Sameer Rau

Final SBIR Grant - The Data Visualization Project

Project Summary

It has become an increasingly difficult problem for data scientists and the like who deal with large data sets daily with the intent of discovering correlations between several variables at a time while displaying such relationships both computationally and graphically. The issue has grown so large that dealing with large and complex data sets has created a new term called "big data". One way to cut down this problem to a degree would be to interpret findings for several related data sets at once, and furthermore automate the process as much as possible which would cut down the time it would take to look at each individual spreadsheet. As part of the Data Visualization Project, a customer would simply input the data sheets that they wish to be analyzed. The corresponding output would describe to them the nature of the variables that they are dealing with, how these variables are related to one another, and a visual representation of these factors at hand.

This Small Business Innovation Research Phase I project would enable the aforementioned ability to have as many spreadsheets be studied as possible, although a potential challenge would be size limitations as computing capacity constraints should be taken into consideration as well. This project must overcome other such technical hurdles such as the formatting of the spreadsheets and the number of sheets being looked at spontaneously. An assumption early on may be to have formatted data with variable names in the first row and the corresponding values in the each column below the name to make it easier for the computer to differentiate between text and numerical values. It may also be a challenge later on to ensure that the computer can differentiate between categorical (nominal) variables, binary variables, and continuous variables. An incorrect reading of the variable could lead to a suggestion for the wrong test to be conducted.

The goal of the research and development stage is to come up with potential constraints that cannot be avoided (such as computing power) as well as errors that could be faced later down the line so when a certain component of the project has been reached a possible work around or solution will have already been developed. For each of these steps, it may be beneficial to just look at many data sets to get a sense for format, size, and the types of variables that are generally studied. It may also be helpful to ensure that files are being read in correctly and no internal errors exist when importing the spreadsheet.

If the Data Visualization project were to be commercialized, it would be a tool that would not only be used by data scientists, but in all fields. It would be helpful for an analytics team, for example, to just be able to obtain the results of the data at hand and express the significance of the findings to other teams at a company. This will effectively cut down the amount of time each person is required to look at the data sheet in order to find trends among the variables at hand. Furthermore, the decision-making of what variables to study in order to find a relationship would no longer be a necessary task for employees as those decisions would be made by the computer itself. With so many data sets at hand to study in a finite amount of time, the efficiency with which the spreadsheets would be looked at would increase and there would therefore be more opportunities provided for conclusions to be drawn and the time spent by people would be better spent by deciding what subsequent actions need to be taken. The project would enable further technological understanding and potentially open up the idea of using computers to optimize time spent on looking at data thus freeing up employees to allocate that time elsewhere. Societal needs can be better met and gauged as more conclusions are drawn from looking at a greater number of data spreadsheets.

Elevator Pitch

For the proposed Data Visualization Project, the potential customers consist of all people, companies, and organizations whose focus is in the field of data analytics, and more specifically, who analyze several such spreadsheets on a daily basis. With the constantly increasing rate at which data is becoming available to us, there is a great need for a tool to be able to look at as much data as possible at one time. For example, given three spreadsheets that focus on data related to sales transactions from TicketMaster (one focused on pure sales, one consisting of adwords included in ticket ads on the website, and one focusing on the likelihood that tickets will be bought based on venue and location), an analyst would have to go through each spreadsheet individually, looking at all the possible combinations of variables that may be related to each other in a single spreadsheet as well as in multiple spreadsheets. It is arduous to use a statistical package such as SAS or R to analyze one data file at a time using this current process. Furthermore, if the goal is to look at correlation among several spreadsheets there will be an extra task in merging spreadsheets as well.

As mentioned in the previous paragraph, the direct benefit of the Data Visualization Project for the interested customer is that they will just simply have to attach the spreadsheets that they want to be analyzed. The automated process helps shift the focus of the customer from deciding what variables need to be looked at to what conclusions can be drawn from the results of the data tool. Automation is perhaps one of the most important effects of having such a powerful tool to look at so much data at once, and the combination of automated analysis and capacity of data that can be handled that makes this Data Visualization Project a key differentiator from any of its previous predecessors that look at similar data sets.

The Data Visualization Project is an innovation of pre-existing scripting statistical packages such as SAS and R as well as more graphical packages such as SPSS and Tableau. This project will have the ability to compute summary statistics, including the measures of central tendency, as well as various regressions as well the entire process will be automated. In SAS, the code is written as a series of PROC statements before the results of a test of significance is returned. In R, the code is written as a series of commands to obtain various results, whether numerically or graphically. Even if the user is aware of the commands to get certain output, they also have to manually add the library from which those commands originate, and take the time to install the package of it does not already exist on the server. In addition to that, there is not a direct way in which null values can be separated when looking at a numeric variable. Typically, a statistical package like R treats a value such as NA or a blank cell as a value in itself which can mess up how a variable is studied. The goal is for the Data Visualization Project to account for outliers and other data values that may have been misinputed. In terms of layout, the Data Visualization Project will be split up into three sections: the summary statistics for variables in the datasets, potential tests that would be appropriate for analysis (including correlation tests, t-test for independent values, and linear regression), and graphical representations of the key variables contained within the spreadsheet.

Commercial/Societal Impact

The Data Science market as of 2010 is worth roughly 100 billion dollars, a number that has increased by roughly ten percent each year within the past decade. Therefore, The Data Visualization Project falls under a division in analytics that is in need of an innovation that is able to look at a greater amount of data simultaneously that would otherwise fall to the wayside to other data sets or spreadsheets

that may be deemed to be more important. Sure, there are currently existing statistical packaging platforms that do a reasonable job of getting the data analyzed at a respectable rate, however each tool comes with its own shortcomings. For example, Stata, SAS, and R requires an analyst to enter in statements or commands before the actual analysis itself is carried out. Moreover, all three have space constraints, so the size of the spreadsheet can not be overly large, and only one spreadsheet can be looked at any time (it is possible to merge the spreadsheet, but again the final spreadsheet can go over the space limits, and merging spreadsheets can lead to undesired consequences). With the market continuously growing at an exponential rate, the expectation is that similar innovations will be developed, however currently no such product has been formulated in the marketplace. The Data Visualization Project combines automated analysis with aesthetically pleasing graphical components that makes it a unique tool in a field that has several products that do similar forms of analysis in the same way.

The intended market that The Data Visualization Project aims to target include all analysts, from those working in smaller companies or startups to larger companies to educational purposes for students who may be tasked with working with larger data. The basic features of The Data Visualization Project are proposed to be free, so it would directly benefit those working in smaller companies where the funding may be needed to be allocated elsewhere, or students attending college that may not have the money to pay for an expensive tool. The main market driver for this innovation is the premise that the analysis is automated. As mentioned previously, in order to use one of the currently existing statistical packages, one has to have a working knowledge as to how to produce a graph or run a regression. Not only will the demand be high to have a data tool that looks at as much data as possible, but it will greatly benefit those who do not know a great deal of commands in any of the platforms. Most of the time spent learning new libraries and doing the analysis can instead be shifted towards drawing conclusions and making decisions as to what steps to take next.

There are several potential risks in bringing the innovation to a larger stage. One risk would be handling malicious input if one were to import a spreadsheet that has unusual values in a column that would not be expected, and another would be the format of the spreadsheet itself. Currently, for The Data Visualization Project to function as intended, the data has to be placed in a certain way; the columns of the variable names go at the top with all the corresponding values going below each respective column. However, if the data itself is not presented in such a way but the numbers are all over the place, the analysis may still go through and the results returned may not be of use or meaningful. In addition to any formatting issues that may occur, space limitations will still exist in how much data can be seen at any point. A database is being used to store multiple spreadsheets as a means of looking at as much data at one time to discover trends that may exist within spreadsheets. Unfortunately, importing too many datasets or file sizes that get way too large can crash the analyzing process, resulting in the product being unusable. Any one of these factors, or a combination of a couple may result in the individual or company utilizing the tool to abandon it in favor of another currently existing tool that does not have the same flaws.

The revenue potential for The Data Visualization Project is very high if marketed the correct way. In the past year, SAS raked in over three billion dollars worth of revenue, largely due to the greater familiarity that many companies have with the tool over R, even though the latter is free and does not require a renewed license to continue using over the course of several years. Therefore, the tool needs to be directly catered to all companies doing analysis big and small. One way to carry this out would be to purchase Google Adwords, a tool which allows a product to be advertised on the search engine when certain keywords are entered. A certain amount of money is paid to Google each time an individual clicks on the link, however in the long run the reach that Adwords would provide would lead to a greater number of people using the product, and therefore lead to a large amount of revenue. Another way of earning money would be to allow a company to advertise on the website. The Data Visualization Project is expected to operate on a web-based platform rather than a software application. Therefore, another

company can increase their reach when various analysts visit the website, and at the same time money can be earned in order to maintain the Visualization tool. A final proposition of potential revenue would be the added benefits of a membership. If a certain sum of money is paid, the analyst can get an upgrade with more features at their disposal. Possible features could include importing spreadsheets of any format to including more graphics to support the data that has been analyzed. The membership could be charged on either a monthly or yearly basis, depending on how often the individual using The Data Visualization tool. Even though the plan is for the tool to earn potential revenue down the line, the preliminary focus is to establish a large number of users who utilize The Data Visualization tool, and therefore the core tools such as automated summary statistics and basic graphs will always be free.

The commercial opportunity that The Data Visualization Project provides benefits all fields of society. In the field of healthcare, a large amount of data is available regarding the effectiveness of several drugs that cure illnesses. Analyzing the results of the data quickly can lead to doctors definitively prescribing a certain therapy to their patient. In business and marketing, quickly finding trends in a large amount of data can help possibly describe target customers that may be interested in a certain product that a company sells. Similarly, handling a large amount of data relating to education can help determine what engages elementary and middle school students, and what makes certain students perform better academically than others. In sports, conclusions can be drawn as to what enhances athletic performance. In short, large amounts of data are coming in at a rapid rate in all fields, and having a tool such as The Data Visualization Project can help make decision making easier and help shape society for the better.

There are no major environmental or health issues associated with the project. As mentioned earlier, the tool aims to solve health issues by making it easier in the healthcare field to make decisions of how to cure illnesses by looking at correlations of a large sample size of data over time. It is appropriate for people of all ages (since there is no need to know how to do the actual analysis itself) with the only necessity being having a computer, an internet connection for which to import the files to the website, and a passion for finding patterns in large sets of data. Although it can be technically used for analyzing demographics and trends in crime, for example, there is a very small chance that someone would exploit the tool to create chaos. Therefore, there is no expectation currently that the product would need regulation from the government. The only warning to take away, as with all tools analyzing data, is to not make decisions and believe everything based on what the data tells alone. Data can be skewed and may not always be accurate, not to mention the fact that the data may not take into consideration all factors that go into making a final decision.

Although it is expected that there will be other competitors in the industry seeking to do similar work, the product could potentially be used unethically so that one could steal the entire framework and market The Data Visualization tool as a new product in itself. Hacking the server in order to reap the benefits of obtaining a membership could be another possible way that the product could be used unethically. The tool could also be used as a means of introducing bias into the data being looked at. For example, if a company was analyzing the amount of hours that each person worked per week for a full year the spreadsheet could be edited beforehand to show a different conclusion, and the Data Visualization tool could be blamed for misinterpreting the data. Perhaps the biggest unethical decision that can be made is not giving credit for when the analysis has been done using The Data Visualization tool. Such unethical practices can not be easily solved, but basic integrity is expected when using the website. In summary, there are large expectations for The Data Visualization tool to make the lives of analysts everywhere easier and improve society for the better.

R&D Plan

There are a couple of technical challenges that may be faced during the time of bringing the project to market. One main challenge is providing the customer the forms of analysis and visualization that they want to see so that the results will be entirely beneficial to them during the automation process. One workaround to this approach would be to provide multiple options of analysis that the user can select and they will receive the corresponding results, however, the user may be unsure of what they want to view from the dataset that was provided. The correlation statistics, for example, that the user gets may be of use to the person to determine relationships that may be seen between variables, which may be more helpful if they are looking at linear or logistic regression (the latter in order to make any transformations to the variable) but not entirely helpful otherwise if they selected correlation as an option. Of course, the point is moot for clients with a strong statistical background, but it would become a legitimate problem for those without such skills. Regarding some of the technical aspects of the project, identifying which set of factors determine the appropriate type of graph is a challenge. Moreover, some distributions will be easier to determine the correct visualization for a variable and others may have several factors that may cancel each other out, resulting in as accurate a weighting system needed as possible. The system will likely increase in accuracy through machine learning. In order to test out the rules for determining the appropriate graphic, a training set must also be constructed to see whether the different factors are tested and the weightage system works as intended. The annotations that appear on the graph must also be arranged in a clear and concise way, otherwise the comments will clutter the visual and make it unreadable. A similar algorithmic weightage system must be in place to determine how useful a markup on the graph will be.

In order to complete this phase of research, there are several key objectives to keep in mind. First, the algorithmic component should operate in an efficient way. In a project that is planned to read in as many spreadsheets as possible and produce meaningful statistics, a slow moving algorithm will increase the time taken to analyze the several datasets and make the tool not highly desirable to use by the consumer. This applies to both the determination of the correct graph as well as the appropriate annotations that go with them. An effective algorithm will have essential characteristics of the graph predetermined to essentially perform a pattern match with the spreadsheet at hand. The annotations must be catered to the specific person that is receiving the statistics. They should be in spots that correctly indicate the mentioned phenomena and not clutter the graph as mentioned previously. It may be necessary to have greater user interaction with the system to determine what statements they wish to be notified about. The final objective is the creation of the website. The site should be well-formatted and user-friendly so that the person can easily navigate it and know where to attach the datasets to be read into the server. The module should be incorporated smoothly with the other working parts of the project (namely, the numeric analysis and its visual component). In the future, the expectation is that the website be used as a means of raking in profit. This may be done by allowing third-party companies to advertise on the domain, or give membership benefits to analysts working in companies that make use of the tool often. Such benefits may include a more comprehensive list of annotations to be displayed on the graph, and greater variety of graphs to be returned to them, or an even more sophisticated summary statistics regarding their dataset. Thus, there are many things that need to be kept in mind as the project progresses to its final state.

There are several innovations that make The Data Visualization Project stand out from its competition. First, the tool converts raw data contained in multiple spreadsheets into statistical analysis automatically without the user having to have a strong working knowledge of a package such as R or SAS (such statistical packages also cannot handle looking at more than one dataset at a time). Next, the visual mediums that are returned to the user are based on the distributions of data between the variables. For example, if a variable is composed of percentages of nominal factors, the tool makes the decision to present the information in the form of a pie chart, as opposed to a line graph or bar graph. Although there is no foolproof strategy that would ensure the most optimal chart is guaranteed to be returned, the algorithm that makes this decision is based on a weighting system. If a certain number of factors that

make up the characteristics of a certain form of chart are seen, the more points will go towards that graph being shown to the user in the results section. As the computer continues to gain knowledge of what the general makeup of a dataset looks like and patterns between variables within each sheet are found, the goal is that the choice of graph made will become more accurate.

Another added benefit the customer gets by using the tool is first-hand feedback as to what factors they should keep in mind with the spread of certain variables in graphs with added annotations. The annotations will also be based on a similar weightage system to determine how important the note will be of use to the user. For instance, if a variable is seen to have a relative strong cluster of observations around a particular value and there are two extraneous outliers, the user will first be notified with arrows pointing to the extreme values and a description of what the observation values are and how far they deviate from the general mean. If a scatterplot is being looked at and there are high leverage points but no outliers (a secondary check with a smaller weight than an outlier), the customer will be similarly notified of the point(s) at hand that are causing the line of best fit to be slightly altered as a result. The same system carries over to bar graphs. The user will be informed of the shape of distribution as it relates to the median and mean of the data. If the data is skewed to the left, the mean is smaller than the median, which signifies a greater number of smaller observations and smaller range of values for a variable. If the data is found to be relatively symmetric, the corresponding mean and variance will be laid out as in a traditional normal distribution. When the number of observations falls under a certain size (currently at 30), a warning is presented to the person letting them know the results may not be entirely accurate which such a small sample. If none of the above are applicable to the data at hand, other factors with smaller weights such as unusual findings of summary statistics will be labeled on the graph. The algorithm also takes into consideration how many labels are currently being presented on the diagram. If the graph is starting to look fairly clustered (at a certain threshold), no more generated statements are produced. Moreover, the labels are placed around the graph so that too much cluttering in a certain area does not become an issue. An example of what the visual component may look like can be seen in figure 1 below.

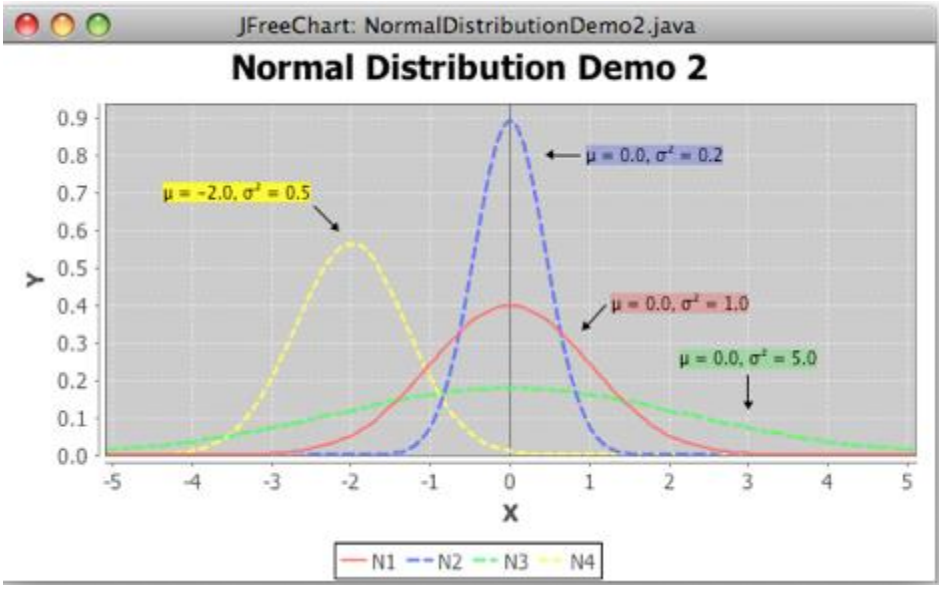


Figure 1: A line graph revealing a symmetric distribution between four variables, with a provided guide of the mean and variance.

Timeline

September 2016

- Determine the field in computer science for which the project will be of use.

- Also decide the scopes of the project and preliminary constraints that may be faced when beginning to work on it. In other words, what aspects of the project can realistically be completed over the course of two semesters.

- Look at previous capstone projects of relevance to gauge an understanding of what has been explored, previously explored, and not yet explored.

October 2016

- Continue to do research on how the project will come into fruition, including potential libraries that may be of use for components of the project.

- Begin working on the preliminary modules outlined in the design document, including both the parser for the .CSV files and the integration of the SQL database in Java.

- Determine which fundamental statistic analysis should be included in the tool, and create corresponding methods to evaluate them.

November 2016

- Continue to add more statistical methods as necessary to make the project more comprehensive.

- Begin working in the integration of the visual component. Test to see whether the graphs created in JFreeChart match the numeric analysis observed in the dataset.

- Present to the professors and class of the entrepreneurial aspect of the project (how The Data Visualization Project is projected to earn revenue on the service) as well as the technical components of how all the individual parts work together.

- At the earliest demo, present the spreadsheets being read in and parsed by the ReadCSV library and the arrays exported to the MySQL workbench (with some hard-coding involving the creation of the table that the data will be placed in). Also, show some preliminary statistics being read, including the measures of central tendency (mean, median, and mode) as well as correlation statistics between variables that have a strong relationship.

December 2016

- Present the 30% demonstration to the professor with a more refined version of the preliminary stages of the project. The version will not be hard-coded, and the process will be automatically run through.

- Demonstrate a basic version of The Data Visualization tool functioning with algorithmic components, including determining the appropriate graph to send to the user and the annotated comments on the graph regarding trends that are being seen in the data.

January 2017

- After returning from break, display a more comprehensive version of the project with the algorithm being slightly more sophisticated. The decision of which graph to choose to illustrate the distribution of a variable will be based on predetermined factors that are deemed to be essential and each factor will be given an aforementioned corresponding weight. The annotations will have a larger directory of potential statements that can be printed out on different types of graphs and also use a weighting system. This algorithm will improve through machine learning.

- Begin working on the website module outlined in the design document, with the specified fields to allow the user to choose which statistical analyses they wish to see displayed, and an area for the spreadsheets to be attached.

February 2017

- By the end of the month, have all of the coding parts of the project mostly complete and integrated with the exception of a few tweaks/modifications.

March 2017

- Continuing to make minor changes to the project for all the individual components. Also try to improve the complexity of the algorithm if possible.

- Add new graph types and statistical methods as there is time to work on them.

April 2017

- Have the entire project completed and ready to be presented at the Science and Engineering Hall (SEH).

Revision

After showing my writing to my peers, I have received mostly positive comments content-wise as to what the project covers. There are however a couple of things that they have pointed out that need work. The most common response I have heard is the tendency to have too many things to say in a sentence. This is something that I've been trying to work on. It is much easier to comprehend when sentences are broken up. The other feedback that I've been given is expressing myself in a complicated manner when it can be say in an easier way. One potential fix for this problem would be to reread my writing as if someone else was looking at the topic from the outside. It should be as easy for them to understand and follow along as to what is being said. From the beginning of the semester, I have been trying to organize my thoughts more cohesively. Previously, I had sentences that belonged to one paragraph of emphasis in several different places. I feel that I have improved slightly in that regard. Moving forward, I will try to take the feedback and improve in these other areas.

Design Document I

1.1 Context viewpoint

The Data Visualization Project aims to take several data spreadsheets to be used as input and discover patterns and trends that are found in the variables within the sets automatically without requiring the user to have to provide anything but the spreadsheets themselves, an important step towards solving the big data problem. At the current state of the design, the expectation is that the datasets follow a specific layout: the first row containing the column names that will represent the categorical variables and the rows that follow being the measured values of the said variable. The users and clients of this product are expected to follow this format when inputting the data into the program. This step marks the only real direct interaction that the consumers will have with the software, as the process is largely automated. The only other step in which the software will prompt the user to respond is the methods of analysis they wish to be conducted. The user has the opportunity to select certain forms of statistical analysis that they want to see from the variables in their spreadsheet, and the program returns only those analytics. The interaction between the client and the server can be seen in figure 1 below.

1.1.1 Design concerns

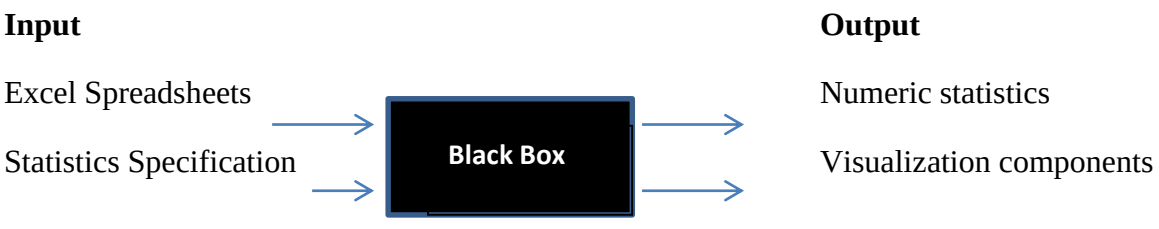
Potential design concerns that may arise when the user interprets the results from their spreadsheets include a potential system overflow from the size of the file the client may attach as input. Although the goal of the project is to be able to look at as many datasets at one time, size limitations will continue to be a constraint moving forward. Currently, a SQL database is being used to store each datasheet in its own table, both to avoid inner joins within files that would lead to a single massive file, and for organizational purposes. Another concern is by asking the user what methods of analysis they wish to see, they may not have an answer to the question. This can be due to a number of reasons, but most likely that the person may not have a strong enough statistical background to understand what each tool does in its calculation, and therefore not being able to comprehend the nature/distribution of the variables at hand to make such a decision (in the case of choosing correlation among variables in a dataset that aren't related to each other). Finally, the project also returns a corresponding visual graphic that represents what is seen in the numeric values. Although a weightage system will be used to allocate points towards choosing the appropriate graph to return to the user in the results section, there is a chance the incorrect graph could be chosen by the computer, especially in instances of distributions that have not been encountered before. The hope is that through machine learning, the decisions made will become increasingly accurate as the familiarity with spreads of variables by the computer improves.

1.1.2 Design elements

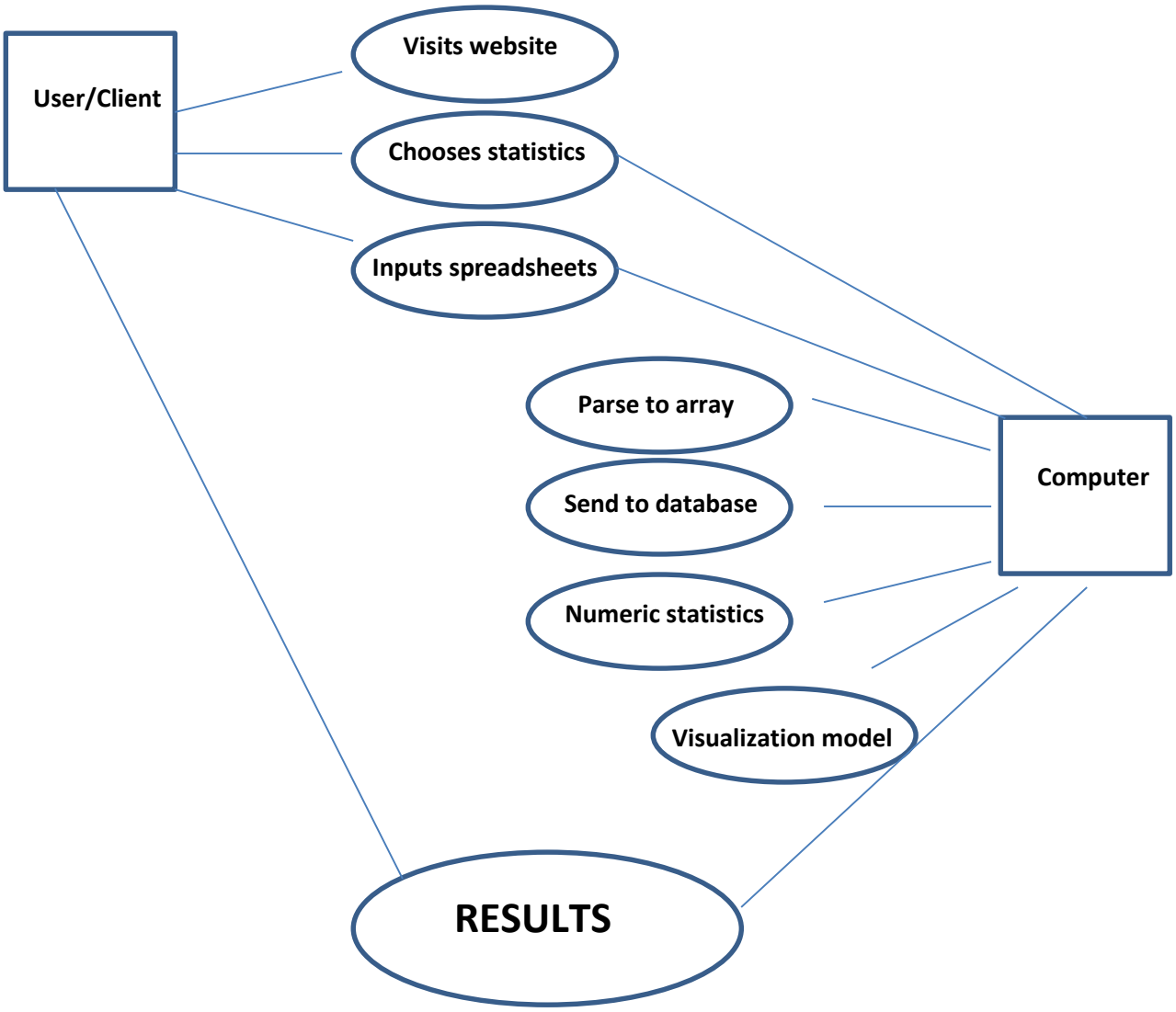
The chronological order in which the software operates is as follows: the user accesses the software by visiting the website domain for which it resides, and enters the appropriate information that they are requested, including the summary statistics that they wish to see. The spreadsheets that the user sends, assuming they follow the correct format, are downloaded and enter the program. The program launches, and begins by parsing each column into its own individual array that will be combined into a single large matrix, with the first row being the names of the columns from the spreadsheet, and the rest of the cells representing the corresponding rows of data values. The two-dimensional array is then sent to the SQL database to be stored. Not only does the SQL database enable more datasets to be stored at a single point, but it makes the retrieval of values easier. The analysis that the consumer wished to see in the final results is then calculated, which may include the measures of central tendency (mean, median, and mode), range, interquartile range, correlation, and outlier values. With the numeric values being calculated, the visualization component is run. If the user chose correlation as one of the statistics to be shown, for example, a scatterplot featuring two variables that were found to be highly related will be shown with annotations reflecting specific observations that are seen (for example, extreme data points or unusual summary statistic findings). If the dataset contains nominal variables with percentages that add up to a whole, a pie chart will be shown if they are all part of a common group. Once both the numeric and visual components have been obtained, the results will be printed on a separate page of the website that the user can then look at and copy for reference. The whole sequence can be seen in figure 2 (shown below).

1.1.3 Example languages (shown through a black box diagram and UML use case)

Black Box Overview (Figure 1)



Use Case Diagram (Figure 2)



1.2 Composition viewpoint

1.2.1 Design concerns

There are a few concerns involved in ensuring all of the submodules are functioning effectively. Since the website requires an online connection, there may be a need further down

the road to make the software an application (with a small payment required to get the application version). This would enable users who frequently use the software to be able to use it whenever they need to without any problems if their internet were to go out. This workaround would also be a potential solution to high internet traffic on the domain. Another concern would be a spreadsheet being attached that does not fit the format guidelines. In such a scenario, the parser would not be able to do its job as intended and values would be processed in an incorrect way, which would also affect all other submodules that interact with it. The values retrieved from the database would be off and throw off the corresponding statistics estimates and final visualization model. A final concern would be the aforementioned exceeding of size constraints when the files are sent to the database. An overly large file or too many files being exported to the database could cause the server to crash or slow down the processing time it takes for the analyzing to get done.

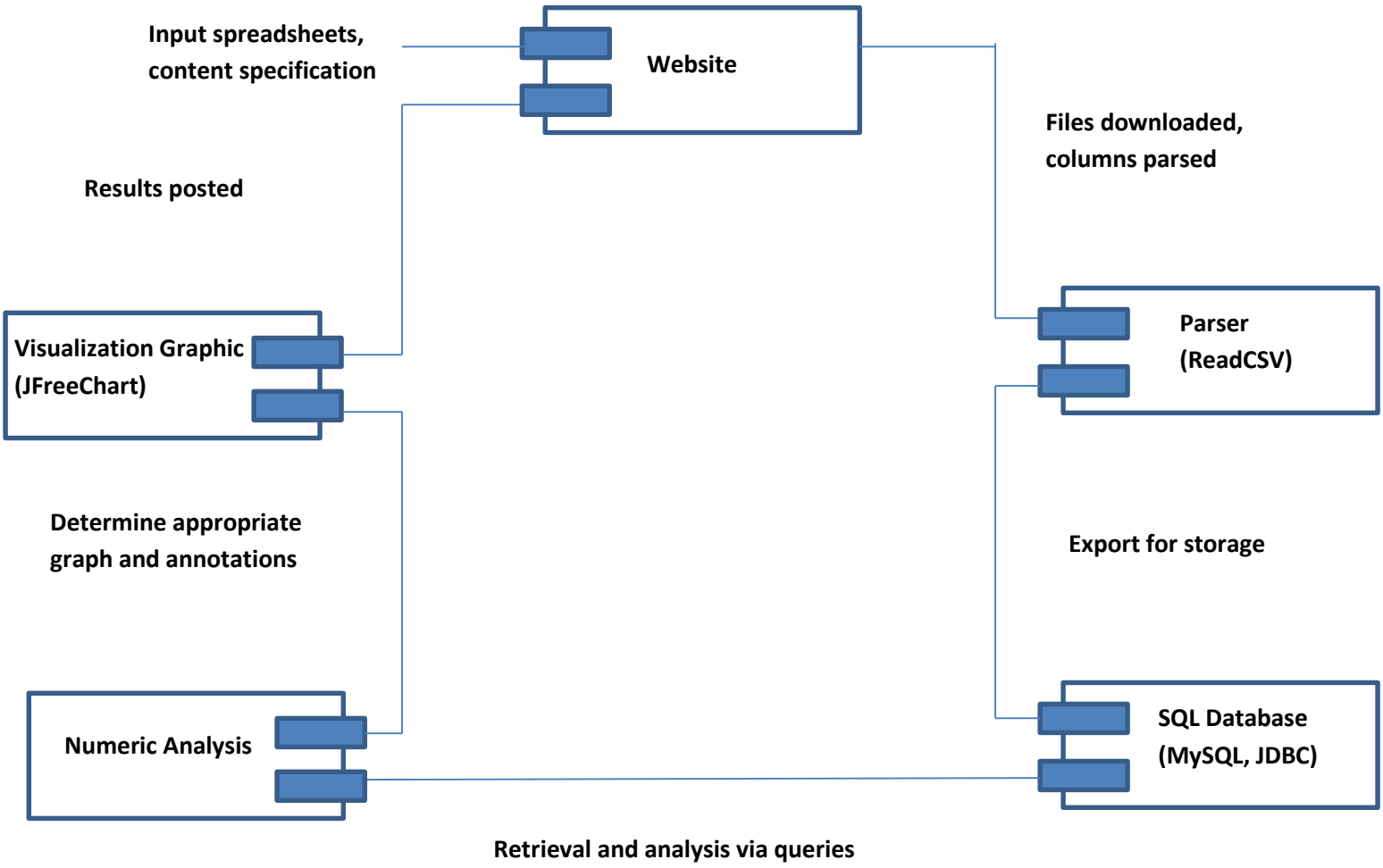
1.2.2 Design elements (including 1.2.2.1 functional attributes and 1.2.2.2 subordinates attributes)

The Data Visualization Project can be divided into five subcomponents that contribute to the final functional model. The first part is the website, which is where the software itself resides. Although the website has not yet been constructed, the layout of the website is expected to include several components traditionally featured in graphical user interfaces. There will be JButtons to allow the user to choose the types of statistics that they want to see in the final results page. There will also likely be multiple text fields and corresponding buttons to allow the user to browse their computer for the desired .txt and .xls files they want to be analyzed. The second component of the project is the parser, which makes use of the ReadCSV library in Java to break down each column contained in the spreadsheet into its own unique array. For non-numeric variables, the string arguments are converted into doubles in order to do the analysis. These arrays are also used to determine the datatype of each variable, which allows the third component to work fully, the SQL database. The SQL database is the intermediary piece of the entire process. It is being used as a place to store the data in one compact place, but also easily retrieve only the necessary values to do the numeric analysis (the fourth component) using various SQL queries through the JDBC driver. The final component is the visualization tool, which is constructed with the help of the library JFreeChart. In making the decision as to which graph is the most optimal to choose to represent the distribution of a variable, the JFreeChart enables several different varieties of graphs to be constructed. The relationship between all modules can be seen in figure 3 below. Currently, the parser and SQL database have completed the development, testing, and integration stage. The numeric analysis has been tested and integrated but has not completed development. This is expected to be finished by late December. The development of the visualization graphics has recently commenced, and is expected to be

completed with the decision-making algorithms in place by late February/Early March. The website will begin being worked on around January and continue until the end of February.

1.2.3 Example languages (shown through a UML component diagram)

UML Component Diagram (Figure 3)



1.3 Logical viewpoint

1.3.1 Design concerns

In some of the predefined libraries being used for the project (such as ReadCSV, JDBC, and JFreeChart), abstraction has been used to implement certain functions whose method

summary has already been created. For example, in the class that constructs graphs in JFreeChart, private methods are used to protect the information contained within them (such as creating the dataset content used to build the graph). However, in the other classes public methods have been used to call functions in other classes (i.e. the JDBC driver calling the class automate to get the server connection). In these methods, the code is much longer and there is a concern that as modifications are made to the code and more content is added, there is some repetition in how each row of data is being read. Therefore, it may be logical moving forward to break more of the code contained in the methods within ReadCSV into a greater number functions (perhaps marking some of them as private) to make the code more concise and compact.

1.3.2/1.3.3 Design elements and Example languages (using a UML class diagram)