**Summary Report**

These are the inferences I could draw after doing histograms, box plots, bar charts, stacked bar charts, scatter plots, correlation tests, corrgrams, chisq tests, t-tests and finally after fitting a linear regression model. Also, **I implemented a logistic regression model for predicting if students will get placed or not.**

The below given inferences are best understood by running the R code sequentially given in Sonal Airline Analysis.R and following the outputs of the commands given there.

Inferences drawn -

1. The salary histogram suggests that around 60% students didn’t get placed while less than 3% got salaries more than 2 lakhs as starting salary. Mostly who got placed got the starting salary around 1 lakh.
2. Box plot of salary shows that more than 60% people are getting salary as 0 as they are not placed.
3. If we make a box plot of salary with only placed students, we see that the median of salary for the placed people is 1 lakhs.
4. It can be seen the distribution of age in the dataset is unimodal and right skewed which suggests that there are more people in their 20s who are enrolled than who are 30 +.
5. Box plot of gmat total score of all students clearly follows a normal distribution with 620 as the median score.
6. Histogram of gmat total percentiles shows the distribution is left skewed suggesting there are far more people with 80 + percentiles than lower.
7. Box plot of work experience (years) shows the distribution has a lot of outliers suggesting generally people with 2-4 years join MBA programs while some have a lot more experience (over 20 years too).
8. Bar chart of satisfaction level with the MBA Program with people who filled responses shows that of those who fill out responses for satisfaction levels of MBA Program gave the course a 6 or had high satisfaction levels.
9. Histogram for Spring Average (s\_avg) shows unimodal and normal distribution for the spring averages with median being 3.
10. The fall average distribution too follows a normal curve with the median being around 3.25 (excluding whose GPA is 0).
11. Bar chart of the quartile ranking of students has the distribution looking pretty uniform with students falling in each of the quartiles pretty uniformly.
12. Bar chart of the first language of people enrolled in the course shows that most students enrolled have English as their first language.
13. A correlation test between s\_avg and salary gives results which confirm our observation that they have a weak correlation between the two variables. Also, the p-value is 0.3065 which is above 0.05 and thus we fail to reject the null hypothesis that salary and spring average affect each other.
14. The regression line for work\_yrs and salary shows an upward trend in the starting salaries and work experience (yrs), which was confirmed by doing a correlation test. We can see that the p-value = 1.403e-06 which falls much below 0.05 and hence we can say that salary and work experience are correlated and the correlation value being 0.45 shows that they both are positively correlated and their relationship has moderate strength.
15. Box plots for salaries by quartiles shows that quartiles does not affect much the distribution of salaries.
16. Box plot for salary by sex shows that there seems to be some gender disparity in place when comparing salaries by gender. Median salary for females seems to be comparatively lower than that of males.
17. While English speakers salary follows a normal distribution with some outliers on both end, none of the students whose first language is not English and are placed got a salary below 90000.
18. From the correlation test between first language and salary, the obtained p-value suggests that first language and salaries are correlated to each other as it lies below 0.05 but the correlation coefficient is 0.27 which although shows a positive correlation but the strength of the relation is weak in nature.
19. The corrgram clearly shows that work experience and age has a strong positive correlation with salaries while first language and spring average also are positively correlated with moderate strength to salary.
20. The barchart shows that females whose first language is English tend to get lower salaries than their male counterparts while it’s the opposite for the females whose first language is not English.
21. Upon performing Chi sq test of independence between quartiles and sex, we found that since the p-value is not below 0.05, we fail to reject our null hypothesis that quartiles and sex are independent of each other.
22. Upon performing Chi sq test of independence between quartiles and first language, we found that since the p-value is not below 0.05, our null hypothesis that quartiles and first language are independent of each other is not rejected.
23. For t-tests, articulating hypothesis as - H1 = Work experience does have an effect on salary Running t-test to test our hypothesis (H0,H1) –

Here, the null hypothesis (H0) is that work experience does not have an effect on salary. The results are as under – since the p-value is < 2.2e-16, we can reject our null hypothesis in favour of our alternate hypothesis (H1) i.e. work experience does have an effect on starting salary.

1. Articulating hypothesis as - H1 = Males have a higher mean starting salary than females

Running t-test to test our hypothesis (H0,H1) - Here, the null hypothesis (H0) is that Males and females have equal mean starting salaries. Since the p-value is not below 0.05, our null hypothesis that Males and females have equal mean starting salaries is not rejected.

1. Let's now compare the above regressions for picking the best one –

Among all the above models, model\_3 has higher Multiple R-square as well as higher adjusted R square as compared to other models, hence I would pick model\_3 with explanatory variables as work\_yrs,s\_avg,sex,gmat\_tpc as the best fit model.

According to model\_3, explanatory variable work\_yrs is statistically significant with its p-value as 1.28e-05 which is way below the level of significance (alpha) which is 0.05.

Then, the overall model's p-value is 3.036e-05 which is also way below alpha 0.05, hence we can accept this model to predict the variations in salary as it means that our model is doing better than the intercept model.

1. Our model's equation is as follows : y = b0 + b1\*x1 + b2\*x2 + b3\*x3 + b4\*x4 + e where y is salary, x1 is work\_yrs and b1 is its coefficient, similarly for the rest of the explanatory variables added in this model. e denotes the rror term or the residuals and b0 is the intercept (when no regressors are added). Our final equation for the model becomes :

salary = 104867.5 + 2497.9\*work\_yrs + 2674.4\*s\_avg -5282.2\*sex - 147.0\*gmat\_tpc + e

Thus, for every year increase in years of experience, the salary will go up by 2497.9 and similarly, we can conclude for other variables too.

1. While comparing those who didn't get a job with those who did, we ran a couple of chi sq tests above but none of them suggests that we reject the null hypothesis as for all the tests, the obtained p-value is not statistically significant.
2. No strong correlations are found between the salary variable and other variables.
3. Logistic Regression -here, we would run this regression to predict if a student will get placed or not. In the logit model the response variable is log odds: ln(odds) = ln(p/(1-p)) = a\*x1 + b\*x2 + … + z\*xn.
4. We found here that work\_yrs and s\_svg are statistically significant explanatory variables with p values less than 0.05. As for the statistically significant variables, s\_avg has the lowest p-value suggesting a strong association of the spring average GPA of the student with the probability of getting placed.
5. We can interpret this model as : a male would reduce the log odds by 0.0264893 while a unit increase in s\_avg would increase the log odds by 1.4842058.
6. Of the two models, the second one has got lesser AIC, which is desirable and thus we choose that for further analysis. Also, we can do ANOVA test on the models, the greater the deviation between null and residuals, the better. The drop in the deviance upon dropping a couple variables suggests that logit\_model\_1 is better among the two.
7. ROC and AUC curve for performance measures : As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5. Our model’s AUC is not so good , it is just 0.37.