CSCI 4243W Writing 4

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 weightage 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 in to 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 racking 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 weightage 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.

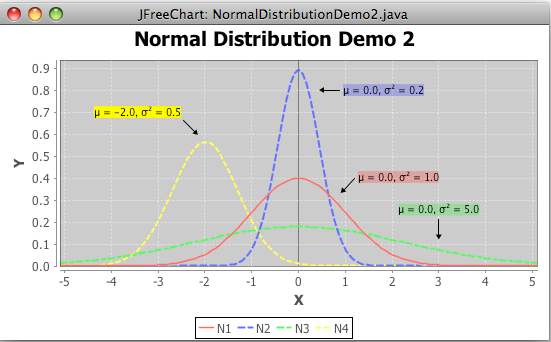
Another added benefit the customer gets by using the tool is firsthand feedback as to what factors they should keep in mind with the spread of certain variables in graphs with added annotations. The annotations will also be based on a similar weightage system to determine how important the note will 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 weightage 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).