Jhon Dela Cruz, Abhimanyu Sachdeva, Rishi Jain
jdelacru@umd.edu, asachde8@terpmail.umd.edu, rishijn@terpmail.umd.edu

INFM 600 – Project Summary

Topic: Stack Overflow Data Analysis
Data Source: Kaggle
Data Source Link: https://www.kaggle.com/
Data URL : https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey

Motivation:

Through the entire fiscal year large Multinational Corporations (MNCs) keep hiring professionals to balance, create and manage their Business Models, IT Infrastructure, products and services. Technical professionals play a crucial role through the creation, management and support of the organization's IT infrastructure, product design and simulation, risk protection, cybersecurity measure and so on. We choose this topic for our research analysis to provide statistical analysis over a huge developer community across the globe as well as provide an optimized solution to the companies for accurate and productive hiring for right candidates as also some extended analytics to help provide organizations opportunities to expand and optimize further.

Dataset:

This dataset contains the survey results of approx. 20000 developers having information about their favorite technologies to their job preferences. This dataset has 129 columns and 98,855 rows of qualitative data about the developers. This 2018 Developer Survey results are organized on Kaggle in the following 2 tables:

1. **survey_results_public** contains the main survey results.
2. **survey_results_schema** contains each column name from the main results along with the question text corresponding to that column(METADATA).

Source:

Kaggle is an online platform to compete with others in competitions which are based on data sources, datasets, data manipulation over machine learning tasks. The dataset we derived from Kaggle presents Annual Developer Survey results for the Stack Overflow Users (developer community) getting specifics about their background, skillsets, knowledgebase, technology expertise, job satisfaction/expectations and income statistics. The dataset has 129 columns and 98,855 rows of qualitative data about the developers.

Target Audience:
For our project, our target audience are decision makers at companies, such as CEOs, hiring managers, or investors. The types of companies that we are targeting are mostly tech companies or any company in need of coders or are looking to hire out of the country workers.

However, we performed analysis to address few business research questions, the results can be utilized by a large variety of users and customers for various different purposes.

Jhon Dela Cruz, Abhimanyu Sachdeva, Rishi Jain
jdelacru@umd.edu, asachde8@terpmail.umd.edu, rishijn@terpmail.umd.edu

Research Questions:

Based on our scope of vision through the dataset, we were able to address the below mentioned business research questions and the results and interpretation follows:

1. Does job satisfaction come with more salary compensations/expectations?
2. Is there a correlation between years coding professionally and job satisfaction?
3. If large Multinational Corporations want to offshore their workload, which countries can be suitable in terms of logistics and workforce?
4. Does the number of years in a job affect the workers' willingness to switch?
5. Does a person's age affect their willingness to switch jobs?
6. Do people use sites like stack overflow to find new jobs?

Results, Interpretations, and Recommendations:

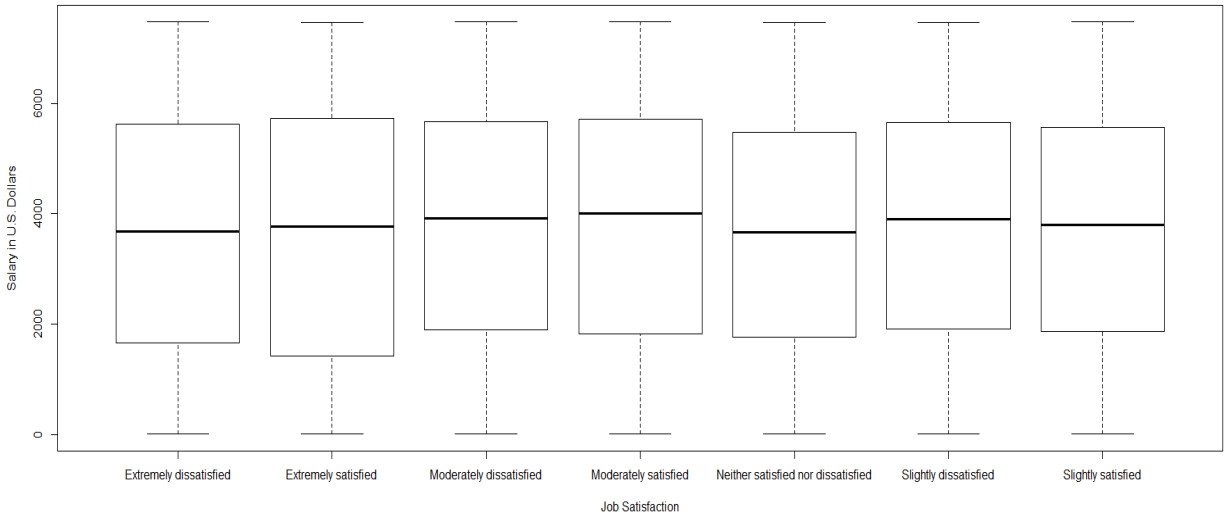
Question 1: Does job satisfaction come with more salary compensations/expectations?

Why this Question:
Compensations play a key role in determining the hiring and attrition of employees. More employees intend to stay if they achieve job satisfaction while working in an organization. Through this research question we try to determine if salary compensation and salary expectations drive the job satisfaction level between the employees who work as coders or developers.

Approach:
We performed a plot for analysis of variable comparison. Also, we performed a correlation test to determine how predictor variable affects the response variable Data Attributes we ran Analysis with:

- Job Satisfaction (Categorical but turned into Numeric)
- Converted Salary (Numeric)

- **Plot:**



- **Results:** The correlation test determined that there is very low correlation between job satisfaction and salary being offered as compensation.

- **Interpretation:** This means that making a lot of money can make individuals happy in their job. There are other factors that may affect how satisfied a person is in their job such as position, workload, and benefits.
- **Recommendation:** Instead of just offering more and more money to workers, companies, can focus on other factors such as giving workers longer breaks, having work activities, or having celebrations and awards for employee and company accomplishments.

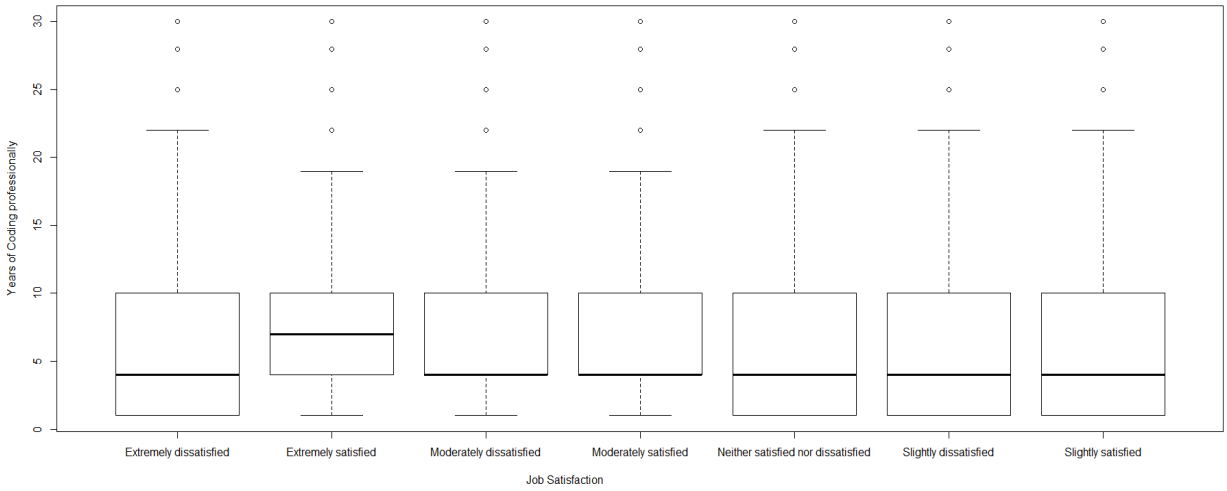
Question 2: Is there a correlation between years coding professionally and job satisfaction?

Why this Question:
Further investigation to our previous question, we had a deeper dive to determine how to find relations to increase job satisfaction between the employees. The longer duration through a work profile and an organization brings in a person to become a better professional, increases stability and sense of responsibility.

Approach:
To analyze this, we needed to use the variables job satisfaction and the number of years that they have been coding professionally (converted to numerical). We performed a plot for analysis of variable comparison. Also, we performed a correlation test to determine how predictor variable affects the response variable.
Data Attributes we ran Analysis with:

- Job Satisfaction (Categorical but turned into Numeric)
- Years Coding Professionally(Categorical but turned into Numeric)

- **Plot:**



- **Results:** Correlation Test determined that there yet again a low correlation between job satisfaction and the number of years a person has been coding professionally.
- **Interpretation:** This means that just because an individual has been coding professionally for a long time, it is less likely to interpret that they are more likely to be satisfied with their jobs. For hiring managers this might mean, not putting too much emphasis on years coding professionally (except for experience purposes), when looking for worker that will fit in and be satisfied in the company.
- **Recommendation:** Other confounding factors must also be analyzed in parallel to reach to a conclusion about the same.

Jhon Dela Cruz, Abhimanyu Sachdeva, Rishi Jain
jdelacru@umd.edu, asachde8@terpmail.umd.edu, rishijn@terpmail.umd.edu

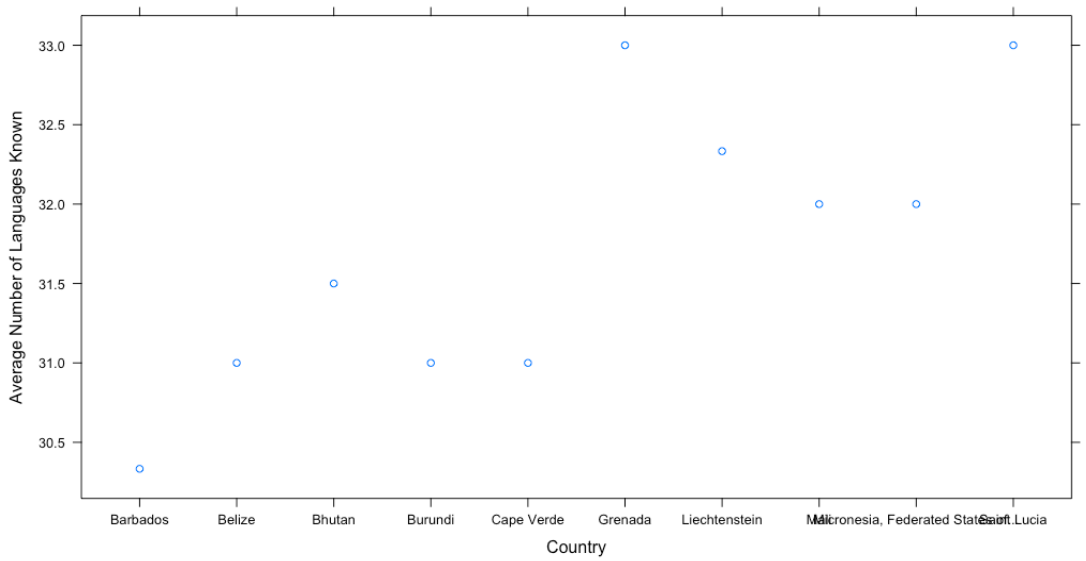
Question 3: If large Multinational Corporations want to offshore their workload, which countries can be suitable in terms of logistics and workforce?

Why this Question:
Growing and emerging companies need to expand their base of operation to cater to the business and market needs. In order to do so, they offshore their workload to other office in the same country location or different geographic locations based on various factors in terms of logistics, capitals, skillsets, cost, return on investment, taxation and law rules and regulations etc.

Approach:
In order to determine the countries with the most skilled coders, he had to make the list of languages knows in to a numeric value and average it per country. We averaged the number of languages known and plotted the results. Filtered for countries with more than 30 since the plot was too crowded. Data Attributes we ran Analysis with:

- Number of languages known (numeric)
- Country (nominal)

- **Plot:**



- **Results:** Granada and Saint Lucia are the top locations with a n average of 33 Languages known
- **Interpretation:** This means that people in these countries, on average, know a greater number of programming languages and platforms than the other ones and the top countries as mentioned. The country having known a greater number of these specifics are suitable and apt for recruiting and the quality of employees could be always a better bet.
- **Recommendation:** Based on the results, is a company is planning to offshore their work, it might be better to open an operations office in Granada or Saint Lucia. (Does not factor labor cost)

Question 4: Does the number of years in a job affect the workers' willingness to switch?

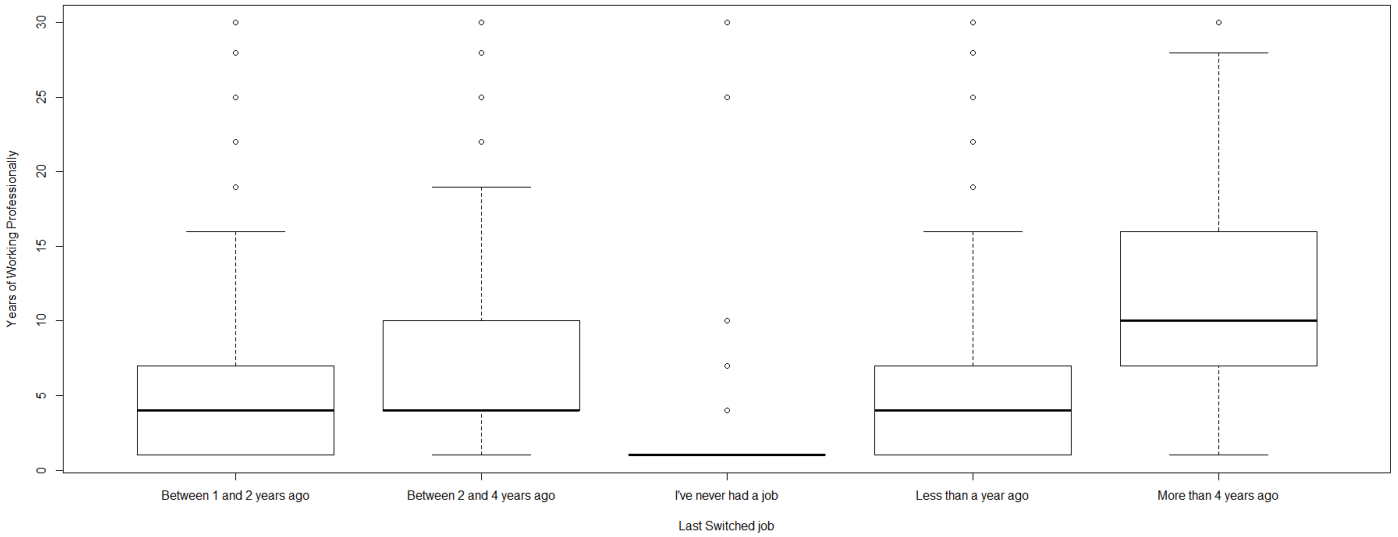
Why this Question:
Stability of employees is an important factor for a company for its recourse planning and management. Stable employees are no less than assets to an organization and they contribute the most in the growth of the organization

and the business. We choose this question to determine stability and loyalty of customers in terms of willingness to switch to

Approach:
In order to determine if work experience or number of years working professionally affect the willingness to switch. Attributes taken into consideration:

- Last job switched
- Years of working professionally

- **Plot:**



- **Results:** Correlation Test determined a positive moderate correlation between the attributes. (.397) This means that as the number of years of coding professionally increases, the probability of switching decreases
- **Interpretation:** This means that there are likely chances for employees to reduce the probability of switching if they have already worked for a longer duration in the industry, this allows to determine if the pre-existing resources are likely to switch or stay for a longer duration, allowing companies to manage even properly.
- **Recommendation:** For companies trying to retain their employees, during the hiring process it would be better to take an applicant who has been coding professionally longer since he or she will be less likely to quit and switch jobs.

Question 5: Does a person's age affect their willingness to switch jobs?

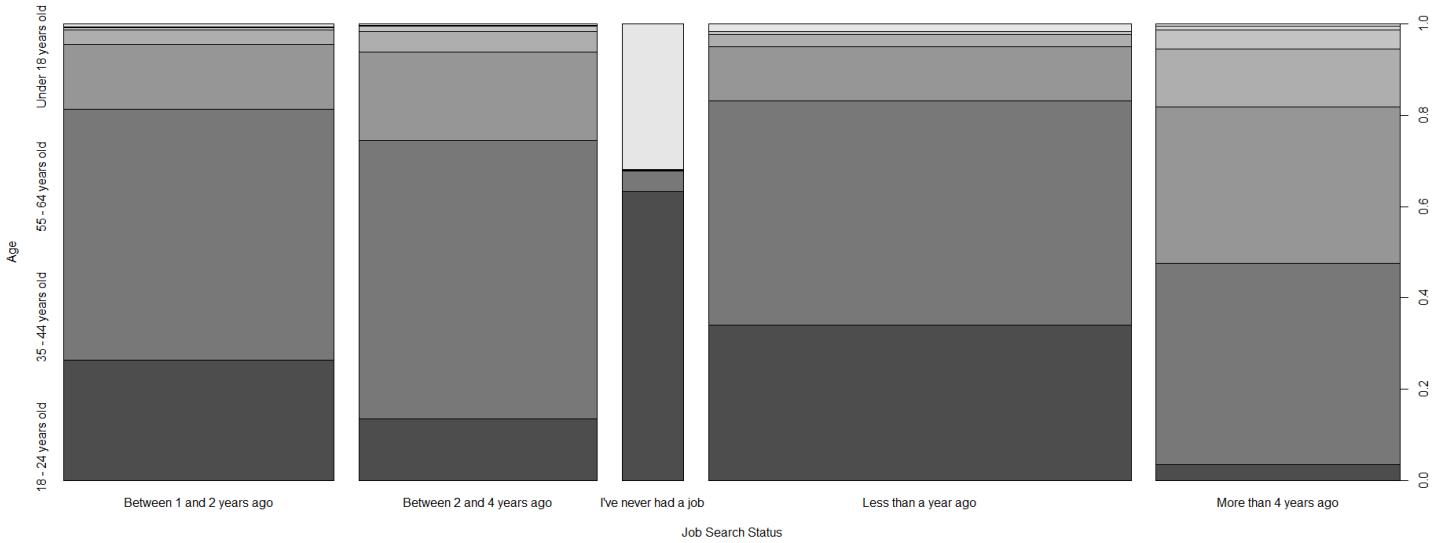
Why this Question:
Job switch affects the organization as a whole and they need to find another replacement to fulfil that position and responsibilities. This is a continuous process which goes on and prediction of job switches if recognized can help organizations improve attrition as also well planning could be done for resource management.

Approach:

- In order to determine this, we use Age and last new job used through a correlation test to determine if the variables have a relationship. We performed a plot for analysis of variable comparison. Also, we performed a correlation test to determine how predictor variable affects the response variable. Data Attributes we ran Analysis with:

- Last New Job (Categorical turned to Numeric)
- Age (Numeric)

- **Plot:**



- **Results:** There is moderate correlation between age and willingness to switch jobs. This means that as age increases, the probability of a person switching their job gets lower.
- **Interpretation:** This aligns with the previous research question inference, dignifying the stability of a person in an organization increasing with the age of the person and in turn reducing the probability of the employee switching into another firm.
- **Recommendation:** For companies this means that their older employees are less likely to leave their jobs. This may mean focusing more employee retention efforts for the younger employees who are more likely to leave.

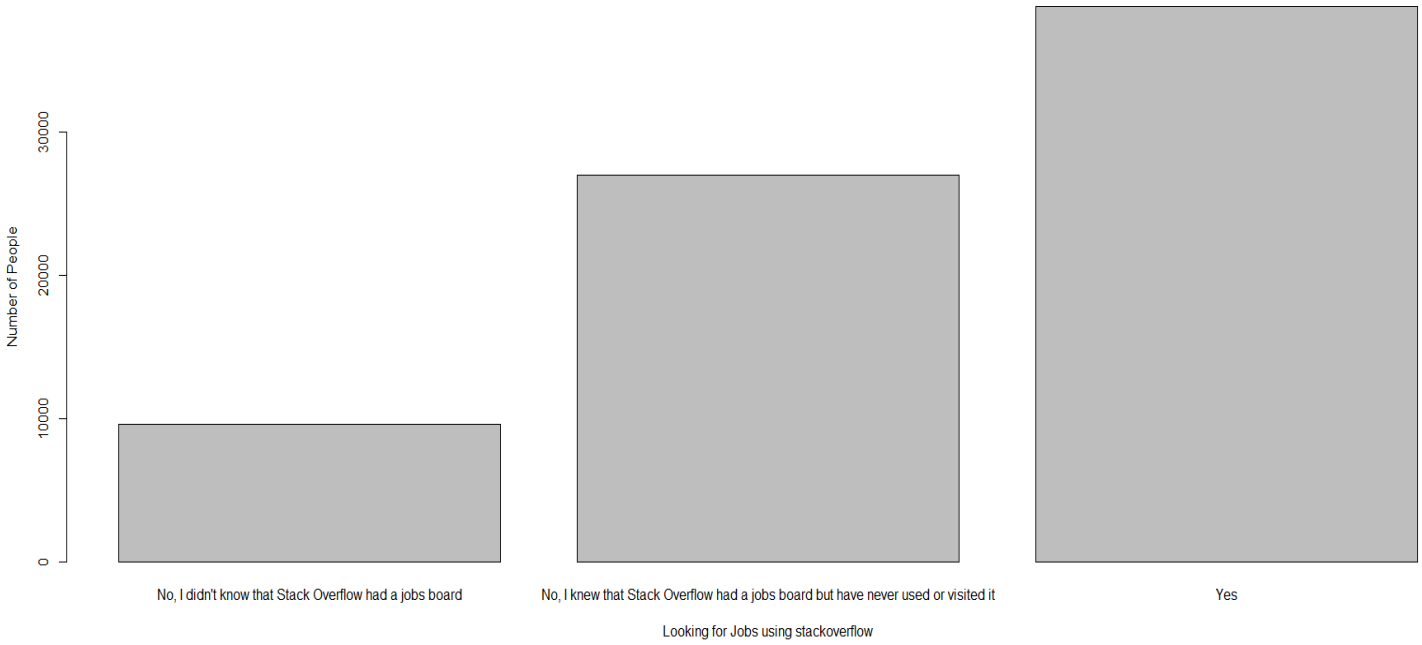
Question 6: Do people use sites like stack overflow to find new jobs?

Why this Question:
Finding jobs is an ask for all sets of people from all communities throughout and always. Be it a graduate fresher or an experienced professional, this is a continuous process which people come influx and map themselves to organizations. Since stack overflow is a community of programs, we try here to determine if programs use such website platforms to find jobs.

Approach:
To determine if users use Stack Overflow to find jobs, we plotted a histogram to determine what the users were saying in regard to this survey question. Also, we performed a plot for analysis of variable comparison. Data Attributes we ran Analysis with:

- Country(Categorical)
- Average Language Known(Numeric)

- **Plot:**



- **Results:** 38727 users said yes, and they use the site to find jobs.
- **Interpretation:** This means that people belonging to the coding and programing community do use such platforms as stack overflow to find jobs available in the industry basis their interests and expectations. As also, this can be equivalently important to firms to fetch these data in order to manage their recruitment processes and resource planning.
- **Recommendation:** Companies looking to hire coders can take advantage of the job listings on Stack Overflow and view some of the projects that the coders looking for jobs had worked on.

Jhon Dela Cruz, Abhimanyu Sachdeva, Rishi Jain
jdelacru@umd.edu, asachde8@terpmail.umd.edu, rishijn@terpmail.umd.edu

Issues Encountered with data:

1. There were a lot of unidentified characters and symbols like â€™. These are the data chunks which are having very indifferent ASCII values than the alphanumeric characters as also they provide no use while data analysis.
2. The unidentified characters and symbols act as a roadblock and hindrance for data cleaning, data delimiting and data formatting since they intervene and result to null data points in the fields.
3. There are a lot of NULL data observed through the dataset. The NULL data plays a crucial role in a dataset in order to identify values holding no values or zero values or junk values. The null values are hard to play with since they have to be kept at some places to determine the count of observation and at the same time need to be removed or eliminated in order to run statically analysis.
4. Open ended text data are meant for observational analysis since no statistics or data analytics can be performed over them. Moreover, these kinds of inputs vary from each observation and variability is huge resulting in nil or least correlations, hence they cannot be grouped or analyzed.
5. The open-ended text needs to be ignored at places where mathematics or grouping is required, as also it can be extremely useful if we run a text analysis or sentiment analysis to determine the nature of input.

Remediation to issues encountered

There are various ways how each issue which is evident in the beginning and which rises up over the time through analysis is analyzed, interpreted and worked upon. Below are the actions taken by us for remediation of the issues encountered:

1. The unidentified characters through the data are delimited through the recognized symbols and are split into cells and then using R functions, those rows with the symbols are eliminated or kept separately in other column, unperturbed, for any further utility if required. “Symbols might always have a meaning”.
2. The NULL data is replaced by values of zeros where count was required along with mathematical importance. However, using R functions, they can be ignored while performing any statistical analysis while they do not disturb our data frame or results.
3. Open ended texts can be recoded to be given likert values for grouping and better understanding over their essence which would not only help us understand their meaning better but also run some analysis and work upon them.
4. NULL values from the primary key are removed since they hold no importance and shall not be included into the analysis.
5. There were a few observations which were misplaced through the matrix which were fixed by using indexing values to the attributed and identifying points of deviations or conflicts.

----------END OF FILE----------