**Capstone Projecthorizontal line**

Final Project

Project By:

Aakash Mishra

Mohit Kewlani

Kewal Jani (kj2062)

**1. Data Set Description**

First of all, we looked for many datasets around and took the data set about the stack over flow website survey from Kaggle. Stack over flow is the website that is used by most of the students and by many Employers. It is one of the best websites created for programmers (Everyone should agree on this or we can do hypothesis testing).

Data Set: https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey

This Data set is a type of survey in which users are asked 130 different questions and the responses are recorded for individual user. This dataset is collected for around 100K users. The rows states the number of user and the columns has the questions asked to the user. These questions can be found in survey\_results\_schema.csv file and the data in stackoverflow\_data.csv file.

We were interested in this particular data set because it states various important features of a programmer. The data set contains information about salary, if the user used to code before his current job, weather he is satisfied with his current job, etc. The dataset will not only help us to get facts but also to get information about where we will be standing in the future.

**The Challenge:**

**The challenge we faced with the dataset was, the data was not cleaned and there were many columns in which we had to apply ---- and generate more columns. Also, as the datatype for certain columns were in characters and string format. we need to classify data column wise and make the data usable. We converted the data in numeric format and it was ready to use.**

**After the data cleaning part, we had to handle the NA values. We handled the NA values row wise, column wise or fill forwarding, According to the requirement of the question that we were supposed to answer**

We decided to answer all the interesting questions that we found and made sure that we implement all the algorithms and techniques that we learned in the Intro to Data science course for getting the answers.

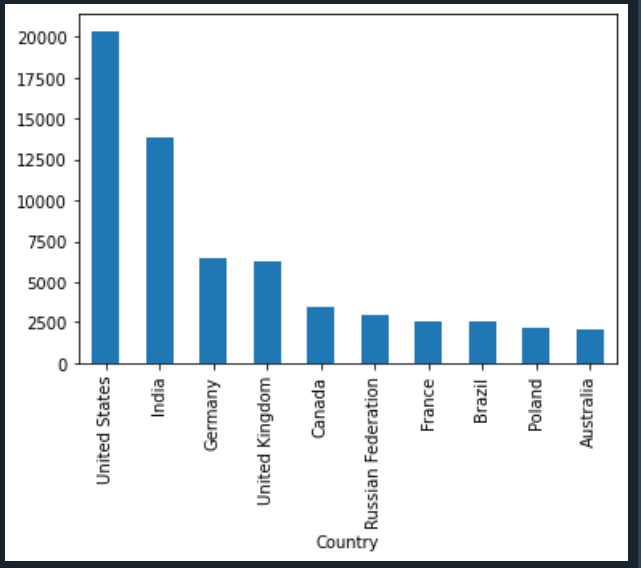
**1 hypothesis Testing:** For comparing the data sets to check the significance for data-set for eg: Weather the salary is gendered biased , Is there any difference in satisfaction of the user that stays in different countries, Which country code for Hobby most , Highest Degree done by Developers, Coders happy to code

**2 Prediction:** Reduction in the dimensionality of the question asked to the user for survey. For example, which question contribute more and provide more information

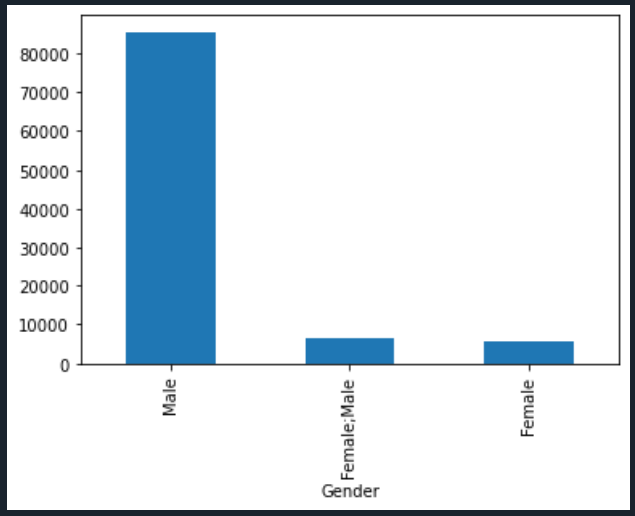
**3 Machine learning :** Training the model with the columns like …… and predict the salary of the user.

**1. Hypothesis Testing**

**Q1 Hypothesis for distribution of the User satisfaction according to the country wise**



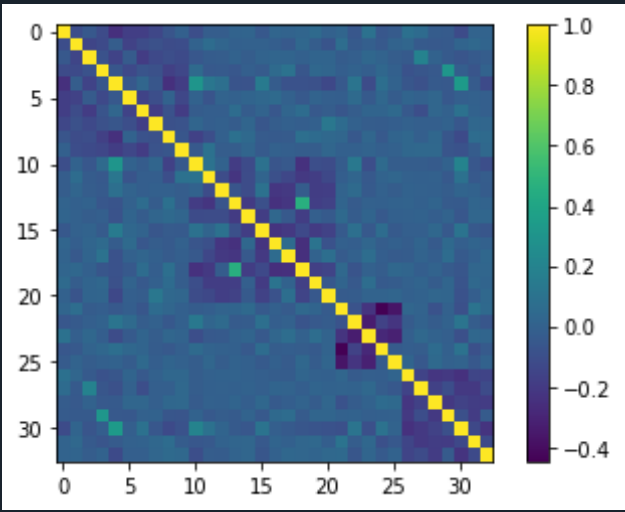
**Q1.2 Is the salary gendered bised in Us?**



**2. Dimensionality reduction :**

*Dimensionality Reduction. We had a data set of around 100K users and We thought that if would be great if we are able to reduce the dimension of the data by removing the data that are corelated to each other and keep those data that contribute to the Dataset.*

There were certain questions who had numeric answer so we decide to apply PCA on it. There were 33 Questions which had numeric answers and we applied correlation matrix on it to check which columns are more correlated to each other. So the result we got for the questions is as shown below



By applying

**3. Prediction**