**FINAL REPORT FOR MBA SALARIES**

From the individual analysis a few observations can be made –

* 25-30 is the average range of age of people who took the exam
* 206 were male and 68 were female among whom this analysis is done
* Most number of people scored between 600-650 and least scored between 750 and 800
* Verbal had a median percentile of 80 and hence the total percentile had a median of 90
* The second quarter has the maximum number of people, and the rest of the quarters have the same amount of people in them
* Many people worked for around 5 years and there are outliers upto 20 years of work in this data set.
* Many people knew language 1 better than language 2

Scatterplot Analysis-

* Relationship between various variables can be seen here- with the pairs command we see a relationship between gmat score, age and gender and we can infer that younger people scored more in gmat. If 1 is male and 2 is female , then more men performed better than the female population.
* A distribution of the verbal and quants percentile with respect to the male and female population is done – here we can see that the white dots are for male and the red for female. The outliers for male is more than female population.
* With the comparison of the language and the total score, those who knew language 1 scored better than those who didn’t have language 1 as their first language

Corrgram Analysis-

* It show that work in years and age,gmat\_vpc and gmat\_tot,gmat\_tot and vpc and gmat vpc and qpc are highly positively correlated whereas the red boxes tell us the factors that are highly negatively correlated – which includes s\_avg and quarter with the highest red saturation.

Among those placed –

* Sex 1 has a median of 54900 and sex 2 has a median of 68200 as their salaries. This is only an approx. from the data.
* The most number of 17 people scored a score of 630 and very few in the outliers.
* More number of people who had first language as 1 were placed
* Most number of people who were placed had a work experience of 2 years and sex 1 earned much more than sex 2
* Most people in the first quarter only had language 1 as their first language.
* Chi square tests were done between a number of parameters for both placed and not placed students. This is done to check the contingency and frequency of the tests. Out of the ten done, only p value of age ,gmat percentile and quarters and genders had values close to the significance. Most of the values were above 0.05, which doesn’t help to reject the null hypothesis.

Regression models-

* These models were done by analysing a few parameters- firstly salary- out of which age and statis only had significance. R squared value of 0.4888. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.
* Then with total scores in gmat and gmat verbal,quants percentile and f\_avg are the significant factors with R square of 0.9946.
* Then with work years- with age,gmat verbal,quants,total and first language as the significant values. R squared is 0.7366

Challenge –

* As we analysed the data set, we observed that there are no missing values and that the model is best fitted there.
* Only gmat\_qpc ,gmat\_vpc and first language are significant
* After enhancing the model with anova() age and sex parameters have gained significance.