The data set is imported using appropriate R commands and the summary of the data set is carried out.

Each variable (column) is analyzed and their means, SD’s are notes to find out their behavior.

Boxplots and bar plots are used for each of the columns present in the data set. Corrgram is also drawn.

Analyses and inferences regarding their distribution can be drawn.

Then the variables are analyzed pair-wise to note any correlation present.

For the first set of problems, a subset of only people with a job(a salary) is created, as asked.

**Hypothesis 1:**

**Males get a higher starting salary compared to females.**

Using the aggregate command, average salaries of males and females is found out. Males are found to have a higher salary.

Therefore the following hypothesis are stated before performing the t test

H0: Females and males have the same salary. Difference between the mean salary of females and mean salary of males is 0

H1: Males have a higher salary compared to the females. Difference between the mean salary of females and mean salary of males is not 0.

Due to very low p value we reject the null hypothesis.

Now we perform a chi square test with the following hypothesis

Null hypothesis states that the salary and sex are independent of each other.

Alternate hypothesis states that there is a dependency between salary and sex

P value is higher than 0.05 therefore we fail to reject the null hypothesis

**Hypothesis 2:**

**People who have English as their first language earn a better salary than other people.**

Using the aggregate command, average salaries of people with English as first language and other languages as first language is found out. People who have English as first language are found to have a higher salary.

Therefore the following hypothesis are stated before performing the t test

H0: People who have English as first language as well as people who have other languages as first language have the same salary. Difference between the mean salary of both of them is 0

H1: People who have English as first language have a higher salary compared to the people who do not have English as their first language. Difference between the mean salary of both of them is not 0.

Due to very low p value we reject the null hypothesis.

Now we perform a chi square test with the following hypothesis

Null hypothesis states that the salary and sex are independent of each other.

Alternate hypothesis states that there is a dependency between salary and sex

P value is higher than 0.05 therefore we fail to reject the null hypothesis

Let us run a chi square test on the salary and first language.

Null hypothesis states that the salary and first language are independent of each other.

Alternate hypothesis states that there is a dependency between salary and first language

P value is found to be less than 0.05. Therefore we reject the null hypothesis.

**Regression models**

3 regression models were built.

1. Dependency of salary on age+work experience

**Salary= [388.8(work experience)]+ [2413.8(age)] + 36967.5**

1. Dependency of salary on age

**Salary= [2728.8(age)] + 29962.6**

1. Dependency of salary on work experience

**Salary= [2699(work experience)]+ 93101**

**Which is the best model?**

The higher the R-squared, the better the model fits your data.

Model 1: R squared was 0.2356

Model 2: R squared was 0.2422

Model 3: R squared was 0.1989

**So model 2 was the best model that fits the data.**

For the next set of problems, we have to analyse the subsets of people having a job and those who don’t have a job.

Therefore we eliminate entries where people’s salaries are 998 and 999.

clean <- mba[mba$salary!=998 & mba$salary!=999, ]

To simplify the dataset, we add another variable to the data set called job which is a categorical variable and it says 1 if the person has a job(has a salary) and 0 if does not have a salary(does not have a job)

job <- ifelse(clean$salary==0,0,1)

clean <- cbind(clean,job)

**CONTINGENCY TABLES**

**Does gender(sex) play a role in getting/not getting a job?**

Referring to the analysis I have conducted, we can conclude that

Almost 70% of the jobs have gone to males and only 30% of the jobs have gone to females.

51% of males have a job while 57% of females have a job

Let us conduct a chi square test to better understand the situation at hand

Null hypothesis is that the job and the sex are independent

Alternate hypothesis is that the job and the sex are not independent

We get a relatively high p value therefore we fail to reject the null hypothesis.

**Does English as first language play a role in getting/not getting a job?**

Referring to the analysis I have conducted, we can conclude that

93% of people who have got jobs have English as their first language

Almost 54% of people who have English as their first language have got jobs

Let us run a chi square test where

H0: Job and first language of the person are independent of each other

H1: Job and first language of the person are not independent of each other

The p value is relatively high so we fail to reject the null hypothesis.

**Logistic Regression**

Logistic regression is conducted when the dependent variable is a categorical variable(it has a fixed set of values). The independent variable may be a categorical, continuous or a mixture of both.

We split the data into two chunks: training and testing set. The training set will be used to fit our model which we will be testing over the testing set.

According to the example on the titanic data set, almost 90% of the data set is used for training and the remaining is for testing.

Here we have 193 rows so 90% of 193 is almost 174.

train <- clean[1:174,]

test <- clean[175:193,]

We consider age and work experience as the independent variables with job( 0 or 1) being the dependent variable.

We use the glm command for the regression model.

We find out that age is statistically significant with a low p value.

Now after training, we try testing the model and report an accuracy of 0.684210526315789.