

5-1-2012

Analyzing Financial Data through Interactive Visualization

Fizi Yadav
fyadav@purdue.edu

Follow this and additional works at: <http://docs.lib.purdue.edu/techmasters>



Part of the [Categorical Data Analysis Commons](#), [Corporate Finance Commons](#), and the [Graphics and Human Computer Interfaces Commons](#)

Yadav, Fizi, "Analyzing Financial Data through Interactive Visualization" (2012). *College of Technology Masters Theses*. Paper 67.
<http://docs.lib.purdue.edu/techmasters/67>


This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Fizi Yadav

Entitled Analyzing Financial Data through Interactive Visualization

For the degree of Master of Science 

Is approved by the final examining committee:

David Whittinghill _____
Chair

James Mohler _____

Nicoletta Adamo-Villani _____

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): David Whittinghill

Approved by: Marvin Sarapin _____ 04/23/2012 _____
Head of the Graduate Program Date

**PURDUE UNIVERSITY
GRADUATE SCHOOL**

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:
Analyzing Financial Data through Interactive Visualization

For the degree of Master of Science 

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Executive Memorandum No. C-22*, September 6, 1991, *Policy on Integrity in Research*.*

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Fizi Yadav

Printed Name and Signature of Candidate

04/23/2012

Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/c_22.html

ANALYZING FINANCIAL DATA THROUGH INTERACTIVE VISUALIZATION

A Thesis

Submitted to the Faculty

of

Purdue University

by

Fizi Yadav

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2012

Purdue University

West Lafayette, Indiana

para el mundo y toda la alegría representa.

To my parents who have loved me unconditionally

To my little brother who brings joy to all and sundry

To the two huskies that I will have in the future

And to the girl who broke my heart, as that was the best thing you could've done for me

ACKNOWLEDGEMENTS

The author would like to thank Prof. Mohler and Prof. Adamo-Villani for their valuable insight and Prof. Whittinghill for his unconditional guidance. A very special mention for Mariya Pylypiv. Thank you for all your help and support without which this research wouldn't have been possible.

TABLE OF CONTENTS

	Page
ABSTRACT.....	vi
CHAPTER 1 INTRODUCTION.....	1
1.1 Background.....	1
1.2 Scope.....	3
1.3 Significance.....	4
1.4 Statement of Purpose.....	5
1.5 Research Question.....	5
1.6 Assumptions.....	6
1.7 Limitations.....	6
1.8 Delimitations.....	7
1.9 Definitions.....	7
1.10 Overview of the Study.....	9
1.11 Summary.....	9
CHAPTER 2 REVIEW OF RELEVANT LITERATURE.....	10
2.1 Visual Analysis.....	11
2.2 Types of Data Analysis.....	11
2.3 History of Visual Analysis.....	12
2.4 Current Challenges.....	13
2.5 Techniques for Visual Analysis.....	14
2.6 Research Purpose.....	16
CHAPTER 3 METHODOLOGY.....	18
3.1 Hypothesis.....	18
3.2 Population.....	19
3.3 Variable.....	20
3.4 Procedure.....	21
3.5 IRB Application.....	23

	Page
CHAPTER 4 FRAMEWORK.....	24
4.1 Questionnaire.....	24
4.2 Treatment Tool.....	25
4.2.1 Interface.....	26
4.2.2 Function.....	31
4.3 Control Tool.....	31
CHAPTER 5 DATA ANALYSIS.....	34
5.1 Sample Description.....	35
5.2 Nomenclature.....	35
5.3 Power Analysis.....	36
5.4 Data Summaries.....	38
5.4.1 Primary Summaries.....	39
5.4.2 Secondary Summaries.....	43
5.4.3 Per Question Summaries.....	46
5.5 Data Relationship.....	48
5.6 Test for Significance.....	52
5.6.1 Duration.....	53
5.6.2 Score.....	55
5.6.3 Survey Results.....	56
5.6.4 Per-Question Score Significance.....	58
CHAPTER 6 CONCLUSIONS.....	67
6.1 Final Thoughts.....	68
LIST OF REFERENCES.....	70
APPENDICES	
Appendix A.....	73
Appendix B.....	75
Appendix C.....	77

ABSTRACT

Yadav, Fizi. M.S., Purdue University, May 2012. Analyzing Financial Data through Interactive Visualization. Major Professor: David Whittinghill.

This investigation explored the role data visualization in the scientific community and its effect on the cognitive ability of an individual. The research attempted to answer the question, “ Whether appropriate data visualization helps in better understanding of corporate financial data?” and used quantitative methods to gain an insight. Participants in the study were divided into two groups of 30 each, with one group receiving the treatment devised for the research and the other using the common prevalent method. Both the groups were subjected to a test, based upon the analysis of which a conclusion was derived. The subjects were tested on their cumulative scores on the test as well as the time taken to complete the test. The resulting analysis concluded that visualization did influence the scores and time taken although the difference was more prominent for the scores than for the time. This meant that the particular visualization impacted the accuracy more than the speed. Although this could be disputed by the inverse relationship shared by accuracy and speed, it is in line with the purpose of a true visualization which should aid in better understanding of the underlying data and its relationships.

CHAPTER 1 INTRODUCTION

This chapter establishes a basis for the financial data visualization research study and provides an overview to the document. It establishes significance to the already existing problem within the field of investment and how visualization is ubiquitous within the context of this research. Also, the question posed by the research is clearly outlined and then further distinguished through the identification of boundaries within which the subsequent study is performed. Finally, this chapter concludes with a brief overview of this project.

1.1 Background

Informed decision-making is essential to the success of an individual or organization. In order to make these decisions, we rely on correct analysis of the data. As the amount of data available to siphon increases, so does the importance of data visualization. This is because effective data visualization allows the decision makers to examine large quantities of data, identify trends and correlations, and make informed decision about different aspects of the data. Yadav (2012) relates to this theme on his blog entry:

A majority of the population today invests in stocks as a security for future. This also includes a number of people with no financial background who only look at the major indicators of financial health of a company namely stocks and financial ratios. More often than not, this information is insufficient and sometimes even overwhelming, to provide a clear picture to an untrained eye. As such the research aims to deliver a more complete outlook on the company's finances by providing the numerical data in a format that is much more comprehensible and intuitive than the current setup. The researcher believes that such an approach would lead to better understanding of the financial data which will in turn lead to more informed decision and smarter investments.

It is to be noted, however, that the research only aims to investigate the relation between data visualization and its effect on lucidity and is not intended to provide a tool for investment.

While the question itself is geared toward a very narrow audience, its application can be reflected in other relative areas as well. Today, a large quantity of data is generated in almost all the fields due to the current sophisticated nature of processes. While the human mind can process a significant amount of data, there remains a limit to its perceptive capabilities. Ultimately, exposure to this excessive amount of material can lead to information overload, which can be defined as the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information (Yang, Chen, & Kay, 2003).

While the field of data visualization is still in its infancy, it does provide a remedy to most of the problems afflicting business processes. Data visualization aids in reducing

the risk of information overload through illustrations and interfaces that are intuitive to the user. Furthermore, exploratory data analysis using visuals has the potential to speed up businesses by showing the information that is most relevant to the user while hiding unnecessary clutter. This increased efficiency in terms in information retrieval reduces the overhead and leads to better time management.

Due to its enormous potential and cross-field usability, the research in this field is starting to pick up pace abetted by the fact that it is not encumbered by expertise in one specific knowledge-area. Almost all professions from artists to engineers can find their own particular niche to pursue in data visualization. As such, numerous techniques are being developed on a constant basis to reproduce data in a unique way, each more beautiful and advanced than the other. There, however, remains a need to develop a methodology for successfully implementing these techniques to a particular cause. While there might be more than one way to represent data, there always is one optimal solution amidst all and this is one area that the research explored as its secondary objective.

1.2 Scope

The question posed by the research is whether appropriate data visualization helps in better understanding of the financial data. In a stricter sense, the research tends to focus on a very specialized area of financial undertakings namely the income reports and balance sheets issued by major institutions at regular intervals of time. These reports are a measure of the company's standing over that time period and are meant to give the shareholder's a better view of the company. More often than not, they tend to be muddled

with a lot of numerical data, which is hard to apprehend for a person with no financial background.

While the research question is very narrow in its outlook, the scope of this research extends beyond the boundaries of financial statements. The aim of this research is to find a method to reduce the amount of excess information and gauge the capability of a company through a comparison of certain parameters; which in this case happens to be income statements. It can be applied to other reports that the companies provide such as balance sheets and cash flow statements. Even beyond finance, data visualization techniques discussed in this research should be applicable to any field that requires measuring a parameter to derive a conclusion. The advantage of basing a research on data visualization is that it encompasses a variety of fields while essentially using the same principle of retrieving, analyzing and outputting the data in a graphical format. The key is how to interpret and visualize the data and that can be done in a number of ways.

1.3 Significance

The problem is significant in that a large number of populations invest in businesses without reading the fine print. A case in point is applying for a credit card where people sign up with a company without going over the small but significant details that might otherwise affect their choice. The same is true in stocks and trading. Multibillion-dollar hedge funds invest hours of research before making a decision while the common person typically just tends to look at the brand reputation, current stock value, and the income statements to reach a conclusion. In this type of a scenario, it is important that the information about a company is provided in the most concise and clear

form. The problem of information overload is very relevant today with the advent in Internet technology and storage resources. So it is imperative that the research focuses on an area that would help reduce the number of attention seeking elements and promote faster analysis from the user.

1.4 Statement of Purpose

The purpose of this research is to explore the value of visualization in representing data and whether it benefits an individual in grasping a given concept in a better way. Understanding the comparative benefits of graphical aids for data representation may lead to the development of better visualization techniques and a methodology for selecting them for appropriate tasks.

1.5 Research Question

The primary question for undertaking this research was:

- Does appropriate data visualization help in better understanding of corporate financial data, specifically annual financial statements?

The secondary questions related to this research were:

- Is an appropriate visual more effective than textual data in relaying financial information to people with little or no knowledge of finance?
- What is the method for selecting an appropriate visual technique to relay a specific kind of data?
- Can the discussed technique be extended to fields other than finance?

1.6 Assumptions

The following assumptions were inherent to the pursuit of this study:

- Subjects were chosen randomly from a select student population.
- Subjects were assigned randomly and equally to either treatment or control group.
- Subjects answered all their questions honestly and to the best of their knowledge.
- The questionnaire provided was able to elicit the type of responses that best represent a subject's understanding of the concept.
- The number of subjects chosen was sufficient to infer the results of the study to larger population.
- Both the treatment and control group were provided the same questionnaire with no constraints on the amount of time taken to complete the test.

1.7 Limitations

The following limitations were inherent to the pursuit of this study:

- The study was limited to the number of volunteer subjects available at the West Lafayette campus of Purdue University.
- The study was limited by the cooperation of the participating students and their availability.
- For the purpose of the study, the data visualization tool was developed exclusively as a web-based applet. However, stand-alone versions for windows, mac and linux systems were also developed.

- Internet connectivity was a requirement in order to complete the study.
- The browser used by the participants of the study had to be java-enabled.
- The data visualization tool was only meant to visualize income statements and balance sheets of companies.
- The study was meant to examine visual analysis techniques and its effect on comprehension as related to financial data.

1.8 Delimitations

The following delimitations were inherent to the pursuit of this study:

- The data visualization tool does not track cash and flow or stock values of businesses.
- It does not track all the parameters listed in the income statements and balance sheets.
- A period of one semester was allotted to build the application and conduct the research study.
- The study was not meant to provide a concrete application for investment analysis.

1.9 Definitions

Data visualization is the study of the visual representation of data, meaning "information which has been abstracted in some schematic form, including attributes or variables for the units of information" (Friendly & Dennis, 2009).

Income statement is a company's financial statement that indicates how the revenue (money received from the sale of products and services before expenses are taken out) is transformed into the net income (the result after all revenues and expenses have been accounted for). It displays the revenues recognized for a specific period, and the cost and expenses charged against these revenues, including write-offs and taxes (Helfert, 2001).

Balance Sheet, prepared as of specific date, records the categories and amounts of assets employed by the business and the offsetting liabilities incurred to lenders and owners. It is also called the *statement of financial condition* or *statement of financial position* (Helfert, 2001).

Information graphics or *infographics* are graphic visual representations of information, data or knowledge. These graphics present complex information quickly and clearly, such as in signs, maps, journalism, technical writing, and education (Newsome & Haynes, 2004).

Application Programming Interface (API) is a particular set of rules and specifications that software programs can follow to communicate with each other. It serves as an interface between different software programs and facilitates their interaction; similar to the way the user interface facilitates interaction between humans and computers (Orenstein, 2000).

Processing is an open source programming language and integrated development environment (IDE) built for the electronic arts and visual design communities with the purpose of teaching the basics of computer programming in a visual context, and to serve as the foundation for electronic sketchbooks.

1.10 Overview of the study

The study follows a quantitative approach to gather a conclusion. The result was determined by the significance level achieved by the data. This data was, in turn, gathered from the scores attained by the subjects on a questionnaire. Each of the questions addressed in this research was intended to further the use of data visualization and design in learning. The researcher expected a positive impact of visual designs on human perception.

1.11 Summary

This chapter has provided an insight to the research project, including background, significance, purpose, research questions, assumptions and scope definitions. The chapter has also concluded with an overview of the study and this document. The next chapter outlines the history of data visualization research including major developments, techniques used, limitations in research, as well as current and future directions of the area

CHAPTER 2 REVIEW OF RELEVANT LITERATURE

This review is meant to provide a strong case for the research, which poses the question whether appropriate data visualization helps in better understanding of financial data. Visual analysis has taken a very important role in today's times. While not an independent field of study in itself, it has gained a lot of significance with the huge amount of data being generated and stored in various fields of work in academia as well as industry. This reveals the need to present the information to the user in such a way so as to prevent overloading. It is difficult to analyze the emerging trends within the data because of its enormity. The user has to exert additional cognitive effort that would, in turn, decrease productivity. According to Weber (1993), visualization helps in gaining a better understanding of problem simply because humans comprehend information most intuitively through visual senses. A visual tool that has the ability to extract data will do part of the analysis for the user by finding the association between the figures and making an appropriate graphic for it. This leads to faster work processes as has been evident in financial, medical, defense and countless other fields where visualization has been applied and where a user generally deals with large quantity of data. The use of visual analysis to view trends in a dataset has also been aided by the advancement in storage technologies. Now, companies maintain huge repositories of data that is used to gain insight into past behavior, current trends or future approximation.

2.1 Visual Analysis

Because visual analysis has such an incredible breadth, it has been difficult to define its constraints as well as evaluate its effectiveness in a given scenario. Almost every person has a slightly different view of the practice depending on his or her area of work. Thomas and Cook (2005) provided the most appropriate definition. According to the authors, Visual analytics can be described as “the science of analytical reasoning facilitated by interactive visual interfaces” (p. 3). It involves gathering and processing vast amounts of data and representing it in a way so as to provide clear insight to the user. As Keim, Mansmann, Schneidewind, and Ziegler (2006) mentioned, “it is an integrated approach that combines visualization, human factors and data analysis” (p. 2).

2.2 Types of Data Analysis

The research focuses both on Exploratory Data Analysis (EDA) as well as Confirmatory Data Analysis (CDA) through the use of visualization. One of the key benefits of using visual cues for information gathering is that the type of data analysis can be affected by the presentation method used. Keim et al. (2006), again point out this aspect of visualization while describing its goals:

There are three major goals of visualization, namely a) presentation, b) confirmatory analysis, and c) exploratory analysis. For presentation purposes the facts to be presented are fixed a priori, and the choice of the appropriate presentation technique depends largely on the user. The aim is to efficiently and effectively communicate the results of an analysis. For confirmatory analysis, one or more hypotheses about the data serve as a starting point. The process can be

described as a goal oriented examination of these hypotheses. As a result, visualization either confirms these hypotheses or rejects them. Exploratory data analysis as the process of searching or analyzing databases to find implicit but potentially useful information, is a difficult task (p. 6).

In other words, visual analytics is not just about presenting data but presenting it in such a way so as to maximize its potential value to the user. It can be used to either confirm an existing hypothesis or suggest a hypothesis to test. While CDA is the most common approach in any research pertaining to data visualization, EDA is harder to implement and there is little research that focuses on this aspect of analysis. Fernholz and Morgenthaler (2000) define it as “an approach to analyzing data for the purpose of formulating hypotheses worth testing, complementing the tools of conventional statistics for testing hypotheses” (p. 79). So, the researcher wants to design a tool that would not only help a novice in gaining an insight into the business of a company but also suggest a hypothesis to test for someone with significant financial proficiency. It would show enough associations to allow the user to hypothesize about an emergent behavior and make decisions based on it. Therefore, in terms of this research, the analysis of data rests solely on user’s expertise and intent.

2.3 History of Visual Analysis

The origins of visual data analysis can be traced to the fields of information and scientific visualizations. These related fields of study were an early precursor to the advent of hypothesis testing and value estimation through the use of visual means. Information visualization relates to summarizing the processes in an organization into a

graphic. Its need arose due to companies implementing complex processes that spawned different geographical locations. This made it harder to track operations accountability. As pointed out earlier, a combination of interactive information interfaces and a deep understanding of organizational decision processes can lead to improvement in strategic decision making on part of the organization. In this case, the study was conducted on a 'Research and Development' wing of a pharmaceutical company (Shen-Hsieh & Schindler, 2002). Scientific visualization on the other hand relates to tracking the values generated in a scientific experiment and providing relevant analysis to an observer. The developments in these fields led to better visualization techniques, which were extended to a number of other applications in diverse scenarios like experimental designs and simulation tracking.

2.4 Current Challenges

One of the major issues in visualization is determining the best-possible way to represent the data. There are a number of statistical algorithms for sorting through the data and plenty of techniques to visualize it. The challenge is to find the appropriate method that the user would most relate to. The case in point is NodeXL, which is an extendible toolkit implemented as an add-on to MS Excel 2007 spreadsheet software (Smith, Shneiderman, Frayling, Rodrigues, Barash, Dunne, Capone, Perer, & Gleave, 2009). Because its main focus is on mapping social networks, it relies on a network of points and nodes to show the association. It is a very intuitive way to make generalization based on the type of graphic created by this tool. Similarly, the most perceptive way to present financial information is through time series line charts.

So it is evident that a data set can be presented in a variety of ways but the key is to find the method that best summarizes the concept. A visual analytics system must gather and understand the data. It should then output the most relevant information and hide the details. The user then has the choice to dig deeper into the data set by focusing on a particular aspect of the illustration. This is in accordance with the visual data exploration mantra by Shneiderman (1996): “Overview first, zoom and filter, and then details-on-demand” (p. 3). The need is to create an interactive environment where the scope of information presented is dependent on the user. A visual analytics problem generally deals with large data sets. The system analyses the data but the user derives reasoning from it. Conversely, a well-defined problem that doesn’t require an interactive medium to gather an analysis should not be treated as a visual analysis problem as it would be a waste of resources and not as efficient. So for any given dataset, it is important to come up with an idea of how to best present the information. This is the whole premise of a visual analytics system as it helps examine and gain conclusion from a hypothesis by channeling the information into a single path and avoiding the abstract.

2.5 Techniques for Visual Analysis

There are, now, a number of techniques to present the data. The efficiency of these methods depends on the type of data to be displayed. As pointed out by Kein (2002), these may fall into one of the following categories:

- One-dimensional
- Two-dimensional
- Multi-dimensional

- Text and hypertext
- Hierarchies and graphs
- Algorithms and soft wares

This research will focus primarily on two-dimensional and textual data because it is looking specifically to gain insight from the income sheets of various businesses. One of the ways to do that is by clustering similar elements in a 2-d or 3-d space using spatial analysis. This is done by grouping elements that possess same characteristics into groups, which are then shown on the screen in a variety of formats. The user can dig deeper into a particular group at which point the interactive visual will again analyze the data set in real-time and display the results. So, it is quite evident that building such a system requires a suitable engine that can extract and differentiate relevant information and adapt itself to the level of information that the user desires. The system also demands different levels of detail so as to not overload information on the user and also due to space and environment restrictions. Any type of visual will be limited by the medium it is displayed on be it a hand-held device, a portable laptop or larger projection screens. The researcher aims to target the common masses and as such the focus is on first two media. Much of today's information can be conveyed through mobile technology so any intended application should be developed keeping in mind the constraints posed by accessing information through such a medium. These limitations can be small screens, bad lighting and color gamut etc. Noirhomme-Fraiture, Randolet, Chittaro and Custinne (2005) offer some recommendations for presenting time series visualizations on small screens. It is important to simplify the information while presenting it on the smaller screens. This

again relates to the author's research in that the amount of information needs to be reduced through innovative analysis before being presented to the end user.

As with visualization techniques, there are various ways of clustering similar content as well, two of which are galaxies and themescapes. The key here, again, is to choose a method that best compliments the resulting analysis depending on the type of data and expertise of the user. A galaxy display uses a 2-d scatterplot to cluster similar objects. It can be applied to this particular research by clamping together companies that show similar traits in their income statements. A themescape on the other hand, shows the relations and interdependencies between the data in the form of a 3-d landscape much like a terrain map. The terrain is intended to relay important information about the theme of the body, which in this case would be the income statements of the corporations. According to Wise, Thomas, Penock, Lantrip, Pottier, Schur, and Crow (1995), a themescape visual representation has several advantages such as utilizing innate human abilities of pattern recognition and spatial reasoning and displaying most of the complex contents of a database.

The author intends to use these methods to initially bundle the companies together based on the kind of income statements that they possess. A themescape seems the more likely option, as then the application will be able to relate not just the income statements but also the type of business, cash flow and balance sheets.

2.6 Research Purpose

This research requires deep incorporation of statistics and visualization to explore a data set. It is to be noted that the tool itself does not provide a solution to the problem,

rather it aids the user in coming up with an appropriate solution by highlighting the relationships between the data elements which otherwise would not have been apparent. There are a number of applications in the market today such as KrackPlot (Krackhardt, Blythe, & McGrath, 1994), Pajek (De Nooy, Mrvar, & Batageli, 2005), NetDraw (Borgatti, 2007) and SocialAction (Perer, & Shneiderman, 2008) that are designed to help analysts through the visual representation of information. However, there are few tools that are specifically designed to relay financial information in a way that assists a layman to make the best possible decision on an investment. The research aims to fill that void by providing such an aid. Also, the tool is meant to assist the users with varying levels of expertise and this gives it a greater breadth than other alternate solutions. This is achieved by focusing on both EDA and CDA, as opposite to just one, which is generally the case with the aforementioned tools. The aim of the researcher is to tightly integrate statistical algorithms used to ascertain the most relevant information with intuitive visualizations. The algorithm for the interface would update in real-time so that the users can filter information according to their liking. Therefore, the tool would lay great emphasis on interaction in order for the user to have full control over the information being displayed.

The works cited in this section highlight the significance of data visualization and the apparent need for developments in this field. The problem that the research focuses on is very relevant and demands a solution in the immediate future. While the researcher does not expect the proposed tool to be all-inclusive, it will still be a step in the right direction and will pave the way for future developments in this area.

CHAPTER 3 METHODOLOGY

The aim of this research was to measure the effectiveness of using interactive visuals in place of traditional textual data. To this end, the researcher chose financial data in the form of income statements and balance sheets of two multi-national organizations that had relative setup in terms of market, organizational structure and company values. Financial information is predominantly composed of numbers that would make little sense to a layman and as such it presents an opportunity to tone down the level of information and present it in a better way. Also, financial data is easily accessible as most companies make them available online or through print media. Lastly, this is a sector that would benefit most from data visualization tools, as there is a demand in the sector from people who wish to gain a greater insight into the market for educational or investment purposes. Having said that, the study has the potential to use data of similar kind from other resources and so it represents a very broad area of research

3.1 Hypothesis

The study involved the following hypothesis for deduction:

H1₀: There is no significant difference in scores between the tests undertaken by the subjects in treatment and control group

H1_α: There is a significant difference in scores between the tests undertaken by the subjects in treatment and control group.

H2₀: There is no significant difference in the time taken by the subjects in treatment and control group to complete a test.

H2_α: There is a significant difference in the time taken by the subjects in treatment and control group to complete a test.

H3₀: Data visualization does not significantly aid in understanding the financial statements.

H3_α: Data visualization significantly aids in understanding the financial statements.

H4₀: There is no comparative significance to the ease of using data visualization over textual data.

H4_α: There is a comparative significance to the ease of using data visualization over textual data.

3.2 Population

This research was conducted at Purdue University, a land-grant university situated in West Lafayette, Indiana. The subjects were selected exclusively from the student population at Purdue. The students were randomly selected and normally distributed. They represented a wide spectrum of learning abilities, age, disciplines and context-awareness and as such were well suited for the task. The chosen students were equally

divided into two distinct groups, which were then subjected to different treatments as decided by the researcher.

3.3 Variable

The following variables were considered to have an effect in the statistical analysis:

Independent Variable:

1. Interactive financial applet (treatment)
2. Financial Statements (control)
3. Timeline Graphs (control)

Dependent Variable:

1. Treatment scores: score attained by the subject in treatment group.
2. Control scores: score attained by the subject in control group.
3. Treatment duration: time taken by the subject in treatment group to complete the questionnaire.
4. Control duration: time taken by the subject in control group to complete the questionnaire.
5. Treatment survey score: feedback provided by the subject about using the applet
6. Control survey score: feedback provided by the subject about using financial statements.

The researcher aims to see the results of the treatment variable on score obtained and time taken and hence the relationship. These are quantitative variables, as they vary in degree and amount of a phenomenon.

3.4 Procedure

The research was designed to analyze the effect of an interactive graphical aide in understanding a particular concept. To this end, the researcher developed an interactive visual interface using 'Processing API for Java'. This tool was used as a treatment for the subjects in order to substantiate the hypothesis proposed by the researcher.

The subjects were chosen randomly from the student population at Purdue University and assigned equally into a treatment group and a control group. Power analysis was used to assess the probability of successfully rejecting the null hypothesis. The alpha level for the test was set at 0.05 while the effect size was set at 0.5 in order to correctly deduce the significance of the research. A sample size of students was set based on these two parameters. Both the groups were provided a separate URL. The treatment URL contained the applet created by the researcher and a web-based questionnaire that the subject had to complete with the aid of the information provided in the applet. The control URL had links to the web forms that contained data presented in the income statements and balance sheets of the two companies. The subject had to complete the same set of questions as was provided to the treatment group. After a subject had submitted the questionnaire online, the data along with a timestamp was automatically filed into a spreadsheet that was only accessible to the researcher.

The researcher then compiled all the scores and the time taken to complete the questionnaire by the participants from the two groups. This data was then further analyzed to reveal trends using STATA statistical analysis tool.

An independent group t-test was performed between the scores, duration and the two survey variables of both the groups to see if the relationship between the dependent and independent variables was significant enough prove the stated hypothesis. In addition, the researcher also performed per question t-tests to examine the given hypothesis for each individual question. All the questions that composed the questionnaire were designed to test a subject's understanding in different ways. As such, per question analysis was meant to provide further insight into a subject's understanding of the concepts.

Besides this, a number of other analysis techniques such as data summaries, box plots, bar-plots, kernel densities, relationship matrices, scatter graphs etc. have been performed to get an understanding of the data from different perspectives. The aim was to describe the trends and results of the data set in a holistic manner and these tests were meant to aid in this approach.

The researcher believed that the comparison of the two groups would show better test scores and lower duration for the treatment group. Also, there was to be a significant difference in the scores, time and survey responses from both the groups, which would be equal to or greater than the effect size threshold.

3.5 IRB Application

The study was approved by IRB as a non-exempt research study that required a consent form to be made available to the students. For this purpose a link to a pdf copy of the consent form was placed on both the treatment and control web pages and the users were encouraged to read the form. A copy of the form is attached in the appendices.

All necessary precautions were taken to protect the privacy of the subjects with no personal information being extracted. The researcher had sole access to the responses provided by the subjects and they did not contain any identifiers that could be traced back to the subjects.

CHAPTER 4 FRAMEWORK

This section details the framework of the study as well as the tools required to gauge the performance of the control and the treatment group. Both groups were alike in that they had to undergo the same post-test under similar conditions. But they were different in the use of treatment that formed the basis of their distinction. One group used the applet developed by the researcher, which is discussed in greater detail in the following section. The other group used the standard medium of retrieving information through data sheets.

4.1 Questionnaire

Each of the groups received exact copies of a questionnaire after being exposed to the treatment of the research. The questionnaire was based on the content provided in the financial statements and was designed to test the attained knowledge of the students on the subject. It was web- based and consisted of seven close-ended questions and two survey questions with likert scale to denote the degree. It also had two text-inputs that did not have a bearing on the data analysis but were instead designed to get an insight into the subject group. While subjects did not have any restrictions on answering the questions or the time taken to complete the test, it was expected that they participated to the best of

their abilities and the time was recorded through a timestamp in the form. The questionnaire can be found in Appendix A.

4.2 Treatment Tool

The applet for the treatment group was developed in java programming language and was used for the specific purpose of showing relevant financial information of pre-selected companies. It was developed bearing in mind all the specifications that are deemed necessary to make a good, compelling visualization. These specifications were formed through the review of existing literature and are as under:

- Design of the graphic should always follow the data presented in the graphic meaning the rendering of a visual should make the information more effective to the user. (Walker & Lev, 1953)
- Visualize all the elements if possible. That means the emphasis should be on converting all numerical information on the display into a visual component. Display numbers only to provide major story points.
- Display all the data sources.
- Avoid comparisons between data sets that use different methodologies. (Fayyad, Grinstein, & Wierse, 2001)
- Avoid renderings that might provide a false bias towards a certain set of information. (Zimmons & Panter, 2003)

- Use minimalistic renderings with soft colors and provide contrast to make information stand out. Keep the visual as simple and uncluttered as possible. (Shermer, 2005)
- Relate all the data points over a timeline if possible while keeping the data intervals to a minimum. (Catmull & Clark, 1978)
- Find comparisons between the different data groups and choose a rendering that best supports these comparisons. (Baldassi, Megna, & Burr, 2006)
- Provide only the most relevant information to the user while keeping intact the data relationship and visual conformity.
- Provide as much user-control as possible. The goal is to aid the user in analyzing the data rather than enforcing it.
- Use information that would appeal to the user. Visualization without proper substance or providing irrelevant, useless data serves no purpose.

4.2.1 Interface

The interface is composed of a main menu that lists the companies. Future iterations of the software can include greater number of companies that can be perused further on a much larger scale. The home screen in Figure 4.1 gives directions about the interface along with buttons that list the companies. For the purpose of this research, the two companies in question are Google and Yahoo who have gone in opposite directions in terms of performance, reputation and net wealth over the last ten years. Users have the

option to click on either of the two buttons and get an in-depth look at the company's finances.

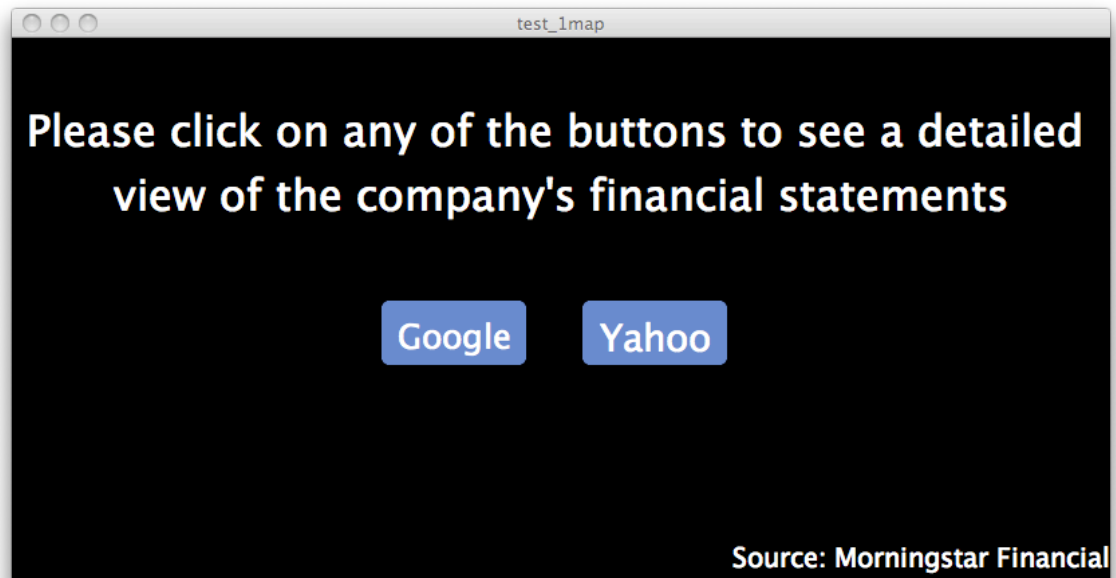


Figure 4.1. The home screen.

Once the user clicks on either of the buttons the screen changes to reveal a blank graph with rectangular buttons, as shown in Figure 4.2. A tool-tip, that appears when a user points the mouse cursor on the button, reveals the purpose and use of a particular button. For this research the three parameters used are Net Income, Operational Costs and Total Assets, which are one of the more important indicators of a company's financial health.

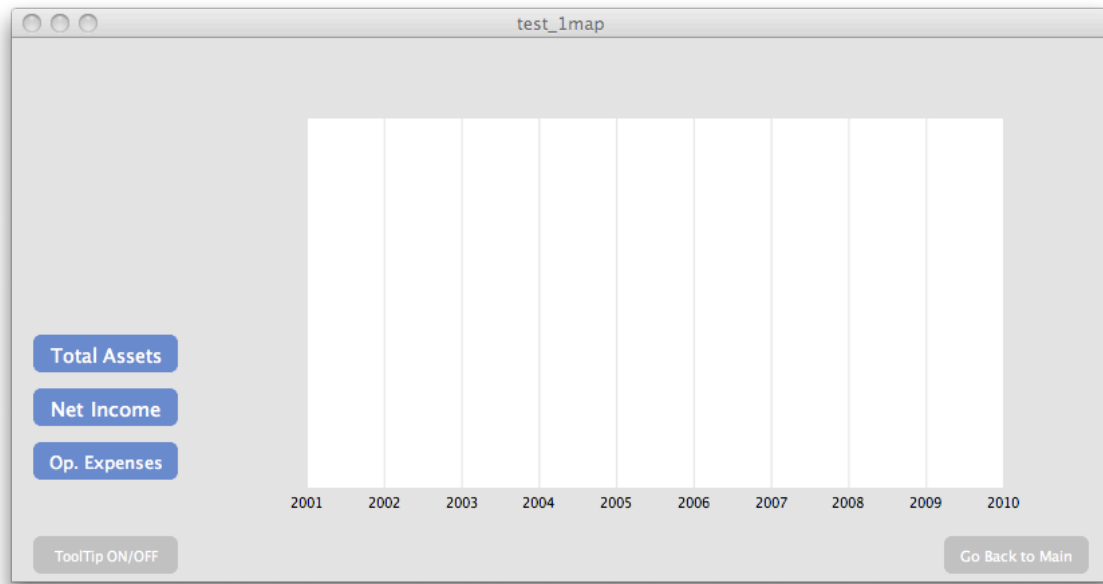


Figure 4.2. Initial graph.

Once the user clicks on any of these categories the graph changes to reveal the performance of the company over a ten-year period with respect to that particular parameter. Vertical grey lines that intersect the graph differentiate the years that form the x-axis on the graph. Several other options such as displaying the graph as points or area covered are also available to the user at this point.

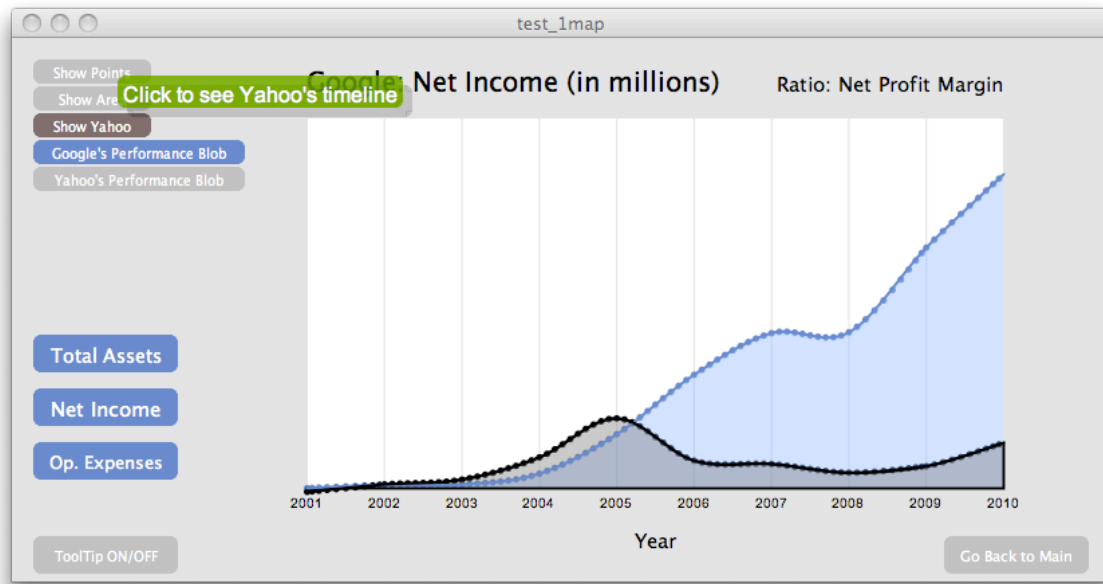


Figure 4.3. Graph with points, area, comparative timelines turned on.

Such selective display enables the interface to avoid information overload and provide only the most relevant information. The y-axis is conspicuous by its absence and is done so for the purpose of reducing redundancy. The y-values are shown at the top of the graph when the cursor moves across the x-values. The application interpolates the values between the points so there is a smooth transition from one value to the other. Another important aspect of the application is being able to compare more than two parameters. This is done with the aid of *performance blobs* that are translucent circles that appear over any point in the graph with their color indicating the company. The size of these circles makes it easy to compare the performance of the two companies as the user is able to perceive the contrast of the two companies just by comparing the radius of the two performance blobs that appear over an x-y point.

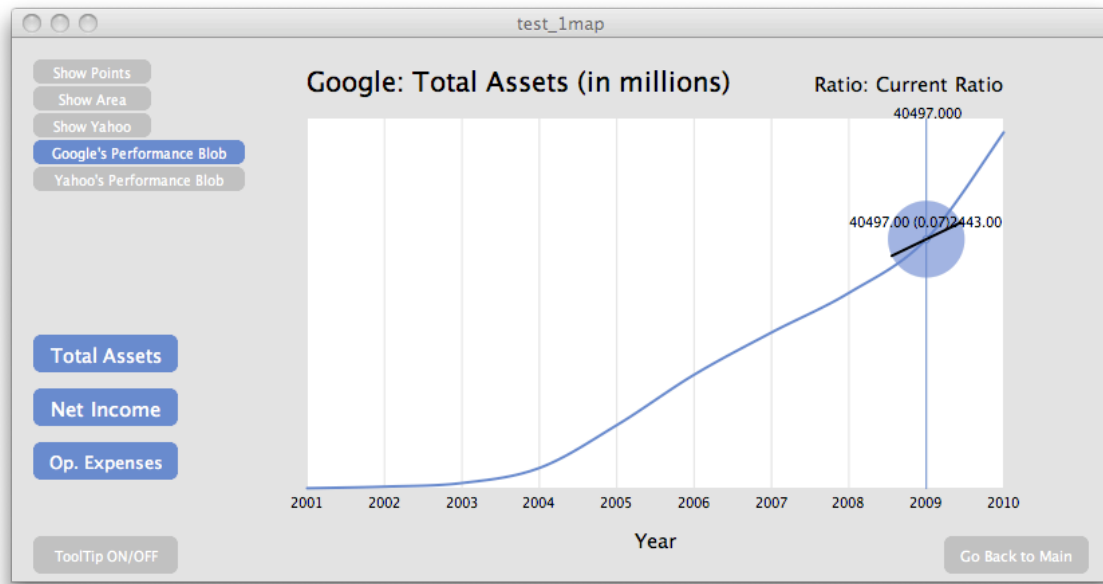


Figure 4.4. Graph showing default performance blob.

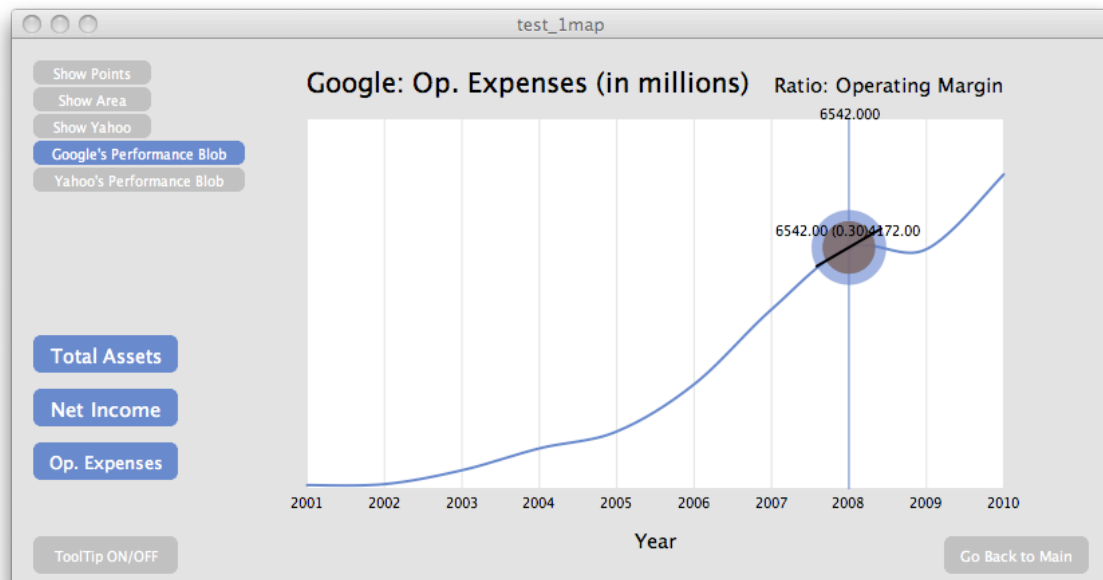


Figure 4.5. Graph showing blobs of two companies.

The speed of rotation of these blobs also reveals another parameter, which in this case is the financial ratio. So in principle, the interface enables us to compare up to five different parameters at any point in the graph. For the purposes of this study, we have restricted it to only three factors.

4.2.2 Function

The data used in the application is collected from Morningstar database. This is the same information that is provided to the control group in the form of web-based data sheets. The data is stored in a comma-separated file, which is then parsed, grouped, interpolated, mathematically processed and displayed in a particular format by the program at the behest of the user. Aside from the three selections available to the user, the program also calculates an attached financial ratio for each of the selection. These ratios are not stored on the file but rather calculated in real-time once the user has made a selection. The parameters and ratios provided to the user are shown in Table 4.1 and 4.2.

4.3 Control Tool

The control group had to rely on the standard financial statements of the two companies in order to correctly respond to the presented questions. These statements contained the parameters used in the treatment tool as well as a standard set of charts. These statements were downloaded from the Morningstar Financial database and then converted to web forms. Following four financial statements were included as web forms:

- Google Balance Sheet
- Google Income Statements

- Yahoo Balance Sheet
- Yahoo Income Statements

Table 4.1

Major Financial Indicators and their Definitions

Financial Indicator	Definition
Total Operational Expenses	A type of expenditure incurred by the business due to its standard commercial operations. Since businesses in separate industries tend to have different spending outlay, companies can choose how they list their expenses. It is found on a company's balance sheet.
Net Income	It is an indicator of the total earnings of a company within a given time-period. It is calculated by compensating the costs of business expenses, interests and taxes from the total revenue. It provides information about the profitability of a company and can be found on the income statement.
Total Assets	An indicator of the resources owned or controlled by the business that have certain economic value. These resources can be tangible or intangible but must have the potential for monetary rewards in the future. The total assets are again found on the company's income statements.

Table 4.2

Major Financial Ratios and their Definitions

Financial Ratio	Definition
Operating Margin	Operating margin is an indication of a company's revenue after deducting variable cost of production such as wages, infrastructure, raw materials etc. It is calculated by dividing operating income by net sales. A sound operating margin is necessary for a company in order to pay for fixed costs on operations.
Net Profit Margin	This value is generally expressed as a percentage and is calculated by dividing net profit by net revenues. It indicates the effectiveness of a company in converting revenue into actual profit. This is a good way of comparing companies in the same industries as such companies are generally subject to similar business constraints. It also gives a fair measure of the cost-control capabilities of a company.
Current Ratio	This ratio provides information about the company's ability to meet short-term debt obligations. It measures whether the company has enough resources to pay its short-term debts and is indicative of a company's market liquidity. Its value is determined by dividing current assets by current liabilities. The higher the ratio, the better the short-term financial strength of a company.

CHAPTER 5 DATA ANALYSIS

This chapter provides description of the methodology that was employed to collect data and summary and analysis of analytical sample. It also details the inference gained from the results of said methods. The primary quantitative method used was independent group t-test (Bruin, 2006). It is a parametric statistical test that compares performance of two different samples of participants, indicating whether two samples perform so similarly that we can conclude that there are no differences between treatments, or they perform so differently that we conclude that there was significant difference between two treatments. The test assumes that variances for the two populations are the same and therefore the subjects are randomly selected from a larger population of subjects. The interpretation for p-value is the same as in other type of t-tests.

This statistical measure is a suitable choice also because of the robustness of the t procedure against non-normality of the population. Larger samples improve the accuracy of probability and critical values from the t distributions. As determined through power analysis, a sample size of 30 is able to achieve adequate power for even clearly skewed distribution. This proves useful in cases where the subject population might be normal but the sample distribution follows a non-normal pattern.

Additionally, a number of other secondary methods were used to gain more insight into the data set and find any relationships that might exist and influence the hypothesis.

5.1 Sample Description

The subjects of the study were solely comprised of the student population at Purdue. In total there were 60 volunteers who participated in the study. They were randomly and equally divided into two groups of 30 each that formed the treatment and control group. Due to the random nature of selection, the researcher did not delve into the influence of gender or age of the subjects on the study but instead chose to focus solely on the effect of the treatment on a sample group. Furthermore, the sample size and the major independent variable weren't large enough to use the personal traits of a subject for statistical significance through regression analysis. Each subject in either group had to undergo a test on their understanding of the treatment provided. The questions on the test were designed to accommodate different types of cognitive processes and type dynamics.

5.2 Nomenclature

Since this chapter analyzes a number of parameters with the aid of statistical software, the researcher had to use certain naming conventions to annotate these parameters as variables inside the software. The nomenclatures used to describe these variables are provided in Table 5.1 and 5.2 and will be followed throughout this section.

5.3 Power Analysis

Power of a test is the probability of rejecting H_0 when H_a is true, which is also the probability of not committing a Type II error. For the purpose of this research a power analysis was performed to determine a sample size that would be adequate to remove the possibility of a Type II error. The sigma value was 0.5 while the alpha level was set at 0.05. A combined sample size of 30 equally divided into two groups was enough to generate 80% power and this was considered a cut-off for this study. The final sample size of 60 generated 96.77% power (Lenth, 2006).

Table 5.1

Primary Variables and their Descriptions

Primary Variable	Description
T-Score	Test scores of the subjects comprising the treatment group.
C-Score	Test scores of the subjects comprising the control group.
T-Duration	Time taken by the subjects comprising the treatment group to complete the test.
C-Duration	Time taken by the subjects comprising the control group to complete the test.

Table 5.2

Secondary Variables and their Descriptions

Secondary Variable	Description
Ease of tool	The apparent ease of using the treatment tool (applet) given as a score between 1 and 5 with 1 being extremely difficult and 5 being very easy
CEase of tool	The apparent ease of using the control tool (financial statements) given as a score between 1 and 5 with 1 being extremely difficult and 5 being very easy
Understanding data	The knowledge attained during the treatment session given as a score between 1 and 5 with 1 being no perceivable knowledge and 5 being considerable understanding of data provided.
CUnderstanding data	The knowledge attained during the control session given as a score between 1 and 5 with 1 being no perceivable knowledge and 5 being considerable understanding of data provided.

Table 5.3

Questions Provided to the Subjects

#	Treatment/Control Questions
Q1	Which company has greater assets in the year 2008?
Q2	In which of the following years is the difference between the assets of Yahoo and Google the greatest?
Q3	Google posted a greater 'Net Profit Margin' in 2008 than in 2010?
Q4	During which financial year does Google overtake Yahoo in terms of total Operational Expenses?
Q5	In which year did Yahoo post its greatest Net Income returns?
Q6	Based on the answer to the above question, what was the highest Net Income return of Yahoo during the period from 2001-2010?
Q7	Based on the answer to the question above, what was the corresponding Net Income return of Google in the year Yahoo posted its highest returns?

5.4 Data Summaries

A summary statistics is used to summarize a set of operations. They are provided to present a clear display of data, and present a condensed, comprehensive summary of what we have found out. It is good way to relay the information about data set in very simplistic terms. This section details the summary statistics and their inference for the primary and secondary variables. These summaries contain the following measures to describe the data:

- Observations: The total number of records in a data set

- Mean: Describes the central tendencies of the data set. Taken as the sum of records divided by the total number of records.
- Standard Deviation: A measure of statistical dispersion. It describes the spread of records around the average or mean.
- Min Value: The minimum value within a given data set
- Max Values: The maximum value within a given data set.

5.4.1 Primary Summaries

The summary statistics for the two groups relative to the primary dependent variables are given in Table 5.4.

Table 5.4

Data Summaries of Primary Variables

Variable	Observations	Mean	Standard Dev.	Min	Max
T Score	30	5.766667	1.546594	2	7
C Score	30	3.666667	1.971055	1	7
T Duration	30	5.866667	3.093189	1	13
C Duration	30	8.066667	7.016893	1	38

Looking at the data summaries it is inherently clear that there is marked improvement in the mean scores of subjects in the treatment group when compared with control group while the mean duration of the treatment has also decreased considerably. Since the test score had a maximum value of 7 it is possible to compute the increase in mean performance, which comes out to be 31.43% approx. Another important statistical

observance is the contrast in standard deviation of the parameters for the two groups. There is less deviation for the treatment group in both categories, which indicates a lower spread. It also indicates the possibility of an outlier in the testing duration of control group. The bar-graph comparisons of the means are provided in Figure 5.1 and 5.2.

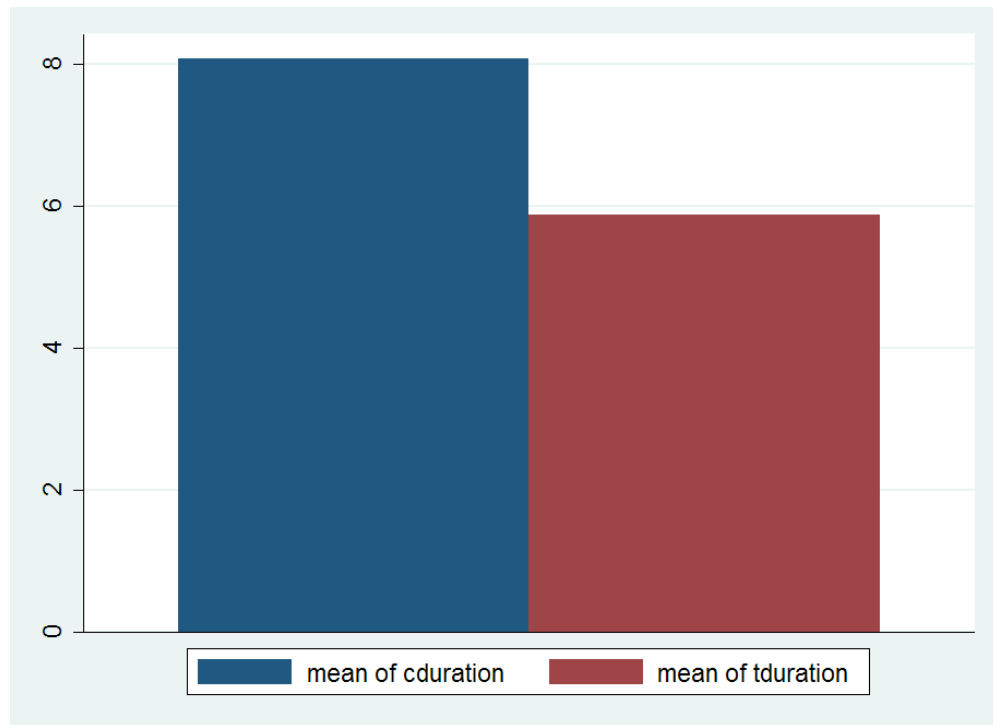


Figure 5.1. Bar graph of duration

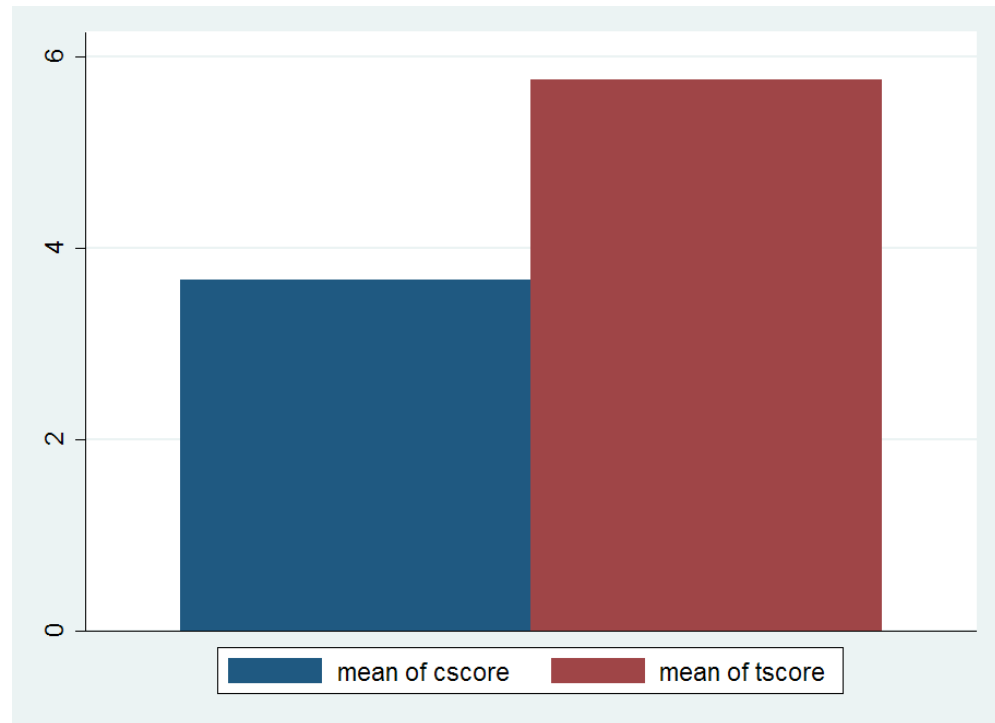


Figure 5.2. Bar graph of score

Also shown in Figure 5.3 and 5.4, are the box-plots for the dependent values. According to Ireland (2010), A box plot is common graphical technique used to depict the variability of a set of samples and allows a rapid visual comparison of some of the statistical characteristics of the sample. The line in the middle of the box separates 50% of the values in the sample. The length of the box is the range from the lower (q1) to upper quartile (q4). The width of the box represents the sample size. The dashed line displays the median, while the whiskers represent presents of extreme observations. While dots outside those whiskers are considered extreme outliers. The width of the box often indicates the relative sample size.

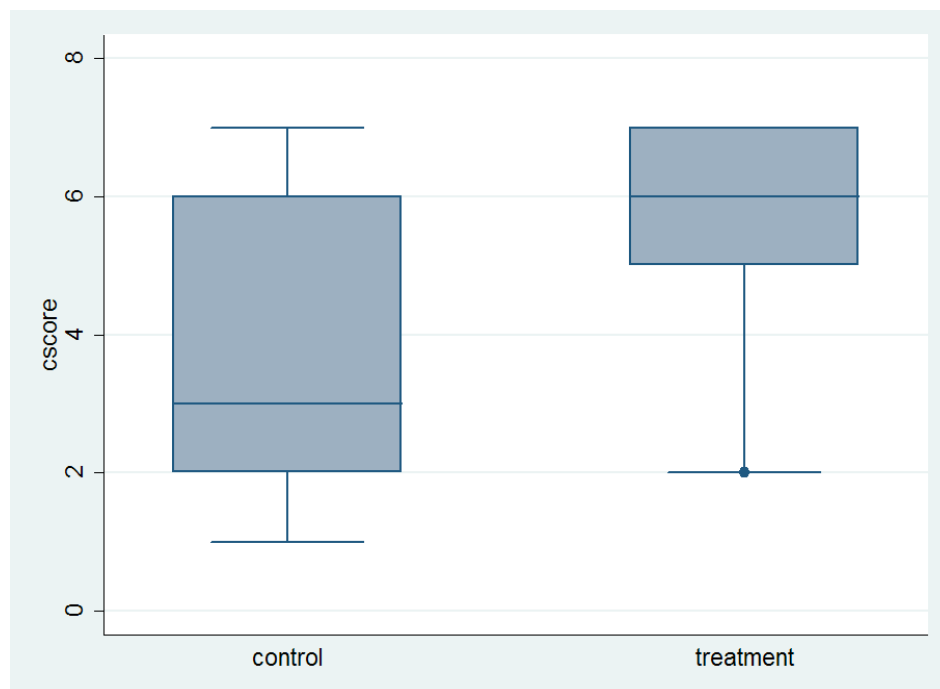


Figure 5.3. Box plot of score

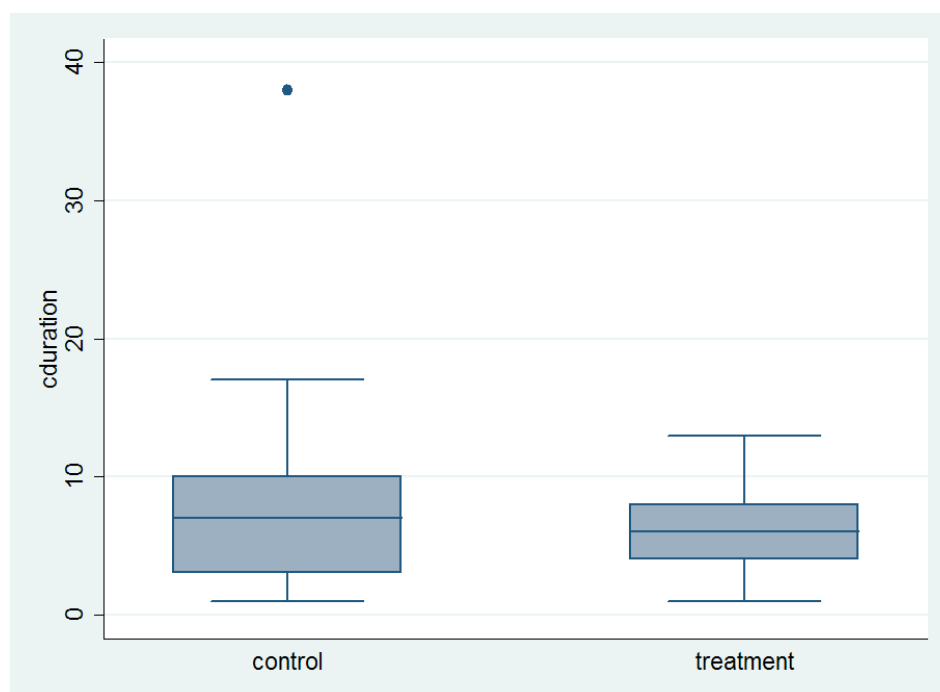


Figure 5.4. Box plot of duration

The box-plots confirm the outlier in testing duration for control group. It also clearly illustrates the different quartiles of the data set. It is evident that the scores and duration of the treatment group are more balanced relative to the placement of mean and inter-quartile range. This means that the treatment provided had a balanced effect on the scores and the duration without any underlying aberrations.

5.4.2 Secondary Summaries

The answers to the two survey questions comprise the secondary variables. Their summaries are presented in Table 5.5.

Table 5.5

Data Summaries of Secondary Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Ease_of_tool	30	3.666667	1.184187	1	5
Cease_of_tool	30	2.466667	1.357821	1	5
Understanding_data	30	3.4	1.132589	1	5
CUnderstanding_data	30	2.466667	1.306043	1	5

The summary statistics shows that the means of the two treatment variables are greater than their counterparts with less standard deviation. This means that the subjects in the treatment group gave a higher score on the survey questions than their counterparts in control group. This, in turn, leads the researcher to speculate that the data visualization tool did impact the performance and understanding of the participants. The individual data summaries of the 4 variables are presented in Table 5.6, 5.7, 5.8 and 5.9.

Table 5.6

Individual Summary of Survey Question I for Treatment Group

Variable: Ease of tool (Treatment)

Score	Frequency	Percent	Cumulative
1	2	6.67	6.67
2	3	10.00	16.67
3	6	20.00	36.67
4	11	36.67	73.33
5	8	26.67	100.00
Total	30	100.00	

Table 5.7

Individual Summary of Survey Question I for Control Group

Variable: CEase of tool (Control)

Score	Frequency	Percent	Cumulative
1	11	36.67	36.37
2	4	13.33	50.00
3	7	23.33	73.33
4	6	20.00	93.33
5	2	6.67	100.00
Total	30	100.00	

Table 5.8

Individual Summary of Survey Question II for Treatment Group

Variable: Understanding data (Treatment)

Score	Frequency	Percent	Cumulative
1	3	10.00	10.00
2	3	10.00	20.00
3	6	20.00	40.00
4	15	50.00	90.00
5	3	10.00	100.00
Total	30	100.00	

Table 5.9

Individual Summary of Survey Question II for Control Group

Variable: CUnderstanding data (Control)

Score	Frequency	Percent	Cumulative
1	10	33.33	33.33
2	5	16.67	50.00
3	8	26.67	76.67
4	5	16.67	93.33
5	2	6.67	100.00
Total	30	100.00	

The individual summaries clearly highlight the difference in opinions of the subject for the use of applet relative to financial statement. For 'Ease of tool' (Treatment) 36% or less stated that tool was hard to understand, while this statistic was close to 73% for 'CEase of tool' (Control). Similarly, approximately 60% of the subjects in the treatment group thought the information provided was very easy to understand while this figure dropped to only 23.34% for the control groups. These figures lead the researcher to the assumption that the treatment did have a positive affect the subject's perception. It also conveys that the specifications charted to develop an interactive application do impact its ability to interest the user.

5.4.3 Per Question Summaries

This section details the individual summaries of the questions that comprised the questionnaire prepared by the researcher. For each correct answer the subject received a point. No points were awarded for incomplete, incorrect or no response. These summaries are provided in Table 5.10 and 5.11.

Table 5.10

Response Summaries of Individual Questions in Control Group

Control Summary

Que. No.	Obs	Mean	Std. Dev.	Min	Max
Q1	30	.9666667	.1825742	0	1
Q2	30	.8	.4068381	0	1
Q3	30	.5333333	.5074163	0	1
Q4	30	.1	.3051286	0	1
Q5	30	.5333333	.5074163	0	1
Q6	30	.3666667	.4901325	0	1
Q7	30	.3666667	.4901325	0	1

Table 5.11

Response Summaries of Individual Questions in Treatment Group

Treatment Summary

Que. No.	Obs	Mean	Std. Dev.	Min	Max
Q1	30	.9666667	.1825742	0	1
Q2	30	1	0	0	1
Q3	30	.7	.4660916	0	1
Q4	30	.5666667	.5040069	0	1
Q5	30	.9333333	.2537081	0	1
Q6	30	.8	.4068381	0	1
Q7	30	.8333333	.379049	0	1

5.5 Data Relationships

This section details the relationships between the dependent variables. An overall condensed mapping of relationship is shown in Figure 5.5.

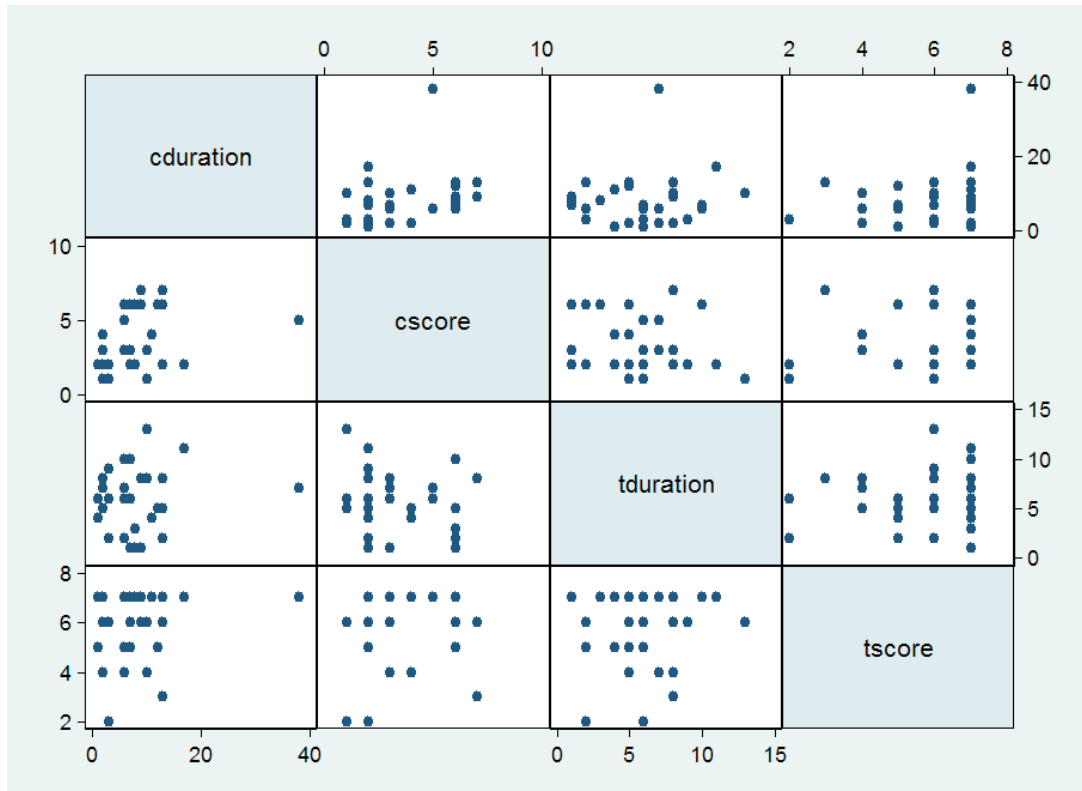


Figure 5.5. Scatter plot of 4 dependent variables

The matrix above provides the relationship between the scatter plots of each variable on a 2-dimensional axis. While certainly plausible, there appears to be no linear pattern between the duration sets because it is not possible to chart a line of best fit through the points. We will also not be able to detect a linear pattern in score since it's an ordinal variable. In other words, it is not possible to form a mathematical equation that will give the points on the graph for any two variables.

Figure 5.6 and 5.7 provide the kernel density estimation of the data sets. It is a statistical technique to estimate the probability density function of a random variable that does not require the data to fit a particular probability distribution. According to Rosenblatt (1956), the mathematical equation for estimating the shape of a function f is shown in equation 5.1.

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad (\text{eqn.5.1})$$

where $K(\bullet)$ is the kernel which can be defined as a symmetric function that integrates to 1. These estimates are closely related to histograms but provide continuity to the data. As is evident from the graphs, the distribution for control duration shows a very high positive skewness while the treatment duration is approximately symmetric. This is in line with the estimates of the researcher as the subjects in the treatment group follow a normally distributed pattern for the time taken to complete the test while the subjects in the control group have the tendency to take more time than their counterparts.

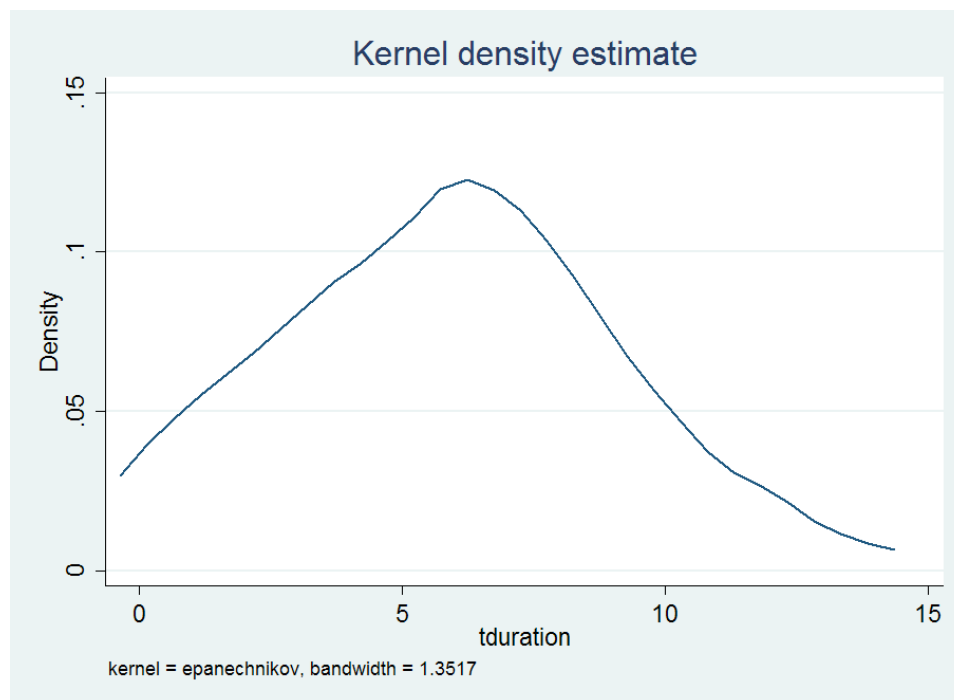


Figure 5.6. Density estimate of treatment duration

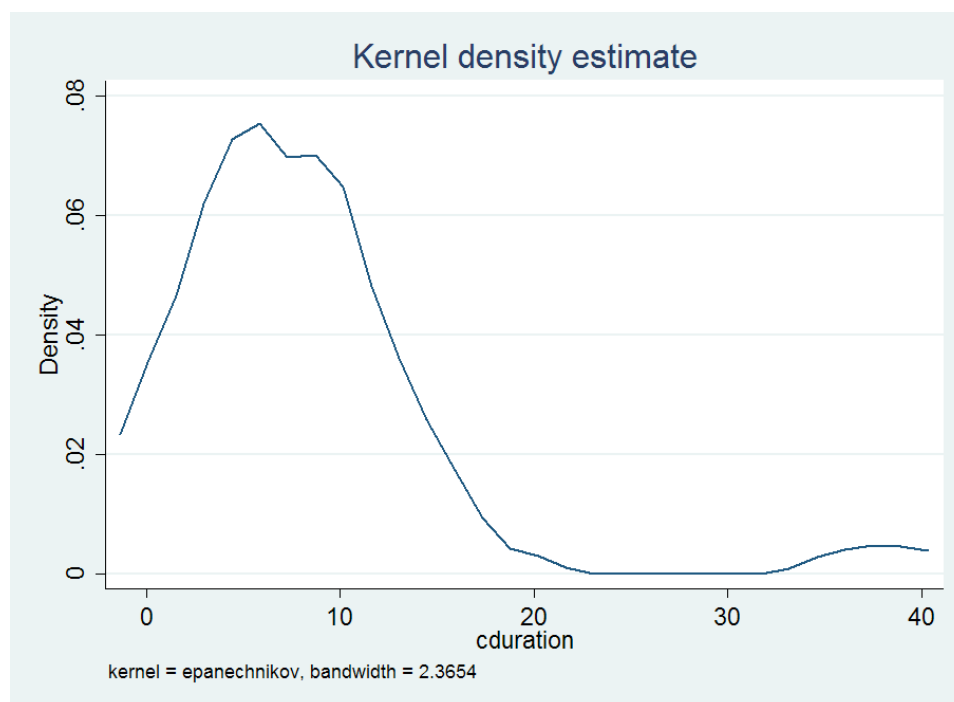


Figure 5.7. Density estimate of control duration.

The distribution for the scores in the two groups also differs in a significant way and brings to light some interesting observations. The density estimate for treatment group in Figure 5.8 shows a beta-distribution to the left that means that a bulk of the subjects achieved higher grades. The slight dips in the graph at regular intervals reveal that the subjects tended to avoid attaining a certain score while following a more regular pattern at other score intervals. This happened due to the design of the questionnaire that contained some relative questions. This means that a person incorrectly answering a question had a high probability of answering a related question incorrectly as well. This might also explain the bi-modality of the control group distribution in Figure 5.9. The two nodes are symmetric with contrasting peaks. The inference here is that the scores of the subjects fall within two defined ranges with 4 being the demarking score. The scores within the two ranges follow a symmetric pattern.

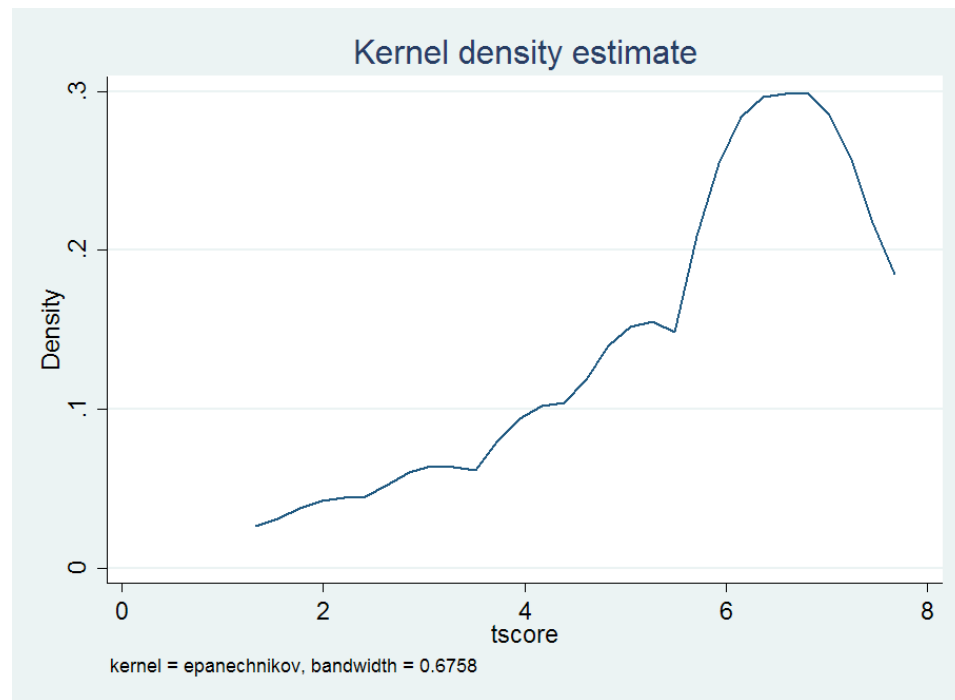


Figure 5.8. Density estimate for treatment score

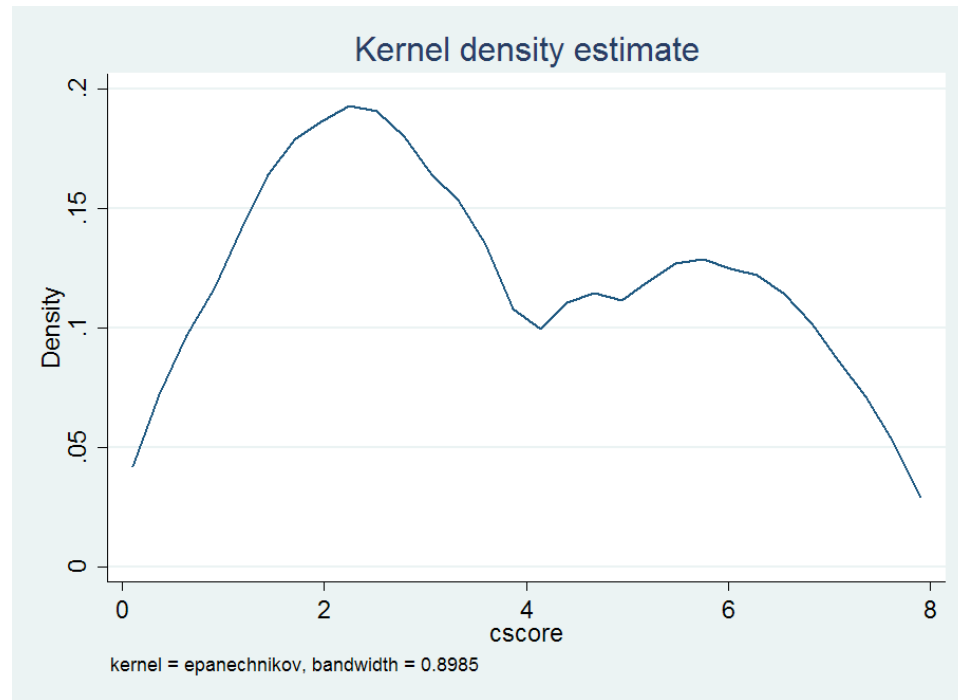


Figure 5.9. Density estimate for control score

5.6 Test for Significance

As stated earlier, the researcher performed an individual group or unpaired t-test with the aid of STATA statistical analysis tool. Then, an F-test was performed on the two data sets to check for equal variance and validate the t-test. If the resulting probability was not significant enough to reject the null hypothesis of equal variance then the t-test was applicable. Otherwise, the researcher would have to account for the difference in variance. This step is necessary since the unpaired t-test is used under the assumption that the two sample sizes are equal and that they have equal variance. According to Hildebrand, Ott, and Gray (2005), the t statistic to calculate whether the means of the two groups in question are different can be calculated as shown in equation 5.2.

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_{pool} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2), \text{ where}$$

$$s_{pool}^2 = \frac{\sum (y_{1i} - \bar{y}_1)^2 + \sum (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (\text{eqn.5.2})$$

The degrees of freedom is given by $2n-2$, where n is the number of subjects in a group. The t-statistic is used to compare the means of data in order to test against a statistical hypothesis. A statistical hypothesis is composed of the null (H_0) and alternative hypothesis (H_a). A null hypothesis is the default assumption about a parameter being tested while the alternative is an assumption that is true if the null hypothesis is rejected (Groebner, Shannon, Fry, & Smith, 2008).

5.6.1 Duration

Two hypotheses that were tested:

- (1) $H_0: \overline{cduration} = \overline{tduration}$ vs. $H_a: \overline{cduration} \neq \overline{tduration}$
- (2) $H_0: \overline{cduration} = \overline{tduration}$ vs. $H_a: \overline{cduration} > \overline{tduration}$

The results for duration are presented in Table 5.12.

We need to check if the two data sets have equal variance. This is achieved by using the F-test that finds the cumulative probability associated with an f value. The F statistic is $5.14 = 7.0168^2 / 3.0931^2 \sim F(29, 29)$. The degrees of freedom of the numerator and denominator are 29 ($=30-1$). The p-value 0.9999 does not reject the null hypothesis of equal variances. Hence, we can use the current t-test that checks for equal variance.

The results show that the probability of the difference of the two means being greater than zero falls above our alpha range of 0.05. But it is still lower than an alpha range of 0.1, which is also a credible value for testing significance. The analysis here falls short of achieving the significance for the set alpha range by a very small margin which can be attributed to low sample size or the irregularity in the amount of effort put in by the subject to complete the test. Since there was no obligation to answer the question, some subjects may have chosen to submit without attributing enough time to the questions, which may have influenced the findings. But in statistical parlance, the results are still credible to be reported as part of the findings.

Table 5.12

Test of Significance for Time Taken to Complete the Test

.ttest duration, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	8.066667	1.281103	7.016893	5.446516	10.68682
treatment	30	5.866667	.5647364	3.093189	4.711651	7.021682
combined	60	6.966667	.7086897	5.489487	5.548582	8.384751
diff		2.2	1.400055		-.602514	5.0022514
diff = mean(control) – mean(treatment)					t = 1.5714	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) = 0.9392		Pr(T > t) = 0.1215		Pr(T > t) = 0.0608		

5.6.2 Score

Two hypotheses that were tested:

(1) $H_0: \overline{cscore} = \overline{tscore}$ vs. $H_a: \overline{cscore} \neq \overline{tscore}$

(2) $H_0: \overline{cscore} = \overline{tscore}$ vs. $H_a: \overline{cscore} > \overline{tscore}$

The researcher then performed the same test on the scores attained from the two groups, the results of which are provided in Table 5.13.

Table 5.13

Test of Significance for Scores Attained on the Test

.ttest cscore, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	3.666667	.3598637	1.971055	2.930663	4.402671
treatment	30	5.766667	.2823682	1.546594	5.189159	6.344174
combined	60	4.716667	.2647797	2.050975	4.186844	5.24649
diff		-2.1	.4574207		-3.015627	-1.184373
diff = mean(control) – mean(treatment)				t = -4.5910		
Ho: diff = 0				degrees of freedom = 58		
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
*Pr(T < t) = 0.0000		Pr(T > t) = 0.0000		Pr(T > t) = 1.0000		

Again we perform the F-test to check if the variances are equal. In this case, $F = 1.9710^2 / 1.5465^2 = 1.6253$. This gives a cumulative probability of 0.90, which is not sufficient to reject the null hypothesis.

It is clearly evident from the table that the result has statistical significance. The probability here is equal to 0.00 that is below the set alpha level of 0.05. Therefore, we can reject the null hypothesis that there is no significant difference in scores between the test taken by the subjects in treatment and control group. This in turn, validates researcher's assertions in the summary analysis that interactive visualization has a positive effect on the perceptive abilities of subjects in comparison to numerical data.

5.6.3 Survey Results

This part details the significance testing of the two survey questions. The first question asked the participants about the apparent ease of using the applet developed by the researcher. The results are given in Table 5.14.

$F = 1.3578^2 / 1.1841^2 = 1.3149$. Therefore, cumulative probability $P(F < 1.3149) = 0.77$. So the t-test for individual pairs can be used in this case. The probability of 0.0003 falls under the set alpha value. Hence the null hypothesis stating that there is no significant ease of using visualization when compared to textual data can be rejected. This is also proved by the confidence interval not covering 0 that means the population mean difference cannot be 0.

The second question asked the participants about their understanding of content provided in either the treatment group or the control groups. This was, in effect, checking the comparative efficiency of using visual content over textual. The results are provided in Table 5.15.

$F = 1.3060^2 / 1.1325^2 = 1.329$. Therefore, cumulative probability $P(F < 1.329) = 0.78$. So, the t-test for individual pairs can be used in this case. This test is also significant

in a considerable way and as such the researcher can again reject the null hypothesis that visualization does not significantly aid in the understanding of financial data. The testing substantiates researcher's claims that there is an effect on the perception of a subject by using different mediums to relay information.

Table 5.14

Test of Significance for Ease of Using the Testing Tool

.ttest ease_of_tool, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	2.466667	.2479031	1.357821	1.959648	2.973685
treatment	30	3.666667	.216202	1.184187	3.224484	4.108849
combined	60	3.066667	.1808121	1.400565	2.704862	3.428471
diff		-1.2	.3289365		-1.858438	-.541562
diff = mean(control) – mean(treatment)					t = -3.6481	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0			Ha: diff != 0		Ha: diff > 0	
*Pr(T < t) = 0.0003			Pr(T > t) = 0.0006		Pr(T > t) = 0.9997	

Table 5.15

Test of Significance for Understanding of the Data Presented to Subjects

.ttest understanding_data, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	2.466667	.2384497	1.306043	1.978982	2.954351
treatment	30	3.4	.2067816	1.132589	2.977084	3.822916
combined	60	2.933333	.1678489	1.300152	2.597468	3.269198
diff		-.9333333	.3156214		-1.565118	-.3015485
diff = mean(control) – mean(treatment)					t = -2.571	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0		Ha: diff != 0			Ha: diff > 0	
*Pr(T < t) = 0.0022		Pr(T > t) = 0.0045			Pr(T > t) = 0.9978	

5.6.4 Per Question Score Significance

Reported below are the t-statistics that test differences between two groups on per question basis. Since the questions were designed to be different in their outlook, they would force the subjects to use different approach to solve for each question. So the results will provide insight as to whether the proposed corollaries hold true for each questions. The results for each question are provided in Table 5.16 to Table 5.22.

Table 5.16

Test of Significance for Question 1

.ttest google_vs_yahoo_Q1, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	9.666667	.0333333	.1825742	.8984923	1.034841
treatment	30	9.666667	.0333333	.1825742	.8984923	1.034841
combined	60	9.666667	.0233696	.1810203	.9199042	1.013429
diff		0	.0471405		-.0943619	.0943619
diff = mean(control) – mean(treatment)					t = 0.0000	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0		Ha: diff != 0			Ha: diff > 0	
Pr(T < t) = 0.5000		Pr(T > t) = 1.0000			Pr(T > t) = 0.5000	

$F = 0.1825^2 / 0.1825^2 = 1$. Therefore, cumulative probability $P(F < 1.329) = 0.5$. So, the t-test for individual pairs can be used in this case.

Table 5.17

Test of Significance for Question 1

.ttest year_fortheg_d_Q2, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	.8	.0742781	.4068381	.6480842	.9519158
treatment	30	1	0	0	1	1
combined	60	.9	.0390567	.30225317	.8218478	.9781522
diff		-.2	.0742781		-.3486838	-.0513162
diff = mean(control) – mean(treatment)					t = -2.6926	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0			Ha: diff != 0		Ha: diff > 0	
*Pr(T < t) = 0.0046			Pr(T > t) = 0.0093		Pr(T > t) = 0.9954	

$F = 0^2 / .4068^2 = 0$. Therefore, cumulative probability $P(F < 1.329) = 0.78$. So, the t-test for individual pairs can be used in this case.

Table 5.18

Test of Significance for Question 3

.ttest true_false_Q3, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	.5333333	.0926411	.5074163	.343861	.7228057
treatment	30	.7	.0850963	.4660916	.5259585	.8740415
combined	60	.6166667	.0632976	.4903014	.4900084	.743325
diff		-.1666667	.1257925		-.4184677	.0851344
diff = mean(control) – mean(treatment)					t = -2.6926	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0			Ha: diff != 0		Ha: diff > 0	
Pr(T < t) = 0.0952			Pr(T > t) = 0.1904		Pr(T > t) = 0.9048	

$F = .5074^2 / .4660^2 = 1.1855$. Therefore, cumulative probability $P(F < 1.1855) = 0.68$. So,

the t-test for individual pairs can be used in this case.

Table 5.19

Test of Significance for Question 4

.ttest overtake_Q4, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	.1	.0557086	.3051286	-.0139369	.2139369
treatment	30	.5666667	.0920187	.5040069	.3784674	.7548659
combined	60	.3333333	.0613716	.4753827	.210529	.4561377
diff		-.4666667	.107568		-.6819875	-.2513459
diff = mean(control) – mean(treatment)					t = -4.3383	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0			Ha: diff != 0		Ha: diff > 0	
*Pr(T < t) = 0.0000			Pr(T > t) = 0.0001		Pr(T > t) = 1.0000	

$F = .5040^2 / .3051^2 = 2.7288$. Therefore, cumulative probability $P(F < 2.7288) = 0.99$. So,

the t-test for individual pairs can be used in this case.

Table 5.20

Test of Significance for Question 5

.ttest net_income_return_year_Q5, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	.5333333	.0926411	.5074163	.343861	.7228057
treatment	30	.9333333	.0463206	.2537081	.8385972	1.02807
combined	60	.7333333	.0575717	.4459485	.6181326	.848534
diff		-.4	.1035759		-.6073297	-.1926703
diff = mean(control) – mean(treatment)					t = -3.8619	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0			Ha: diff != 0		Ha: diff > 0	
*Pr(T < t) = 0.0001			Pr(T > t) = 0.0003		Pr(T > t) = 0.9999	

$F = .5074^2 / .2537^2 = 4$. Therefore, cumulative probability $P(F < 4.0) = 0.99$. So, the t-test for individual pairs can be used in this case.

Table 5.21

Test of Significance for Question 6

.ttest yahoo_return_Q6, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	.3666667	.0894855	.4901325	.1836482	.5496852
treatment	30	.8	.0742781	.4068381	.6480842	.9519158
combined	60	.5833333	.064184	.4971671	.4549014	.7117652
diff		-.4333333	.1162966		-.6661263	-.2005404
diff = mean(control) – mean(treatment)					t = -3.7261	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0			Ha: diff != 0		Ha: diff > 0	
*Pr(T < t) = 0.0002			Pr(T > t) = 0.0004		Pr(T > t) = 0.9998	

$F = .4901^2 / .4068^2 = 1.4514$. Therefore, cumulative probability $P(F < 1.4514) = 0.84$. So,

the t-test for individual pairs can be used in this case.

Table 5.22

Test of Significance for Question 7

.ttest google_return_Q7, by(treatment)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
control	30	.3666667	.0894855	.4901325	.1836482	.5496852
treatment	30	.8333333	.0692046	.379049	.6917941	.9748726
combined	60	.6	.0637793	.4940322	.472378	.727622
diff		-.4666667	.1131235		-.693108	-.2402253
diff = mean(control) – mean(treatment)					t = -3.7261	
Ho: diff = 0					degrees of freedom = 58	
Ha: diff < 0			Ha: diff != 0		Ha: diff > 0	
*Pr(T < t) = 0.0001			Pr(T > t) = 0.0001		Pr(T > t) = 0.9999	

$F = .4901^2 / .3790^2 = 1.6722$. Therefore, cumulative probability $P(F < 1.6722) = 0.91$. So, the t-test for individual pairs can be used in this case.

The test for significance fails to reject null hypothesis for questions 1 and 3, which means that there was no apparent advantage of providing the treatment. The result was expected, as to answer these questions the user has to compare only two values and as such is not challenged in a significant way to affect his/her accuracy. This in turn, renders the use of visualization redundant.

For the questions 2 and 4, the test was statistically significant in a considerable way. The thing to note here was that, the user now has to compare several possible

options that will be better served through visualization than through calculating the values over several years. In such a scenario, it is far easier to reach a conclusion through comparison rather than calculation and as such will benefit greatly through visualization. This explains the high significance for these questions.

Questions 5, 6 and 7 are related to each other and as such the researcher expected to receive similar results for the three questions. As the results show, they are all statistically significant to reject the current null hypothesis. Out of the three, only question 5 forces the user to compare over an entire row of data while the other two request additional information about the same record. Therefore, if the subjects fail to find a correct value for the first question then they would simultaneously provide erroneous results for the next two related questions. This also explains the bi-nodal behavior of the score distribution in the control group.

CHAPTER 6 CONCLUSIONS

The objective of this study was to investigate the educational benefits and cognitive qualities of using visualization. Apart from this, the researcher also wanted to analyze the characteristic features of a good visualization and what separates it from the bad ones. Visualization is, in essence, composed of two parts: the graphic that the user sees and the data that forms the basis for the graphic. The key for a good visual is to convey the underlying data as sincerely as possible but there are no standard methodologies to do so. Data can be represented in a variety of ways and more often than not the designer is more concerned about the appearance of the graphic rather than its integrity. This might result in false bias or incorrect interpretation of results. Therefore, great care must be given into developing a visual that is designed in a manner so as to fulfill its purpose of relaying information to the user. This transfer of information can again be divided into several categories. Sometimes a visual is meant to provide a holistic view while at other times it is used to simplify the data. The researcher took into account all these factors while designing an application to convey financial data. The goal was to develop a tool that fulfilled all the requirement of a suitable visualization. In order to validate the characteristics that formed the basis of the tool, a study was designed to test the effectiveness of the tool in comparison to the usual medium of information. Since

the purpose of the tool was to relay financial information the study would test a subject's cognitive abilities as well as their predisposition to use the tool instead of financial statements.

The results of the study were fairly conclusive as stated during the data analysis. While the study could not categorically reject the null hypothesis that the visualization has no bearing on the time taken to complete the test, the result was significant enough to be stated which meant that there was some effect on the duration of the control and treatment group. The study was able to reject every other null hypothesis initially proposed by the researcher, which claimed that visualization had no impact on the understanding, scores or inclination of the subjects.

Overall the results concluded that the subjects received better grade with visualization than without it. They also understood the content better and were more comfortable with using the visualization rather than financial statements.

6.1 Final Thoughts

The research provides substantial insight into the makings of a good visualization and what impact it has on the user. It is able to correlate the effect of graphical analysis and perception through extensive quantitative study. While it does not categorically state the methodology for designing better visuals, it does lay out guidelines that are influential in this respect. These guidelines were developed through review of existing literature as well as the researcher's experiences while creating and testing the tool. The results of the study further endorsed the researcher's claim. To sum up, while there is no existing matrix for visual systems it is still possible to maximize their effectiveness by following

simple strategies. Every data is unique in its interpretation and structure but is still adaptable in its purpose. Ultimately, a visual should be construed in such a way so as to have the maximum impact on the user in terms of understanding and relevance. Since data extraction and analysis is a common theme in almost every field of study, the scope of visualization is quite large. It can be applied to almost any setting that deals with large amount of data and its significance increases as the depth of information systems increases. So it is applicable in areas beyond those discussed in this study.

LIST OF REFERENCES

LIST OF REFERENCES

- Baldassi, S., Megna, N., & Burr, D.C (2006). Visual clutter causes high-magnitude errors. *PLoS Biol.*, 4(3), 56. doi: 10.1371/journal.pbio.0040056.
- Borgatti, S. (2007). Netdraw 2. *Analytic Technologies*.
- Bruin, J. (2006). newtest: command to compute new test. *UCLA: Academic Technology Services, Statistical Consulting Group*. Retrieved from http://www.ats.ucla.edu/stat/stata/output/ttest_output.htm.
- Catmull E., & Clark, J. (1978, November). Recursively generated B-spline surfaces on arbitrary topological meshes. *Computer Aided Design*, 10(6), 350–355.
- De Nooy, W. Mrvar, A., & Batageli, V. (2005). *Exploratory social network analysis with pajek*. Cambridge: Cambridge University Press.
- Fayyad, U., Grinstein, G. & Wierse, A. (2001). *Information Visualization in Data Mining and Knowledge Discovery*. Waltham, MA: Morgan Kaufmann.
- Fernholz, L. T., & Morgenthaler, S. (2000). A conversation with John Tukey and Elizabeth Tukey. *Statistical Science*, 15, 79–94.
- Friendly, M., & Denis, D.J. (2009). Milestones in the history of thematic cartography, statistical graphics, and data visualization. *National Sciences & Engineering Research Council of Canada*. Retrieved from www.math.yorku.ca/SCS/Gallery/milestone.
- Groebner, D.F., Shannon, P.W., Fry, P.C., & Smith, K.D. (2008). *Business Statistics A Decision-Making Approach*. New Jersey: Pearson Education Inc.
- Helfert, E. A. (2001). The nature of financial statements. In E. A. Helfert (Ed.), *Financial analysis: Tools and techniques: A guide for managers* (pp. 33-44). New York: McGraw-Hill. doi: 10.1036/0071395415
- Hildebrand, D. K., Ott, R. L. & Gray, J. B. (2005). *Basic Statistical Ideas for Managers*. Belmont, CA: Thomson Brooks/Cole.

- Ireland, C. (2010). *Experimental Statistics for Agriculture and Horticulture*. Oxfordshire, UK: CABI.
- Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). Challenges in visual data analysis. *IEEE International Conference on Information Visualization*, 9-16. doi: 10.1109/IV.2006.31.
- Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 1-8. doi: 10.1109/2945.981847.
- Krackhardt, D., Blythe, J., & McGrath, C. (1994). KrackPlot 3.0: An improved network drawing program. *Connections*, 17(2), 53-55.
- Lenth, R. V. (2006-9). Java Applets for Power and Sample Size [Computer software]. Retrieved from <http://www.stat.uiowa.edu/~rlenth/Power>.
- Newsom, D. A., & Haynes, J. (2004). *Public relations writing: Form and style*. (pp. 236). Belmont, CA: Wadsworth Publishing.
- Noirhomme-Fraiture, M., Randolet, F., Chittaro, L., & Custinne, G. (2005). Data visualizations on small and very small screens. *Proceedings of Applied Stochastic Models and Data Analysis*. Retrieved from <http://asmda2005.enstbretagne.fr/>.
- Orenstein, David . (2000). QuickStudy: Application Programming Interface (API). *Computerworld*. Retrieved from http://www.computerworld.com/s/article/43487/Application_Programming_Interface.
- Perer, A., & Shneiderman, B. (2008). Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis. *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, 8, 265-274. doi: 10.1145/1357054.1357101.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27, 832-837. doi:10.1214/aoms/1177728190.
- Shen-Hsieh, A., & Schindler, M. (2002). Data visualization for strategic decision making. *American Institute of Graphic Arts Experience Case Study Archive*, 1-17. doi: 10.1145/507752.507756.
- Shermer, M. (2005, April). Tufte-Feynman Principle. *Scientific American*, 38.
- Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *IEEE Symposium on Visual Languages*, 336. doi: 10.1109/VL.1996.545307.

- Smith, M.A., Shneiderman, B., Frayling, N.M., Rodrigues, E.M., Barash, V., Dunne, C., Capone, T., Perer, A., & Gleave, E.(2009). Analyzing (social media) networks with NodeXL. *C&T' 09 Proceedings of the fourth international conference on Communities and Technologies*, 255-264. doi: 10.1145/1556460.1556497.
- Thomas, J., & Cook, K. A. (2005). Illuminating the path: Research and development agenda for visual analytics. *IEEE Computer Graphics and Applications*, 26(1), 10-13. doi: 10.1109/MCG.2006.5.
- Yadav, Fizi. *Explaining financial data through visualization*. Retrieved April 12, 2012, from <http://www.gardensandmachines.com/AD61600/page/3/>
- Yang, C.C., Chen, & H., Hong, K. (2003). Visualization of large category map for Internet browsing. *Decision Support Systems*, 35 (1), 89–102. doi:10.1016/S0167-9236(02)00101-X.
- Walker, H. M., Lev, J. (1953). *Statistical inference*. New York: Holt, Rinehart & Winston. doi: 10.1037/11773-000.
- Weber J. (1993, April). Visualization: Seeing is Believing. *Byte*, 18(4), 121-128.
- Wise, J. A., Thomas, J. J., Penock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the nonvisual: Spatial analysis and interaction with information from text documents. *Proceedings from Information Visualization Symposium*, 51-58. doi: 10.1109/INFVIS.1995.528686.
- Zimmons, P. & Panter, A. (2003). The influence of rendering quality on presence and task performance in a virtual environment. *Proceedings of IEEE VR*. doi: 10.1109/VR.2003.1191170

APPENDICES

Appendix A Questionnaire

Please answer all the questions provided below with the help of the data links provided on this page. Please note that speed is of importance and as such try to find the answers in as less time as possible without hampering your accuracy.

What is your field of study/major?

Which company has greater assets in the year 2008?

Google

Yahoo

The assets are equal

In which of the following years is the difference between the assets of Yahoo and Google the greatest?

2006

2001

2010

2003

Google posted a greater 'Net Profit Margin' in 2008 than in 2010

True/Cierto

False/Falso

During which financial year does Google overtake Yahoo in terms of total Operational Expenses?

2001-2002

2004-2005

2006-2007

2005-2006

In which year did Yahoo post its greatest Net Income returns?

2004

2005

2006

2010

Based on the answer to the above question, what was the highest ever Net Income return of Yahoo during the period from 2001-2010?

Give the Net Income returns of Yahoo for the year in the previous question

Based on the answer to the question above, what was the corresponding Net Income return of Google in the year Yahoo posted its highest returns?

Give the Net Income returns of Google for the same year as in the previous question

Please tell us about the ease of using the financial statements in the current format?

Very Easy **1 2 3 4 5** Extremely Hard

Please tell us about the perceived understanding of data from the financial statements?

Very Easy **1 2 3 4 5** Extremely Hard

Appendix B Academic Backgrounds

List of Majors in Treatment Group	
Electrical & Computer Engineering	Industrial Engineering
Accounting	Finance
Computer Graphics Technology	Management
Nursing	Mechanical Engineering
Hospitality and tourism management	Biology
Architectural Engineering	Criminal Justice
Human Services	Mathematics
Environmental Engineering	Psychology
Philosophy and Religions	Medicine
Aeronautics & Astronautics Engineering	Financial Planning
Architectural Engineering	French
Mechanical Engineering	

List of Majors in Control Group	
Computer Science	Fine Arts
Health Sciences	Finance
Computer Graphics Technology	Electrical & Computer Engineering t
Nursing	Mechanical Engineering
Clinical Psychology	Marketing

Architectural Engineering	Mechanical Engineering
Material Science and Engineering	Petroleum Engineering
Biomedical Engineering	Physics
Industrial Engineering	Law and Corporate Communication
Architectural Engineering	

Appendix C IRB Approval Letter

To: DAVID WHITTINGHILL
 KNOY
From: JEANNIE DICLEMENTI, Chair
 Social Science IRB
Date: 03/16/2012
Committee Action: **Approval**
IRB Action Date 03/09/2012
IRB Protocol # 1202011906
Study Title Impact of Visualization in Understanding Statistical Data
Expiration Date 03/08/2013

Following review by the Institutional Review Board (IRB), the above-referenced protocol has been approved. This approval permits you to recruit subjects up to the number indicated on the application form and to conduct the research as it is approved. The IRB-stamped and dated consent, assent, and/or information form(s) approved for this protocol are enclosed. Please make copies from these document(s) both for subjects to sign should they choose to enroll in your study and for subjects to keep for their records. Information forms should not be signed. Researchers should keep all consent/assent forms for a period no less than three (3) years following closure of the protocol. Revisions/Amendments: If you wish to change any aspect of this study, please submit the requested changes to the IRB using the appropriate form. IRB approval must be obtained before implementing any changes unless the change is to remove an immediate hazard to subjects in which case the IRB should be immediately informed following the change. Continuing Review: It is the Principal Investigator's responsibility to obtain continuing review and approval for this protocol prior to the expiration date noted above. Please allow sufficient time for continued review and approval. No research activity of any sort may continue beyond the expiration date. Failure to receive approval for continuation before the expiration date will result in the approval's expiration on the expiration date. Data collected following the expiration date is unapproved research and cannot be used for research purposes including reporting or publishing as research data. Unanticipated Problems/Adverse Events: Researchers must report unanticipated problems and/or adverse events to the IRB. If the problem/adverse event is serious, or is expected but occurs with unexpected severity or frequency, or the problem/event is unanticipated, it must be reported to the IRB within 48 hours of learning of the event and a written report submitted within five (5) business days. All other problems/events should be reported at the time of Continuing Review. We wish you good luck with your work. Please retain copy of this letter for your records.