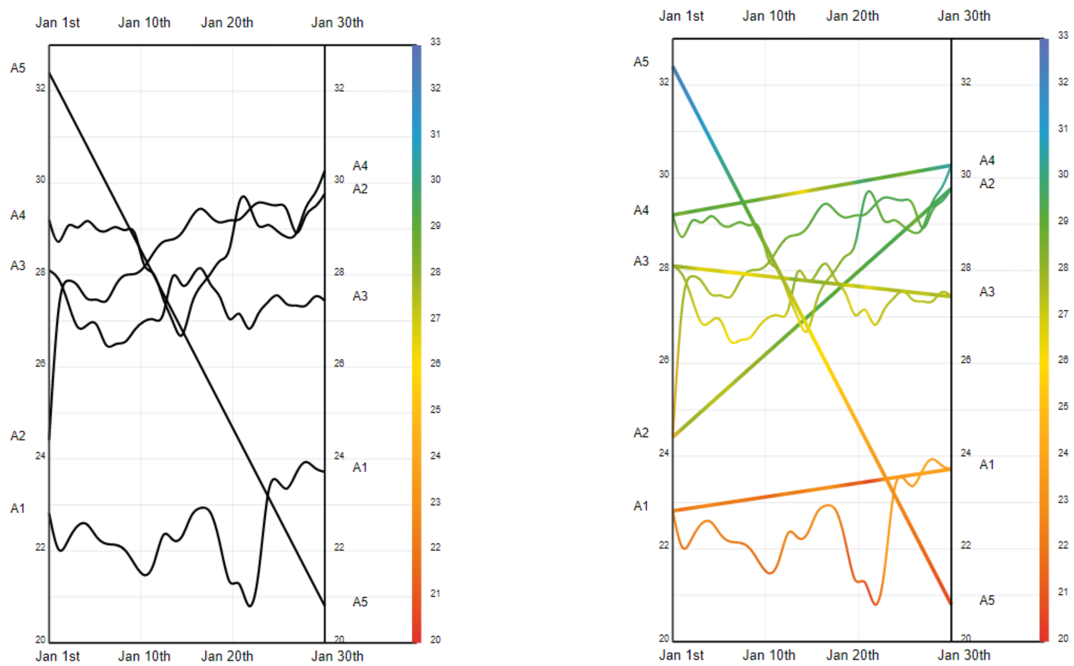


Figure 4.5: Interpretation of how a line graph is changed to a colored line and then how it is visualized as a colored slope graph.

4.4 Algorithm Design

Pseudo Code:

1. Let m be the number of time-series curves to be visualized
2. Let n be the number of temporal attributes.
3. Sort the m time-series curves based on the first temporal attribute.
4. Find the minimum value in the whole data, i.e. the minimum value in m*n data points.
5. Find the maximum value in the whole data, i.e. the maximum value in m*n data points
6. Construct the color transform function to encode the temporal attributes based on color.
    a. Let k be the number of color stops in the color transform function.
    b. Set the first color stop at the minimum value in the data using a linear scale and map it to # e0332c.
    c. The total number of color stops is nine. The percentage at which the color stop occurs is calculated by (Max Val – Min Val)/ (k - 1)

d. Set the k-2 intermediate color stops to the data using a linear scale and map it to the corresponding hex code.

e. Set the final color stop at the maximum value in the data using a linear scale and map it to # 4f7db8.

7. For each time-series data, at every temporal attribute, get the color from the color scale and construct the linear color gradience along the direction of the line.

8. After the generation of the linear gradience, plot the slope line with the linear gradience.

9. Repeat step 6 for all the m time-series data.

This visualization algorithm was developed using d3.js version 5 and HTML5/CSS/JavaScript.

4.5 Summary

This chapter outlines the design of the proposed visualization, algorithm along with the pseudo-code of the visualization technique, the color scale used for the visualization and the color transform function used for the visualization and the technology used to create the visualization. This gives the reader the chance to replicate and develop the color slope graph.

CHAPTER 5: EVALUATION

This chapter provides a detailed overview of the framework and the methodology used in this research study to evaluate the color slope visualization based on the benchmark of other predominant visualizations for time-series data such as line graph, spark area graph, and heatmap.

5.1 Other Predominant Visual Design

As a benchmark for the evaluation of the colored slope graph, the predominant time-series visualizations were selected. As mentioned by Few, the main visualizations for time-series data analysis are line graph, sparkline and heat map. These three visualizations are developed using d3.js version 5 and HTML5/CSS/JavaScript. Design consistency was followed by making sure the background color, font family and size was the same among all the visualizations.

5.1.1 Line graph

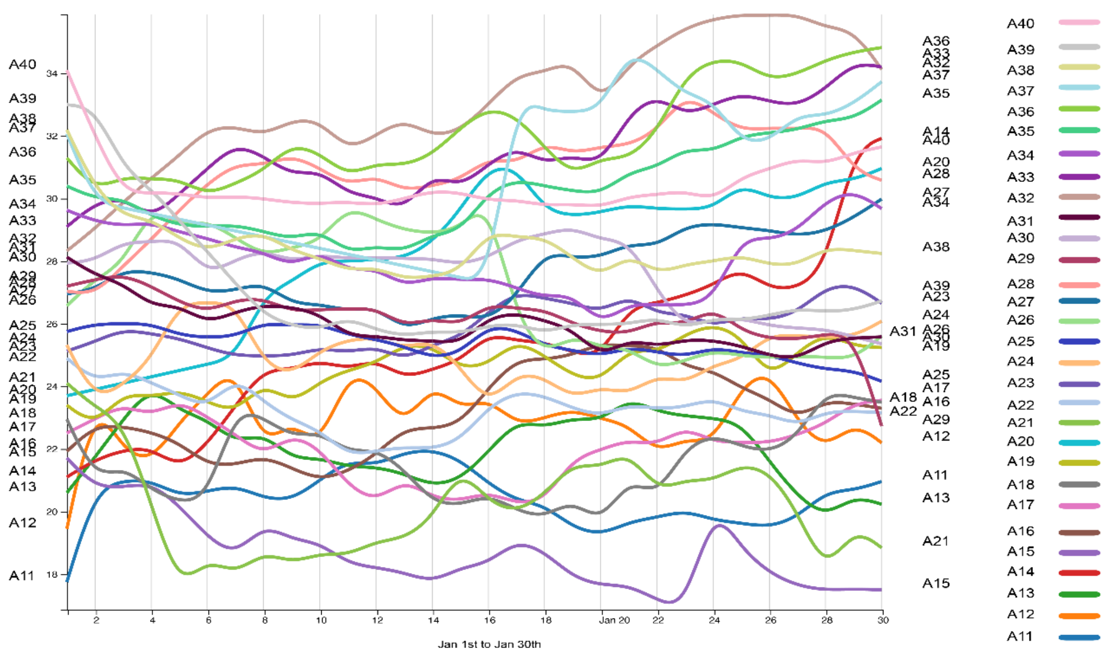


Figure 5.1: Line graph Visualization.

The line graph is the traditional graph predominantly used to represent the times series data illustrated in Figure 5.1. The x-axis represents the temporal attributes and the y-axis represents the value of the data points for a given time-series curve. Line graph visualization algorithm for this research study was developed using d3.js version 5 and HTML5/CSS/JavaScript.

5.1.2 Spark area

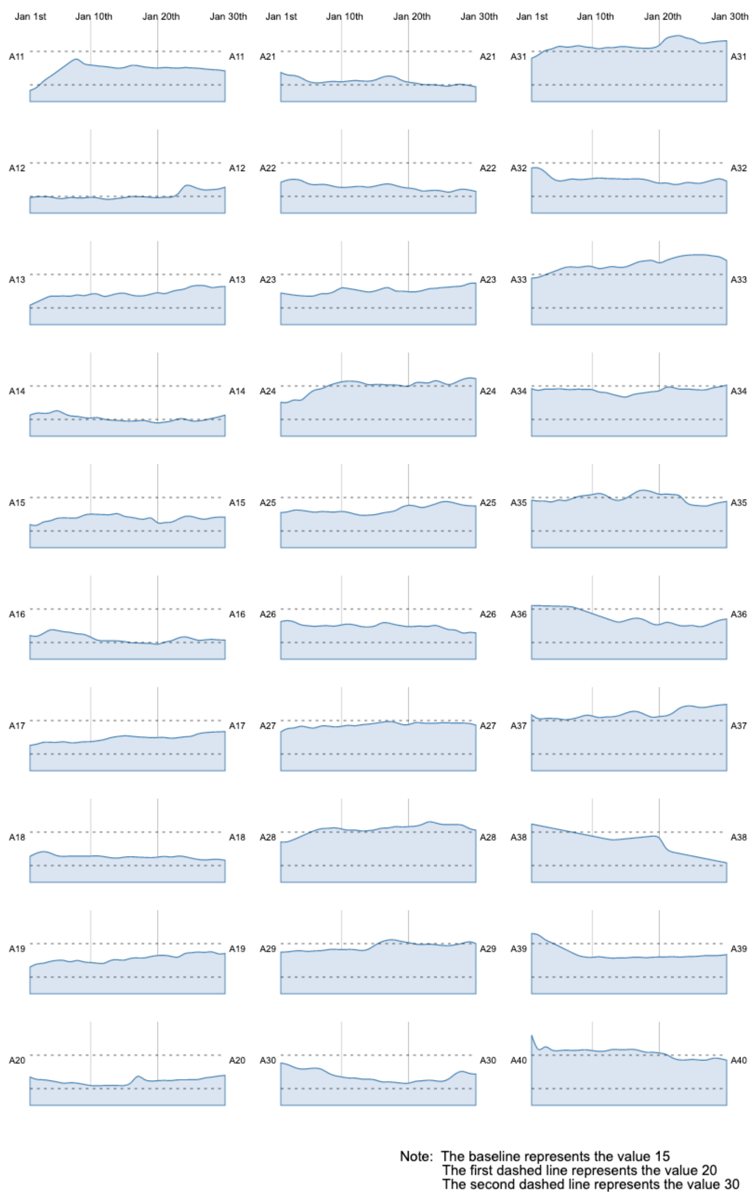


Figure 5.2: Spark area graph Visualization.

Spark area is the visualization that is inspired from the sparklines by Tufte illustrated in Figure 5.2. Each stock is visualized independently however the scale for all the stock in each screen space is constant. The x-axis represents the temporal information and the y-axis represents the value of the data point for the corresponding temporal attribute. Spark Area visualization algorithm for this research study was developed using d3.js version 5 and HTML5/CSS/JavaScript.

5.1.3 Heatmap

The heat map used in the experiments is a slight modification of the traditional heatmaps. Each stock is stacked upon one other in this visualization as illustrated in Figure 5.3. The x-axis represents the temporal attributes. All the data points are represented by color-coding. Color-coding is specified using a color scale that reflects the absolute values of the data. In this visualization, the same color scale which was used for colored slope graph visualization was used. For each line, a linear gradience is constructed based on the color transform function in the horizontal direction. Heat map visualization algorithm for this research study was developed using d3.js version 5 and HTML5/CSS/JavaScript.

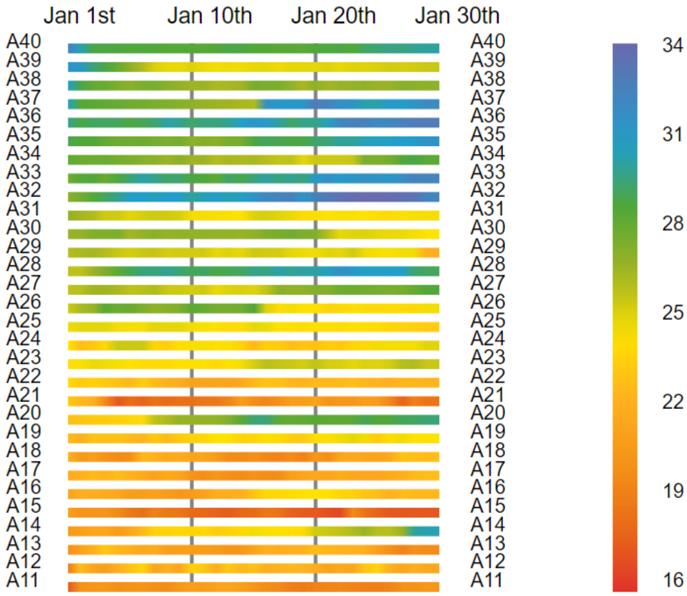


Figure 5.3: Heatmap Visualization.

5.2 Design of Experiments

A quantitative study was adopted for the evaluation of the proposed visualization. This study consisted of a training process, two human subject experiments and followed by a participant's survey. The experiments were designed to measure the efficacy, the effectiveness and the graphical perception of the visualization. In the first experiment, the colored slope graph technique was compared to line graph visualization, spark area visualization and heatmap, which are the most widely used visualization techniques for time-series data analysis. In the first experiment, the correctness of the participant's sketch was calculated. After the experiments, the participants were given a survey. From the experiments and the surveys, the hypothesis listed in Chapter 3 were tested.

5.2.1 Training

At the beginning of the study, participants were briefed about the purpose of the study. The participants were explained how line graphs, heatmaps, spark area graphs are used for the visualization of time-series data. After this, the participants were shown a "Colored slope line". Then the participants were shown the corresponding line graph. The participants were explained how to interpret the color encoding of the colored slope visualization. Then, the participants were provided with colored slope line with a legend stating the value corresponding to each color. The participants were asked to sketch the associated line graph on the given colored slope line. This step aids the participants on how to perceive information based on color and slope. The original line graph was superimposed on the sketched graph and was shown to the participants to show them how well they could perceive the color encoded information on the slope line. It is a crucial step as participants are completely new to the proposed visualization.

5.2.2 Experiment 1

Experiment 1 was mainly designed to focus on the effectiveness of the colored slope visualization compared to the other predominant visualization for the time-series data. The experiment design followed a within-subject design. The main objective of this experiment is to evaluate how well the participants perceive and understand the temporal information in the colored slope graph visualization, heat-map, line graph, and spark-area graph. Based on the four visualizations the four treatment groups are formed. Four datasets were prepared to create the visualizations. For each treatment method, a different dataset was used to avoid the participants from learning the right answers to the questions. The Groups are “Colored Slope A”, “Heatmap B”, “Line Graph C” and “Spark Lines D” where each group corresponds to colored slope visualization, line graph visualization, and sparkline visualization respectively illustrated in Table 5.1. Every participant was randomly assigned to four “Sets”. The four sets are “Human subjects Set 1”, “Human subjects Set 2”, “Human subjects Set 3”, “Human subjects Set 4” illustrated in Table 5.1. Each set helps in deciding the first treatment and the first dataset the participants will be introduced to. Different data set were selected for each group, and 30 stocks were visualized using the heatmap line graph, sparkline graph, and the colored slope visualization for the respective dataset. The participants were required to complete all the questions for each group. Each treatment was presented in the form of a Qualtrics survey which had a total of 7 questions. A set of 4 treatments are given in Figure 5.4, Figure 5.5, Figure 5.6 and Figure 5.7. These questions were designed to test a hypothesis and aim to answer a research question. The choices of the questions are all the stocks used for the corresponding visualization. For the whole experiment, a total of 16 visualizations were used based on the dataset and treatment group. Across all the visualizations, the background color, font type was kept constant to maintain design consistency. For none of the visuals, the values of the stocks were listed. Legend denoting the necessary information was provided as per requirements. The necessary explanation was provided for each question so that the participants had a clear understanding of what was expected to answer each question. The scoring of each choice is described in depth in Chapter 4. This assures the randomization and non-biasedness of the experiment.

Table 5. 1: Experiment Design – Each Participant will be assigned to a “Set”, which decides the first Group the participant will be introduced to.

Sets\Treatment	**Treatment 1**	**Treatment 2**	**Treatment 3**	**Treatment 4**
Human Set 1	ColoredSlope (A,1)	Heatmap (B,2)	Line (C,3)	Spark (D,4)
Human Set 2	Heatmap (B,2)	ColoredSlope (A,3)	Spark (D, 4)	Line (C, 1)
Human Set 3	Line (C, 3)	Spark (D, 4)	Heatmap (B, 1)	ColoredSlope (A,2)
Human Set 4	Spark (D, 4)	Line (C, 1)	ColoredSlope(A,2)	Heatmap (B, 3)

Note: ColoredSlope (A,1) refers to the treatment "A" and Dataset "1" used for the visualization.

Below are the seven questions for each treatment group.

1. Identify the stock which has the highest value on 20th Jan.

   Choices – All the stocks were listed with a single correct answer.

2. Identify the stock which covaries with a certain stock.

   Choices – All the stocks were listed with a single correct answer.

3. Identify the stock which has the highest variability.

   Choices – All the stocks were listed with a single correct answer.

4. Identify the stock which increases the most from 1st Jan to 30th Jan.

   Choices – All the stocks were listed with a single correct answer.

5. Identify the percentage of stocks increasing the percentage of stocks decreasing.

   Choices – [80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80].

6. Identify the stock which has the lowest value on 10th Jan.

   Choices – All the stocks were listed with a single correct answer.

7. Identify the stock which decreases the most from 1st Jan to 30th Jan.

   Choices – All the stocks were listed with a single correct answer.

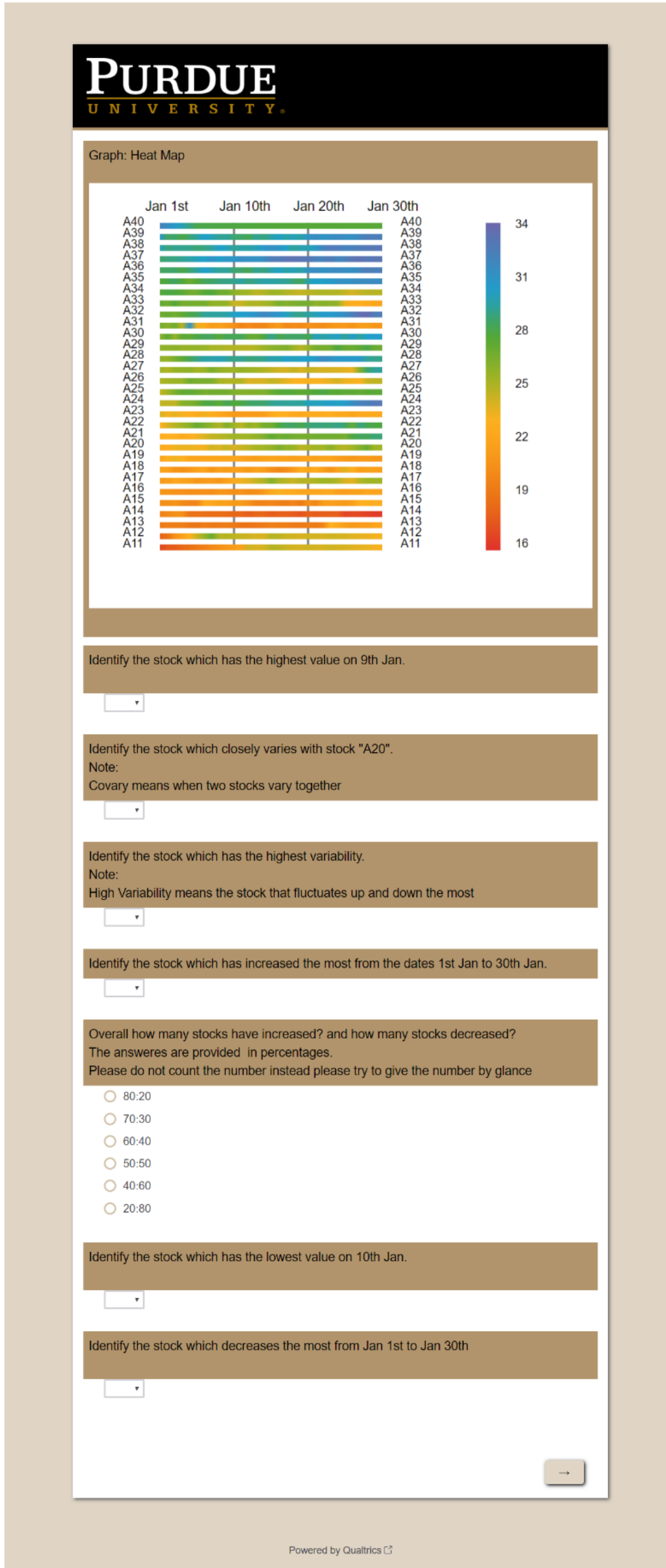
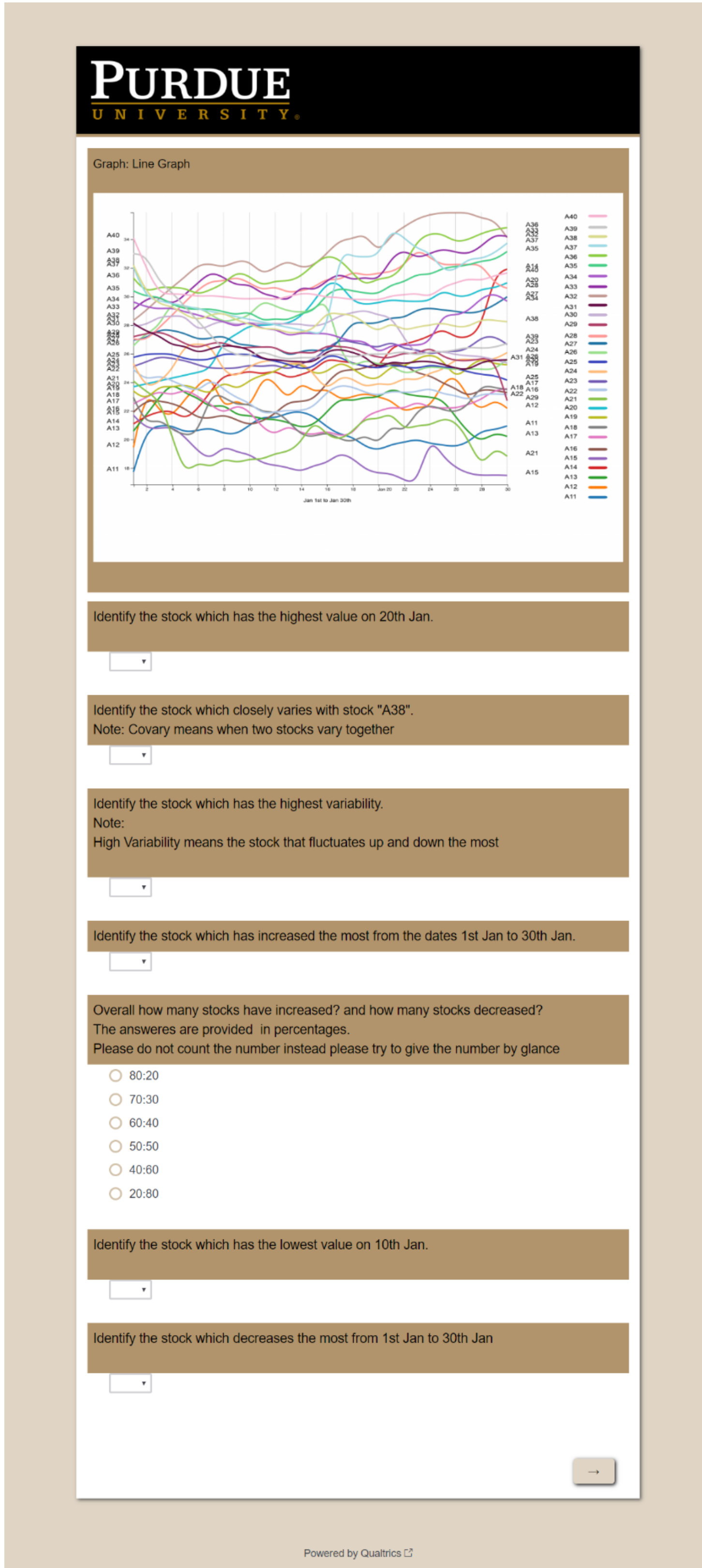


Figure 5.4: Heatmap treatment for experiment 1



PURDUE
UNIVERSITY

Graph: Line Graph

Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A38".
Note: Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from 1st Jan to 30th Jan

→

Powered by Qualtrics

Figure 5.5: Line graph treatment for experiment 1

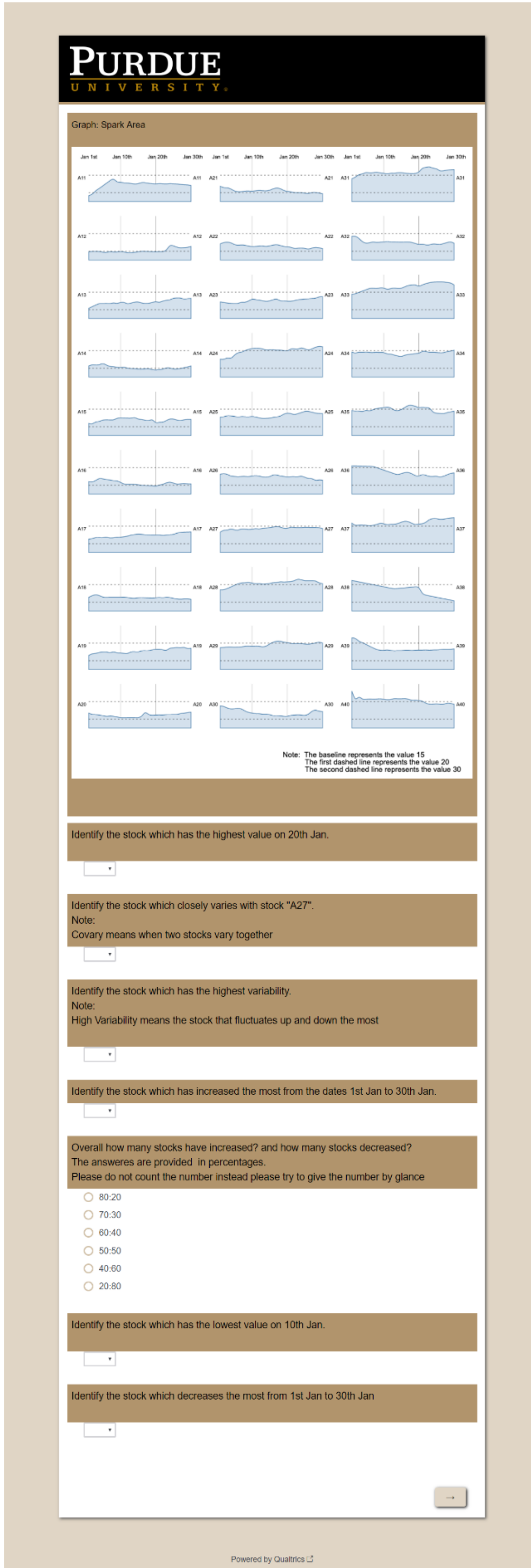


Figure 5.6: Spark area treatment for experiment 1

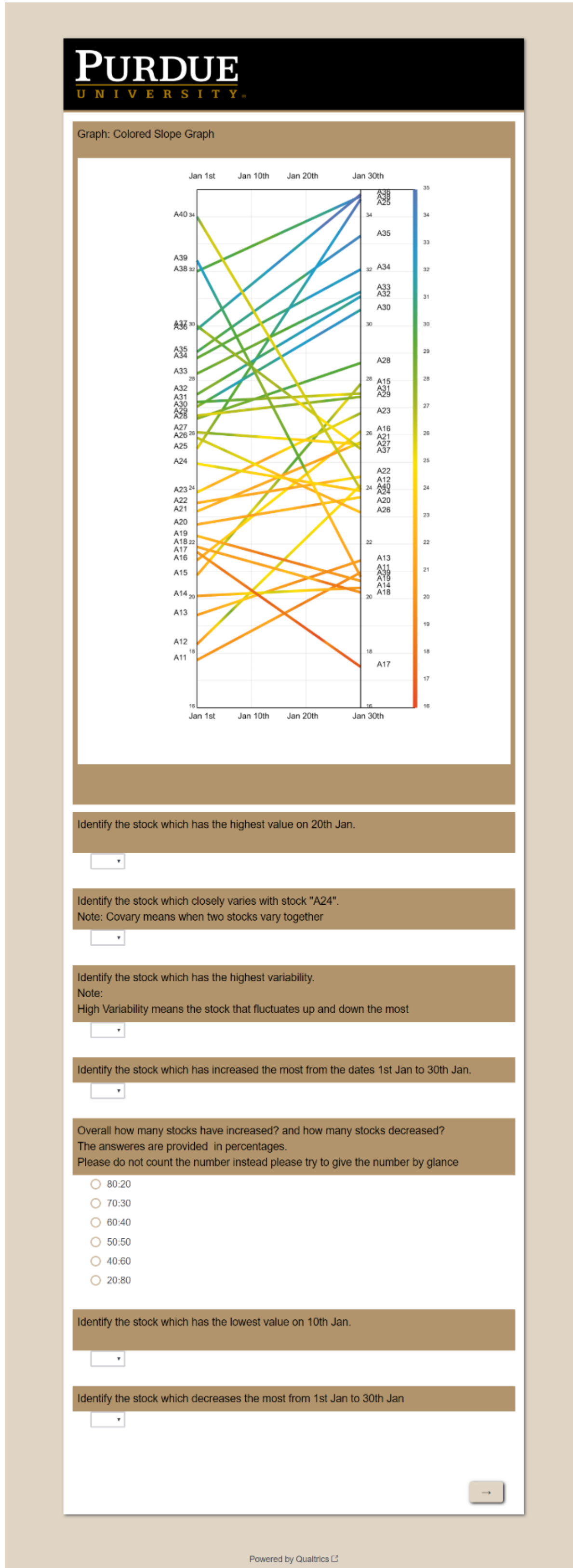
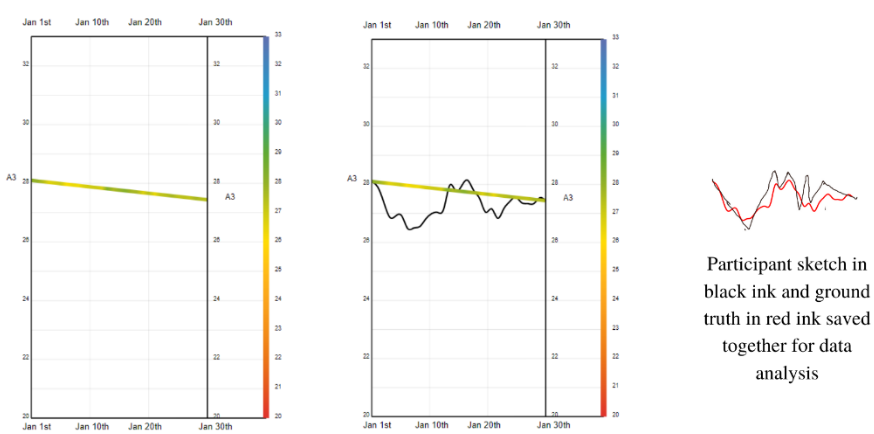


Figure 5.7: Colored slope graph treatment for experiment 1

Based on the participants' choice of the answers, the scores were computed which is explained in detail in Chapter 6. Along with the scores for each question, the time taken to complete each visualization was also recorded. From the scores, statistical inference (mean, median, standard deviation) were calculated and the results were analyzed using the Mann–Whitney U test.

5.2.2 Experiment 2

After conducting the first experiment, the second experiment was mainly designed to understand how well the users could perceive the information. In this experiment, the graphical perception of a slope with encoded color was studied. This helps to quantify the measure of efficacy for the proposed visualization technique. Experiment 2 consisted of one main task. The tasks were conducted using the sketch app on the iPad as shown in Figure 5.8. First, the participants were briefed about how a line graph is visualized in terms of colored slope lines and heat map lines. Then the participants were given 5 colored slope lines as shown in Figure 5.9 and were asked to sketch. The results from this step were not recorded. Once the participants had a clear understanding of the colored slope line, the participants proceeded with the tasks.

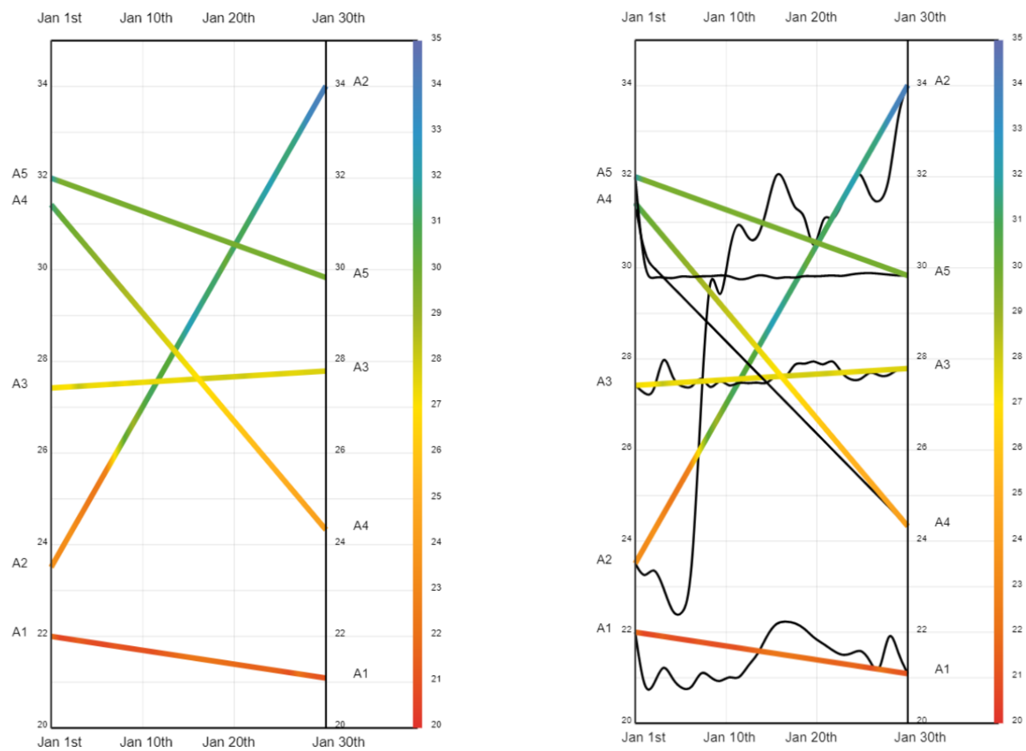


Colroed slope line shown to particiapant to sketch

Participant sketch in black ink

Participant sketch in black ink and ground truth in red ink saved together for data analysis

Figure 5.8: Experiment 2 set up.

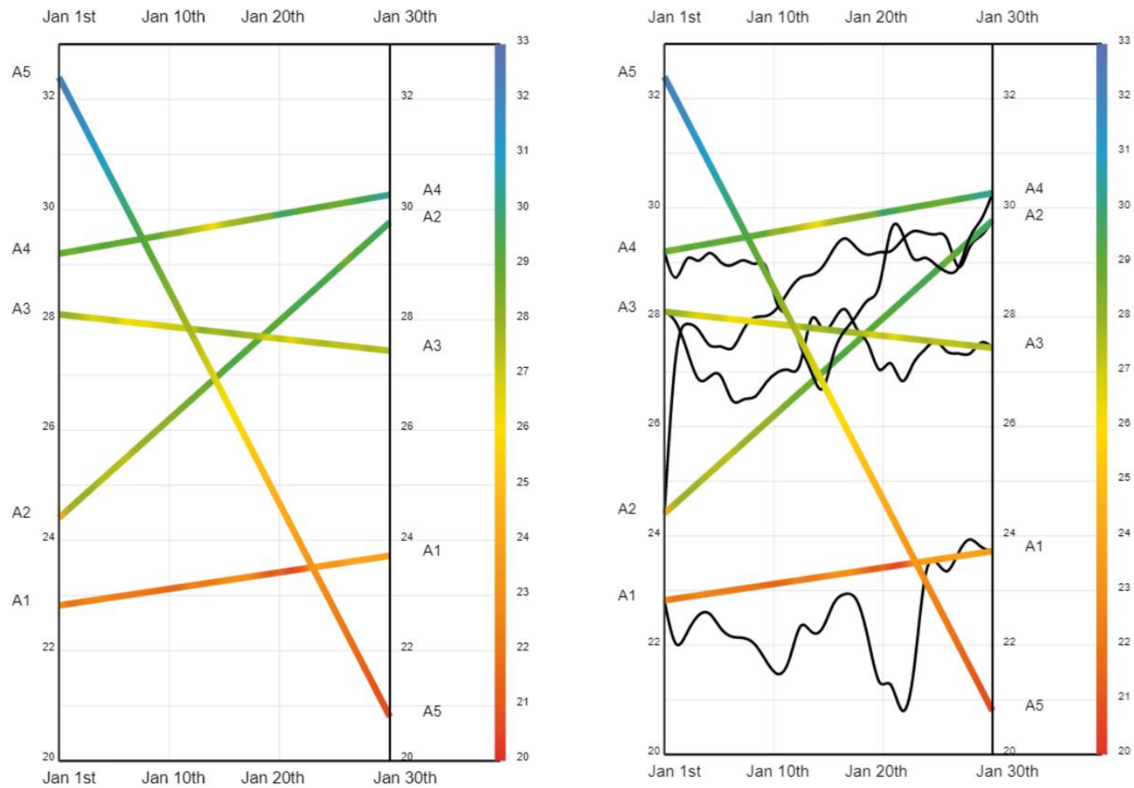


Colored Slope graph

Colored Slope graph with ground truth

Figure 5.9: Colored slope graph used for training participants.

First, the participants were provided with 5 colored slope lines with a legend stating the value corresponding to each color. The participants were asked to sketch the associated line graph on the given colored slope line as shown in Figure 5.10. This task test how participants perceive information based on color and slope. In this task, the original line graph was superimposed on the sketched graph and percentage error will be calculated as shown in Figure 5.8.



Colored Slope graph

Colored Slope graph with ground truth

Figure 5.10: Colored slope graph used for sketching in experiment 2.

5.2.3 Demographic Information

Participant's demographic information such as participant's gender, age, ethnicity, race, the highest level of education, program details were collected. None of the personal information which could identify the participant was recorded.

5.2.4 Participants Survey

A nondiscriminatory survey that is both reliable and pre-validated was used. The survey was used from Zhao, Liu, Guo, Qian, & Chen, (2019) and NASA TLX. As the experiment design is inspired by (Zhao, Liu, Guo, Qian, & Chen, 2019). The participants' survey consisted of the following question. All the choices for each question were randomized.

1. Among all the visuals pick the visual which you prefer for temporal Information
2. Among all the visuals, pick the visual which the easiest to understand the information

3. Among all the visuals, pick the visual which was the hardest to understand the information
4. Among all the visuals, pick the visual which the easiest to perceive the information
5. Among all the visuals, pick the visual which was the messiest or the visual which was very frustrating for you to understand
6. Rank the effectiveness of finding and comparing similar items for each visual.
7. Rank the effectiveness of finding the stocks with a dramatic increase/decrease.
8. Rank the effectiveness of finding the stocks with high variance.
9. Rank the effectiveness of finding extreme(maximum) values.
10. Rank the intuitiveness of the visualization design representation.
11. Rank the aesthetics of the visualization design.
12. Rank the visualizations learnability.

5.3 Summary

This chapter outlines the design of the proposed visualization, the dataset used, the overall design of the research study: evaluation, methods used, experimental collection and demographic sample. This gives the reader the chance to replicate and validate the work if one wishes to.

CHAPTER 6: ANALYSIS

This chapter provides in-depth detail of demographic information, data processing and data analysis of the data collected from the experiments.

6.1 Demographics

There was a total of 35 participants from Purdue University for this research study. The breakdown of the demographics is as follows. Out of the 35 participants, 21 participants are male, and 14 participants are female illustrated in Figure 6.1. For the breakdown of the demographics by the highest level of education is that 9 of them belonged to bachelor's degrees, 10 of the participants are from Ph.D. and 16 participants are master's students illustrated in Figure 6.2. The breakdown of the participants by race is as follows where 9 participants were White, 5 participants were Black, 18 participants were Asian, 2 participants were Hispanic, 1 participant was American Indian or American Alaskan which is illustrated in Figure 6.4

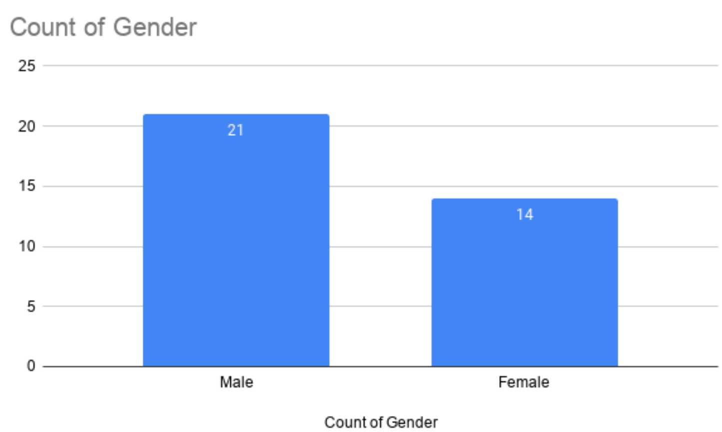


Figure 6.1: Demographic information by Gender.

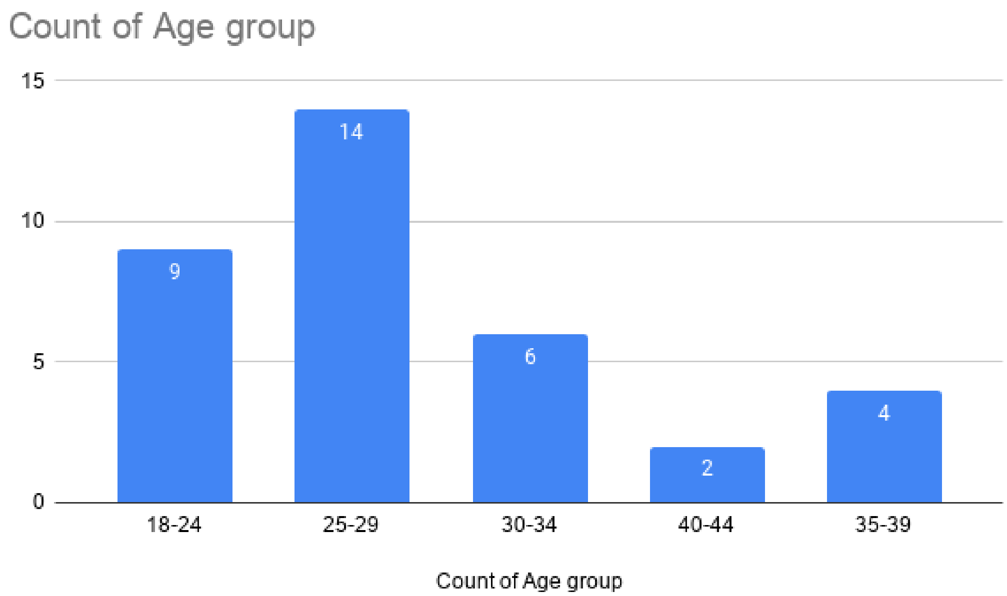


Figure 6.2: Demographic information by age group.

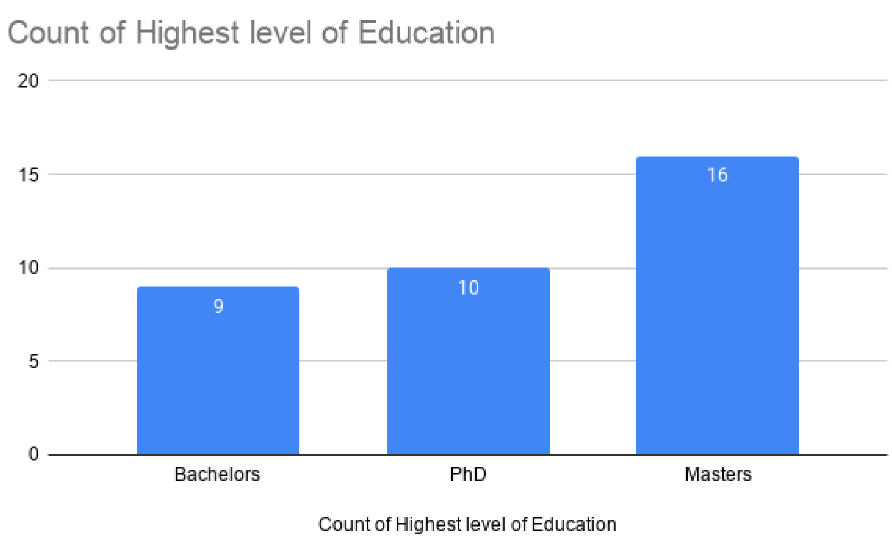


Figure 6.3: Demographic information by level of education.

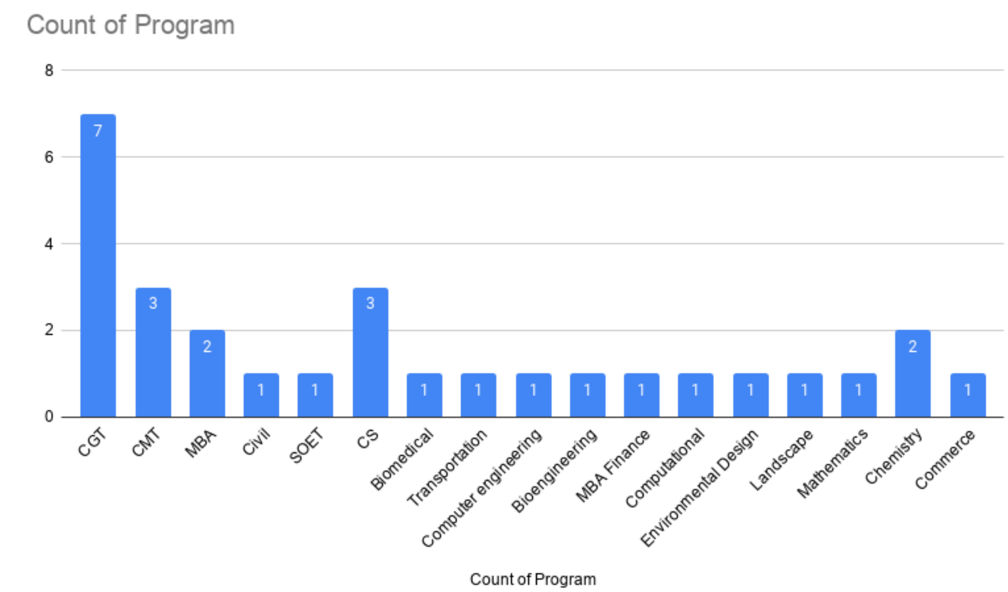


Figure 6.4: Demographic information by the program.

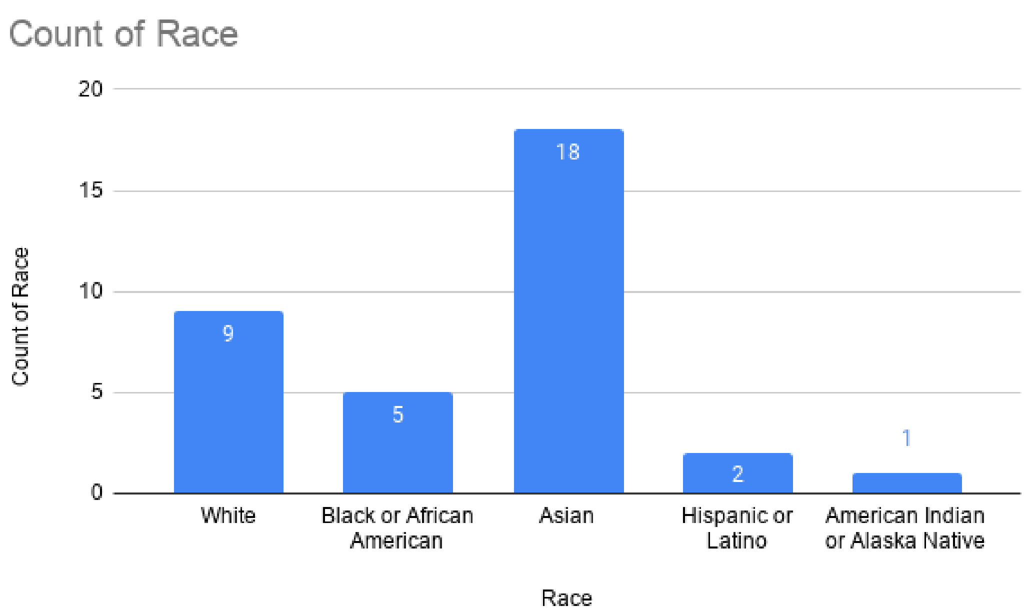


Figure 6.5: Demographic information by race.

For the first experiment, 2 participants were eliminated as the participants were identified as random clickers based on the time recorded and the number of correct answers in all the treatments. For the second experiment, all the 35 participants' sketches were considered for the data analysis.

6.2 Experiment 1

6.2.1 Distance as a similarity and scoring metric

"Euclidean distance or Euclidean metric is referred to as the ordinary straight-line distance between two points in Euclidean space" (Euclidean distance, 2020). Given two points in space, by taking the distance between these two points, the Euclidean distance can be calculated. In machine learning, Euclidean distance is used as metric to compare the similarity index of two points based on the reference point in an n-dimensional space (Measuring Dis/Similarities Between Objects (Cells) In 'n'-Dimensional Space, 2020). Euclidean distance can be used for points if they are in the same scaling system. This general idea was used to score the choices for the questions. In our scoring system, the correct answer is the reference data point. If the participant picks the correct choice, the participant gets the maximum score. However, most of the time, the users pick very close choices and not all choices can be treated equally as all the choices are non-dependent on one other. The scoring of all the choices cannot be uniform. By computing the Euclidean distance between the choice picked by the participant and the correct choice. i.e. the reference point or the ground truth, the similarity or scoring metrics can be calculated. This metric of scoring gives a non-uniform scoring system.

6.2.2 Data Processing

The data collected from the first experiment was scored for each question based on the criteria of the distance of the selected value from the ground truth. The ground truth refers to the value of the right choice. The rationale behind calculating the distance of a data point from the ground truth was that all points on the time-series curves can be considered as random independent points. All the distances calculated are not linearly dependent on each other. This implies that the scoring of all the choices cannot be uniform. Therefore, the score of each choice was evaluated based on the value of the data point minus the ground truth. The maximum score a participant can get is "0" and the minimum score a participant can score is minimum (datapoints value) minus the ground truth.

1. Score: Data point value of the Choice - Ground Truth
2. Max Score: "0"

For each treatment group, once the scores were computed, statistical inferences such as Mean, Standard deviation were calculated. To test the significance of each treatment group, Mann–Whitney U test one-sided was computed.

6.2.3 Data Analysis

How to analyze the results:

For each question, there are two which contain the results of u-test. The one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Question 1:

Identify the stock which has the highest value on 20th Jan.

For each dataset, the stock with the highest value was identified. The highest value was recorded as the ground truth. The value of the participants' choice is recorded. The score for the participant for a treatment group with the corresponding dataset was calculated by subtracting the ground truth from the participants' value. The maximum score a participant can get is zero. To all the scores mean and the standard deviation was calculated and for each treatment pair, Mann–Whitney U test one-sided testing was computed.

Table 6.1: Statistical Inferences of question 1 for each treatment

Treatment	Mean	Standard Deviation
Colored Slope	-2.742	6.374
Heatmap	-4.848	6.592
Line graph	-7.133	9.112
Spark Area graph	-6.951	9.157

Table 6.2: Mann–Whitney U test one-sided results denoting the statistic value from U-test.

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		110.0	108.0	213.5
Heatmap	110.0		534.0	435.5
Line graph	108.0	534.0		435.5
Spark Area	213.5	435.5	435.5	

Table 6.3: Mann–Whitney U test one-sided results denoting the p-value.

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		0.020	0.039	0.007
Heatmap	0.020		0.434	0.289
Line graph	0.039	0.434		0.341
Spark Area	0.007	0.289	0.341	

In the above two tables i.e. Table 6.2 and Table 6.3, the one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Below are the stats of the number of participants who have picked the exact answer.

Table 6.4: Participants with exact answers for question 1 for each treatment

Treatment	No of exact correct answers
Colored Slope	21
Heatmap	12
Line graph	15
Spark Area	11

From the mean and the standard deviation, it can be noted that the colored slope visualization performed better than the rest of the groups. These scores were used to calculate Mann–Whitney U test one-sided (with $\alpha = 0.05$). From the u test, there is a significant difference in the performance of the colored slope visualization. The p-value for colored slope and heat map was $p = 0.02$ with mean and standard deviation being (-2.742, 6.374) and (-4.848, -6.592) respectively. The p-value for colored slope and line graph was $p = 0.039$ with mean and standard deviation being (-2.742, 6.374) and (-7.133, 9.122) respectively. The p-value for colored slope and

Spark area was p = 0.007 with mean and standard deviation being (-2.742, 6.374) and (-6.951, 9.157) respectively. Therefore, the null hypothesis has been rejected for its respective pairs. From the statistical inferences of mean, standard deviation and p-value from u test it can be concluded that colored slope graph performs better than the rest of the visualizations to identify the extreme(maximum) values.

Therefore, from the analysis, it can be concluded that the performance of the colored slope graph is significantly greater compared to the line graph, spark area graph and heat map in identifying the extreme(maximum) value for a given temporal attribute.

Question 2:

Identify the stock which covaries with a certain stock.

This question was scored by calculating the covariance coefficient of two time-series curves. The covariance coefficient between two time-series curves returns the values form [-1, 1]. If the value returned is 1, it means that given two time-series curves perfectly covary. This means that the derivatives at each datapoint of the time-series curves are equal. If the covariance coefficient is in the range [0, 1), it means that the derivative of the data points is in the same direction as the given time-series curve. If the covariance coefficient is in the range [-1,0), it means that the derivatives of the data points for a given time-series curve are in the opposite direction. A direct function from the NumPy module of python was used to calculate the covariance coefficient. Therefore, for each dataset, the covariance coefficient was calculated with the stock listed on the question. After computing the covariance coefficient, the score for each choice is computed by the Covariance Coefficient of the data point minus the ground truth. The maximum score a participant can get is zero. To all the scores mean and the standard deviation was calculated and for each treatment pair, Mann–Whitney one-sided U test was computed.

Table 6.5: Statistical Inferences of question 2 for each treatment

Treatment	Mean	Standard Deviation
Colored Slope	-0.654	0.682
Heatmap	-1.441	1.185
Line graph	-1.525	1.178
Spark Area graph	-1.195	1.097

Table 6.6: Mann–Whitney U test one-sided results denoting the statistic value from U-test

	Colored Slope	**Heatmap**	**Line graph**	**Spark Area**
Colored Slope		389.0	330.5	351.0
Heatmap	389.0		527.0	513.5
Line graph	330.5	527.0		476.5
Spark Area	351.0	513.5	476.5	

Table 6.7: Mann–Whitney U test one-sided results denoting the p-value

	Colored Slope	**Heatmap**	**Line graph**	**Spark Area**
Colored Slope		0.023	0.003	0.007
Heatmap	0.023		0.414	0.348
Line graph	0.003	0.414		0.193
Spark Area	0.007	0.348	0.193	

In the above two tables i.e. Table 6.6 and Table 6.7, the one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Below are the stats of the number of participants who have picked the exact answer.

Table 6. 8: Participants with exact answers for question 2 for each treatment

Treatment	**No of exact correct answers**
Colored Slope	3
Heatmap	2
Line graph	1
Spark Area	1

From the mean and the standard deviation, it can be noted that the colored slope visualization performed better than the rest of the groups. These scores were used to calculate the Mann–Whitney U test (with $\alpha = 0.05$). From the u test, there is a significant difference in the performance of the colored slope visualization. The p-value for colored slope and heat map was $p = 0.023$ with mean and standard deviation being (-0.654, 0.682) and (-1.441, 1.1857) respectively. The p-value for colored slope and line graph was $p = 0.003$ with mean and standard deviation being (-0.654, 0.682) and (-1.525, 1.178) respectively. The p-value for colored slope and Spark area was p = 0.007 with mean and standard deviation being (-0.654, 0.682) and (-1.195, 1.097) respectively. Therefore, the null hypothesis has been rejected for its respective pairs. From the statistical inferences of mean, standard deviation and p-value from u test it can be concluded that colored slope graph performs better than the rest of the visualizations to identify coverability between two time-series curves.

Therefore, from the analysis, it can be concluded that the performance of the colored slope graph is significantly greater compared to the line graph, spark area graph and heat map in identifying coverability between two time-series curves.

Question 3:

Identify the stock which has the highest variability.

A time curve has high variability when the time curve fluctuates. For computing the fluctuations in a time-series curve, the time curve is considered as a signal. The fluctuations in a signal are calculated by extracting the peaks and valleys. These points are extracted based on a sliding window and calculating the local maxima or local minima in the sliding window. Signal.Peaks and Signal.Valleys method from the SciPy module of python was used to extract the peaks and valleys. The ground truth was calculated by choosing the data point which has the maximum number of peaks and valleys. The scores of the participant's choices are calculated by extracting the number of peaks and valleys for the participants' choice minus the ground truth. The maximum score a participant can get is “0”. To all the scores, the Mann–Whitney one-sided U test was computed.

Table 6.9: Statistical Inferences of question 3 for each treatment

Treatment	Mean	Standard Deviation
Colored Slope	-9.758	10.852
Heatmap	-11.636	9.616
Line graph	-12.667	8.940
Spark Area graph	-10.424	10.015

Table 6.10: Mann–Whitney U test one-sided results denoting the statistic value form the u-test

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		468.0	433.0	477.0
Heatmap	468.0		513.5	509.5
Line graph	433.0	513.5		441.0
Spark Area	477.0	509.5	441.0	

Table 6.11: Mann–Whitney U test one-sided results denoting the p-value

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		0.162	0.075	0.191
Heatmap	0.162		0.347	0.328
Line graph	0.075	0.347		0.092
Spark Area	0.191	0.328	0.092	

In the above two tables i.e. Table 6.10 and Table 6.11, the one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Below are the stats of the number of participants who have picked the exact answer.

Table 6.12: Participants with exact answers for question 3 for each treatment

Treatment	No of exact correct answers
Colored Slope	7
Heatmap	1
Line graph	4
Spark Area	2

From the mean and the standard deviation, it can be noted that the colored slope visualization performed slightly better than the rest of the groups. These scores were used to calculate Mann–Whitney one-sided U test (with $\alpha = 0.05$). From the u test, there is a significant difference in the performance of the colored slope visualization. The p-value for colored slope and heat map was $p = 0.162$ with mean and standard deviation being (-9.758, 10.852) and (-11.636, 9.616) respectively. The p-value for colored slope and line graph was $p = 0.075$ with mean and standard deviation being (-9.758, 10.852) and (-12.667, 8.940) respectively. The p-value for colored slope and Spark area was p = 0.191 with mean and standard deviation being (-9.758, 10.852) and (-10.424, 10.015) respectively. Therefore, the null hypothesis could not be rejected for its respective pairs. From the statistical inferences of mean, standard deviation and p-value from u test it can be concluded that there is no significant difference among the groups to identify high variability between two time-series curves.

Therefore, from the u-test analysis, it can be concluded that there is no significant difference in performance among colored slope graph, line graph, spark area graph and heat map in identifying the sticks with high variability. However, from the number of exact answers, the participants have picked and more the statistical inferences of mean and standard deviation it can be said that the colored slope graph performed slightly better than the other visualizations.

Question 4:

Identify the stock which increases the most from 1st Jan to 30th Jan.

For each dataset, the stock with the highest increase in the value from 1st Jan to 30th Jan was identified. The highest increase value i.e. the maximum of the stock value on 30th Jan minus the stock value on 1st Jan was recorded as the ground truth. The value of the participants' choice is recorded. The score for the participant for a treatment group with the corresponding dataset was calculated by subtracting the participants' value from the ground truth. The maximum score a participant can get is zero. To all the scores, the Mann–Whitney one-sided U test was computed.

Table 6.13: Statistical Inferences of question 4 for each treatment

Treatment	**Mean**	**Standard Deviation**
Colored Slope	-2.336	8.212
Heatmap	-7.080	8.936
Line graph	-10.763	8.442
Spark Area graph	-9.822	9.660

Table 6.14: Mann–Whitney U test one-sided results denoting the statistic value from the u-test.

	Colored Slope	**Heatmap**	**Line graph**	**Spark Area**
Colored Slope		295.5	157.5	221.0
Heatmap	295.5		390.5	438.0
Line graph	157.5	390.5		489.5
Spark Area	(221.0	438.0	489.5	

Table 6.15: Mann–Whitney U test one-sided results denoting the p-value and statistics

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		0.000	0.000	0.000
Heatmap	0.000		0.023	0.083
Line graph	0.000	0.023		0.241
Spark Area	0.000	0.083	0.241	

In the above two tables i.e. Table 6.14 and Table 6.15, the one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Below are the stats of the number of participants who have picked the exact answer.

Table 6.16: Participants with exact answers for question 4 for each treatment

Treatment	No of exact correct answers
Colored Slope	27
Heatmap	13
Line graph	5
Spark Area	9

From the mean and the standard deviation, it can be noted that the colored slope visualization performed better than the rest of the groups. These scores were used to calculate the Mann–Whitney U test (with $\alpha = 0.05$). From the u test, there is a significant difference in the performance of the colored slope visualization. The p-value for colored slope and heat map was $p = 0.000$ with mean and standard deviation being (-2.336, 8.212) and (-7.080, 8.936) respectively. The p-value for colored slope and line graph was $p = 0.000$ with mean and standard deviation being (-2.336, 8.212) and (-10.763, 8.442) respectively. The p-value for colored slope and Spark area was $p = 0.000$ with mean and standard deviation being (-2.336, 8.212) and (-9.822, 9.660) respectively. Therefore, the null hypothesis has been rejected for its respective pairs. From the statistical inferences of mean, standard deviation and p-value from u test it can be concluded that colored slope graph performs better than the rest of the visualizations to identify the stocks that most increase.

Therefore, from the analysis, it can be concluded that the performance of the colored slope graph is significantly greater compared to the line graph, spark area graph and heat map in identifying the stock which increases the most from Jan 1st to Jan 31st.

Question 5:

Identify the percentage of stocks increasing the percentage of stocks decreasing.

For each dataset, the percentage of the number of stocks increasing, and the percentage of the number of stocks decreasing was calculated. The percentage of the number of stocks increasing was recorded as the ground truth. The value of the participants' choice is recorded. The score for the participant for a treatment group with the corresponding dataset was calculated by subtracting the participants' value from the ground truth. To all the scores, the Mann–Whitney one-sided U test was computed.

Table 6.17: Statistical Inferences of question 5 for each treatment

Treatment	Mean	Standard Deviation
Colored Slope	-0.303	3.881
Heatmap	-1.818	11.403
Line graph	-5.455	13.946
Spark Area graph	-0.303	12.906

Table 6.18: Mann–Whitney U test one-sided results denoting the statistic value from the u-test

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		476.0	393.5	494.0
Heatmap	476.0		469.0	476.0
Line graph	393.5	469.0		427.0
Spark Area	494.0	476.0	427.0	

Table 6.19: Mann–Whitney U test one-sided results denoting the p-value

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		0.165	0.018	0.237
Heatmap	0.165		0.165	0.232
Line graph	0.018	0.165		0.062
Spark Area	0.237	0.232	0.062	

In the above two tables i.e. Table 6.18 and Table 6.19, the one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Below are the stats of the number of participants who have picked the exact answer.

Table 6.20: Participants with exact answers for question 5 for each treatment

Treatment	No of exact correct answers
Colored Slope	28
Heatmap	10
Line graph	7
Spark Area	10

From the mean and the standard deviation, it can be noted that the colored slope visualization performed better than the rest of the groups. These scores were used to calculate the Mann–Whitney U test (with $\alpha = 0.05$). From the u test, there is a significant difference in the performance of the colored slope visualization. The p-value for colored slope and heat map was $p = 0.165$ with mean and standard deviation being (-1.515, 3.585) and (-9.090, 7.119) respectively. The p-value for colored slope and line graph was $p = 0.018$ with mean and standard deviation being (-1.515, 3.585) and (-11.515, 9.573) respectively. The p-value for colored slope and Spark area was $p = 0.237$ with mean and standard deviation being (-1.515, 3.585) and (-10.0, 8.165) respectively. Therefore, the null hypothesis has been rejected for its respective pairs except for color slope and line graph. From the statistical inferences of mean, standard deviation and p-value from u test it can be concluded that colored slope graph performs better than the line graph to identify the general trend in the data. However, there is no significant difference between colored slope graph, heatmap and spark area in identifying the ratio of the percentage of increase to the percentage of decrease in the number of stocks.

Therefore, from the analysis, it can be concluded that there is no significant difference among colored slope graph, spark area graph and heat map in identifying the percentage of increase to percentage of decrease from Jan 1st to Jan 30th. However, the performance of the colored slope graph is significantly greater than the performance of the line graph in identifying the percentage of increase to the percentage of decrease from Jan 1st to Jan 30th.

Question 6:

Identify the stock which has the lowest value on 10th Jan.

For each dataset, the stock with the lowest value was identified. The lowest value was recorded as the ground truth. The value of the participants' choice is recorded. The score for the participant for a treatment group with the corresponding dataset was calculated by subtracting the participants' value from the ground truth. The maximum score a participant can get is zero. To all the scores mean and the standard deviation was calculated and for each treatment pair, Mann–Whitney U test was computed.

Table 6.21: Statistical Inferences of question 6 for each treatment

Treatment	Mean	Standard Deviation
Colored Slope	-2.272	4.060
Heatmap	-5.239	5.568
Line graph	-5.884	6.697
Spark Area graph	-4.721	4.812

Table 6. 22: Mann–Whitney U test one-sided results denoting the statistic value from u-test

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		329.0	357.0	346.0
Heatmap	329.0		539.0	520.0
Line graph	357.0	539.0		516.0
Spark Area	346.0	520.0	516.0	

Table 6.23: Mann–Whitney U test one-sided results denoting the p-value

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		0.002	0.005	0.004
Heatmap	0.002		0.474	0.380
Line graph	0.005	0.474		0.357
Spark Area	0.004	0.380	0.357	

In the above two tables i.e. Table 6.21 and Table 6.22, the one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Below are the stats of the number of participants who have picked the exact answer.

Table 6.24: Participants with exact answers for question 6 for each treatment

Treatment	No of exact correct answers
Colored Slope	20
Heatmap	9
Line graph	12
Spark Area	11

From the mean and the standard deviation, it can be noted that the colored slope visualization performed better than the rest of the groups. These scores were used to calculate the Mann–Whitney U test (with $\alpha = 0.05$). From the u test, there is a significant difference in the performance of the colored slope visualization. The p-value for colored slope and heat map was $p = 0.002$ with mean and standard deviation being (-2.272, 4.060) and (-5.239, 5.568), respectively. The p-value for colored slope and line graph was $p = 0.005$ with mean and standard deviation being (-2.272, 4.060) and (-5.884, 6.697) respectively. The p-value for colored slope and Spark area was $p = 0.004$ with mean and standard deviation being (-2.272, 4.060) and (-4.721, 4.812) respectively. Therefore, the null hypothesis has been rejected for its respective pairs. From the statistical inferences of mean, standard deviation and p-value from u test it can be concluded that colored slope graph performs better than the rest of the visualizations to identify the extreme(minimum) values.

Therefore, from the analysis, it can be concluded that the performance of the colored slope graph is significantly greater compared to the line graph, spark area graph and heat map in identifying the extreme(minimum) value for a given temporal attribute.

Question 7:

Identify the stock which decreases the most from 1st Jan to 30th Jan.

For each dataset, the stock with the lowest increase in the value from 1st Jan to 30th Jan was identified. The lowest increase value i.e. minimum of the stock value on 30th Jan minus the stock value on 1st Jan was recorded as the ground truth. The value of the participants' choice is recorded. The score for the participant for a treatment group with the corresponding dataset was calculated by subtracting the ground truth from the participant's value. The maximum score a participant can get is zero. To all the scores mean and the standard deviation was calculated and for each treatment pair, Mann–Whitney one-sided U test was computed.

Table 6.25: Statistical Inferences of question 7 for each treatment

Treatment	Mean	Standard Deviation
Colored Slope	-1.366	5.566
Heatmap	-16.732	12.322
Line graph	-16.852	12.148
Spark Area graph	-13.021	12.064

Table 6.26: Mann–Whitney U test one-sided results denoting the statistic value from u-test

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		110	108	213.5
Heatmap	110		534	435.5
Line graph	108	534		434.5
Spark Area	213.5	435.5	434.5	

Table 6. 27: Mann–Whitney U test one-sided results denoting the p-value

	Colored Slope	Heatmap	Line graph	Spark Area
Colored Slope		0.000	0.000	0.000
Heatmap	0.000		0.449	0.081
Line graph	0.000	0.449		0.079
Spark Area	0.000	0.081	0.079	

In the above two tables i.e. Table 6.26 and Table 6.27, the one-sided test is read based on the treatment listed on the left-most column in the table with the treatment listed on the topmost column. The green background color represents the performance of the treatment listed on the leftmost column is significantly greater than the treatment listed in the topmost column for the corresponding cell.

Below are the stats of the number of participants who have picked the exact answer.

Table 6.28: Participants with exact answers for question 7 for each treatment

Treatment	No of exact correct answers
Colored Slope	29
Heatmap	4
Line graph	4
Spark Area	10

From the mean and the standard deviation, it can be noted that the colored slope visualization performed better than the rest of the groups. These scores were used to calculate the Mann–Whitney U test (with $\alpha = 0.05$). From the u test, there is a significant difference in the performance of the colored slope visualization. The p-value for colored slope and heat map was $p = 0.000$ with mean and standard deviation being (-1.366, 5.566) and (-16.732, 12.322) respectively. The p-value for colored slope and line graph was $p = 0.000$ with mean and standard deviation being (-1.366, 5.566) and (-16.852, 12.148) respectively. The p-value for colored slope and Spark area was $p = 0.000$ with mean and standard deviation being (-1.366, 5.566) and (-13.021, 12.064) respectively. Therefore, the null hypothesis has been rejected for its respective pairs. From the statistical inferences of mean, standard deviation and p-value from u test it can be concluded that colored slope graph performs better than the rest of the visualizations to identify the stock that decreases the most.

Therefore, from the analysis, it can be concluded that the performance of the colored slope graph is significantly greater compared to the line graph, spark area graph and heat map in identifying the stocks that decrease the most.

Time Analysis:

Time taken for each treatment was also analyzed. The below table summarizes the average time taken by the participants for each treatment group.

Table 6.29: Statistical Inferences of time analysis

Treatment	Mean	Standard Deviation
Colored Slope	229.303	164.293
Heatmap	247.303	192.598
Line graph	243.64	209.17
Spark Area graph	341.27	465.344

From the mean and the standard deviation of the time taken, participants performed faster in colored slope graph treatment.

6.3 Experiment 2

6.3.1 Data Processing

The participants' sketch was superimposed on the ground truth i.e. the exact line graph for the time-series data. Each ground truth has 30 data points. A python script was written to extract the pixel points of the sketch and the ground truth at 50 time-steps which means that the 30 data points are linearly mapped to the 50 data points. As the participants are not given a reference temporal grid, it was necessary to map the values to smaller intervals with more time steps. Once the pixels were extracted, the difference between the sketch pixel value and ground truth pixel value was calculated. For each of this difference at 50 time-steps, statistical inferences such as Mean, Median, and the standard deviation was calculated. Using the statistical inferences, the line curve was recreated using tableau.

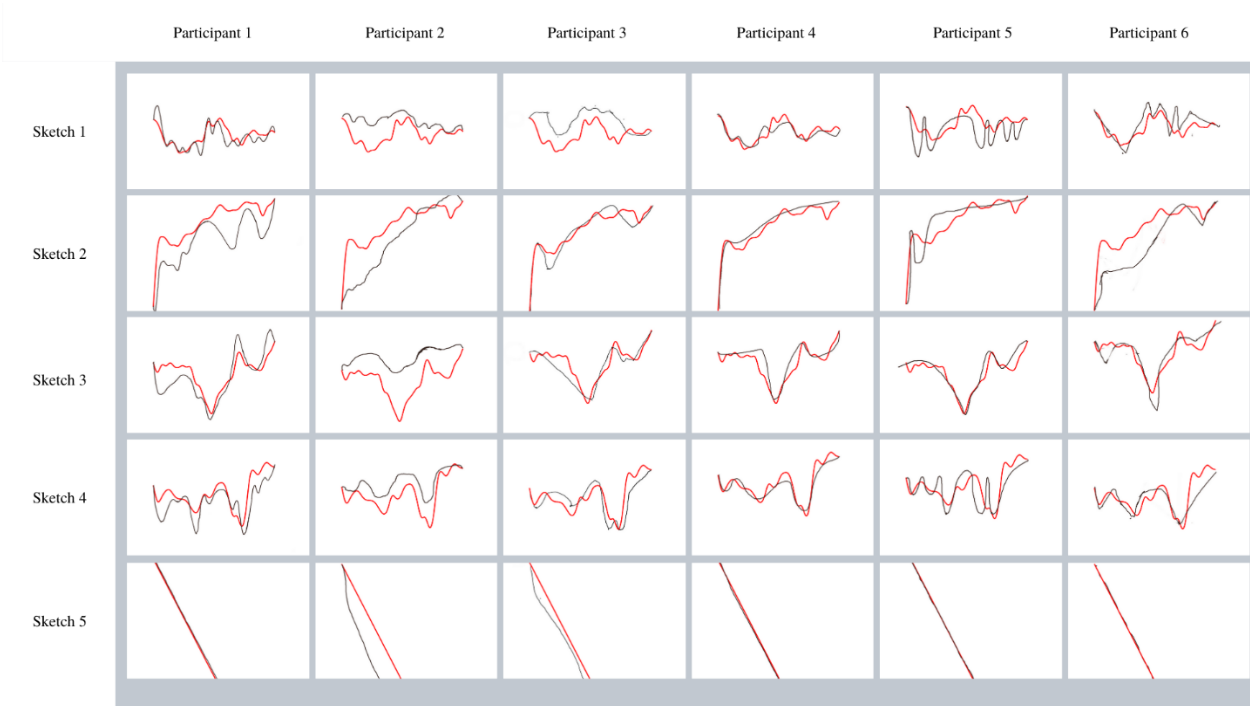


Figure 6.6: Participants' sketches from experiment 2.

6.3.2 Data Analysis

How to analyze the results:

After the analysis of the data, the 5-quartile number summary along 30 temporal points are computed and plotted using a box plot. The 30 median values are connect using a blue line and the red line represents the ground truth value.

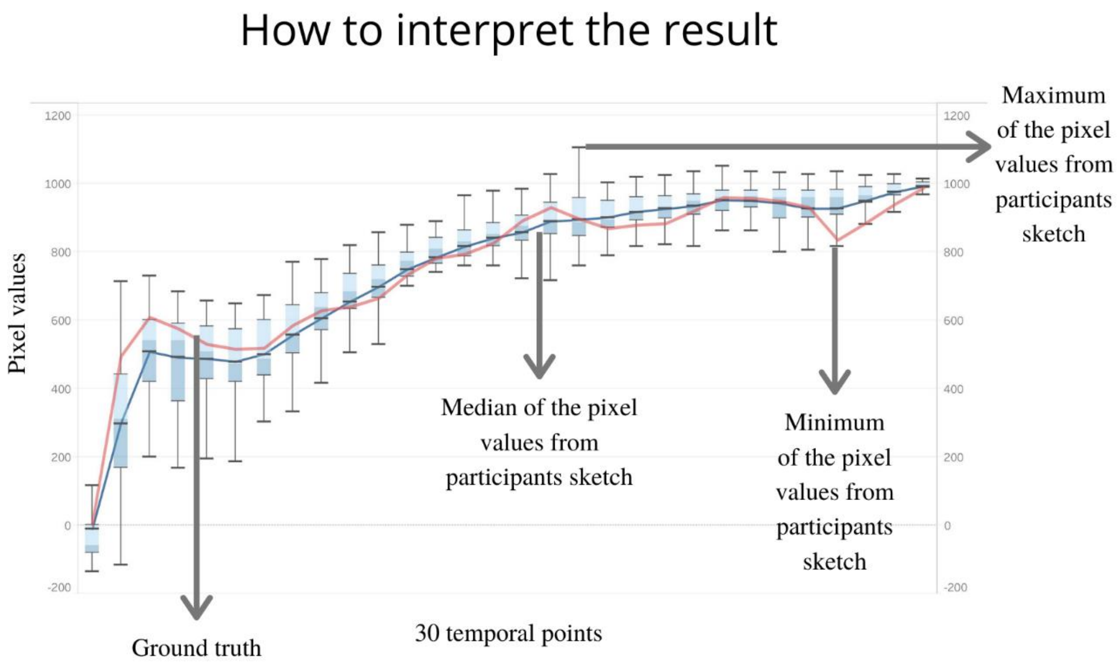


Figure 6.7: How to interpret results from experiment 2.

For each sketch, the pixel values of the ground truth and the participants' sketch are extracted at 30 temporal points. From all the 35 participants' pixel points at each temporal point, statistical inferences mean, maximum and minimum are calculated. This gives 30 means, 30 maximum, 30 minimum values for each task. From the extracted statistical inferences, at each temporal point, a boxplot is plotted which visualizes the mean, maximum value and minimum value. The 30 box plots are connected by a blue line along with the median value of the sketched. The red line illustrates the value of the ground truth. Below are the results of experiment 2.

Result 1

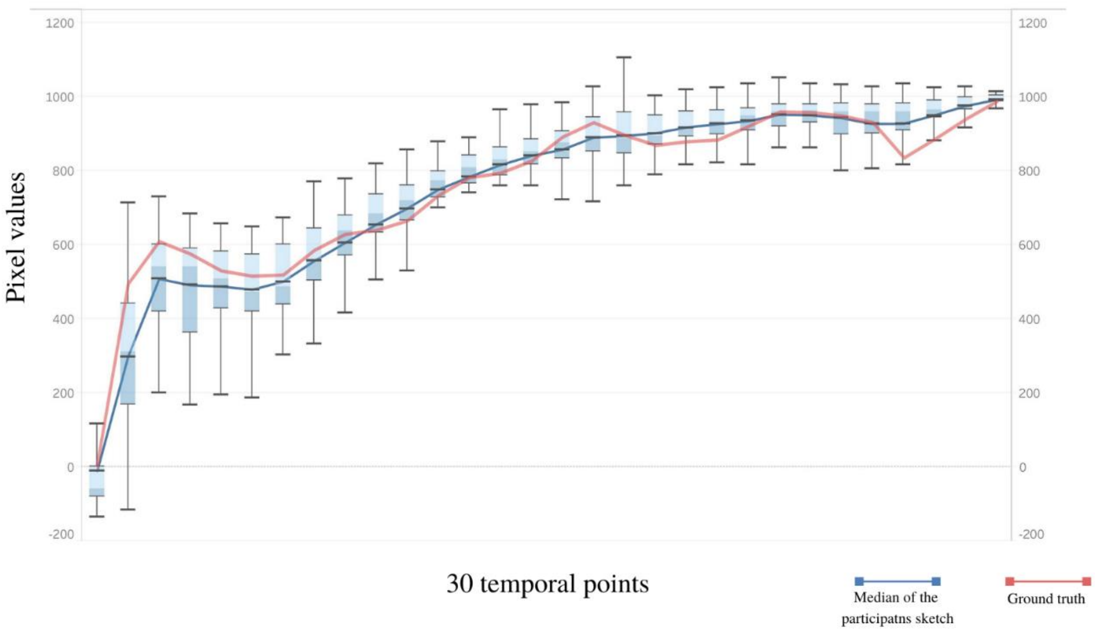


Figure 6.8: Result 1 of sketch 1.

Result 2

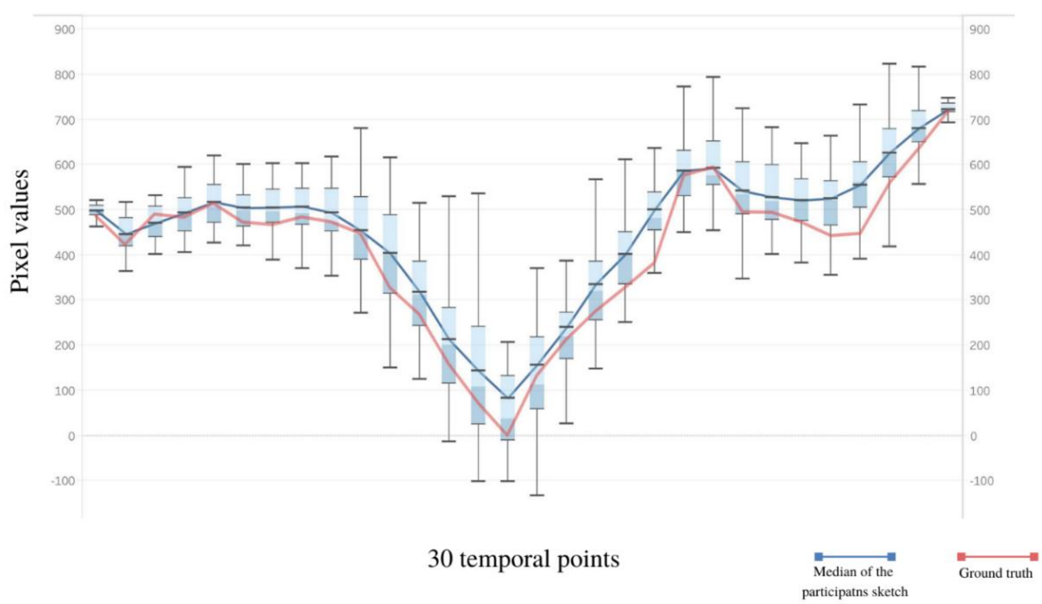


Figure 6.9: Result 2 of sketch 2.

Result 3

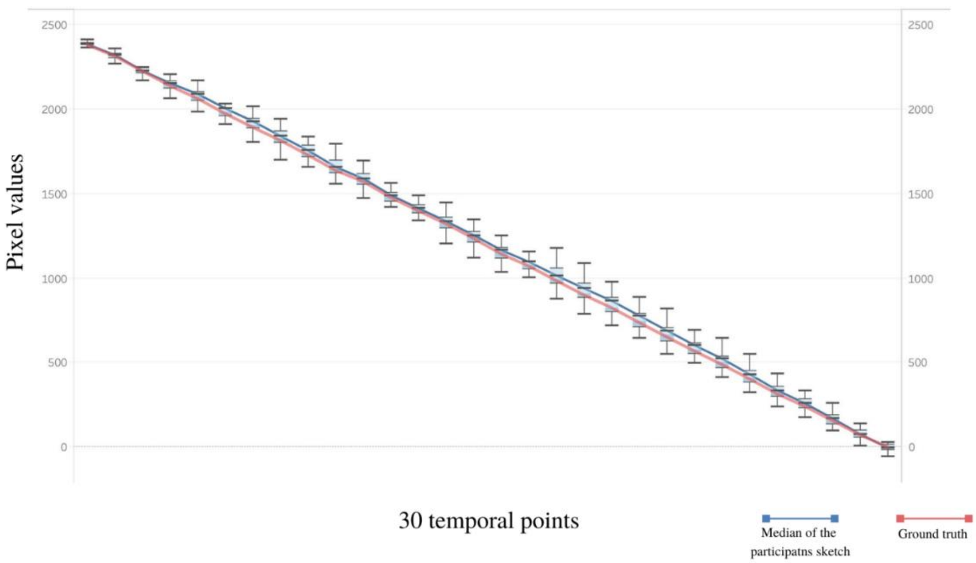


Figure 6.10: Result 3 of sketch 3.

Result 4

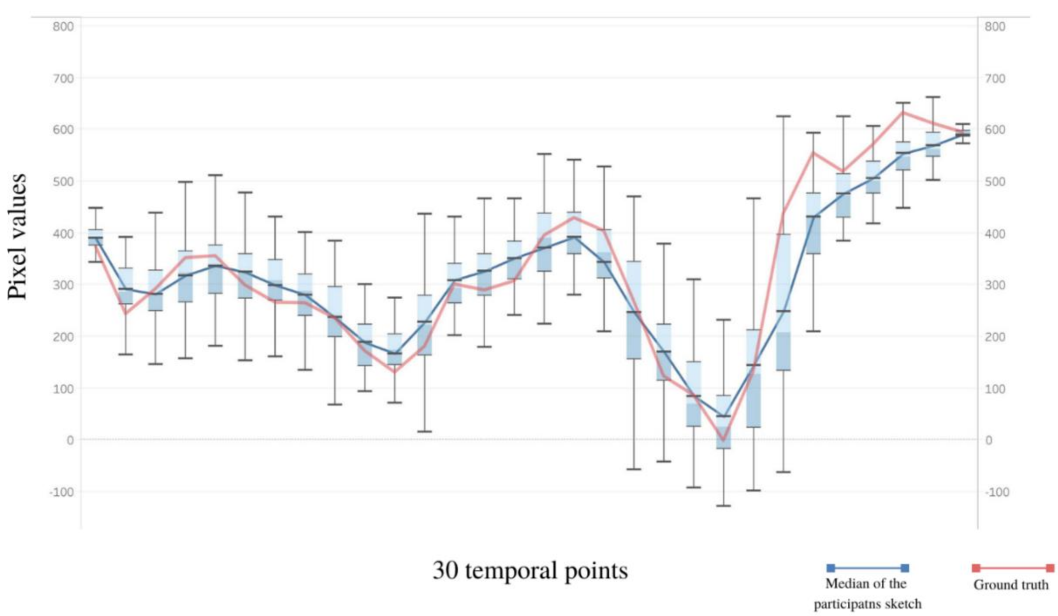


Figure 6. 11: Result 4 of sketch 4.

6.4 Survey

From the participants' survey, the percentage of participants who preferred colored slope visualization and an average score of colored slope visualization was calculated.

1. Among all the visuals pick the visual which you prefer for temporal Information.

   For this question, 54.3% of the participants picked colored slope graph visualization as their preference for visualization for temporal information. As the percentage is more than half, it can be concluded that colored slope visualization can be used for temporal information.

2. Among all the visuals, pick the visual which the easiest to understand the information.

   For this question, 53 % of the participants picked colored slope graph visualization as being the easiest visualization to understand. As the percentage is more than half, it can be concluded that colored slope visualization is a simple and easy visualization to understand.

3. Among all the visuals, pick the visual which was the hardest to understand the information.

   For this question, 41.2 % of the participants picked line graph visualization as being the hardest visualization to understand. It can be concluded that line graph visualization hard to understand when multiple time-series curves are visualized.

4. Among all the visuals, pick the visual which the hardest to perceive the information.

   For this question, 40.0 % of the participants picked line graph visualization as being the hardest visualization to perceive information.

5. Among all the visuals, pick the visual which was the messiest or the visual which was very frustrating for you to understand

   For this question, 51.4 % of the participants picked line graph visualization as being the messiest or the visual which was very frustrating to understand, 22.9% of the participants felt spark area was the messiest, 20% of the participants felt heatmap was the messiest and 5.7% of the participants felt the colored slope graph was the messiest. It can be concluded that line graph visualization is the messiest of the visualizations when multiple time-series curves are visualized.

6. Rank the effectiveness of finding and comparing similar items for each visual.

   The average rank of colored slope visualization for the effectiveness of finding and comparing similar items was 2.

7. Rank the effectiveness of finding the stocks with a dramatic increase/decrease.

The average rank of colored slope visualization for the effectiveness of the stocks with a dramatic increase/decrease was 1.

8. Rank the effectiveness of finding the stocks with high variance.

   The average rank of colored slope visualization for the effectiveness of finding the stocks with high variance was 1.

9. Rank the effectiveness of finding extreme(maximum) values.

   The average rank of colored slope visualization for the effectiveness of finding extreme(maximum) values was 2.

10. Rank the intuitiveness of the visualization design representation.

    The average rank of colored slope visualization for the intuitiveness of the visualization design representation was 2.

11. Rank the aesthetics of the visualization design.

    The average rank of colored slope visualization for the aesthetics of visual design was 2.

12. Rank the visualizations learnability.

    The average rank of colored slope visualization for the visualization learnability was 2.

From the percentages of participants picking colored slope visualization for temporal information as to their choice and from the average rank of colored slope graph for identifying several patterns, it can be concluded that the participants were comfortable working with colored slope visualizations.

6.5 Summary

This chapter outlines the data processing, data analysis, results from the experiments, how to and what has been interpreted from the experiment results.

CHAPTER 7: DISCUSSIONS & CONCLUSION

This chapter provides the discussions, conclusion of the research study along with the future work from the research. The applications, as well as future work, are mentioned in detail.

7.1 Discussion

The key contribution of this research study is the design technique for visualizing multiple time-series data simultaneously. The first phase of the research study focused on the review of the literature to identify the gaps in the predominant time-series visualizations. The literature on graphical perception of color, slope, size, and position of the visualization was also studied. The main bottleneck of visual clutter was identified. Visual clutter is an issue when multiple time-series curves are visualized simultaneously. The predominant visualizations such as the traditional line graph, spark area perform well when the number of line graphs is less than 10. However, once the range of the number of time-series data increases, the traditional techniques fail to perform. The visual clutter makes it difficult to perceive the information visualized. From the literature, it was also noted that users are good at estimating the averages and summary comparisons when a color is a form of visual encoding. After the identification of the issue of visual clutter, a novel visualization technique called colored slope graph visualization was proposed, which was inspired by the existing slope graph. A slope graph is very similar to a line graph where the change in the non-temporal attributes is visualized over a period. The temporal attributes are represented by the x-axis, and the non-temporal attributes are represented by the y-axis, respectively. The first y-axis is the starting value of the stock. The second y-axis is the last data point in a period. Unlike a line graph, a slope graph considers the change only between the two temporal points. Usually, a slope graph visualizes the trend between the temporal points, which helps in handling the overplotting is a visualization. The data points which are in between the selected temporal points are not considered. However, this leads to the loss of information in the visualization. So, in the colored slope graph visualization, in each slope line, the intermediate temporal points were encoded based on a linear color scale, which was developed using a 9 color scale. As this visualization is aimed at visualizing multiple time-series data, a single hue is not the right option to use. It is difficult to identify the estimate with a single hue when the range of data is huge. In this study, the cooler color was used to represent a higher value, and the warmer cooler was used to represent the lower values because stock market data was used as the primary dataset for the visualization. In the financial industry, red represents a decline in the value, and blue represents growth. The algorithm was developed using d3.js.

A series of quantitative experiments were carried out to evaluate the colored slope graph. Line graph, a heat map, and spark area, which were identified from the literature as the predominant visualization was used as a benchmark for the quantitative experiments. This experiment aimed at testing the performance of the colored slope graph when multiple time-series data were visualized simultaneously. As 30 stocks is a large range of time-series curves, this study visualized 30 stocks simultaneously. The results and the findings of this study hold valid only for an optimally large range of time-series curves. A total of seven patterns were tested, the patterns being 1. Identifying maximum value, 2. Identifying Minimum value 3. Identifying high variability 4. Identifying covariance 5. Identifying the ratio of percentage increase to percentage decrease 6. Identifying stocks that increase the most. 7. Identifying stocks that decrease the most. These questions were aimed at testing the seven-hypothesis framed by taking several keynotes from the literature. A quantitative methodology was adopted to identify how effective the graphical perception of the colored slope graph was when multiple time-series data was visualized. From the evaluation, several findings were identified.

7.2 General Findings

The results of this research study hold only for an optimally large number of time-series data visualized simultaneously. From this research study, several findings could be concluded in the study. The colored slope graph has both pros and cons. From the first experiment, it was found that colored slope graph visualization is good at finding the extreme (minimum) value, finding the extreme - maximum value. From the work of Saket, Endert , & Demiralp, (2018), it was mentioned that the line graph is usually not the right visualization to find extreme values. Also, the graphical perception color helps users estimating the value of a data point. Color combined with slope gives an intuitive comparison of change. The slope lines remove the visual cluttered by removing visual detail. From the experiments, we also found that the Colored slope graph is good at identifying the general trend of data. Straight lines make it easy to understand the rate of change. The steeper the line, the higher the change, which signifies a higher rate of change. The Colored slope graph is also good at identifying the coverability between two time-series curves. However, Colored slope graph is not the ideal visualization for identifying variability. Identifying variability is a cognitive intensive task which requires, complete attention that too in a less cluttered display.

7.2.1 Research question 1

From the experiment 1, there were two limitations of the colored slope graph that were identified. The colored slope graph may not be the right visualization for identifying the high variability among the times- series data. However, it was noted that none of the visualizations could perform well for identifying high variability in time-series data when multiple time-series curves are visualized as well. The other limitation of the colored slope graph is that there was no significant difference in the performance of the colored slope graph for identifying the ratio of percentage increase to decrease in stocks from Jan 1st to Jan 30th. However, it should also be noted that none of the visualizations could perform well for identifying the ratio as well.

7.2.2 Research question 2

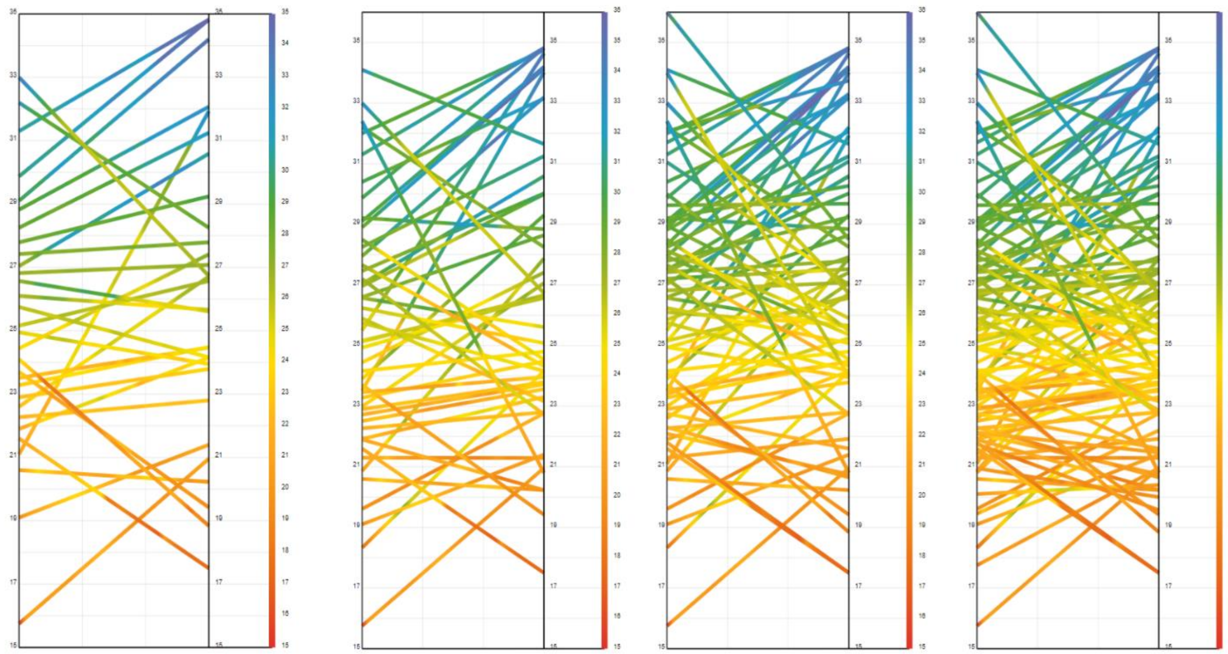
From the second experiment, it was found that participants were correctly able to perceive the information color encoded on the slope lines. From the literature and from the experiments, it can be established that users are correctly able to identify the estimate of the data represented as color encodings on the slope line.

7.2.3 Research question 3

From the survey, it was found that the participants were comfortable working with the colored slope graph visualization. Most participants found the visualization easy to perceive and understand. Most users also stated that it was an aesthetic design. Participants also stated that the line graph was the messiest and frustrating to understand. This survey is more of an understanding of the subjective viewpoint of the participants who participated in the research study. These results cannot be interpolated to the entire population. The results of this survey hold true only for the participants who participated in the study

7.2 Limitations

This research study is limited to an optimally large number of visual marks at which the visualization tends to get visually cluttered. The results of this study will not hold valid for the number of visual marks less than 10. As per the research study by Saket, Endert, & Demiralp, (2018), visual marks less than 50 is an optimal level of numerous visual marks in the visualization at which one can produced visualizations. Users will still be able to perceive the information graphically. Once the visual marks exceed 50, users will not be able to read any visualizations as the visuals are too cluttered. In Figure7.1, it is demonstrated the fact that while the volume increases, the visual clutter problem will still naturally appear. When Colored slope was applied for 30 data points (See Figure 7.1 (a)), 50 data points (See Figure 7.1 (b)), 80 data points (See Figure 7.1 (c)) and 120 data points (See Figure 7.1 (d)), after 50 data points, even the Colored slope graph becomes visually cluttered, which guides back to the work demonstrated by Saket et al. \cite{task-based-effectiveness}. Linked this observation with study, it is obvious that Colored slope graph could perform well when we apply it to an optimally large number of visual marks, taking 30 as an example. However, once we scale the visual marks to more than 50, the cluttering issue re-emerges. As claimed previously, this study was aimed to investigate the capacity of Colored slope. Therefore, it should be noticed that if we visualize a small batch of time series data, like 5 or 10, Colored slope may not dominate the performance compared to other visualizations. Like integrating small multiples with small number of spark area charts, the users can easily eye scan the graphs thoroughly.



(a) 30 data points (b) 50 data points (c) 80 Data points (d) 120 Data points

Figure 7. 1: Colored Slope with different number of lines

7.2 Applications

The main advantage of using colored slope visualization is that one can visualize multiple time-series data simultaneously. One of the major applications of color slope visualization can be applied in the financial industry. Also, colored slope visualization techniques can be applied to several other time-series data such as weather and ranking data. The colored slope visualization can also be a good replacement for heatmap or horizon graphs.

7.3 Future Work

A key research concept that can be studied is the number of taime-series curves visualized simultaneously will result in a visual clutter. In this research study, 30 stocks were visualized simultaneously, which gives a cluttered visualization when applied traditional techniques are used. For identifying the threshold, several quantitative studies can be implemented to segregate the range of time-series curves that can be visualized simultaneously. The second research concept that can be extensively studied is the study of variability in time-series data. From this research study, it was found that there was no significant difference in the performance of the four visualizations in identifying the time-series data with high variability. This opens several research questions that can be further explored. Human subject experiments on how users identify high variability in the data when the range of visual marks varies. Other major future works for this research could be an extensive study of the graphical perception of single and multiple Colored Slope graphs. Also, work on the graphical perception of heatmap can be studied to compare and understand how slope and color can impact how users perceive the information.

REFERENCES

Aigner, W., Miksch, S., Müller, W., & Schumann, H. (2007). Visualizing time-oriented data-a systematic view. *Computers & Graphics*.

Alsakran, J., Zhao, Y., & Zhao, X. (2010). Tile-based parallel coordinates and its application in financial. *International Society for Optics and Photonics*, 753003.

beammeupscoty. (2019). Retrieved from https://beammeupscoty.wordpress.com/

Bertin, M. J. (1967). *Semiologie graphique gauthier villars.* Paris.

Chernoff, H. (1973). The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association*,, 361-368.

Cleveland , W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 828-33.

Cleveland, W. S. (1993). A model for studying display methods of statistical graphics. *Journal of Computational and Graphical Statistics*, 323– 343.

Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of American Statistical Association*, 531-554.

Cleveland, W. S., Diaconis, P., & McGill, R. (1982). Variables on Scatterplots Look More Highly Correlated When the Scales are Increased. *Science*, (pp. 1138-1141).

Color Brewer 2.0. (n.d.). Retrieved from Color Brewer 2.0: http://colorbrewer2.org/#type=diverging&scheme=Spectral&n=11

Correll, M., Albers, D., Franconeri, S., & Gleicher, M. (2012). Comparing averages in time series data. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 1095–1104).

Croxton, F. E., & Stryker, R. E. (1927). Bar Charts versus Circle Diagrams. *Journal of the American Statistical Association*.

Dao, H. T., Bazinet , A., Berthier , R., & Shneiderman , B. (2008). NASDAQ Velocity and Forces: An Interactive. *Dao, H. T., Bazinet, A. L., Berthier, R., & Shneiderman, B. (2008). NASDAQ Velocity and Forces: An Interactive Visualization of Activity and Change. J. UCS, 14(9), 1391-1410.*

Edwards, R. D., Magee, J., & Bassetti, W. H. (2018). *Technical analysis of stock trends. CRC press.* CRC Press.

Elmqvist, N., Dragicevic, P., & Fekete, J.-D. (2008). Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. *Visualization and Computer Graphics* (pp. 1539-114). IEEE.

Euclidean distance. (2020, March 28). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Euclidean_distance

Few, S. (2009). *Now you see it: simple visualization techniques for quantitative analysis.* Analytics Press.

Furnas, G. W. (1986). Generalized fisheye views. *CHI '86 Conference Proceedings*, (pp. 16-23).

Glassner, A. (1994). *Principles of Digital Image Synthesis.* Elsevier Wordmark.

Havre, S., Hetzler, B., & Nowell, L. (2000). ThemeRiver: visualizing theme changes over time. *Information Visualization (INFOVIS)* (pp. 115-123). IEEE.

Heer, J., Kong, N., & Agrawala, M. (2009). Sizing the Horizon: The Effects of Chart Size and Layering. *ACM Human Factors in Computing Systems (CHI).*

Hernandez, J., Paredes, P., Roseway, A., & Czerwinski, M. (2014). Under pressure: sensing stress of computer users. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 51-60).

Huang, W., & Eades , P. (2005). How people read graphs. *Asia-Pacific symposium on Information visualisation*, (pp. 51-58). Sydney.

Inselberg, A. (1985). The plane with parallel coordinates. *Springer*, 69-91.

Javed, W., McDonnel, B., & Elmqvist, N. (2010). Graphical Perception of Multiple Time Series. *IEEE Transactions on Visualization and Computer Graphics*, 927 - 934.

Kaggle. (2019). Retrieved from Kaggle: www.kaggle.com

Kaggle Wikipedia. (2019). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Kaggle

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 1-8.

Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the Information Age: Solving Problems with Visual Analytics,.* Eurographics Association.

Keogh, E., Hochheise, H., & Shneiderman, B. (2002). An Augmented Visual Query Mechanism for Finding Patterns in Time Series Data. *International Conference on Flexible Query Answering Systems* (pp. 240-250). Berlin, Heidelberg: Springer.

Ko, S., Cho, I., Afzal, S., Yau, C., Chae, J., Malik, A., . . . Ebert, D. S. (2016). A Survey on Visual Analysis Approaches for Financial Data. *Computer Graphics Forum*, 599-617.

Koffka, K. (2014). *Principles Of Gestalt Psychology.* Mimesis International.

Kumar, N., Lolla, V., Keogh, E., Lonardi, S., & Ratanamahatana, C. (2005). Time-series Bitmaps: a Practical Visualization Tool for Working with Large Time Series Databases. *Data Mining Conference.*

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Lin, L., Cao, L., & Zhang, C. (2005). The fish-eye visualization of foreign currency exchange data streams. *Asia-Pacific symposium on Information visualisation*, (pp. 91-96).

M, K., E, B., & DA, K. (2011). CloudLines: compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics*.

Measuring Dis/Similarities Between Objects (Cells) In 'n'-Dimensional Space. (2020, March 28). Retrieved from https://hlab.stanford.edu/: https://hlab.stanford.edu/brian/making_measurements.html

MiilIer, W., & Schumann, H. (2003). VISUALIZATION METHODS FOR TIME-DEPENDENT DATA - AN OVERVIEW . *Proceedings of the Winter Simulation Conference*, (pp. 737-745).

Murphy, J. J. (1999). *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications.* New York: Prentice Hall Press.

Peterson , L. V., & Schramm, W. (1954). How accurately are different kinds of graphs read? 178-189.

Saket, B., Endert , A., & Demiralp, Ç. (2018). Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 2505 - 2512.

Simkin , D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 454-467.

Sorenson, E., & Brath, R. (2013). Financial Visualization Case Study: Correlating Financial Timeseries and Discrete. *17th International Conference on Information Visualisation* (pp. 232-238). IEEE.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, 61-77.

Talbot, J., Gerth, J. A., & Hanrahan, P. (2012). An Empirical Model of Slope Ratio Comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 2613–2620.

Time Series Analysis and Forecasting. (2013). Retrieved from http://itfeature.com/time-series-analysis-and-forecasting/time-series-analysis-forecastin

TLX. (2019, October 20). Retrieved from NASA Ames - Home: https://humansystems.arc.nasa.gov/groups/TLX/

Tufte, E. R. (2006). *Beautiful Evidence.* Graphics Press.

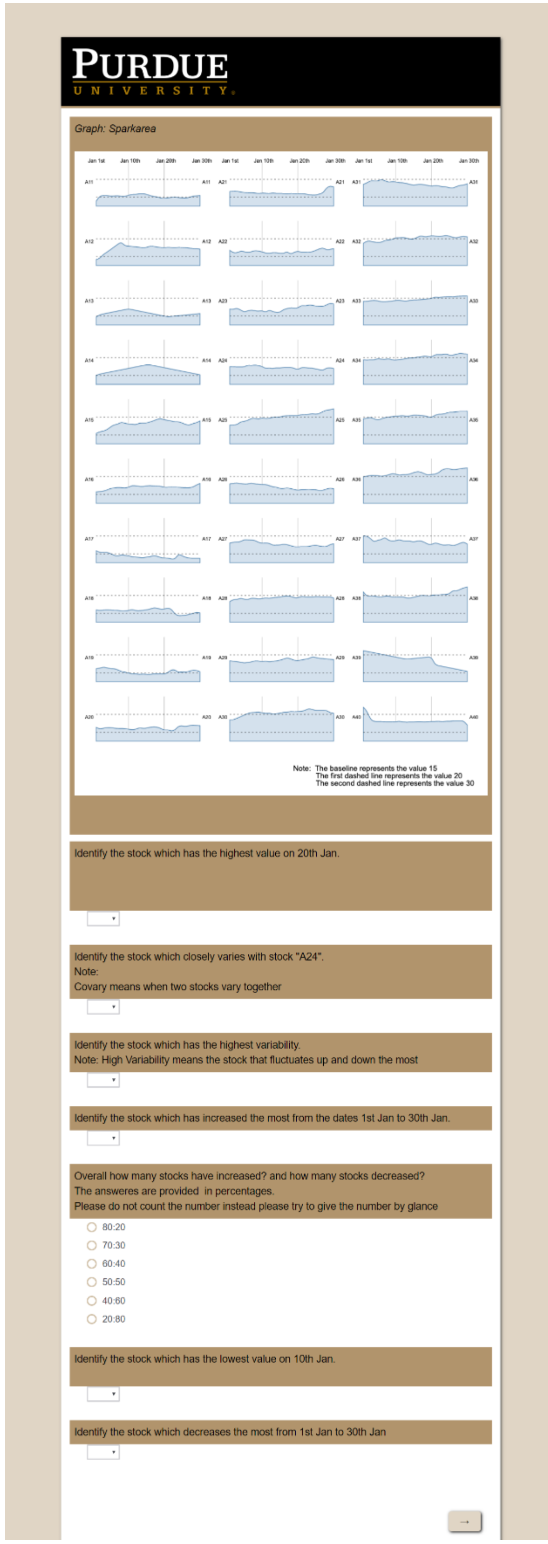
Tufte, Edward R. (1983). *The Visual Display of Quantitative Information.* Graphics Press.

Ware, C. (2012). *Information Visualization: Perception for Design.* San Francisco: Morgan Kaufmann Publishers Inc.

Wattenberg, M. (1999). Visualizing the Stock Market . *ACM*, 188-189.

Zhao, J., Liu, X., Guo, C., Qian, Z. C., & Chen, Y. V. (2019). Phoenixmap: An Abstract Approach to Visualize 2D. *IEEE Transactions on Visualization and Computer Graphics*.

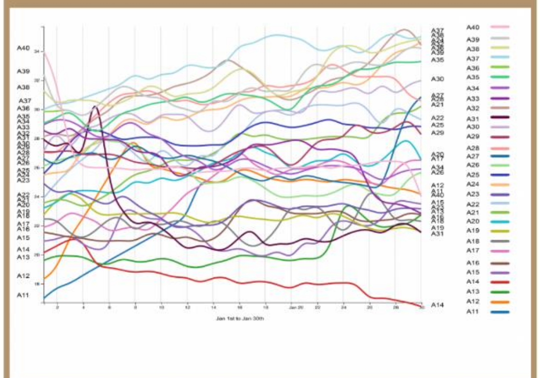
APPENDIX A. SPARK AREA GRAPH DATASET 1



APPENDIX B. LINE GRAPH DATASET 2

PURDUE
UNIVERSITY

Graph: Line Graph



Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A20".
Note:
Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

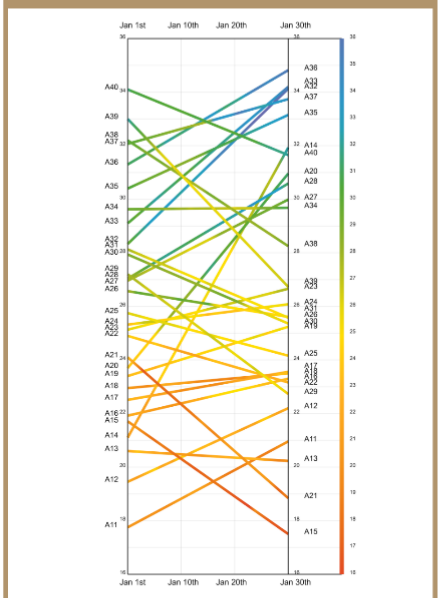
Identify the stock which decreases the most Jan 1st to Jan 30th

→

APPENDIX C. COLORED SLOPE GRAPH DATASET 3

PURDUE UNIVERSITY

Graph: Colored Slope Graph



Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A38".
Note:
Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

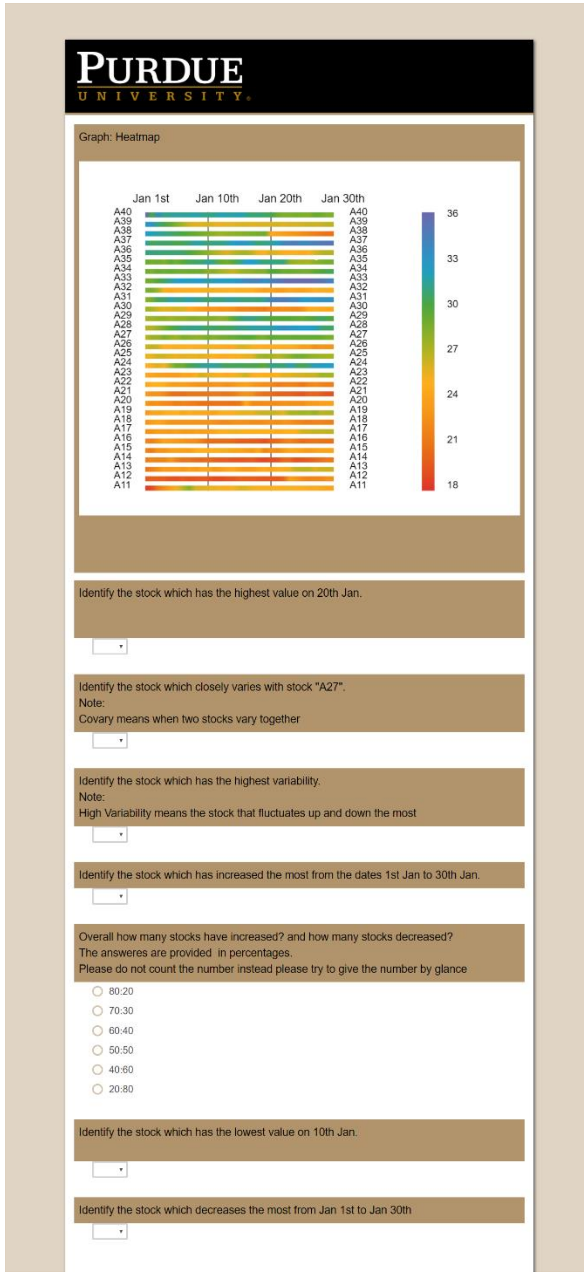
Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from Jan 1st to Jan 30th

APPENDIX D. HEATMAP DATASET 4



PURDUE UNIVERSITY

Graph: Heatmap

Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A27".
Note:
Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

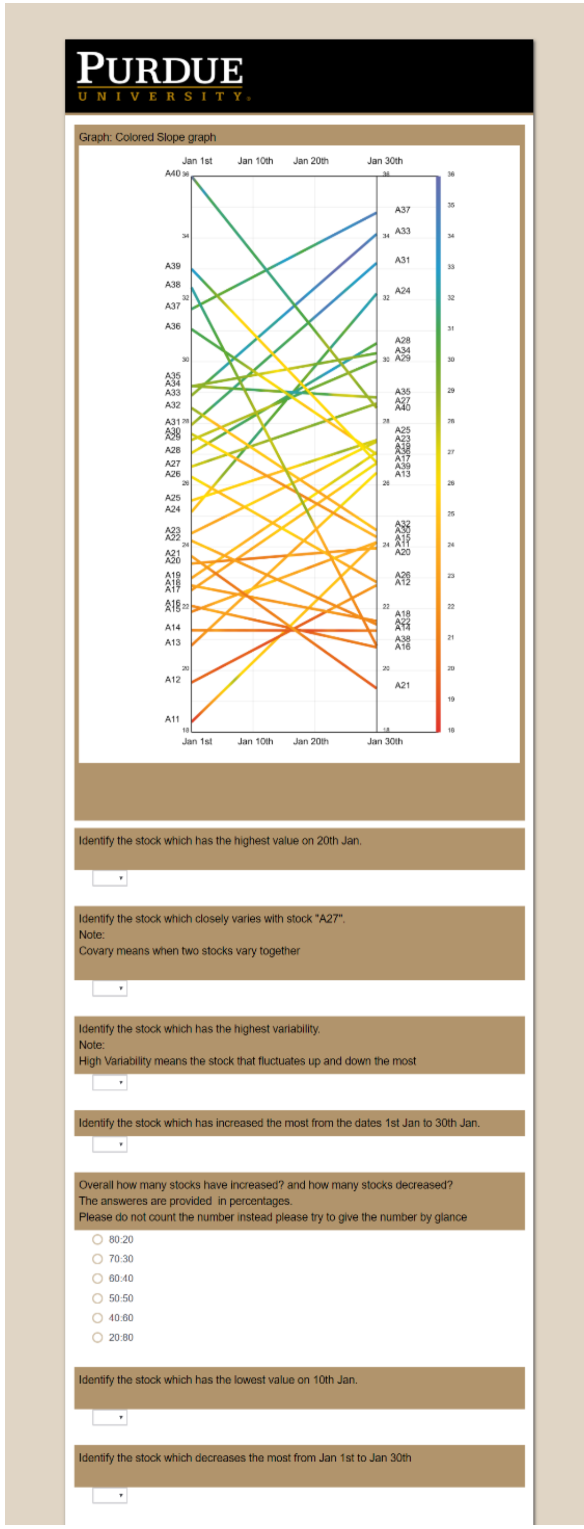
Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

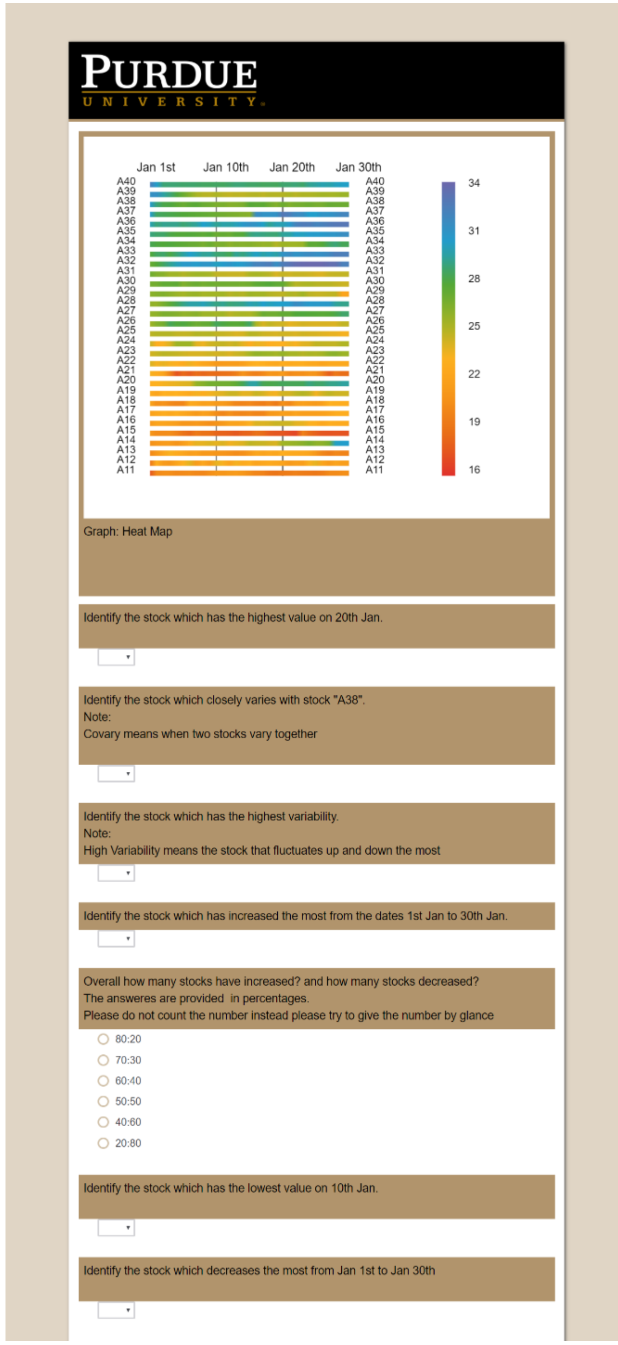
Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from Jan 1st to Jan 30th

APPENDIX E. COLORED SLOPE GRAPH DATASET 1



APPENDIX F. HEATMAP DATASET 2



Graph: Heat Map

Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A38".
Note:
Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

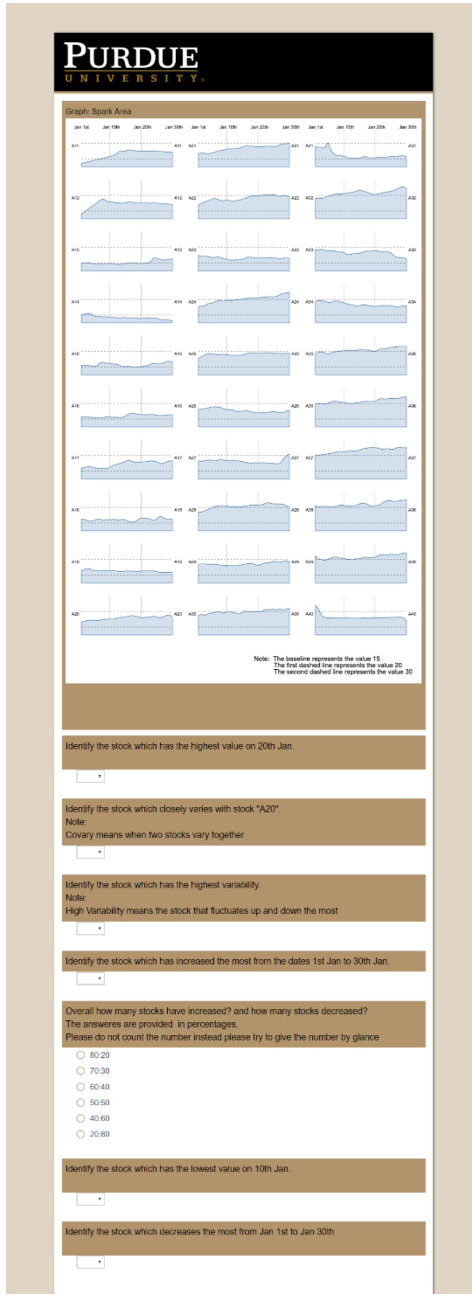
Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

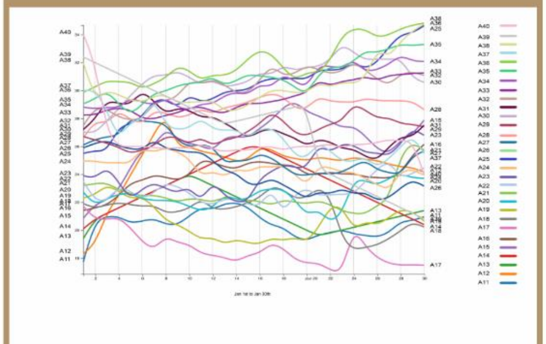
Identify the stock which decreases the most from Jan 1st to Jan 30th

APPENDIX G. SPARK AREA DATASET 3



APPENDIX H. LINE GRAPH DATASET 4

Graph: Line graph



Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A38".
Note: Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

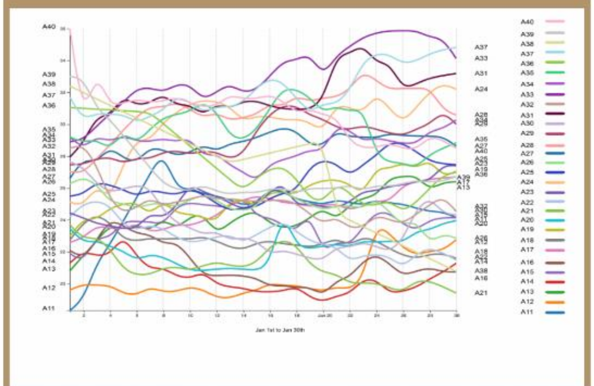
Identify the stock which decreases the most from 1st Jan to 30th Jan

→

APPENDIX I. LINE GRAPH DATASET 1

PURDUE
UNIVERSITY

Graph: Line graph



Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A27".
Note:
Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

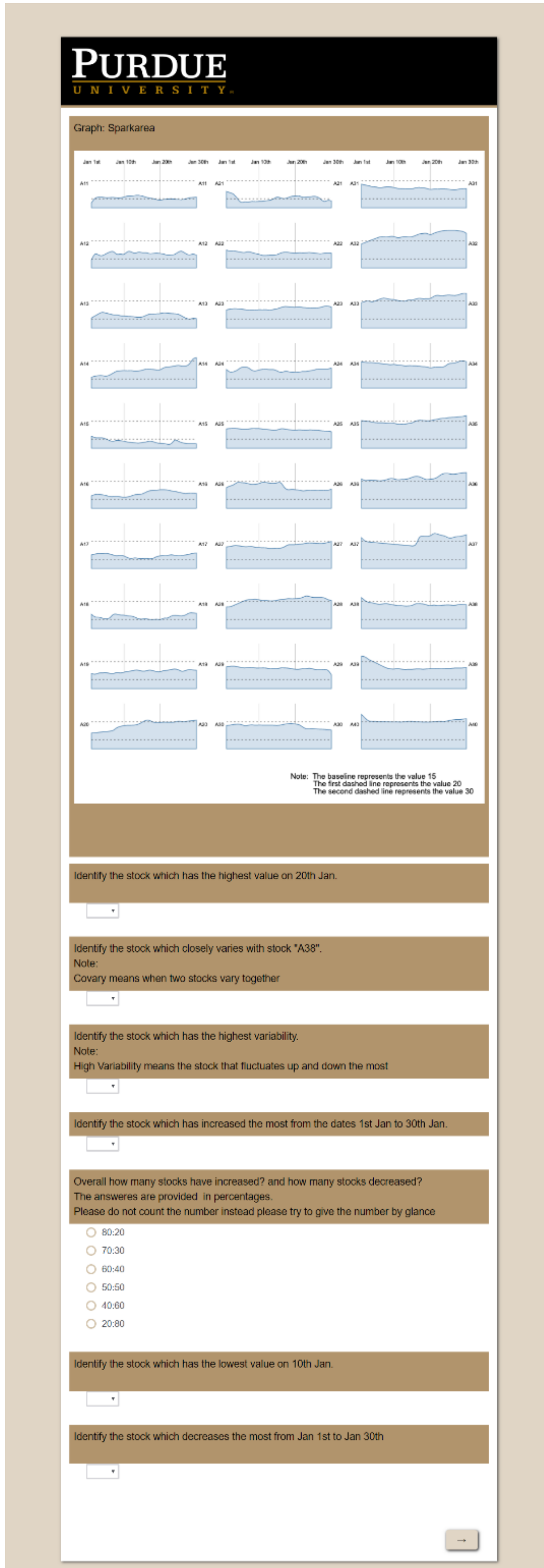
Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

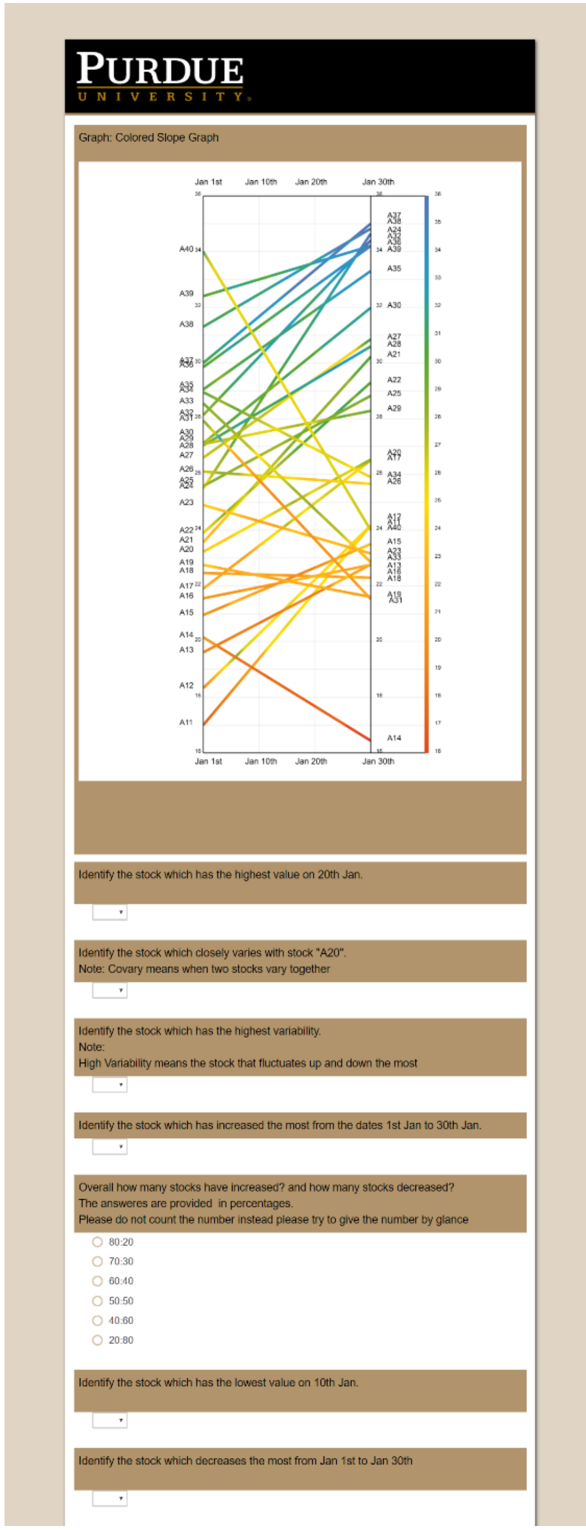
Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from 1st Jan to 30th Jan

APPENDIX J. SPARK AREA DATASET 2



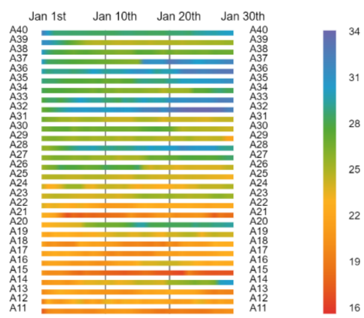
APPENDIX K. COLORED SLOPE DATASET 4



APPENDIX L. HEATMAP DATASET 3

PURDUE
UNIVERSITY

Graph: Heat Map



Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A20".
Note: Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

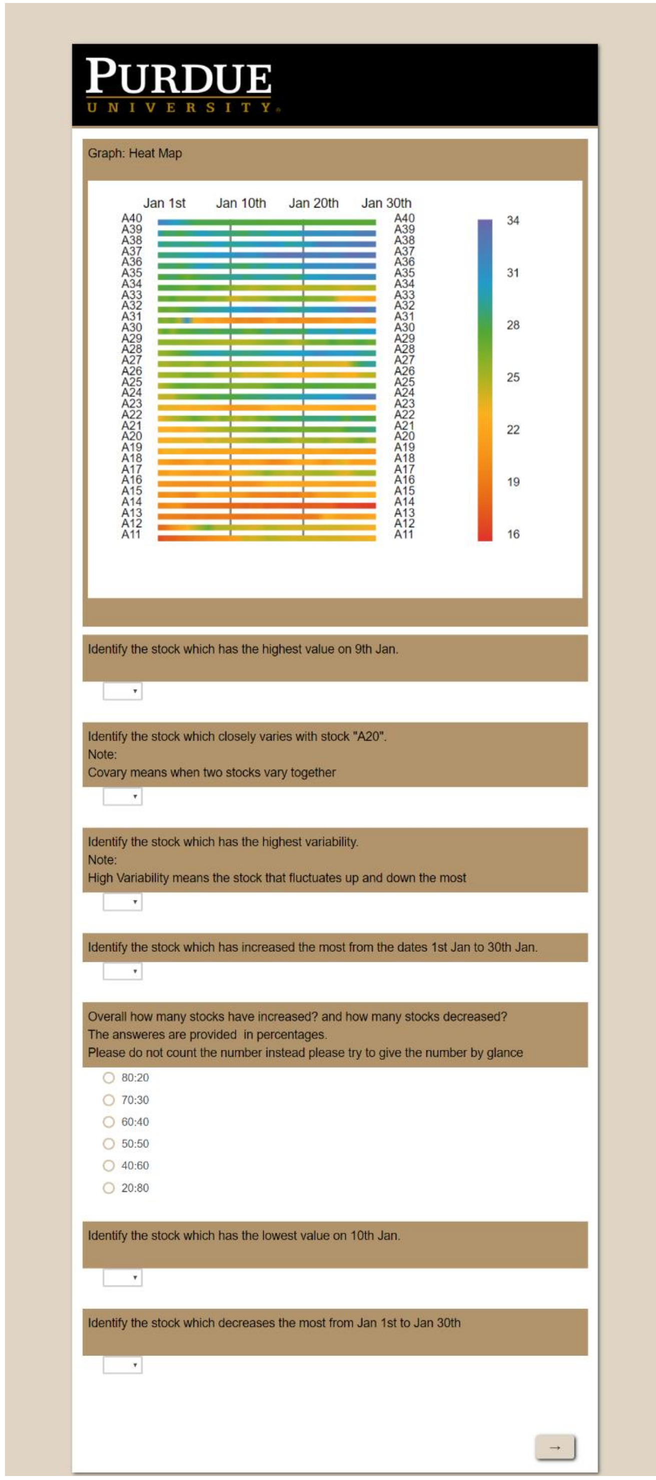
Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from 1st Jan to 30th Jan

→

Powered by Qualtrics

APPENDIX M. HEATMAP DATASET 1



Graph: Heat Map

Identify the stock which has the highest value on 9th Jan.

Identify the stock which closely varies with stock "A20".
Note:
Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

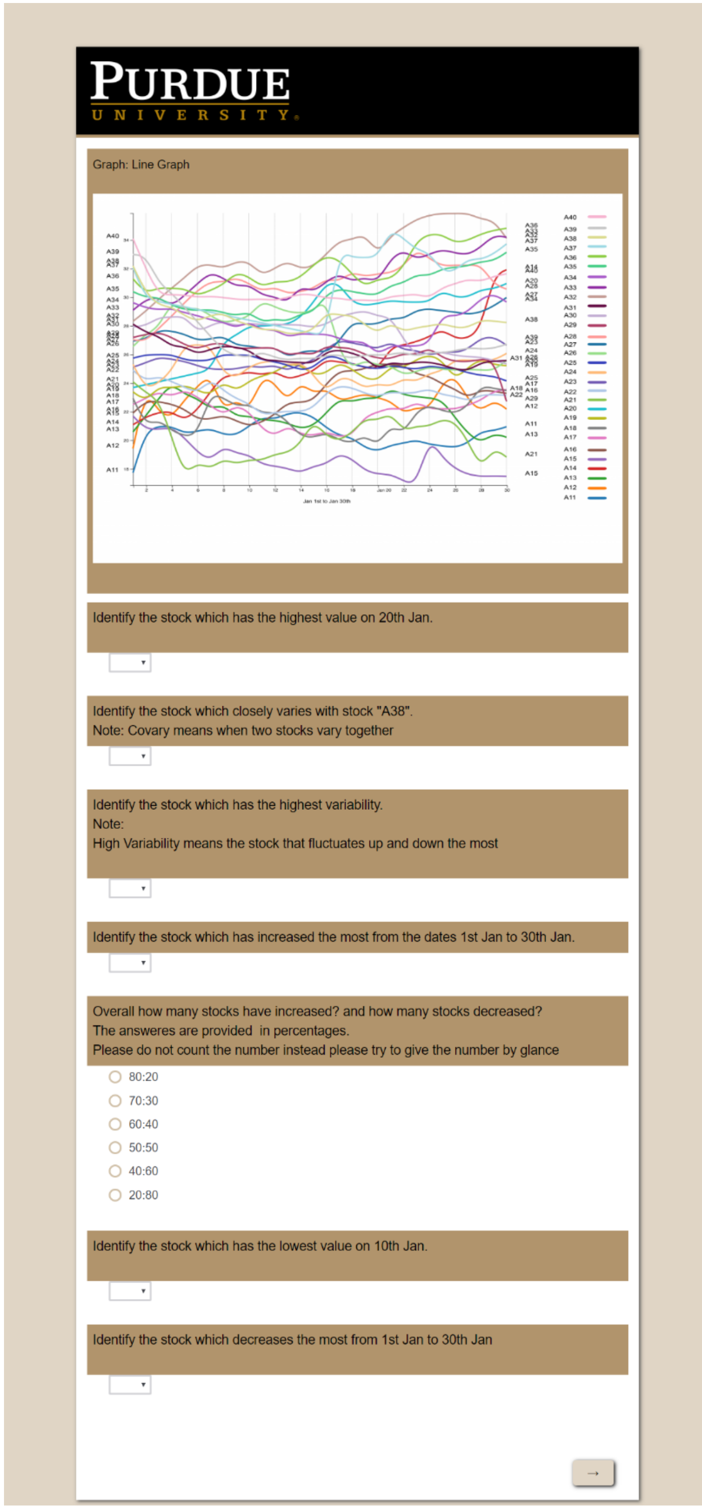
- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from Jan 1st to Jan 30th

→

APPENDIX N. LINE GRAPH DATASET 3



PURDUE UNIVERSITY

Graph: Line Graph

Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A38".
Note: Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

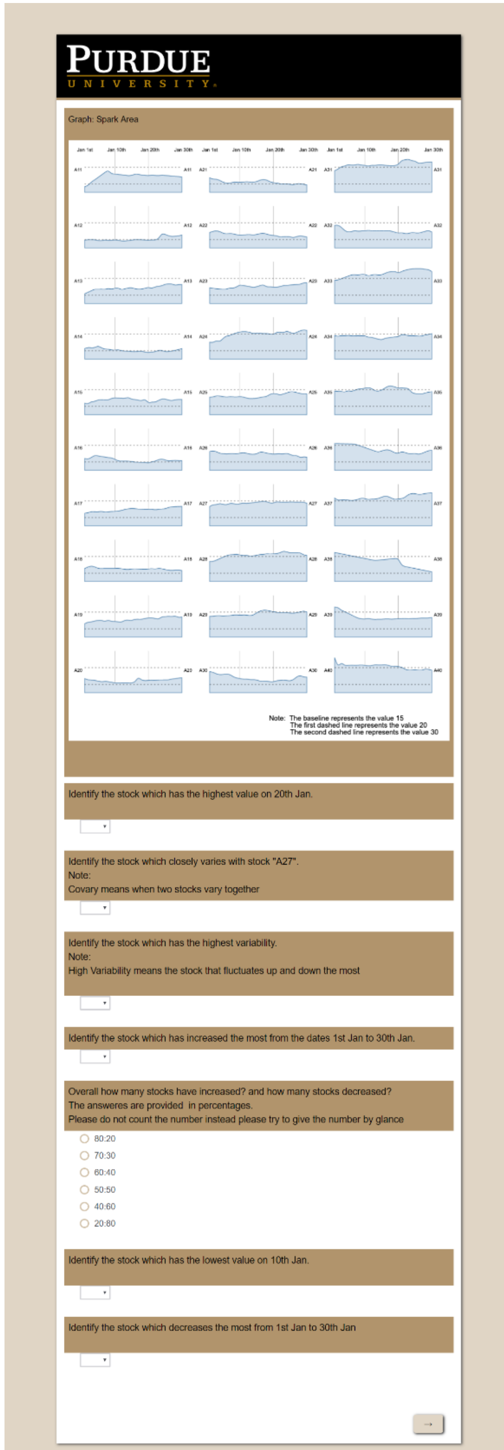
Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from 1st Jan to 30th Jan

APPENDIX O. SPARK AREA DATASET 4



Graph: Spark Area

Note: The baseline represents the value 15
The first dashed line represents the value 20
The second dashed line represents the value 30

Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A27".
Note:
Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

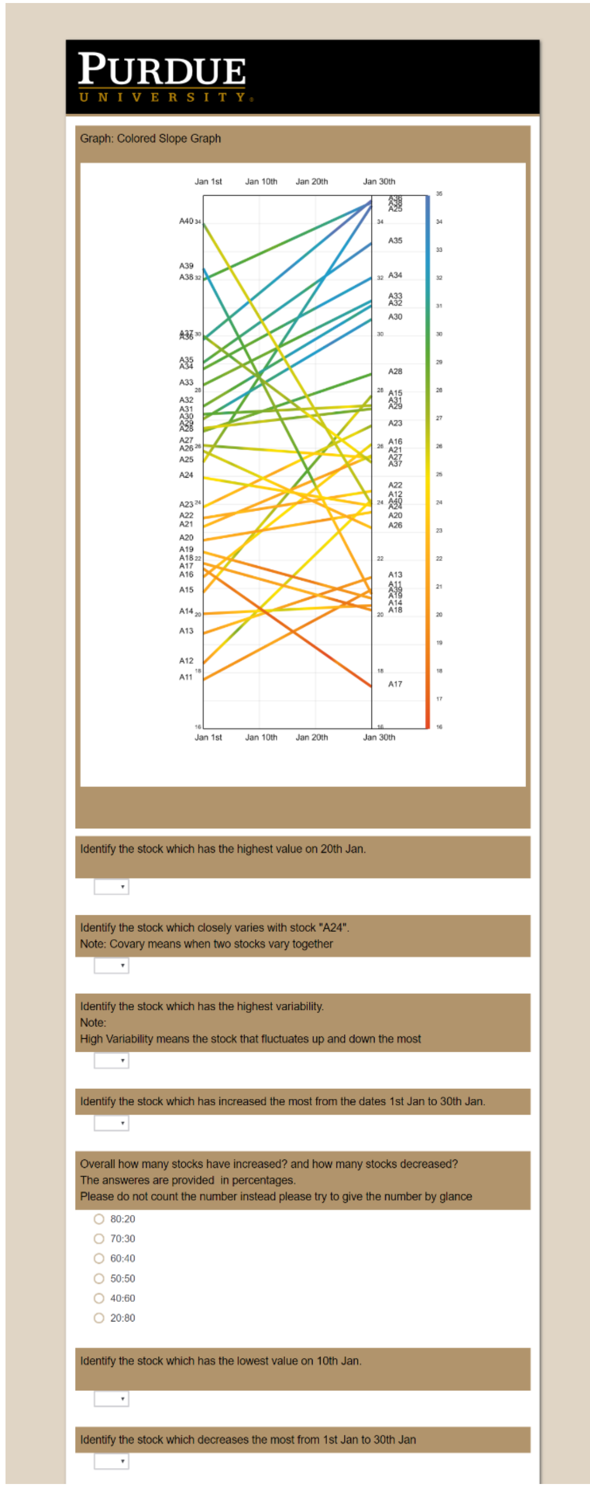
Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from 1st Jan to 30th Jan

APPENDIX P. COLORED SLOPE GRAPH DATASET 1



PURDUE
UNIVERSITY

Graph: Colored Slope Graph

Identify the stock which has the highest value on 20th Jan.

Identify the stock which closely varies with stock "A24".
Note: Covary means when two stocks vary together

Identify the stock which has the highest variability.
Note:
High Variability means the stock that fluctuates up and down the most

Identify the stock which has increased the most from the dates 1st Jan to 30th Jan.

Overall how many stocks have increased? and how many stocks decreased?
The answeres are provided in percentages.
Please do not count the number instead please try to give the number by glance

- 80:20
- 70:30
- 60:40
- 50:50
- 40:60
- 20:80

Identify the stock which has the lowest value on 10th Jan.

Identify the stock which decreases the most from 1st Jan to 30th Jan

APPENDIX Q. DEMOGRAPHIC SURVEY

PURDUE
UNIVERSITY

Gender

- Male
- Female

Ethnicity

- Hispanic or Latino
- Not Hispanic or Latino

Race

- American Indian or Alaska Native
- Asian
- Black or African American
- Hispanic or Latino
- Native Hawaiian or Other Pacific Islander
- White

Age group

- 18-24
- 25-29
- 30-34
- 35-39
- 40-44
- 45-60

Highest level of Education

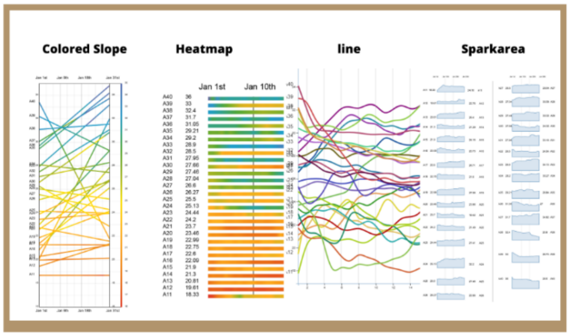
- Bachelors
- Masters
- PhD

Program

→

APPENDIX R. USER SURVEY

PURDUE UNIVERSITY



Among all the visuals, pick the visual which you prefer for temporal information

- Colored Slope
- Heatmap
- Line graph
- Sparkarea graph

Among all the visuals, pick the visual which the easiest to understand the information

- Colored slope graph
- Heatmap
- Line graph
- Sparkarea graph

Among all the visuals, pick the visual which was the hardest to understand the information

- Colored slope graph
- Heatmap
- Line graph
- Sparkarea graph

Among all the visuals, pick the visual which the easiest to perceive the information

- Colored slope graph
- Heatmap
- Line graph
- Sparkarea graph

Among all the visuals, pick the visual which the hardest to perceive the information

- Colored slope graph
- Heatmap
- Line graph
- Sparkarea graph

Among all the visuals, pick the visual which was the most messy or the visual which was very frustrating for you to understand

- Colored slope graph
- Heatmap
- Line graph
- Sparkarea graph

Rank the effectiveness of finding and comparing similar items for each visual,
where 1 being the best and 4 being the worst. Please drag and reorder the list

- Sparkarea graph
- Heatmap
- Colored slope graph
- Line graph

Rank the effectiveness of finding the stocks with dramatic increase/decrease,
where 1 being the best and 4 being the worst. Please drag and reorder the list

- Heatmap
- Line graph
- Colored slope graph
- Sparkarea graph

Rank the effectiveness of finding the stocks with high variance,
where 1 being the best and 4 being the worst. Please drag and reorder the list

- Colored slope graph
- Sparkarea graph
- Heatmap
- Line graph

Rank the effectiveness of finding extreme (maximum) values,
where 1 being the best and 4 being the worst. Please drag and reorder the list

- Heatmap
- Line graph
- Sparkarea graph
- Colored slope graph

Rank the intuitiveness of the visualization design representation,
where 1 being the best and 4 being the worst. Please drag and reorder the list

- Line graph
- Sparkarea graph
- Colored slope graph
- Heatmap

Rank the aesthetics of the visualization design,
where 1 being the best and 4 being the worst. Please drag and reorder the list

- Line graph
- Sparkarea graph
- Heatmap
- Colored slope graph

Rank the visualization's learnability,
where 1 being the best and 4 being the worst. Please drag and reorder the list

- Sparkarea graph
- Colored slope graph
- Heatmap
- Line graph

→