Peter Mattoon

Design Document

The overall plan for this project is to create a data analysis and visualization tool built for people without any real background in statistics or data analysis. This goal means that the project aims to be able to take in data from an excel spreadsheet or csv file and then output to the user an annotated graph that best fits the data provided with the possibility of further analysis on request from the user. There are a few different requirements that need to be fulfilled to make this project possible. The first is obviously to be able to read in data from a spreadsheet or file and turn that into a form that can be analysed. The second is to determine how that data should be presented, meaning which graph should be used. After that the graph will need to annotated to best inform the user and any further analysis will need to be performed and this will all need to be presented to a user in an easily understable form on a easy to use website. The project is trying to bridge the gap for people who have collected data in some form but do not have the knowledge to understand it by themselves and are not willing to pay to have someone else do it for them. The world is getting ever more data driven and this project will help more people understand what it means.

Users:
- Everyone - The website will be available for anyone to use

Target Users:
- Small Businesses
  - Use Case: The website can be used to help understand data collected by a small businesses in an inexpensive and easy to understand manner
- Students
  - Use Case:The website can be used to do easy understandable research for projects and reports
- Amateur researchers/writers:
  - Use Case: The website can help someone do inexpensive and easy to use research without a lot of background knowledge on whatever interests them
- Government
  - Use Case: The website will be able to analysis data sets many of which are created by the government and help them to understand them

Design Components
- File reader
  - Must be able to read and understand excel and csv with a minimal amount of formating
  - Must be able to take data from file and put it into an array or database
- Graphing Algorithm
  - Must be able to determine from characteristics of the data what is the correct way to graph it

- ○ Will need to be able to determine between line, scatter, pie and bar graphs
  - ○ Should be able to determine when it is not approapriate to graph at all
- ● Annotation Algorithm
  - ○ Must determine the best way to annotate the graph
    - ■ Where annotations should be place
    - ■ How many should be placed
    - ■ How annotations differ based off of data and graph type
  - ○ Should determine if further analysis is needed
- ● Analyses
  - ○ Should be able to perform analysis on a given set of data
- ● Display to webpage
  - ○ Needs to be displayed in a understandable and well formatted manner
  - ○ Cannot be overwhelming to the user
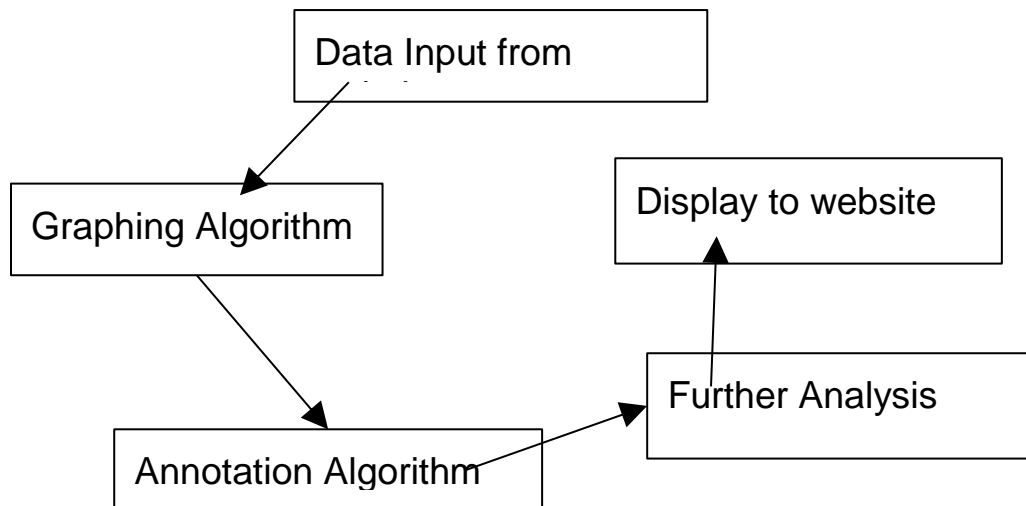
Each of the components will be its own class and they should all run sequentially through the further analysis steps will sometimes be skipped. Also the analysis section will be called by both the graphing and annotation classes to understand the data and to perform their own tasks.

Timeline
- ● September
  - ○ Decide on the focus of the project and research past projects in this field
- ● October
  - ○ Begin creation of project
- ● November
  - ○ Continue to write code for project focusing the most basic features needed to build later innovations
  - ○ Present on the overall plan of the project

- Early December
  - Demonstrate to professor Simha
- Mid December
  - The creation of a rule based version of the graph and annotation algorithms
    - This will be presented to professor Wood
- Mid January
  - A more complete version of both of the algorithms
    - This will be accomplished with many experiments to find the most important factors and how to weight them along with the training data created to teach a machine learning algorithm
    - This will be displayed to professor Wood
- Late January
  - The creation of a basic website
- Mid February
  - Integration of all code
- Till the end
  - Continued refinement of all of the algorithms and features
  - Expansion of current features
  - Addition of new features as time permits
- Final Project Presentation