CSE 584

Summer 2015

41001

CPT Internship Report

Amazon

Sai Kishore Petla

ASU ID: 1207656017

Computer Science

Robert J Moore

rjmoore@us.ibm.com

408-927-1403

--------------------------- ----------------

Mentor’s Signature Date

--------------------------- ----------------

Faculty Chair/Program Chair Signature Date

**Company Information:**

“IBM is a global technology and innovation company headquartered in Armonk, NY. It is the largest technology and consulting employer in the world, with more than 400,000 employees serving clients in 170 countries. IBM offers a wide range of technology and consulting services; a broad portfolio of middleware for collaboration, predictive analytics, software development and systems management; and the world's most advanced servers and supercomputers” [1].

“The IBM Research Accelerated Discovery Lab allows IBM researchers and partners to work together to tackle the most difficult big data analytics problems across a variety of industries and domains. The technology and resources in the lab enable business analysts and scientists to make new discoveries in their fields more easily, and at a more rapid pace, giving IBM researchers the chance to constantly improve the technology used for discovery” [2].

**Background about the problem:**

I have developed a web application which facilitates face to face collaborative data analysis called “DataStroming”. “DataStroming” sessions are also called as data brainstorming sessions where people from diverse backgrounds come together to analysis a specific dataset. It is normally a 2-3 hours long face to face sessions among a group of researches, data analysts and SMEs (Subject Matter Experts).

The general goals of datastorming sessions are data exploration, data analysis, developing business intelligence and fostering cross-disciplinary learning. The analysis that they perform can be exploratory, or seek solution a specific problem. The DataStorming is a generic procedure where we can use any kind of dataset, including structured data, unstructured text, audio-visual recordings or even specimens.

As of now there is no efficient tool which facilitates the datastorming sessions at IBM. There is need for an efficient data analysis tool which has the following features,

* It must enable datastorming participants to explore data and test hypotheses together in real time.
* It should provide quick, easy-to-use visualizations and analysis methods for multiple types of data.
* It should be lightweight and provide quick analysis results.

This web application will facilitate researches, data analysts and SMEs (Subject Matter Experts) at IBM research to get a quick understanding and overview of the dataset and can implement the statistical analysis models on any kind of data to extract knowledge or patterns out of the dataset. It will also be helpful for anyone to get an initial feel and look of the dataset by exploring and generating quick visualizations.

**Focus of the effort:**

Initially I did research in deciding which programing language should be used to develop this tool and after initial analysis I came to understanding that it can efficiently developed in two languages namely, R and python. Both these languages have many packages using which we can develop effective visualizations and implement analysis methods to the data.

But R language in particular has an advantage of providing quick results and because of this features many scientists and researches use R to compute results of their research. It has many packages/libraries which provide interactive visualizations and implement statistical models. Few important packages I have used in my project are,

* ggplot2 (package which is used to plot graphs and charts)
* shiny (package for developing web applications)
* plyr (package for splitting, applying and combining data)
* stats (package for statistical calculations and random number generation.)
* car (package for computing generalized linear models regression)
* cor (package for computing correlation in data )
* googlevis (package which is used to plot google visualizations)
* rCharts (package which is used to generate interactive visualizations)

One of the important package that I have used is Shiny package which is a web application framework for R. Using this framework along with JavaScript and different packages in R I have developed data analysis web application. During this time I got a chance to learn about reactive programing concepts in R shiny and also learnt how to use JavaScript for generating interactive visualizations.

While implementing the tool I have learnt application and implementation of data mining techniques and statistical models like k-means clustering, hierarchical clustering, Correlation, Association rules mining and regression techniques such as logistic generalized linear model, poisson generalized linear model.

The concepts that I have learnt in my graduates courses helped me a lot to develop an effective web application for data analysis. Concepts of Data Visualization course work has enabled me to implement effective interactive visualizations in tool. In this course I have learnt JavaScript and it came handy while developing the front end of the application. I have used the concepts of data cleansing, data preprocessing, transformations and visualization that I have learnt in this course in developing the tool.

Courses such as Data mining has helped me in implementing data mining techniques and also understanding statistical models. It is because of this course I could easily implement K means clustering, Correlation and Generalized Linear Model regression in the tool. Other courses such as Distributed Software Development (DSOD) has helped me in using Rest API concepts and web application development.

**Results:**

In the data analysis web application that I have developed has mainly three components namely,

* Data Exploration :

Using this component users have the capability to see a sample set of records of the dataset and they can also perform search/query on each column individually or on the entire dataset. They can also see a subset of data (few columns) and save that subset of data for future reference.

* Data Visualization:

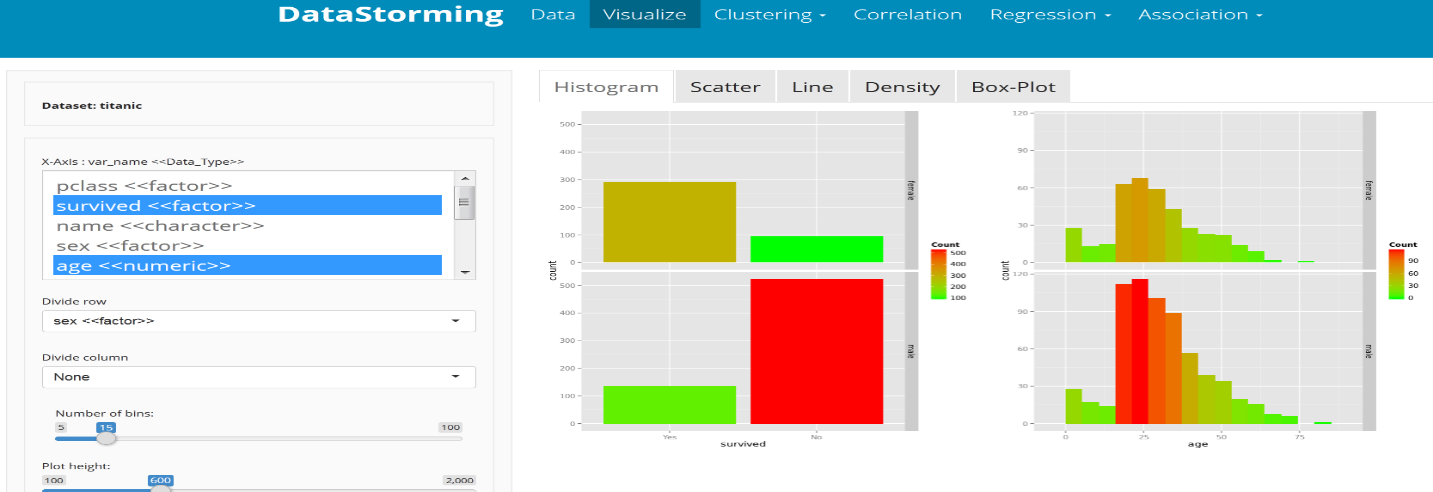
Using this component users can generate easy and quick visualization. These visualizations are interactive and are very helpful in providing a detailed overview of the dataset. Below are few visualizations that we have implemented in the tool,

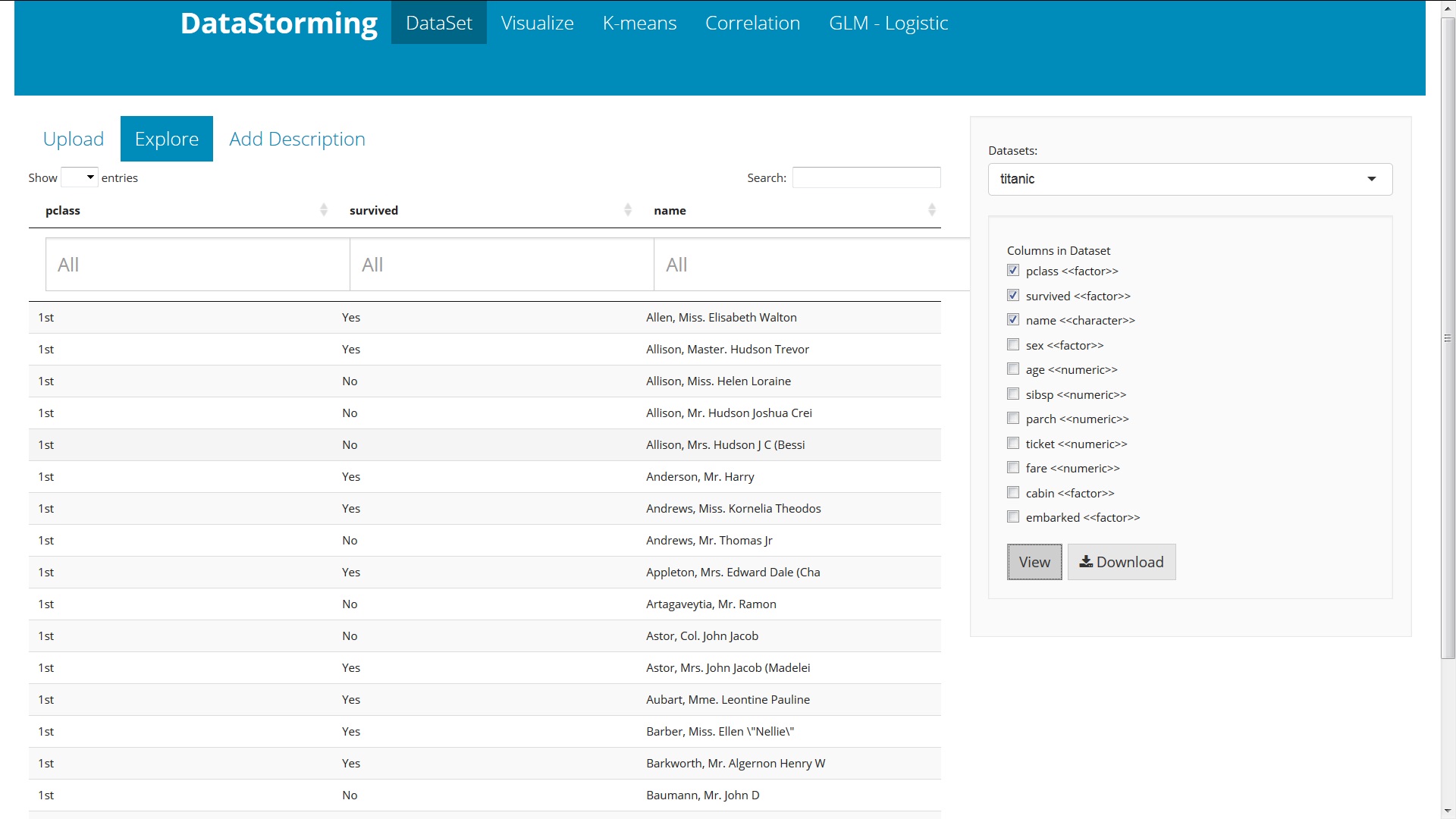
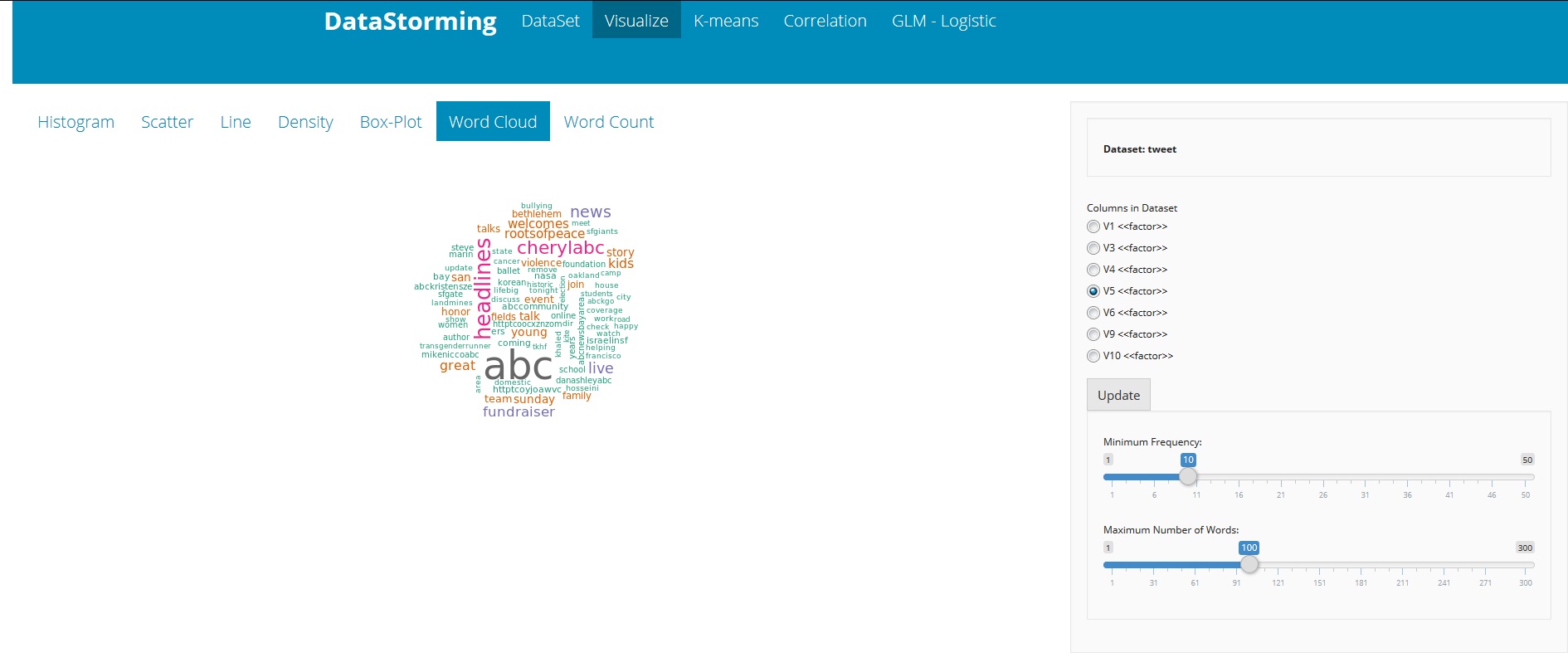
* Histogram
* Scatter Plot
* Line Graph
* Box Plot
* Density Plot
* Word Cloud
* Word Count
* Analysis Methods:

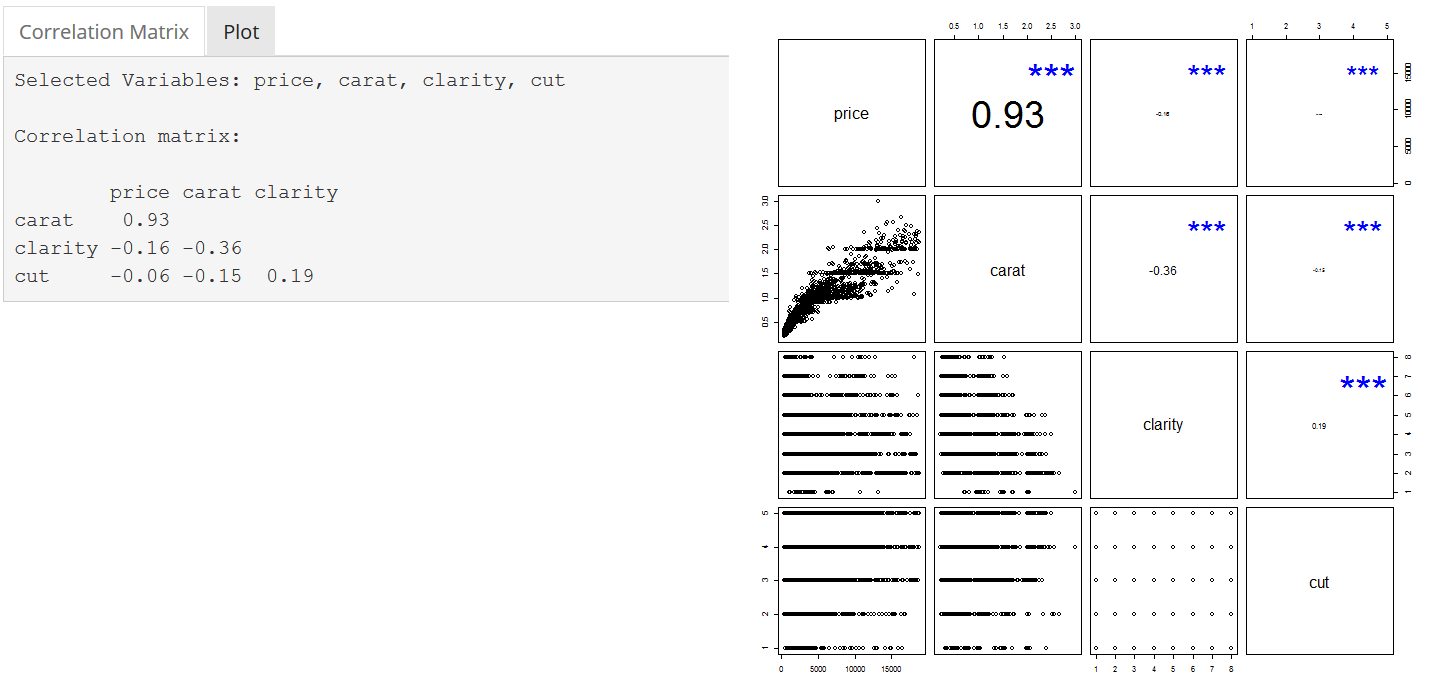
This component gives users of the application the feasibility to apply analysis methods and deduce interesting patterns and extract knowledge out of the data. Below are few Data mining techniques and analysis methods that we have incorporated into the tool,

* K means Clustering
* Correlation
* Logistic GLM Regression

We can see the three main components of the tool in the below screenshots of the web application,







These screenshots are the results of demo sessions where researches at IBM tested the application with datasets of different kinds like structured and unstructured tweets data. During the demo session they found that this tool to be very convenient and easy for them to do data analysis and the feature of providing quick and easy visualizations has helped them in getting the overview of the data without any significant effort. They were also able to apply analysis methods to deduce the patterns and extract knowledge from it.

After development the tool with the above three components we have conducted few demo data brainstorming sessions so that the real users can get a chance to look at the tool and give us the valuable feedback. During these demo sessions we have collected the initial user feedback and according to that we have done modifications to the tool more efficient.

**Learning experience:**

How did the internship contribute to the learning experience?

How did you apply knowledge from courses to the internship?

Did you gain any insights from the internship?

What skills did you develop as part of the internship?

I got great hands on experience on,

* R Programing
* JavaScript
* Statistical Models
* Data Visualizations
* Java J2EE
* RDMS programming
* REST API development
* Experience in Data science, including data integration and ETL (Extract-Transform-Load)

Importance of testing

Clean and efficient way of coding

Team skills

Communication skills

Importance of documentation.

Implementation of Agile methodology

**References:**

[1] https://www.linkedin.com/company/ibm

[2] http://www.research.ibm.com/client-programs/accelerated-discovery-lab/