Individual ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

TYPE INTAKE CODE: APU2F2211CS (CYB)

HAND OUT DATE: 7 FEBRUARY 2023

HAND IN DATE: 13 MARCH 2023

WEIGHTAGE: 50%

LECTURER: Farhana Illiani Binti Hassan

Name: KUAN ZHI HUI

TP NUMBER: TP067195

Table of Contents

Introduction

This research's analysis of the provided dataset aims to find students with undiscovered problems and provide practical decision-making data. The dataset contains 25 columns and 17007 rows for students' placement information and attributions. Along with background details about the students, the information dataset additionally contains information about their families, extracurricular activities, academic records, and placements status. In order to discover the undiscovered problems, the various data exploration, manipulation, transformation and visualization techniques will be utilized. Additionally, it is necessary to investigate additional conceptions that can enhance the impact of retrieval.

1.0 Data import

```
#import data
importData = read.csv("C://Users//Asus//OneDrive//Documents//R//Assignment PFDA//Placement_Data_Full_Class.csv")
importData
```

In the fundamental step to start analysing, we must import the data first. In this analysis, we will use the CSV file type to import "Placement_Dte_Full_Class.csv" which located in the file directory.

```
#install packages
install.packages("ggplot2")
install.packages("plotrix")
install.packages("Hmisc")
install.packages("scales")

#################################
library(ggplot2)
library(Hmisc) #boxplot
library(scales)
library(plotrix)
```

In this analysis programming, i had installed ggplot2, plotrix, Hmisc and Scales for carry out meaningful analysis.

1.1 Data cleaning

```
#data clean
importData <- na.omit(importData)
importData
```

Data cleaning will remove the columns with NA in importData, which makes it easier to analyze clearly. However, we can skip in this analysis because the imported data is clean enough.

2.0 Pre-processing

2.1 preprocessing - Rename column

```
#rename column
names(importData) = c("Student_ID", "Student_Gender", "Student_Age","Student_Address",
                      "Student_Medu", "Student_Fedu", "Student_Mejob", "Student_Fejob",
                      "Family_Support", "Class_Paid","Student_Activities", "Student_Internet",
                      "Secondary_Education", "Secondary_Education_Board","Higher_Secondary_Education", "Higher_Secondary_Education_Board",
                      "Higher_Secondary_Specialisation","Degree %","Degree_Specialisation","Work_Experience",
                      "Employee_Test", "Master_Specialisation","MBA %", "Placement_Status",
                      "Salary")
importData
```

```
> #rename data
> names(importData)
 [1] "Student_ID"                    "Student_Gender"                "Student_Age"                   "Student_Address"
 [5] "Student_Medu"                  "Student_Fedu"                  "Student_Mejob"                 "Student_Fejob"
 [9] "Family_Support"                "Class_Paid"                    "Student_Activities"            "Student_Internet"
[13] "Secondary_Education"           "Secondary_Education_Board"     "Higher_Secondary_Education"    "Higher_Secondary_Education_Board"
[17] "Higher_Secondary_Specialisation" "Degree %"                    "Degree_Specialisation"         "Work_Experience"
[21] "Employee_Test"                 "Master_Specialisation"         "MBA %"                         "Placement_Status"
[25] "Salary"
```

Before analyzing with data visualization, it is important to rename the columns from the importData file, as this will make the data names understandable when analyzing.

2.2 pre-processing -Data structure

```
#data structure
class(importData)
```

```
> class(importData)
[1] "data.frame"
```

```
> str(importData)
'data.frame':   17007 obs. of  25 variables:
 $ Student_ID                       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Student_Gender                   : chr  "M" "M" "M" "M" ...
 $ Student_Age                      : int  23 19 19 21 22 19 19 18 19 21 ...
 $ Student_Address                  : chr  "U" "U" "U" "U" ...
 $ Student_Medu                     : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Student_Fedu                     : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Student_Mejob                    : chr  "at_home" "at_home" "at_home" "health" ...
 $ Student_Fejob                    : chr  "teacher" "other" "other" "services" ...
 $ Family_Support                   : chr  "no" "yes" "no" "yes" ...
 $ Class_Paid                       : chr  "no" "no" "yes" "yes" ...
 $ Student_Activities               : chr  "no" "no" "no" "yes" ...
 $ Student_Internet                 : chr  "no" "yes" "yes" "yes" ...
 $ Secondary_Education              : num  67 79.3 65 56 85.8 ...
 $ Secondary_Education_Board        : chr  "State" "State" "Private" "Central" ...
 $ Higher_Secondary_Education       : num  91 78.3 68 52 73.6 ...
 $ Higher_Secondary_Education_Board: chr  "State" "Central" "Private" "State" ...
 $ Higher_Secondary_Specialisation  : chr  "Commerce" "Science" "Arts" "Science" ...
 $ Degree %                         : num  58 77.5 64 52 73.3 ...
 $ Degree_Specialisation            : chr  "Sci&Tech" "Sci&Tech" "Comm&Mgmt" "Sci&Tech" ...
 $ Work_Experience                  : chr  "No" "Yes" "No" "No" ...
 $ Employee_Test                    : num  55 86.5 75 66 96.8 ...
 $ Master_Specialisation            : chr  "Mkt&HR" "Mkt&Fin" "Mkt&Fin" "Mkt&HR" ...
 $ MBA %                            : int  78 80 77 50 86 63 59 83 51 67 ...
 $ Placement_Status                 : chr  "Placed" "Placed" "Placed" "Not Placed" ...
 $ Salary                           : int  350000 200000 350000 NA 250000 NA NA 300000 350000 NA ...
```

With function class() and str() we can understand that the data type and their information.

3.0 Data exploration

3.1 Data exploration - summary for the data

```
> summary(importData)
   Student_ID    Student_Gender      Student_Age    Student_Address     Student_Medu
 Min.   :    1   Length:17007       Min.   :18.00   Length:17007       Min.   :0.000
 1st Qu.: 4252   Class :character   1st Qu.:19.00   Class :character   1st Qu.:2.000
 Median : 8504   Mode  :character   Median :20.00   Mode  :character   Median :3.000
 Mean   : 8504                      Mean   :20.49                      Mean   :2.513
 3rd Qu.:12756                      3rd Qu.:22.00                      3rd Qu.:4.000
 Max.   :17007                      Max.   :23.00                      Max.   :4.000

  Student_Fedu   Student_Mejob       Student_Fejob      Family_Support      Class_Paid
 Min.   :0.000   Length:17007       Length:17007       Length:17007       Length:17007
 1st Qu.:1.000   Class :character   Class :character   Class :character   Class :character
 Median :2.000   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :2.489
 3rd Qu.:3.000
 Max.   :4.000

 Student_Activities Student_Internet   Secondary_Education Secondary_Education_Board
 Length:17007       Length:17007       Min.   :40.89      Length:17007
 Class :character   Class :character   1st Qu.:61.00      Class :character
 Mode  :character   Mode  :character   Median :72.00      Mode  :character
                                       Mean   :72.44
                                       3rd Qu.:84.00
                                       Max.   :95.00
```

```
 Higher_Secondary_Education Higher_Secondary_Education_Board Higher_Secondary_Specialisation
 Min.   :37.00              Length:17007                     Length:17007
 1st Qu.:61.00              Class :character                 Class :character
 Median :72.00              Mode  :character                 Mode  :character
 Mean   :72.45
 3rd Qu.:84.00
 Max.   :97.70

    Degree %     Degree_Specialisation Work_Experience    Employee_Test   Master_Specialisation
 Min.   :50.00   Length:17007          Length:17007       Min.   :50.00   Length:17007
 1st Qu.:61.00   Class :character      Class :character   1st Qu.:61.00   Class :character
 Median :72.00   Mode  :character      Mode  :character   Median :72.00   Mode  :character
 Mean   :72.39                                           Mean   :72.32
 3rd Qu.:84.00                                           3rd Qu.:84.00
 Max.   :95.00                                           Max.   :98.00

     MBA %       Placement_Status       Salary
 Min.   :50.00   Length:17007       Min.   :200000
 1st Qu.:61.00   Class :character   1st Qu.:250000
 Median :72.00   Mode  :character   Median :300000
 Mean   :72.54                      Mean   :308532
 3rd Qu.:84.00                      3rd Qu.:350000
 Max.   :95.00                      Max.   :500000
                                    NA's   :8265
```

With the function sum() we can understand all the data information clearly.

3.2 Data exploration - Define the size of data.

```
>
> length(importData) #no of column
[1] 25
>
> ncol(importData) #no of column
[1] 25
>
> nrow(importData) #no. of rows
[1] 17007
```

The data exploration allows us to understand the data length, numbers of column and numbers of row.

```
> #number of rows and columns
> dim(importData)
[1] 17007    25
```

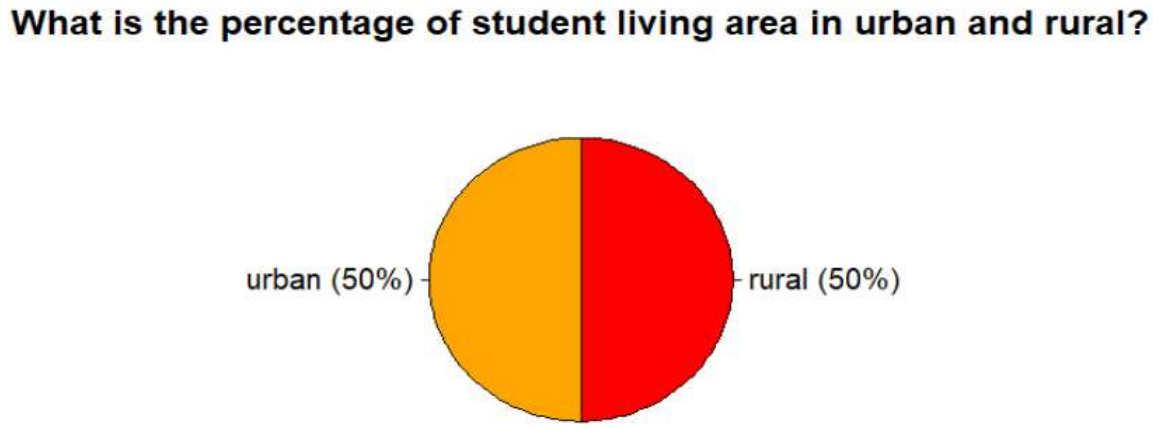
With the function of dim() we can know both data numbers of rows and columns.

Question 1: Does the student living area will affect their placement status?

Analysis 1.1: What is the percentage of student living area in urban and rural?

```
#Analysis 1.1: What is the percentage of student living area in urban and rural
Student_Address = c("urban","rural")
pie(table(importData$Student_Address), Student_Address, main= "what is the amount student living area in urban and rural ",
    col = c("red","orange"),clockwise = TRUE)
#to show the percentage
Student_Address = c("urban","rural")
prop <- prop.table(table(Student_Address)) * 100
pie(prop, labels = paste0(names(prop), " (", round(prop, 1), "%)"), main = "what is the percentage of student living area in urban and rural?",
    col = c("red", "orange"), clockwise = TRUE)
```

The code is created by using the table() and prop.table() functions to calculate the proportions, and the pie() function to create the chart.

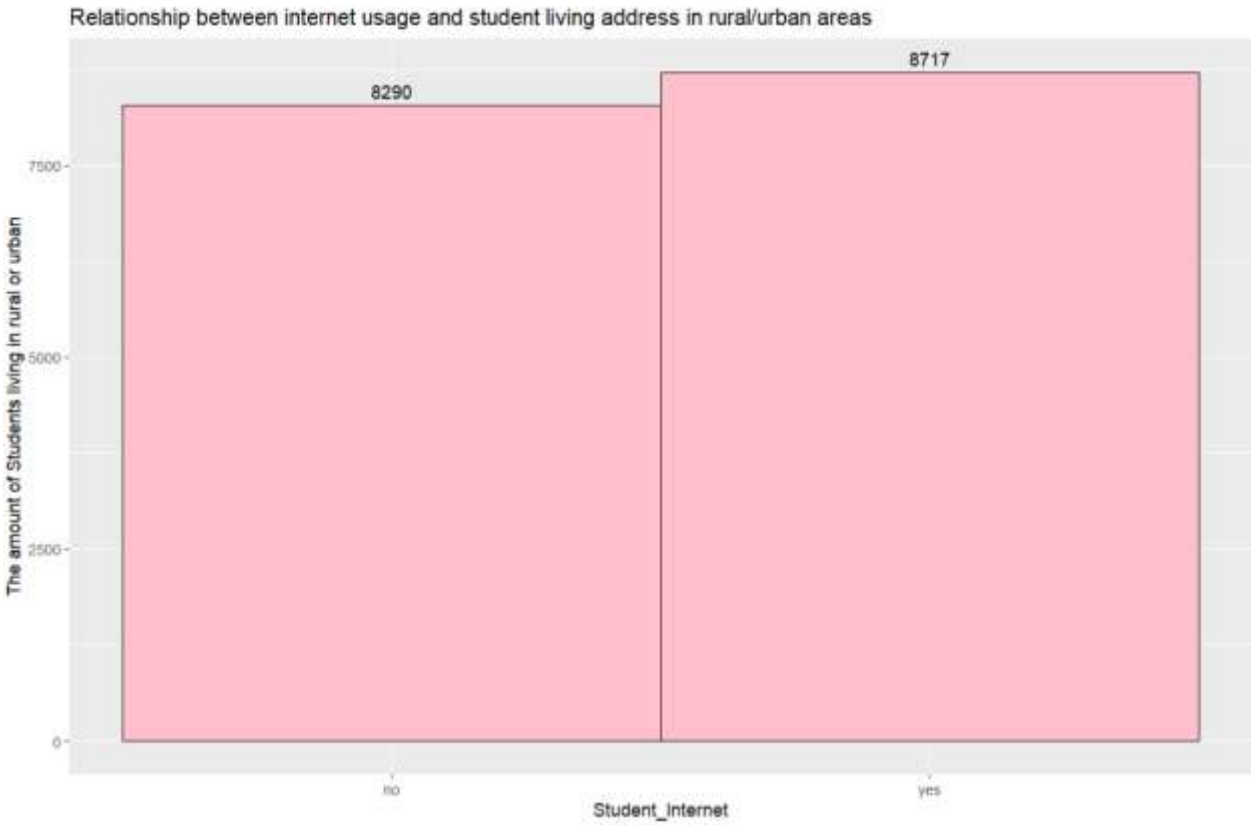


From the graph, we can see that 50% of the students live in both urban and rural areas, so according to the analysis, this does not have much impact on placement status.

Analysis 1.2: What is the relationship between internet usage and student living address in rural and urban area?

```
# Analysis 1.2: Relationship between internet usage and student living address in rural/urban areas
ggplot(importData, aes(x=Student_Internet, fill=Student_Internet))+
  geom_bar(color="black",fill="pink", width = 1)+
  geom_text(stat='count', aes(label=..count..), vjust=-0.5) + # add count labels
  labs(x="Student_Internet", y="The amount of Students living in rural or urban") +
  ggtitle("Relationship between internet usage and student living address in rural/urban areas")
```

The graph is created by function geom_bar() with fill pink color, width 1 and the count label.

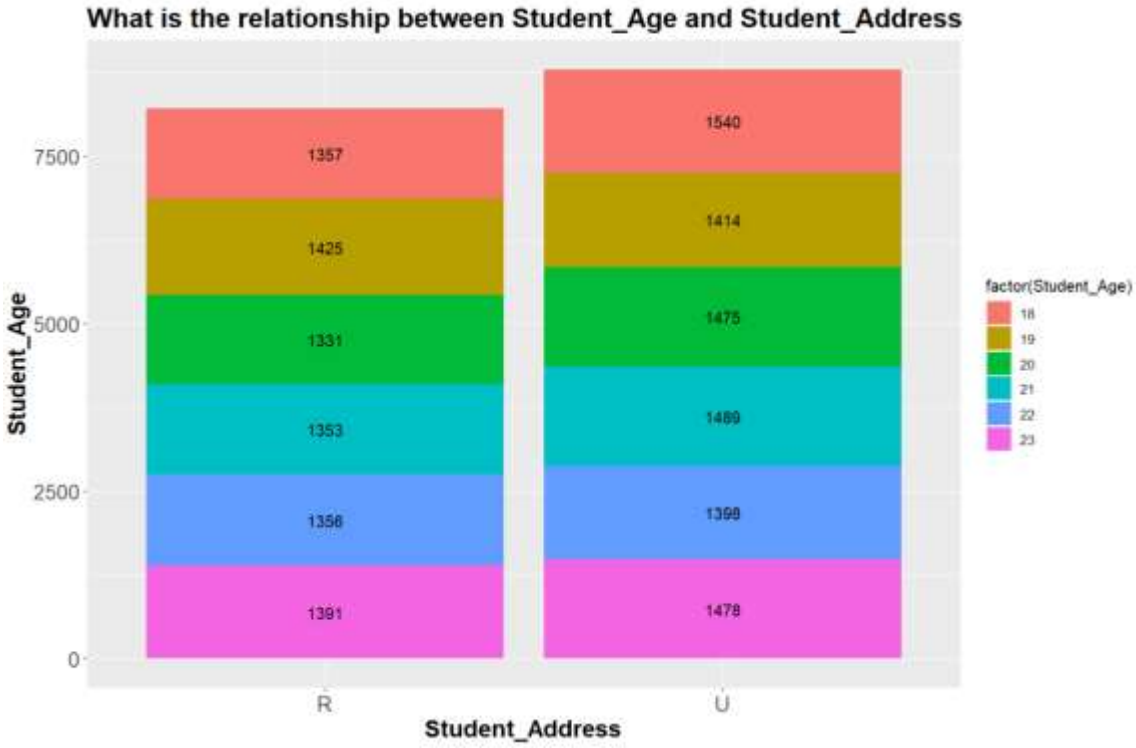


There are more results of yes than no internet usage status in two student living areas, yes status 8717 and no status 8290. It can be understood that urban area students will likely get more job placed status, for their internet usage heavily used.

Analysis 1.3: What is the relationship between student age and student address.

```
ggplot(importData, aes(x=Student_Address, fill=factor(Student_Age))) +
  geom_bar() +
  geom_text(aes(label=..count.., y=..count..),
            stat='count', position=position_stack(vjust=0.5)) +
  labs(x="Student_Address", y="Student_Age",
       title = "Student_Address") +
  ggtitle("What is the relationship between Student_Age and Student_Address") + #with label
  theme(axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title = element_text(size = 16, face = "bold"),
        plot.title = element_text(size = 18, face = "bold"))
```

The graph is created with geom_bar(), geom_text and theme() functions.

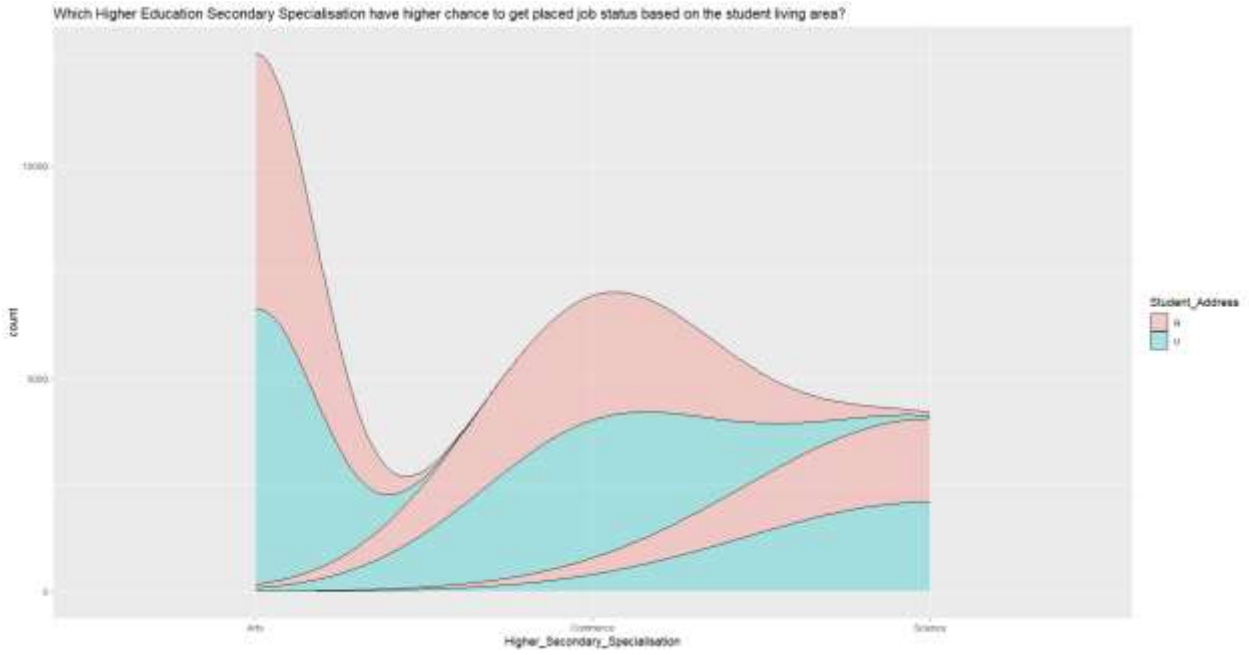


Axis-y is indicating that the age of distribution of each address group, while axis-x is indicating the student address group. The fill color will display different age group, for example, age 18 is red, age 20 is green, age 23 is pink and so on. As can be seen from the graph, there are more students in urban areas than in rural areas, especially 18-year-old students. It can be understood that the job placement status for urban area students have the chances to get in placed status in future.

Analysis 1.4: Which Higher Education Secondary Specialization have higher chance to get placed job status based on the student living area?

```
ggplot(importData, aes(x=Higher_Secondary_Specialisation, stat(count), fill=Student_Address))+
  geom_density(alpha=0.3,position = "stack")+
  labs(x="Higher_Secondary_Specialisation")+
  ggtitle("Which Higher Education Secondary Specialisation have higher chance to get placed job status based on the student living area?")
```

The graph is created from a density map with a transparency of 0.3, with stacked positions counting the frequency of observations. Fill colors will show urban and rural areas to identify graph labels. The axis-x will be showing the higher education specialization.

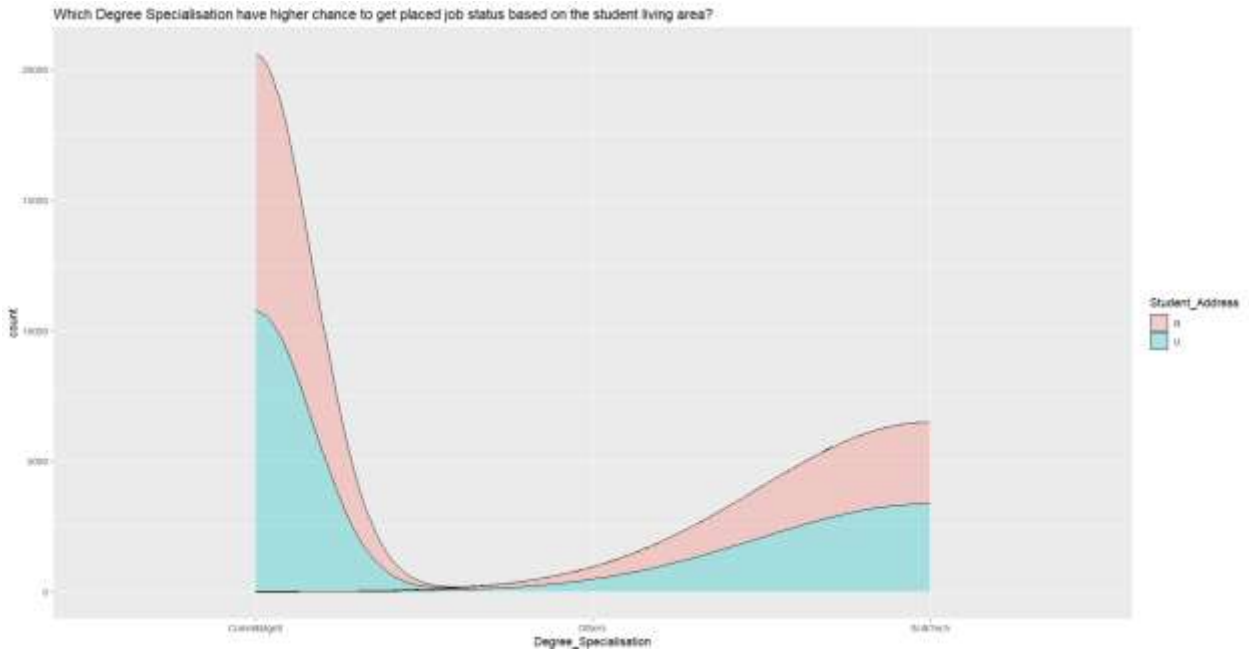


Based on the analysis, it can be assumed that Arts Stream students in Higher Education secondary specialization have higher chance to get job placement status placed in both urban and rural area. Second, students who study Commerce Stream will get job placement status placed. Third, students who study Science Stream will also get job placement status placed. The data for urban and rural areas are nearly identical, making it challenging to predict in which area students will secure employment. The data only provide information on which higher secondary education's specializations are more likely to have higher or lower job placement rates.

Analysis 1.5: Which Degree Specialization have higher chance to get placed job status based on the student living area?

```
ggplot(importData, aes(x=Degree_Specialisation, stat(count), fill=Student_Address))+
  geom_density(alpha=0.3,position = "stack")+
  labs(x="Degree_Specialisation")+
  ggtitle("Which Degree Specialisation have higher chance to get placed job status based on the student living area?")
```

The graph is created from a density map with a transparency of 0.3, with stacked positions counting the frequency of observations. Fill colors will show urban and rural areas to identify graph labels. The axis-x will be showing the student's degree specialization.

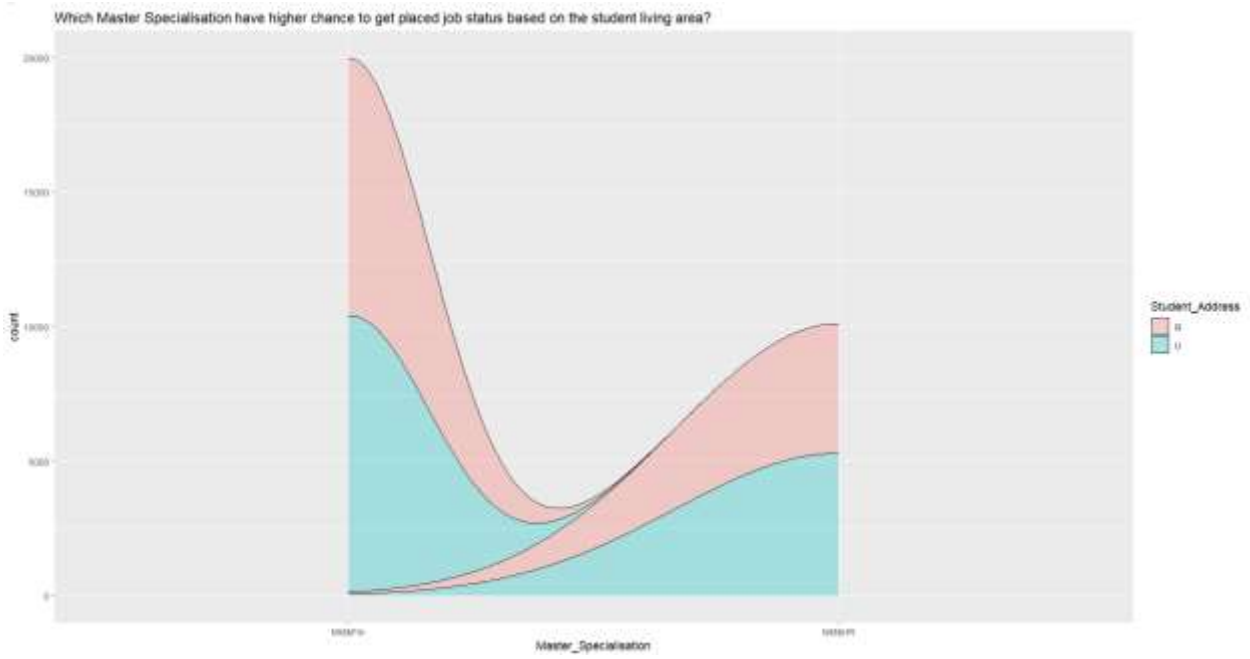


Based on the graph, we can understand that most of the students get job status placed when their degree specialization is Comm&Mgmt or Sci&Tech. Others degree specializations are rarely chosen by students. The data for urban and rural areas are nearly identical, making it challenging to predict in which area students will secure employment. The data only provide information on which degree's specializations are more likely to have higher or lower job placement rates.

Analysis 1.6: Which Master Specialization have higher chance to get placed job status based on the student living area?

```
#analysis 1.6 : Their master specialisation - based on the address
ggplot(importData, aes(x=Master_Specialisation, stat(count), fill=Student_Address))+
  geom_density(alpha=0.3,position = "stack")+
  labs(x="Master_Specialisation")+
  ggtitle("Which Master Specialisation have higher chance to get placed job status based on the student living area?")
```

The graph is created from a density map with a transparency of 0.3, with stacked positions counting the frequency of observations. Fill colors will show urban and rural areas to identify graph labels. The axis-x will be showing the student's master specialization.



Based on the graph, we can understand that most of the students get job status placed when their master specialization is Mkt&Fin. Whereas students who study Mkts&HR master specialization are rarely getting job status placed. The data for urban and rural areas are nearly identical, making it challenging to predict in which area students will secure employment. The data only provide information on which master's specializations are more likely to have higher or lower job placement rates.

Conclusion for Question 1:

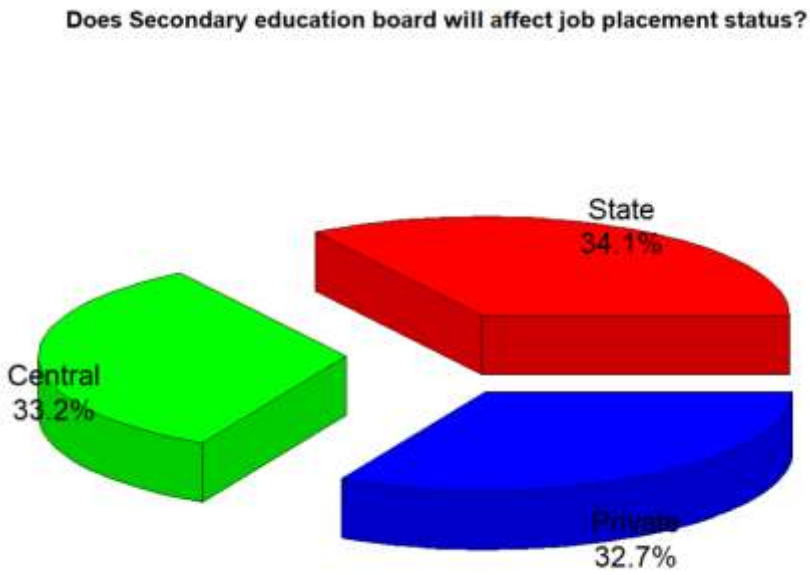
The analysis proved that students with a Master's specialization in Mkt&Fin or Comm&Mgmt have a higher chance of getting job placement status placed, while those studying Mkts&HR have a lower chance. Similarly, students in the Arts, Commerce, and Science streams have a higher likelihood of getting job placement status placed. The distribution of age groups suggests that there are more students in urban areas, particularly 18-year-olds, which may impact their job placement status. Additionally, internet usage is higher among urban area students, indicating a potential correlation with job placement outcomes. Finally, living in either urban or rural areas does not seem to have a significant impact on job placement status.

Question 2: How does the job placement status change based on the level of education in secondary and high schools?

Analysis 2.1: Does Secondary education board will affect job placement status?

```
#Analysis2.1 : Secondary education board
Secondary_Education_Board = c("State", "Central","Private")
pie3D(table(importData$Secondary_Education_Board), labels=Secondary_Education_Board, explode=0.3,
      ggtitle="Does Secondary education board will affect job placement status? ")

# Calculate the percentages
board_counts <- table(importData$Secondary_Education_Board)
board_percentages <- round(prop.table(board_counts) * 100, 1)
board_labels <- paste(Secondary_Education_Board, "\n", board_percentages, "%", sep = "")

# Plot the 3D pie chart with percentages
pie3D(board_counts, labels = board_labels, explode = 0.3,
      main = "Does Secondary education board will affect job placement status?")
```

The chart is created in 3D pie chart with label the pie piece name and show the data with percentage clearly for carry out analysis.

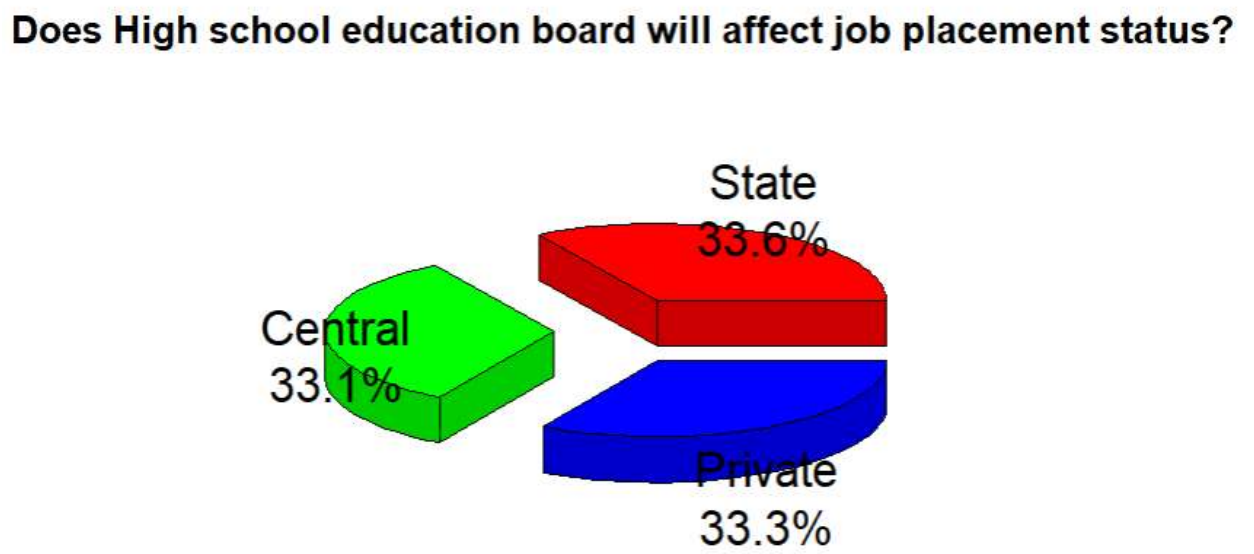


As can be seen from the 3D pie chart, the green pie chart represents the Central Secondary Education Board at 33.2%, the red pie chart represents the State Secondary Education Board at 34.1%, and the blue pie chart represents the Private Secondary Education Board at 32.7%. Based on the analysis, we can understand that the percentage of secondary board education in State will get more job placement status with placed.

Analysis 2.2: Does High school education board will affect job placement status?

```
#Analysis2.2 : High school education board
Higher_Secondary_Education_Board = c("State", "Central","Private")
pie3D(table(importData$Higher_Secondary_Education_Board), labels=Higher_Secondary_Education_Board, explode=0.3,
      ggtitle="Does High school education board will affect job placement status? ")

# Calculate the percentages
board_counts <- table(importData$Higher_Secondary_Education_Board)
board_percentages <- round(prop.table(board_counts) * 100, 1)
board_labels <- paste(Higher_Secondary_Education_Board, "\n", board_percentages, "%", sep = "")

# Plot the 3D pie chart with percentages
pie3D(board_counts, labels = board_labels, explode = 0.3,
      main = "Does High school education board will affect job placement status?")
```

The chart is created in 3D pie chart with label the pie piece name and show the data label with percentage clearly for carry out analysis.

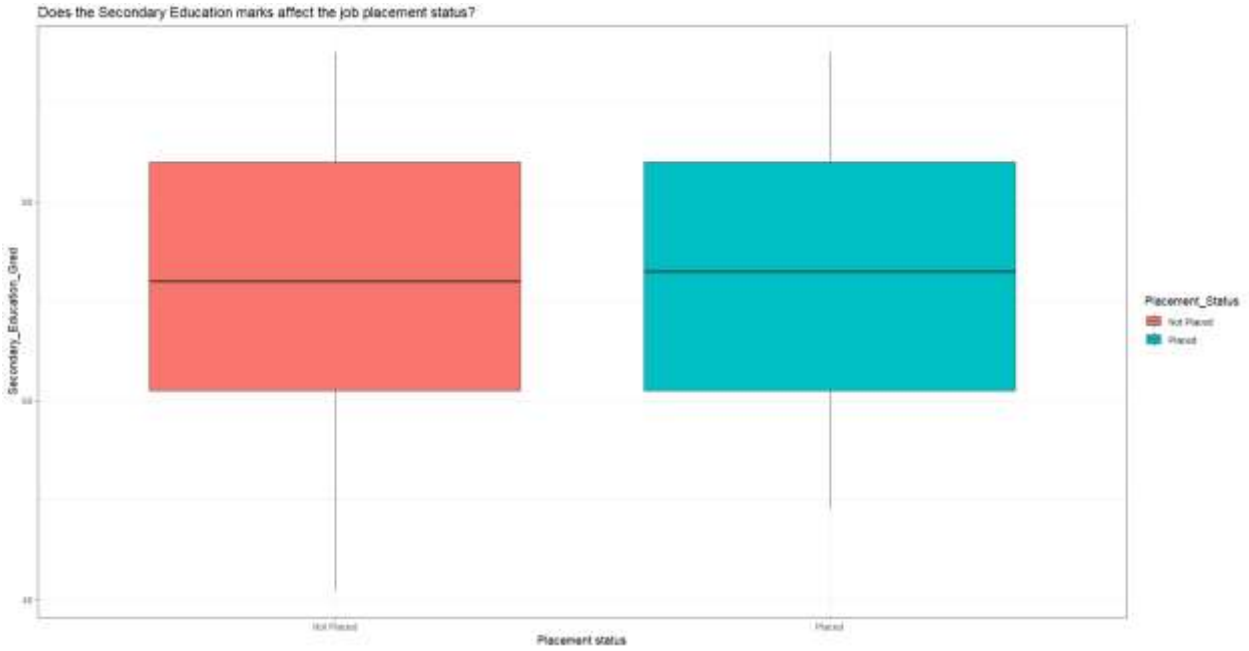


As can be seen from the above 3D pie chart, the green pie chart represents the Central High School Education Board at 33.1%, the red pie chart represents the State High School Education Board at 33.6%, and the blue pie chart represents the Private High School Education Board at 33.3%. Based on the analysis, we can understand that the percentage of secondary board education in State will get more job placement status with placed.

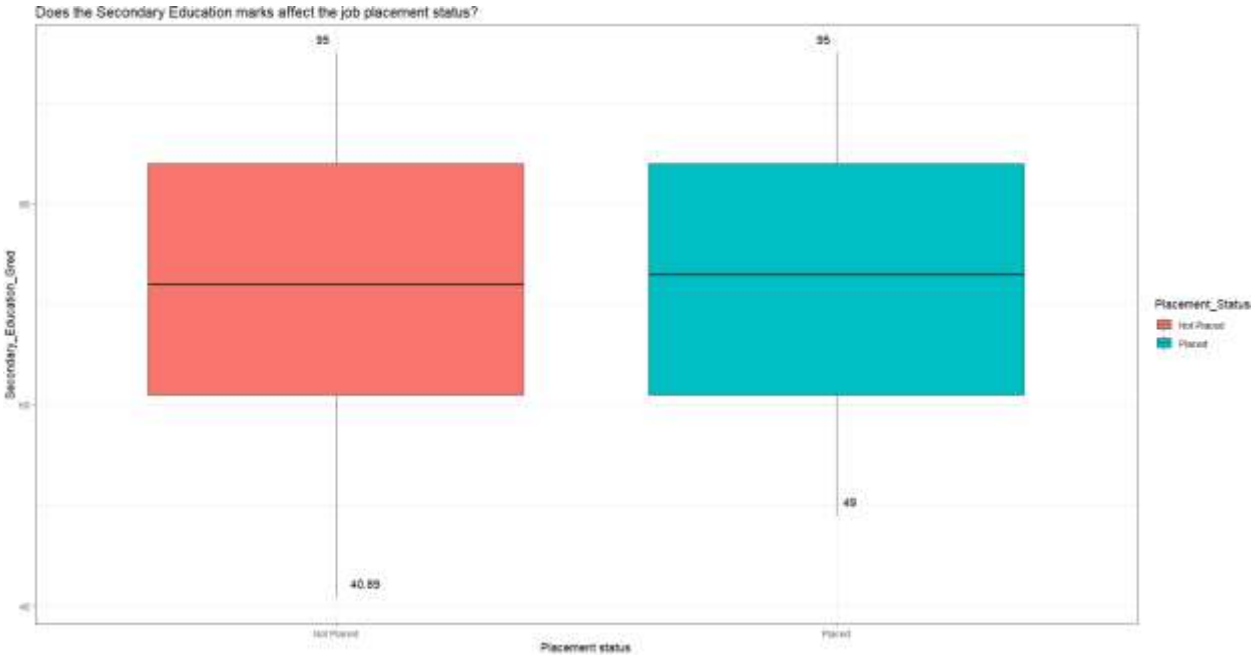
Analysis 2.3: Does the Secondary Education marks affect the job placement status?

```
#Analysis2.3 : Secondary education mark
ggplot(importData, aes(x=Placement_Status, y=Secondary_Education_Gred, fill=Placement_Status))+
  geom_boxplot()+theme_bw()+ggtitle("Does the Secondary Education marks affect the job placement status?")+
  labs(x="Placement status", y="Secondary_Education_Gred")
#add label to define which are the max and min value
ggplot(importData, aes(x = Placement_Status, y = Secondary_Education_Gred, fill = Placement_Status)) +
  geom_boxplot() +
  stat_summary(fun = min, geom = "text", aes(label = round(..y.., 2)), vjust = -1, hjust = -0.5) +
  stat_summary(fun = max, geom = "text", aes(label = round(..y.., 2)), vjust = -1, hjust = 1.5) +
  theme_bw() +
  ggtitle("Does the Secondary Education marks affect the job placement status?") +
  labs(x = "Placement status", y = "Secondary_Education_Gred")
```

The graph is created for showing the data labels representing minimum or maximum values, and the graph type is box plot.



The axis-x is indicating placement status, and axis-y is indicating the secondary education mark.

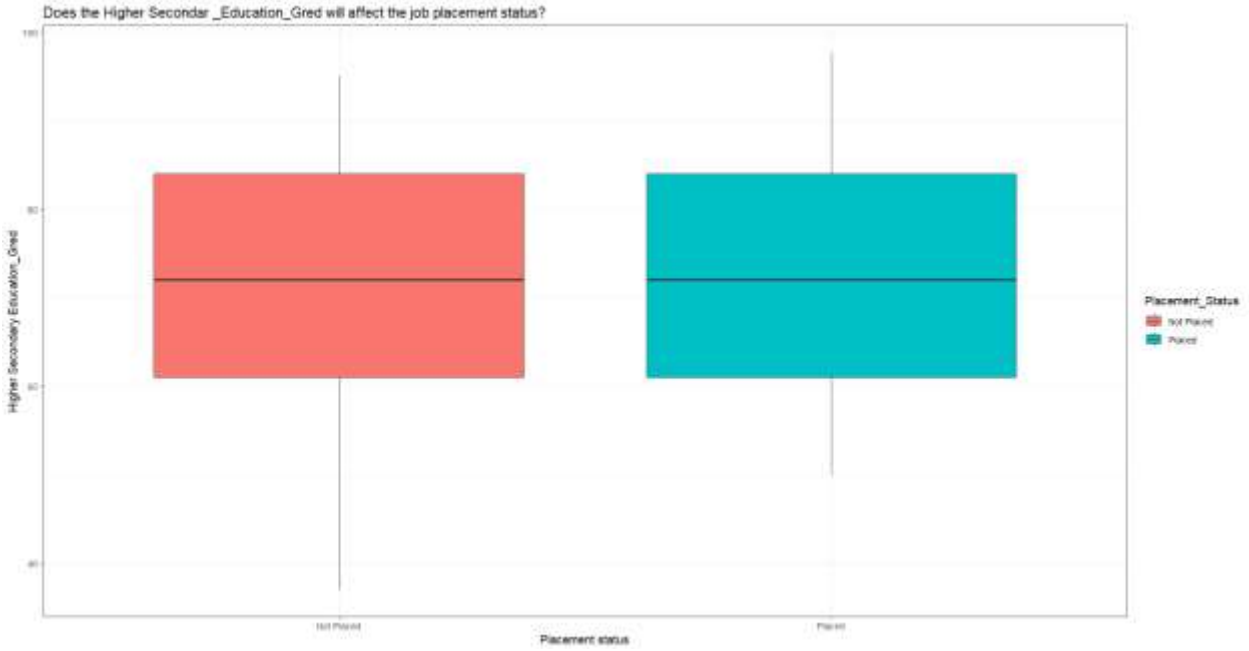


According to the graph, we can know that the highest secondary education mark is 95 and the lowest secondary education mark is 40.89, in not placed status. However, the highest secondary education mark for placement status is same with the not placed status, and the lowest secondary education mark is 49.

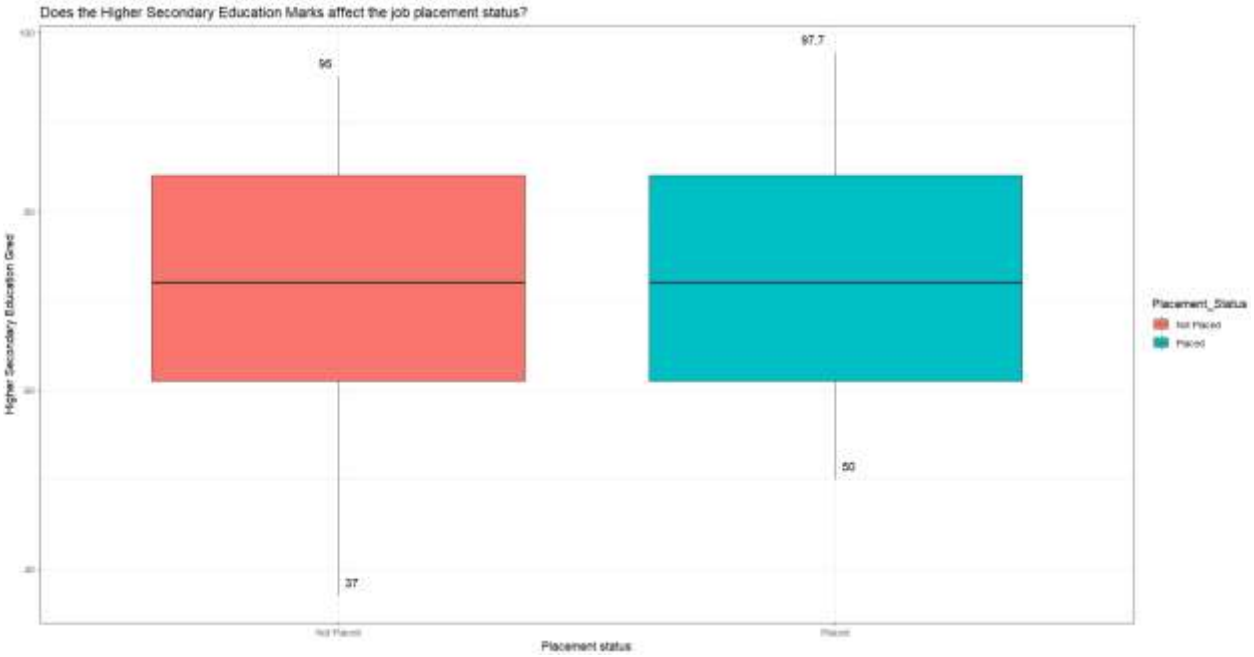
Analysis 2.4: Does the Higher Secondary Education Marks affect the job placement status?

```
ggplot(importData, aes(x = Placement_Status, y = Higher_Secondary_Education_Gred, fill = Placement_Status)) +
  geom_boxplot() +
  stat_summary(fun = min, geom = "text", aes(label = round(..y.., 2)), vjust = -1, hjust = -0.5) +
  stat_summary(fun = max, geom = "text", aes(label = round(..y.., 2)), vjust = -1, hjust = 1.5) +
  theme_bw() +
  ggtitle("Does the Higher Secondary Education Marks affect the job placement status?") +
  labs(x = "Placement status", y = "Higher Secondary Education Gred")
```

The code is created with show the data label maximum and minimum value in the graph, and the graph type is box plot.



The axis-x is indicating placement status, and axis-y is indicating the higher secondary education mark.

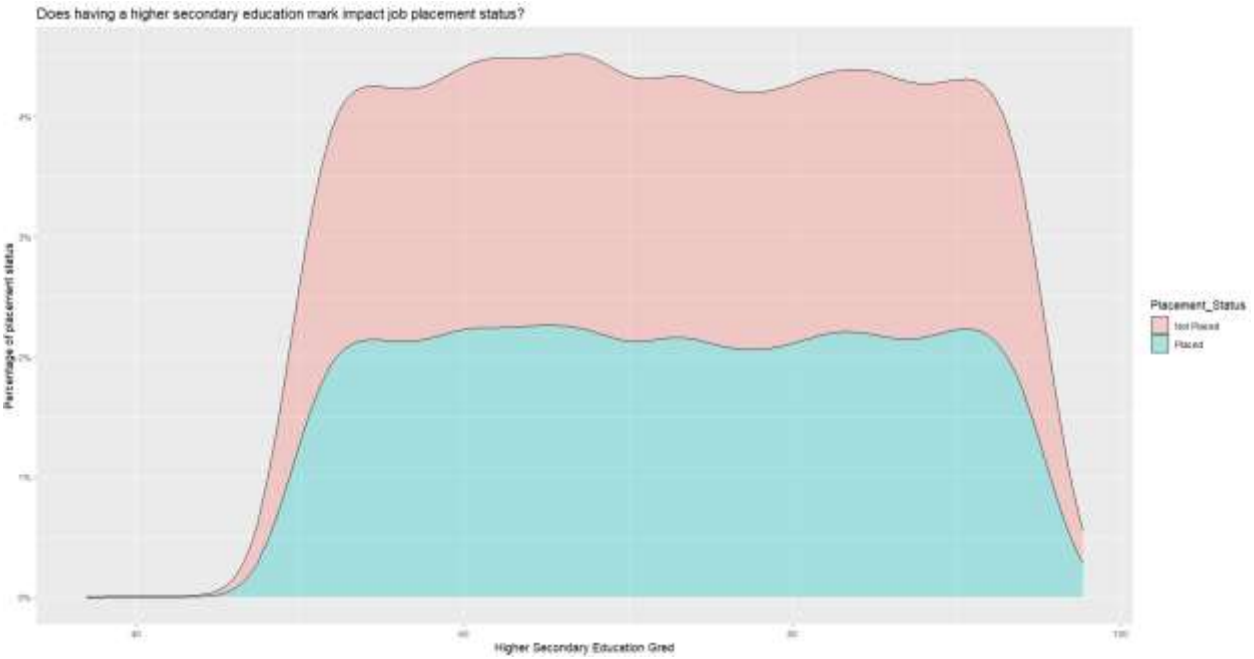


According to the graph, we can know that the highest is higher secondary education mark is 95 and the lowest higher secondary education mark is 37, in not placed status. However, the highest higher secondary education mark for placement status is 97.7, and the lowest high school education mark 50.

Analysis 2.5: Does having a higher secondary education mark will impact to job placement status?

```
ggplot(importData, aes(x=Higher_Secondary_Education_Gred, y=..density.., fill=Placement_Status)) +
  geom_density(alpha=0.3, position = "stack") +
  labs(x="Higher Secondary Education Gred", y="Percentage of placement status") +
  scale_y_continuous(labels=scales::percent_format()) +
  ggtitle("Does having a higher secondary education mark impact job placement status?")
```

The geom_density() function is used to create a density plot, which shows the distribution of data in a continuous manner. The position "stack" is used to stack the density curves for each Placement Status category on top of each other, making it easier to compare the distributions of each category.



The graph shows density of Higher Secondary Education Grade for each Placement Status. Y-axis is scaled to show density, and labels are percentages. According to the analysis, students who achieve higher marks are more likely to receive a job placement status of "not placed" with a percentage of 4. Conversely, students who also achieve high marks but lower than the "not placed" category, have a job placement status of "placed" with a percentage of 2.

Conclusion for Question 2:

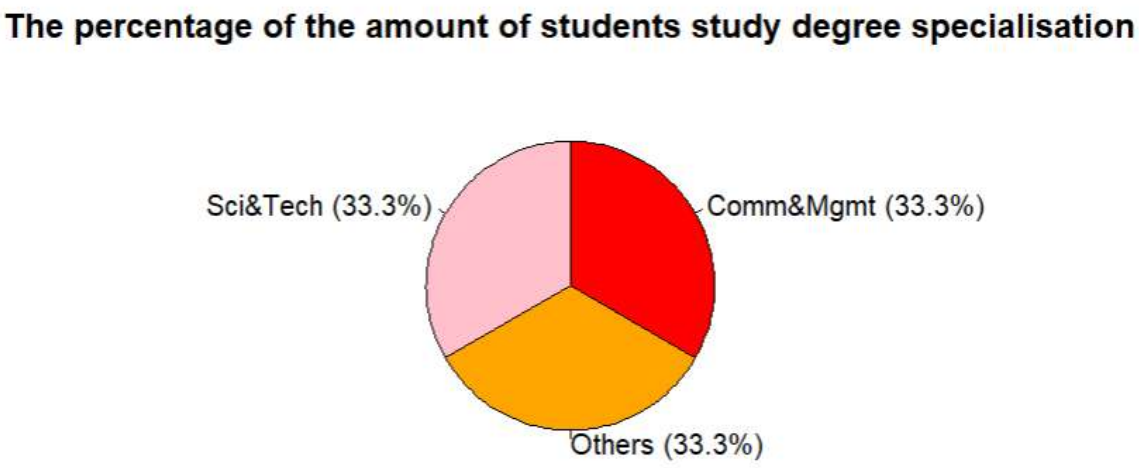
Based on the analysis of the data, it can be inferred that the type of secondary education board does not significantly affect the job placement status. However, when it comes to high school education board, the State High School Education Board has a slightly higher percentage of job placement status with placed. Additionally, the graph indicates that students who score higher marks in secondary and higher secondary education are more likely to receive a job placement status of "not placed" compared to those who score slightly lower marks but still qualify for the "placed" status. This highlights the importance of not only achieving high marks but also meeting the minimum requirements for job placement status.

Question 3: Why did students who specialized in science and technology degrees did not get job placement status placed?

Analysis 3.1: How many students are studying degree course in comm&magtmt, others or sci and tech?

```
Degree_Specialisation = c("Comm&Mgmt", "Others","Sci&Tech")
prop <- prop.table(table(Degree_Specialisation)) * 100
pie(prop, labels = paste0(names(prop), " (", round(prop, 1), "%)"),
    main = "The percentage of the amount of students study degree specialisation",
    col = c("red", "orange","pink"), clockwise = TRUE)
```

The chart is created into a pie chart with label name on the pie piece and show the data label with percentage clearly for carry out analysis.

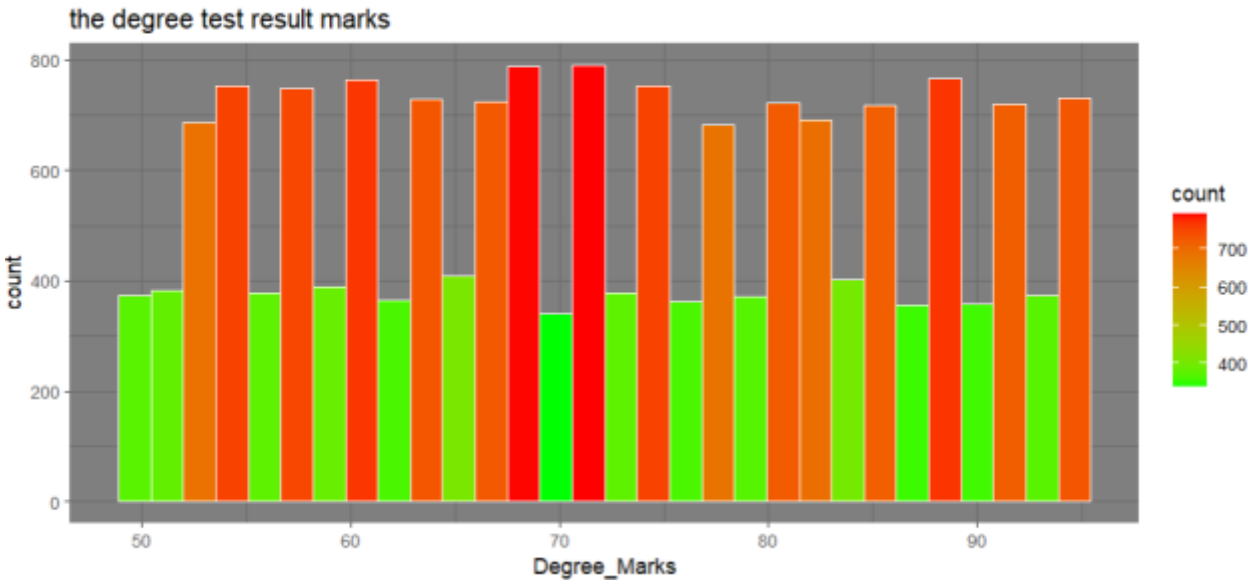


According to the chart, the data shows that the data is the same for all degree specialization, and it can be understood that there may be an oversupply status of graduate's field, which could make it more difficult for students to find job placements. However, it is important to notice that it could be other factors such as individual skills or job search technique.

Analysis 3.2: Analyze the degree marks of students.

```
ggplot(importData, aes(x=Degree_Marks, fill=Degree_Marks))+
  geom_histogram(color="white", aes(fill=..count..))+
  scale_fill_gradient("count", low="green", high="red")+
  labs(title="the degree test result marks")+
  theme_dark()
```

In order to analyze the degree test result marks, the geom_histogram() function will be colored with white on the graph to display the data and scale fill gradient function to show mark value (low in green and high in red).

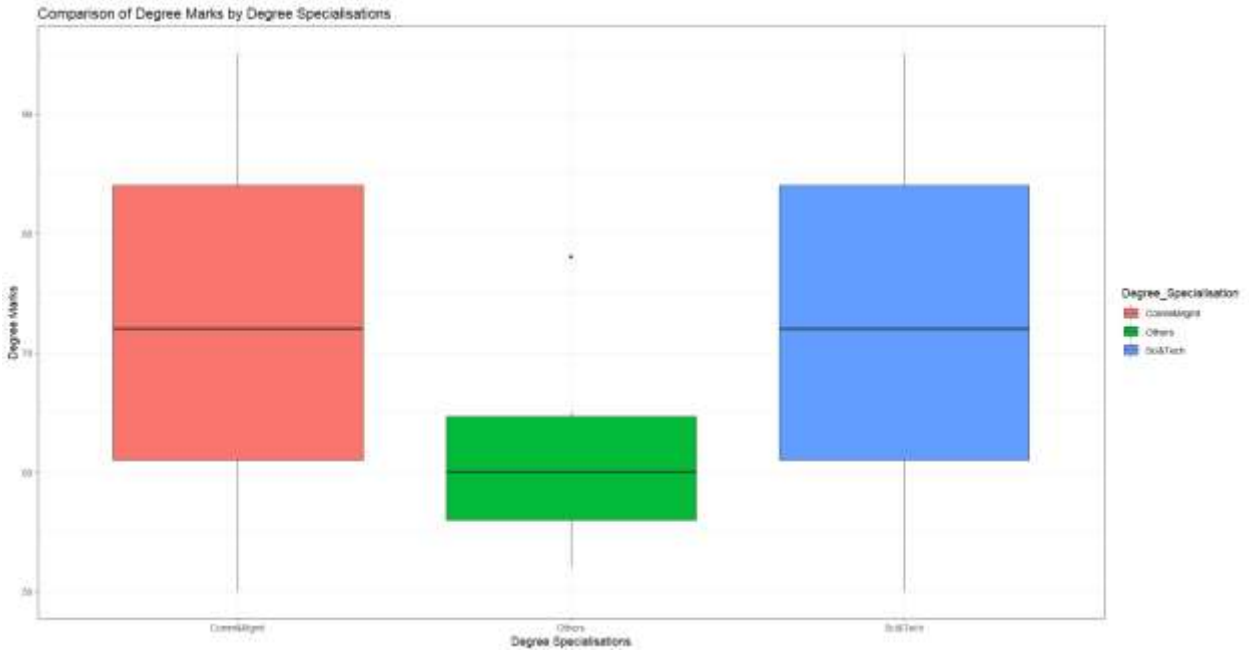


Based on the bar chart showing the degree scores of all students, it can be inferred that the majority of students scored marks above 70%. This issue means that students may face challenges in qualifying for job placement, but this does not necessarily mean that they are unlikely to obtain job placement.

Analysis 3.3: Does degree marks and degree specialization will affect for placement status?

```
ggplot(importData, aes(x = Degree_Specialisation, y = Degree_Marks, fill = Degree_Specialisation)) +
  geom_boxplot() +
  labs(x = "Degree Specialisations", y = "Degree Marks") +
  ggtitle("Comparison of Degree Marks by Degree Specialisations") +
  theme_bw()
```

The graph is created by boxplot to combine two data for analysis with function theme_bw() to returns the theme become black-and-white.



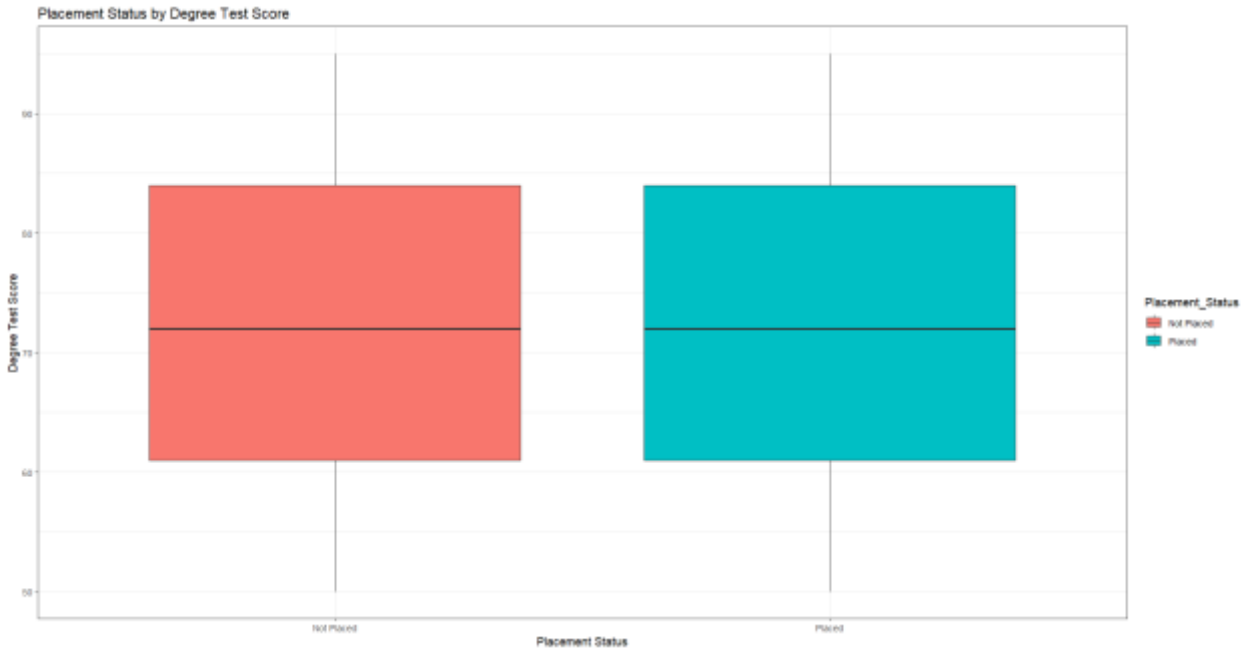
The data combines two data for comparison, the student's degree exam scores and the degree specialization.

It is understood that students in Comm&Mgmt or Sci&Tech specializations may have higher chances of getting job placement status compared to students in other specializations, as they have higher degree exam scores and their marks are almost similar, which are maximum is 95, median is 72 and minimum is 50. However, it should be noted that students in other specializations may still get job placement status despite having lower degree exam scores. As seen from the boxplot label, the others specializations degree marks minimum data is 52, and Comm&Mgmt or Sci&Tech specializations degree marks minimum are lowest than others specialization, which is 50.

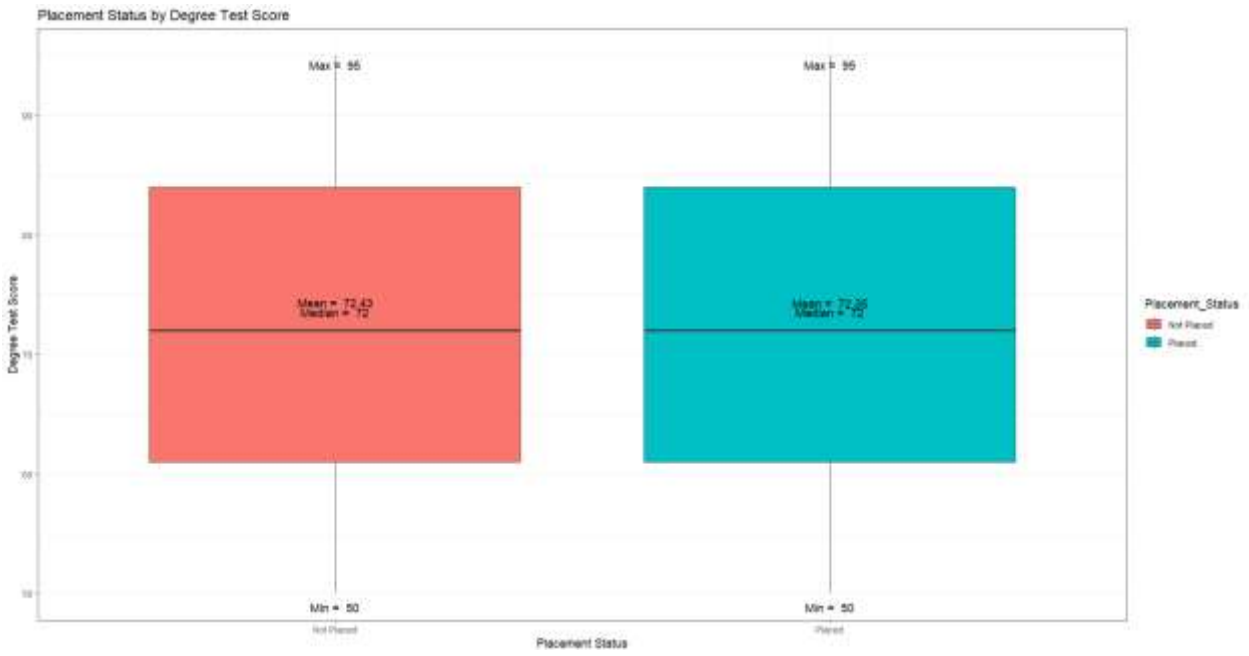
Analysis 3.4: Does placement status have relationship with degree test score?

```
ggplot(importData, aes(x = Placement_Status, y = Degree_Marks, fill = Placement_Status)) +
  geom_boxplot() +
  labs(x = "Placement Status", y = "Degree Test Score") +
  ggtitle("Placement Status by Degree Test Score") +
  theme_bw()
```

The graph is created by boxplot to combine two data for analysis with function theme_bw() to returns the theme become black-and-white.



The purpose of combining these two data for analysis is to find the relationship between them. According to the graph, the job placement status and degree test scores are the same, so the two data are averaged. Therefore, students with low- or high-test scores will also face the problem of placed or unplaced job placement status.

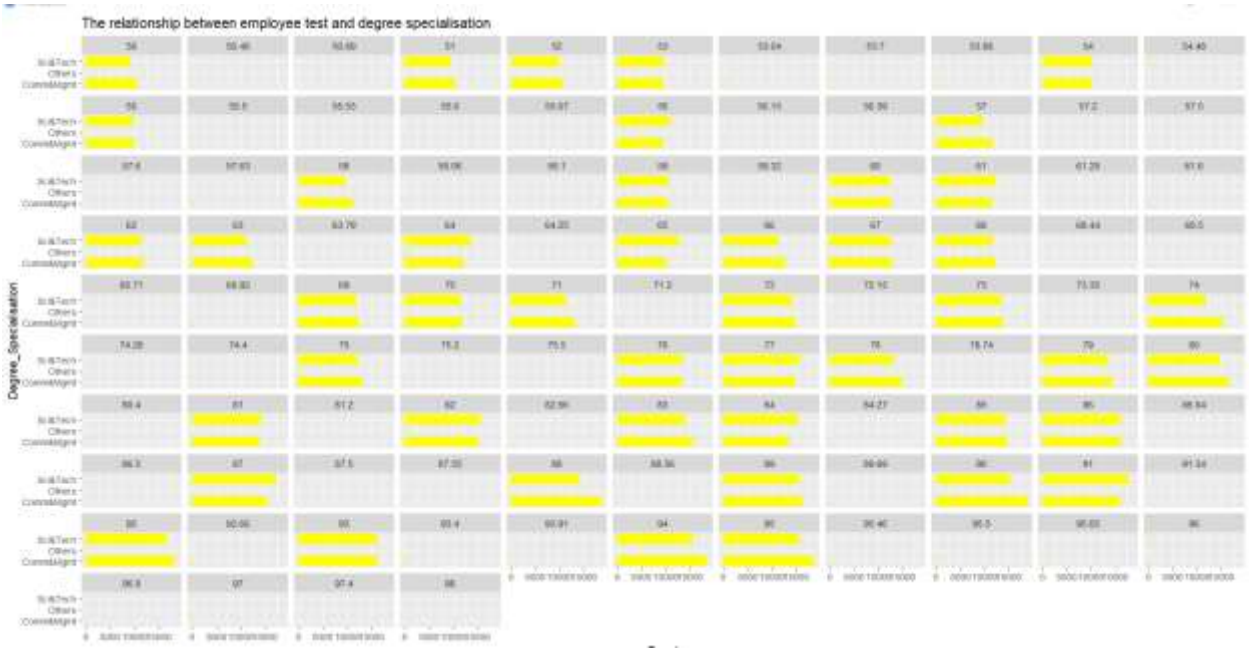


However, after adding the data label it can be clearly seen that placed status degree test score mean is lowest than not placed status. The not placed status degree test score mean is 72.43, and the placed status degree test score mean is 72.35. Both placed and not placed status their minimum, maximum and median have the equal amount.

Analysis 3.5 : The relationship between employee test and degree specialisation

```
#analysis 3.5: employee test and degree specialisation
ggplot(importData, aes(Employee_Test, Degree_Specialisation))+geom_bar(stat="identity", fill="yellow")+
  facet_wrap(~Employee_Test)+labs(x="Count", y="Degree_Specialisation")+
  ggtitle("The relationship between employee test and degree specialisation")
```

The graph is combining two, which are bar and facet wrap by faceting multiple variables automatically. The facets will be separated into multiples through the test marks value from the import data.



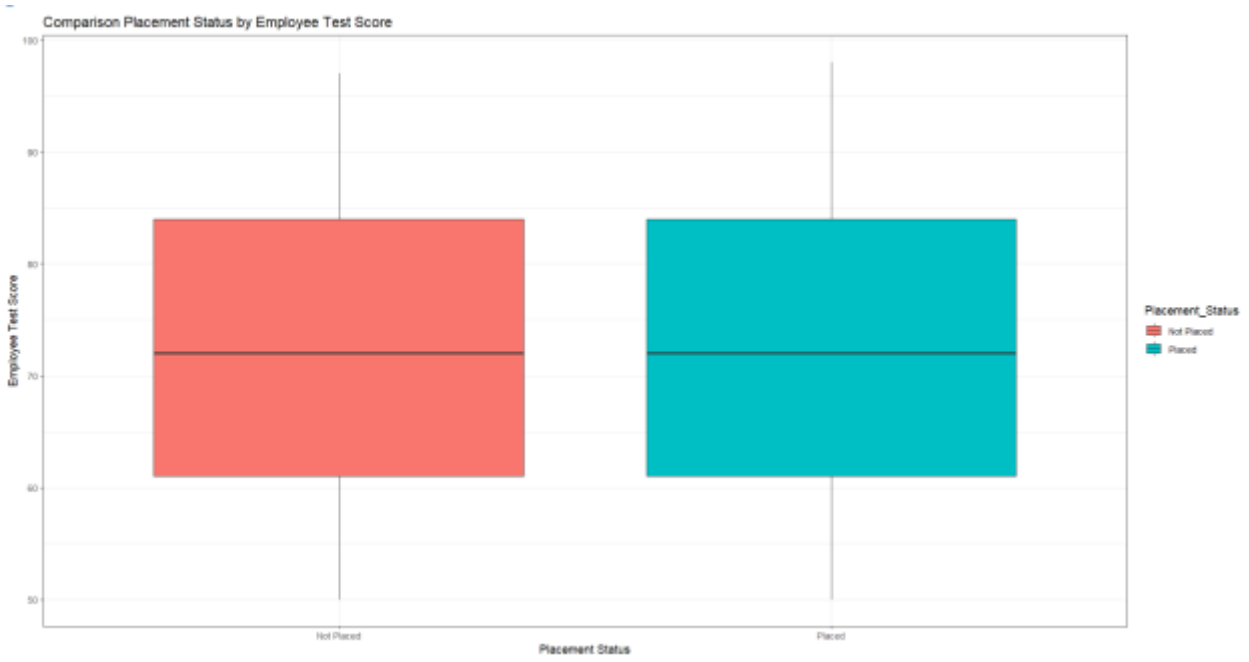

```
> min(importData$Employee_Test)
[1] 50
> max(importData$Employee_Test)
[1] 98
```

The analysis was performed by combining two data, namely employee tests and degree specializations. Based on the analysis, the graph shows multiple employee test results, with no employee tests for other degree specializations, while the employee test for Scie&Tech or Comm&Mgmt with a minimum test score result of 50 and maximum test score result is 98. Therefore, it can be understood that others degree specialization will not get job placed status mostly.

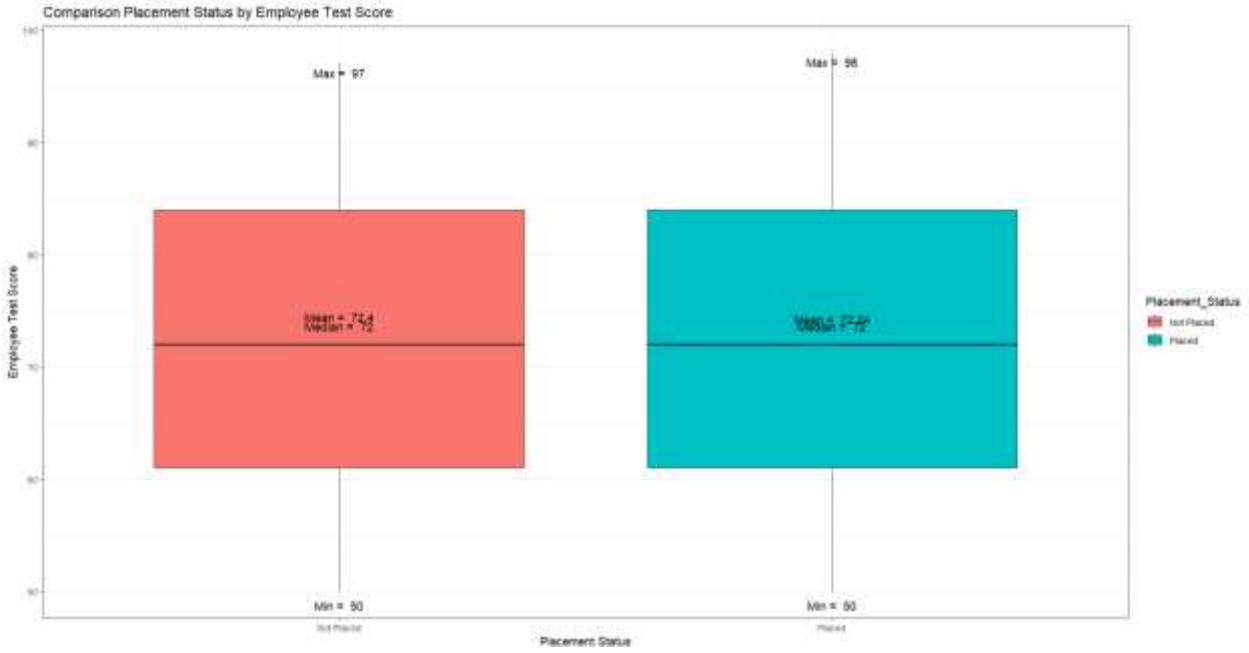
Analysis 3.6: Does placement status have relationship with employee test score?

```
ggplot(importData, aes(x = Placement_Status, y = Employee_Test, fill = Placement_Status)) +
  geom_boxplot() +
  labs(x = "Placement Status", y = "Employee Test Score") +
  ggtitle("Comparison Placement Status by Employee Test Score") +
  theme_bw()
```

The graph is created by boxplot to combine two data for analysis with function theme_bw() to returns the theme become black-and-white.



The purpose of combining these two data for analysis is to find the relationship between them. According to the graph, the job placement status and employee test scores are the same, so the two data are averaged. Therefore, students with low- or high-test scores will also face the problem of placed or unplaced job placement status.



With adding the show data label, we can understand that the employee test score in not placed status maximum is 97, mean is 72.4, median is 72 and minimum is 50. Whereas employee test score in placed status maximum is 98, mean is 72.24, median is 72 and minimum is 50.

Conclusion for Question 3:

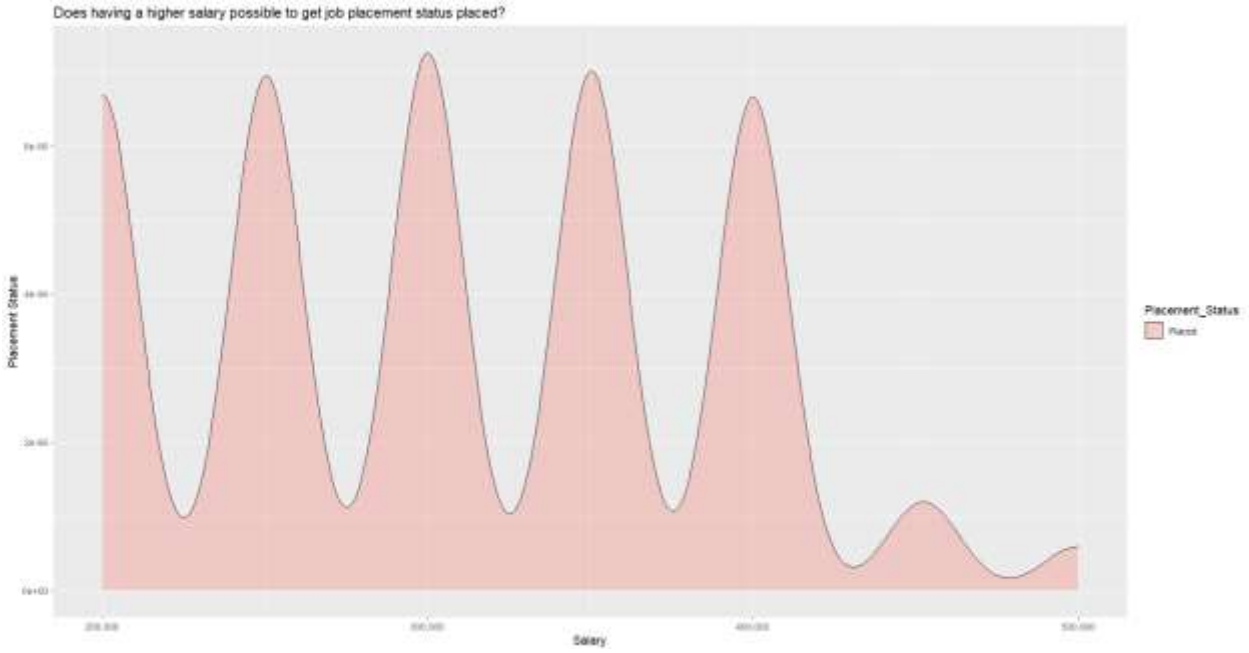
Based on all the analysis, no conclusive evidence that students who study in Sci&Tech get job placed or not placed status due to all the data being similar and average with the minimum, maximum and median values for the degree test marks and employee test marks. One possible reason students will have difficulty placing job status is that high expectations may contribute to job placement difficulties; equal marks imply strong competence.

Question 4: Does students having the highest salary paid because of their master specialization, work experience or MBA test marks?

Analysis 4.1: Does having a higher salary possible to get job placement status placed?

```
ggplot(importData, aes(x = Salary, fill = Placement_Status)) +
  geom_density(alpha = 0.3, position = "stack") +
  scale_x_continuous(labels = scales::comma, limits = c(2e5, 5e5)) + #format the label
  labs(x = "Salary", y = "Placement Status") +
  ggtitle("Does having a higher salary possible to get job placement status placed?")
```

The graph is created in geom_density() with alpha 0.3, position stack and adjust graph the format label on axis-y.

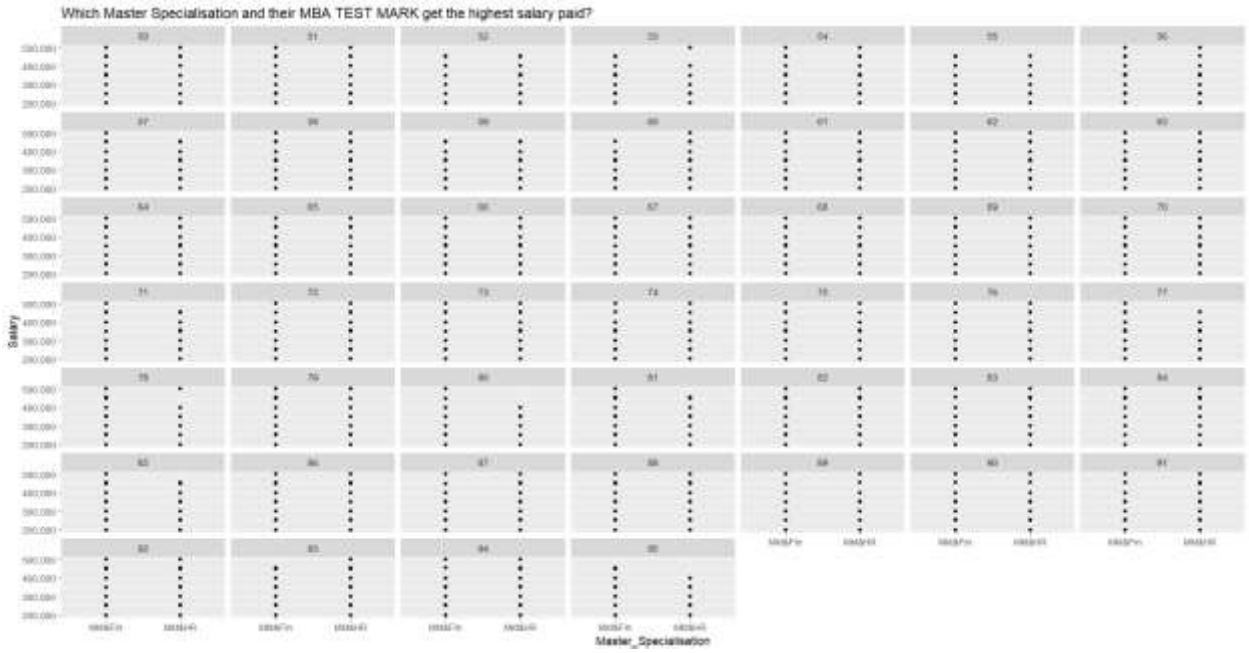


It can be seen from the figure that the maximum salary of $500,000 does not mean that the status of job placement can be placed, and most of them are not placed. While salaries worth between $200,000 and $400,000 may get more job placement status in placed. In fact, the job placement status is placed, and the student's salary will be paid based on the data import.

Analysis 4.2: Which Master Specialization and their MBA test marks will be possibly get the highest salary paid?

```
ggplot(importData, aes(Master_Specialisation, Salary)) +
  geom_point(stat = "identity") +
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  facet_wrap(~Salary) +
  labs(x = "Master_Specialisation", y = "Salary") +
  ggtitle("Which Master Specialisation get the highest salary paid?")
```

The graph is created in geom_point stat identity, scale the axis-y data label and facet wrap test marks into multiple graphs.

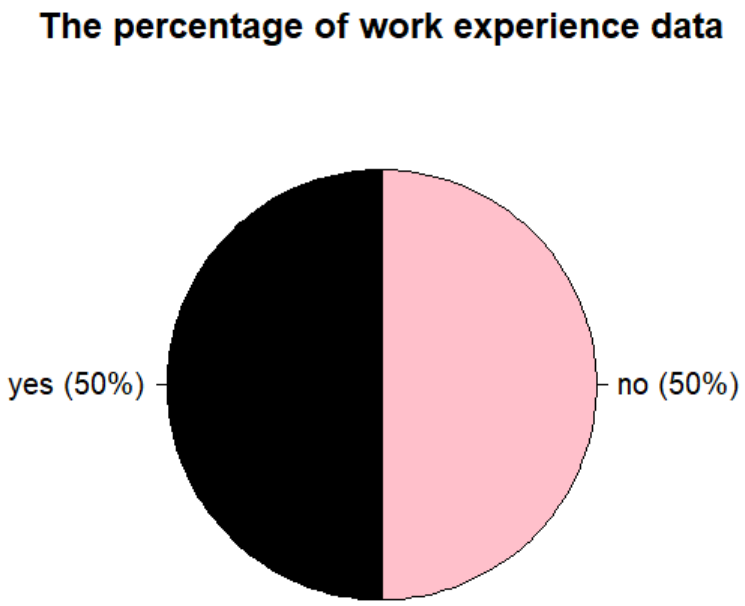


As you can see from the graph, students who study Mkt&Fin and their MBA test mark is 95% their salary amount is the highest, which is $450,000. While students who study Mkt&HR and their MBA test marks with 93% their salary amount is the highest, which is $500,000. Both of master specialization score with test mark 94% and get salary amount are equal, which is $500,000.

Analysis 4.3: The work experience for the students

```
#Analysis 4.3: work experience
Work_Experience = c("yes","no")
pie(table(importData$Work_Experience), Work_Experience, main= "The work experience data",
    col = c("pink","black"),clockwise = TRUE)
#to show the percentage
Work_Experience = c("yes","no")
prop <- prop.table(table(Work_Experience)) * 100
pie(prop, labels = paste0(names(prop), " (", round(prop, 1), "%)"), main = "The percentage of work experience data",
    col = c("pink", "black"), clockwise = TRUE)
```

The chart is created with pie() function with color pink, black and clockwise.

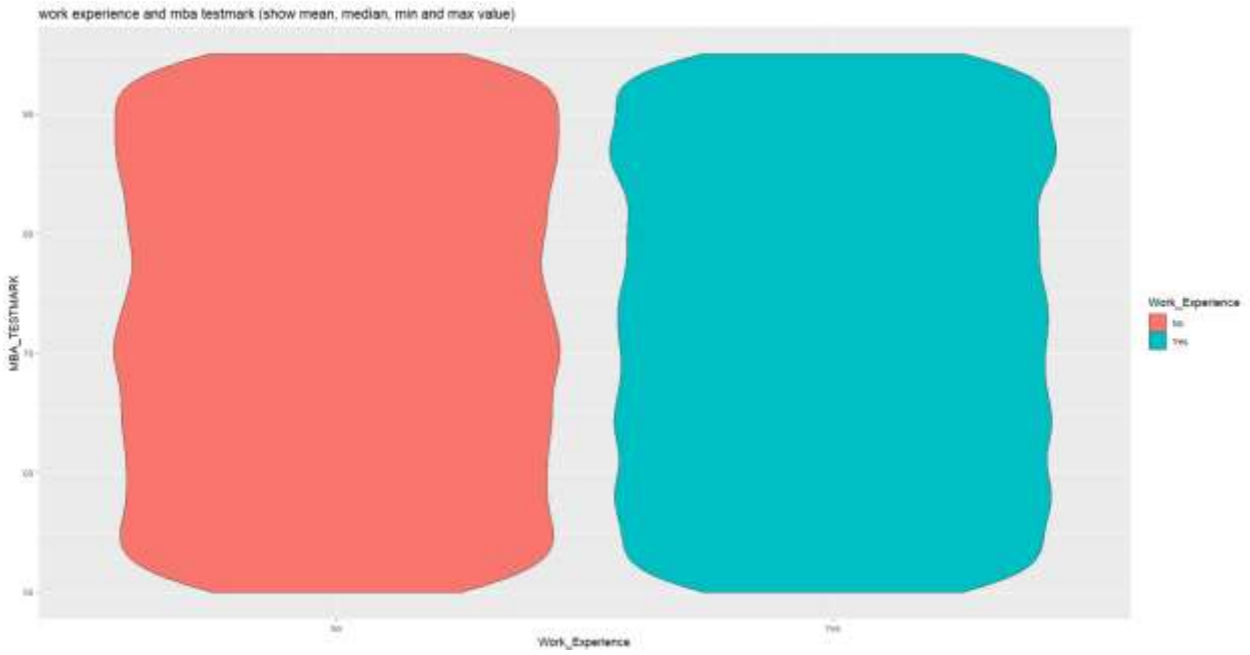


The Pie chart show the data of work experience for the students have 50% Yes and No. The black pie piece is indicating work experience yes status, and pink is indicating work experience no status.

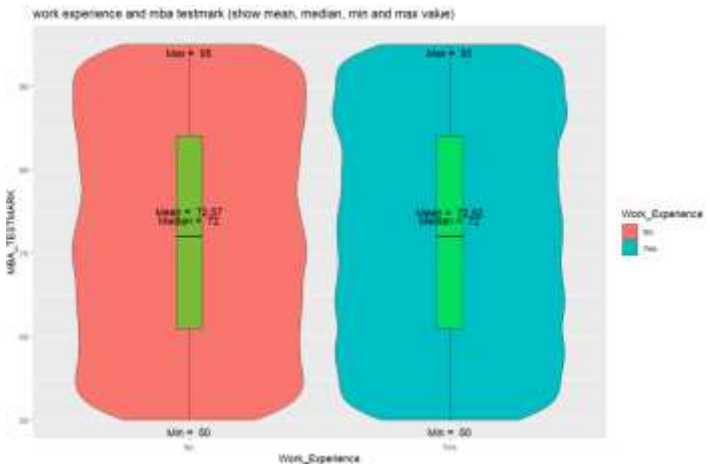
Analysis 4.4: What is the MBA test mark (show min and max value) that have relationship to the work experience?

```
# analysis 4.4 : work experience and mba testmark (show min and max value)
ggplot(importData, aes(x = Work_Experience, y = MBA_TESTMARK, fill= Work_Experience)) +
  geom_violin() +
  labs(x = "Work_Experience", y = "MBA_TESTMARK",
       title = "work experience and mba testmark (show mean, median, min and max value)")

#add the box plot in violist - work experience and mba testmark (show min and max value)
ggplot(importData, aes(x = Work_Experience, y = MBA_TESTMARK, fill= Work_Experience)) +
  geom_violin() +
  geom_boxplot(width=0.1, fill="green", alpha=0.5, outlier.color="white") +
  stat_summary(fun = median, geom = "text", aes(label = paste("Median = ", round(..y..,2))),
               vjust = -1.5, show.legend = FALSE) +
  stat_summary(fun = mean, geom = "text", aes(label = paste("Mean = ", round(..y..,2))),
               vjust = -2, show.legend = FALSE) +
  stat_summary(fun = min, geom = "text", aes(label = paste("Min = ", round(..y..,2))),
               vjust = 2, show.legend = FALSE) +
  stat_summary(fun = max, geom = "text", aes(label = paste("Max = ", round(..y..,2))),
               vjust = 1.5, show.legend = FALSE) +
  labs(x = "Work_Experience", y = "MBA_TESTMARK",
       title = "work experience and mba testmark (show mean, median, min and max value)")
```

The graph is created with geom_violin() and geom_boxplot() with adding width 0.1, fill green, alpha 0.5 and outlier color white.



The axis-x is indicating work experience, and the axis-y is indicating MBA test marks. The fill will show the work experience status yes and no.

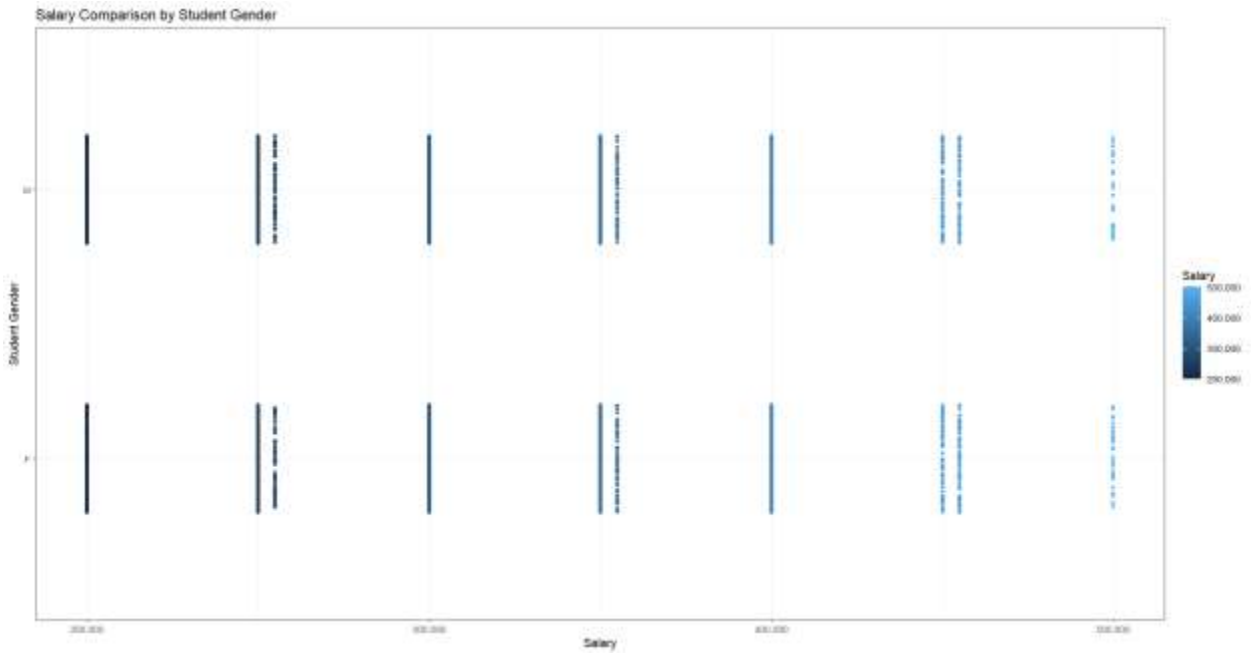


Based on the graph, the data label will be shown with the maximum value, mean, median and minimum. According to the plot, the two are almost similar for the yes or no status of the work experience with minimum = 50, median = 72 and maximum = 95. The only different is to a mean of 72.57 for the no status and 72.52 for the yes status.

Analysis 4.5: Which student gender will get highest salary paid?

```
#analysis 4.5 : Salary Comparison by Student Gender
ggplot(importData, aes(x = Salary, y = Student_Gender, color = Salary)) +
  geom_point(position = position_jitterdodge(dodge.width=0.75, jitter.height = 0.2)) +
  scale_x_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  scale_color_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  labs(x = "Salary", y = "Student Gender") +
  ggtitle("Salary Comparison by Student Gender") +
  theme_bw()
```

The graph is created with geom_point() and scale the data label in axis-y and fill color with theme_bw().

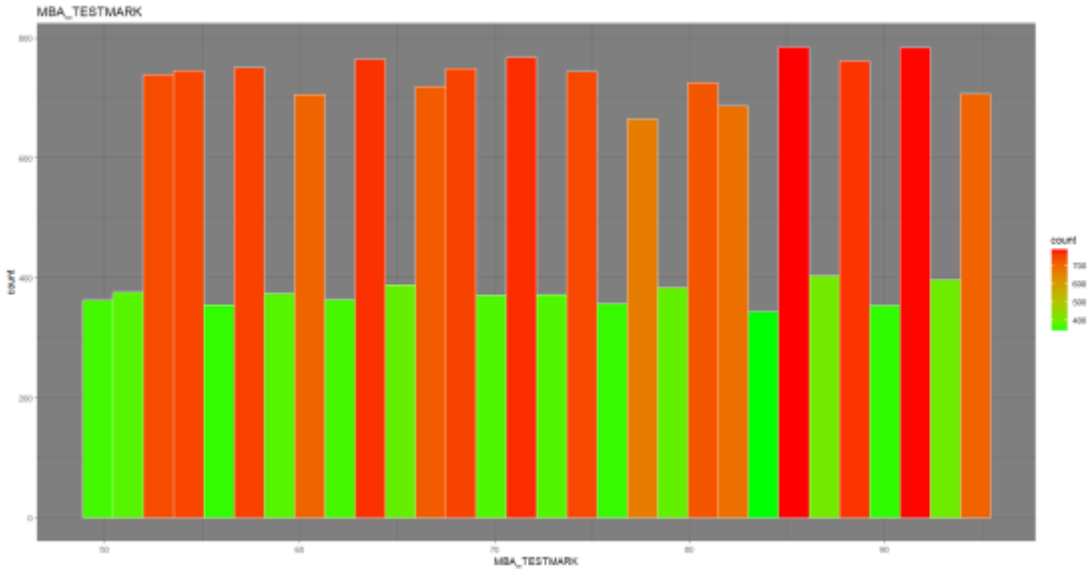


As you can see from the graph, it is difficult to determine which student's gender will get the highest salary because all students will get the same salary paid based on different salary amounts.

Analysis 4.6: Does students score highest MBA test score will get highest salary paid?

```
ggplot(importData, aes(x=MBA_TESTMARK, fill=MBA_TESTMARK))+
     geom_histogram(color="white", aes(fill=..count..))+
    scale_fill_gradient("count", low="green", high="red")+
    labs(title="MBA_TESTMARK")+
    theme_dark()
```

In order to analyze the MBA test result marks, the geom_histogram() function will be colored with white on the graph to display the data and scale fill gradient function to show mark value (low in green and high in red).



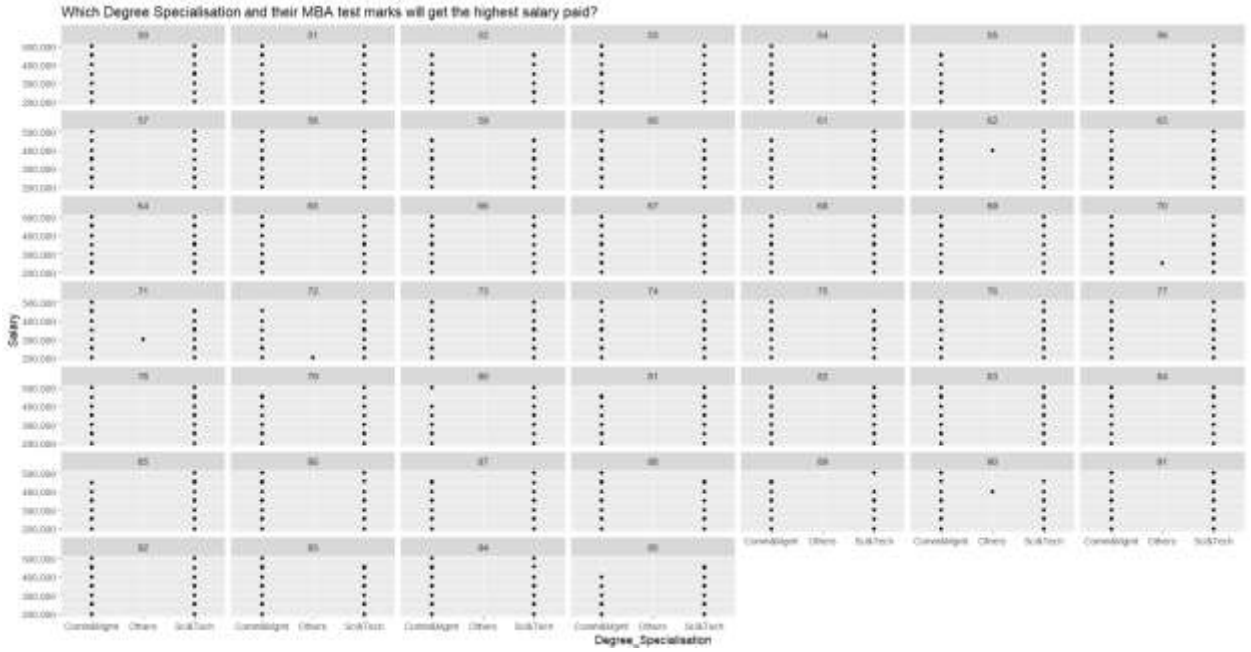

```
> max(importData$MBA_TESTMARK)
[1] 95
> mean(importData$MBA_TESTMARK)
[1] 72.54478
> median(importData$MBA_TESTMARK)
[1] 72
> min(importData$MBA_TESTMARK)
[1] 50
```

The score maximum MBA test marks is 95%, mean is 72.54478%, median is 72%, minimum is 50%. According to the data, the job placement status of students appears to be evenly split between those who have been placed and those who have not. As seen from the graph, most students can score their marks above 60%.

Analysis 4.7: Which Degree Specialization and their MBA test marks will get the highest salary paid?

```
ggplot(importData, aes(Degree_Specialisation, Salary)) +
  geom_point(stat = "identity") +
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  facet_wrap(~MBA_TESTMARK) +
  labs(x = "Degree_Specialisation", y = "Salary") +
  ggtitle("Which Degree Specialisation and their MBA test marks will get the highest salary paid?")
```

The graph is created in geom_point stat identity, scale the axis-y data label and facet wrap test marks into multiple graphs.

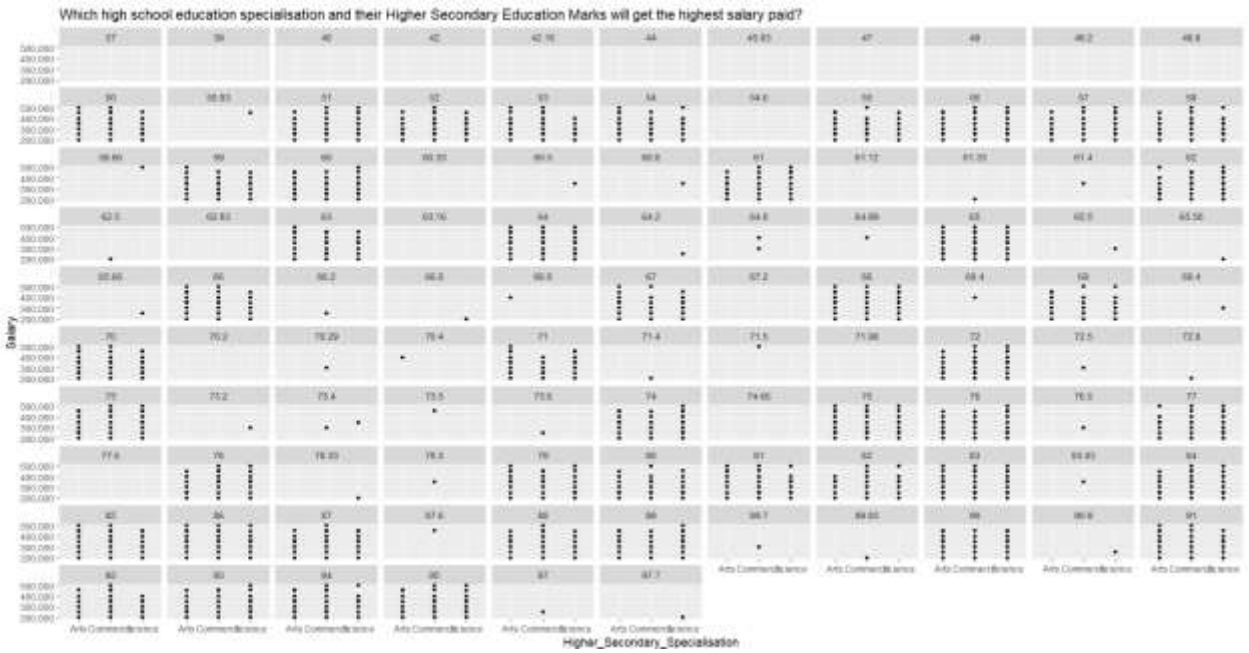


As you can see from the graph, Sci&Tech specialization that score 95% MBA test marks will be possible to get highest salary paid amount $500,000. While others degree specialization that score MBA test marks 90% get highest salary paid amount is $400,000 and Comm&Mgmt score MBA test marks is 93% get highest salary paid amount is $500,000.

Analysis 4.8: Which high school education specialization and their Higher Secondary Education Marks will get the highest salary paid?

```
ggplot(importData, aes(Higher_Secondary_Specialisation, Salary)) +
  geom_point(stat = "identity") +
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  facet_wrap(~Higher_Secondary_Education_Gred) +
  labs(x = "Higher_Secondary_Specialisation", y = "Salary") +
  ggtitle("Which high school education specialisation and their Higher Secondary Education Marks will get the highest salary paid?")
```

The graph is created in geom_point stat identity, scale the axis-y data label and facet wrap test marks into multiple graphs.

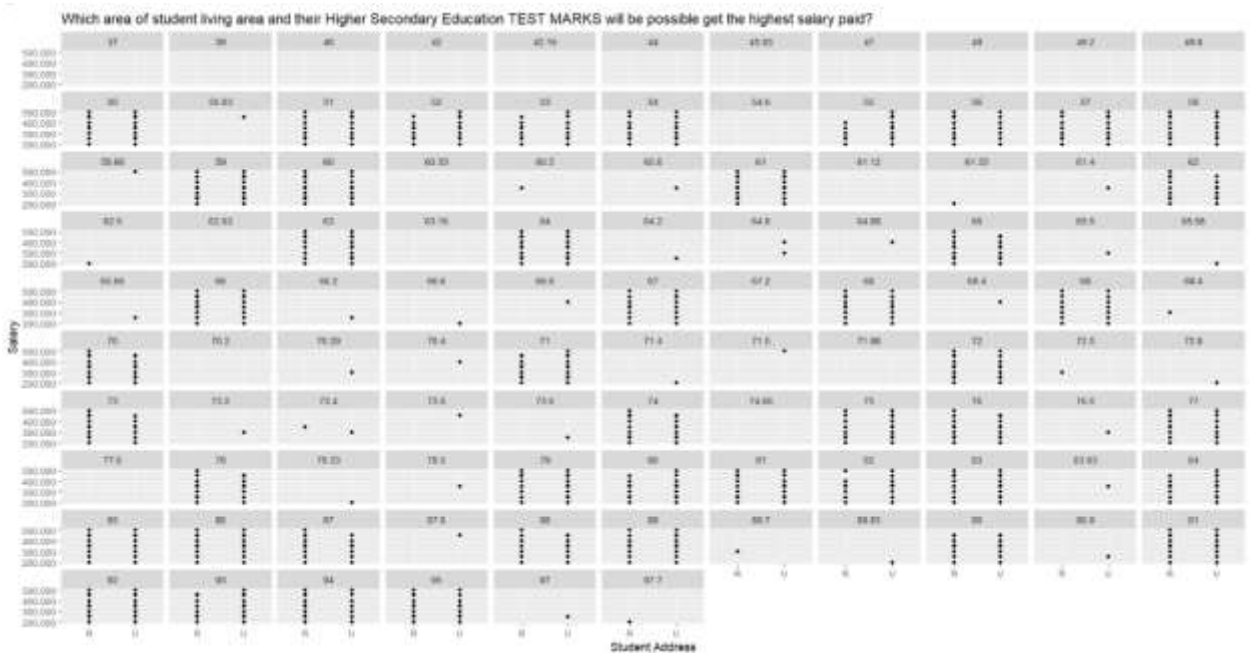


Students who study Science or Commerce Stream with score their test marks with 95% in higher secondary education specialization will be possible to get highest salary paid with amount $500,000. For students who score 97.7% is Science Stream, and their salary paid with amount is $200,000.

Analysis 4.9: Which area of student living area and their Higher Secondary Education TEST MARKS will be possibly getting the highest salary paid?

```
ggplot(importData, aes(Student_Address, Salary)) +
  geom_point(stat = "identity") +
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  facet_wrap(~Higher_Secondary_Education_Gred) +
  labs(x = "Student Address", y = "Salary") +
  ggtitle("Which area of student living area and their Higher Secondary Education TEST MARKS will be possible get the highest salary paid?")
```

The graph is created in geom_point stat identity, scale the axis-y data label and facet wrap test marks into multiple graphs.



Based on the graph, students that score higher secondary education test marks with 97.7% at Rural area their higher salary is $200,000. While students' lives at both rural and urban area with higher secondary education test marks 95% and their highest salary amount is $500,000.

Conclusion for Question 4:

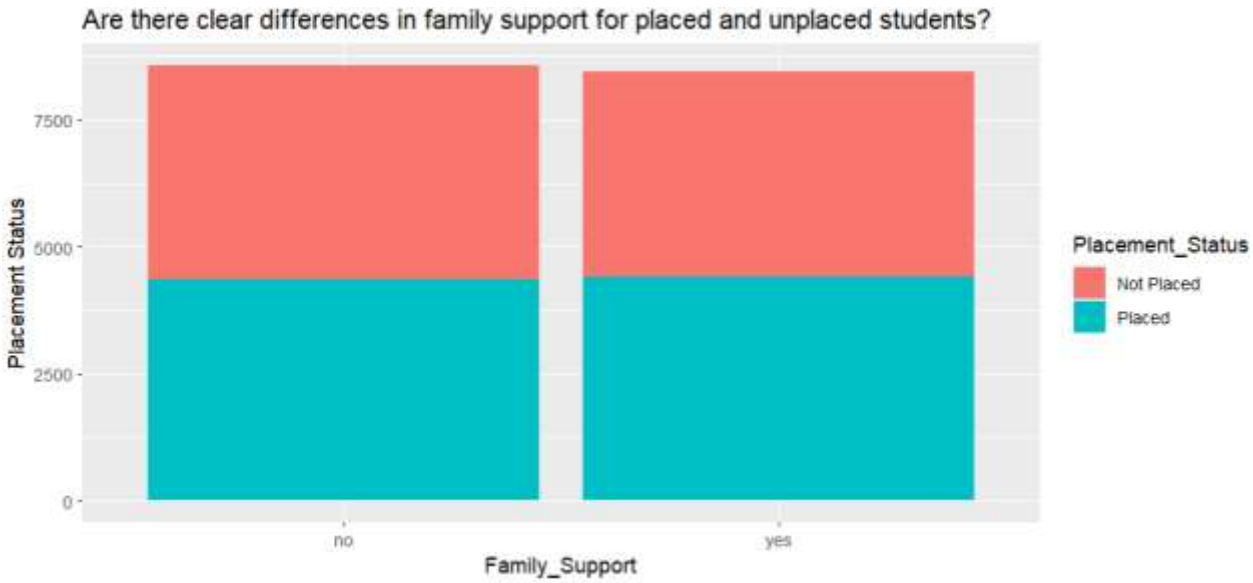
The analysis proven that students with higher family incomes, especially in the range of $200,000 to $400,000, may have a better chance of job placement. However, the highest reported salary of $500,000 does not guarantee placement. The data also shows that education and specialization play a significant role in determining salary amounts, with MBA test marks being the primary factor. Additionally, work experience and gender do not appear to impact salary amounts. Students in rural areas with higher secondary education test marks of 97.7% are likely to receive a salary of $200,000, while those in urban and rural areas with higher secondary education test marks of 95% have the potential to earn the highest salary of $500,000.

Question 5: Does family support important for student job placement and have relationship between students active in participation in extracurricular activities?

Analysis 5.1: Are there clear differences in family support for placed and unplaced students?

```
ggplot(importData, aes(x=Family_Support, fill=Placement_Status)) +
  geom_bar() +
  labs(x="Family_Support", y="Placement Status",
       title = "How many family support") +
  ggtitle("Are there clear differences in family support for placed and unplaced students?")
```

The graph is created with geom_bar() and combine two data to compare the data.

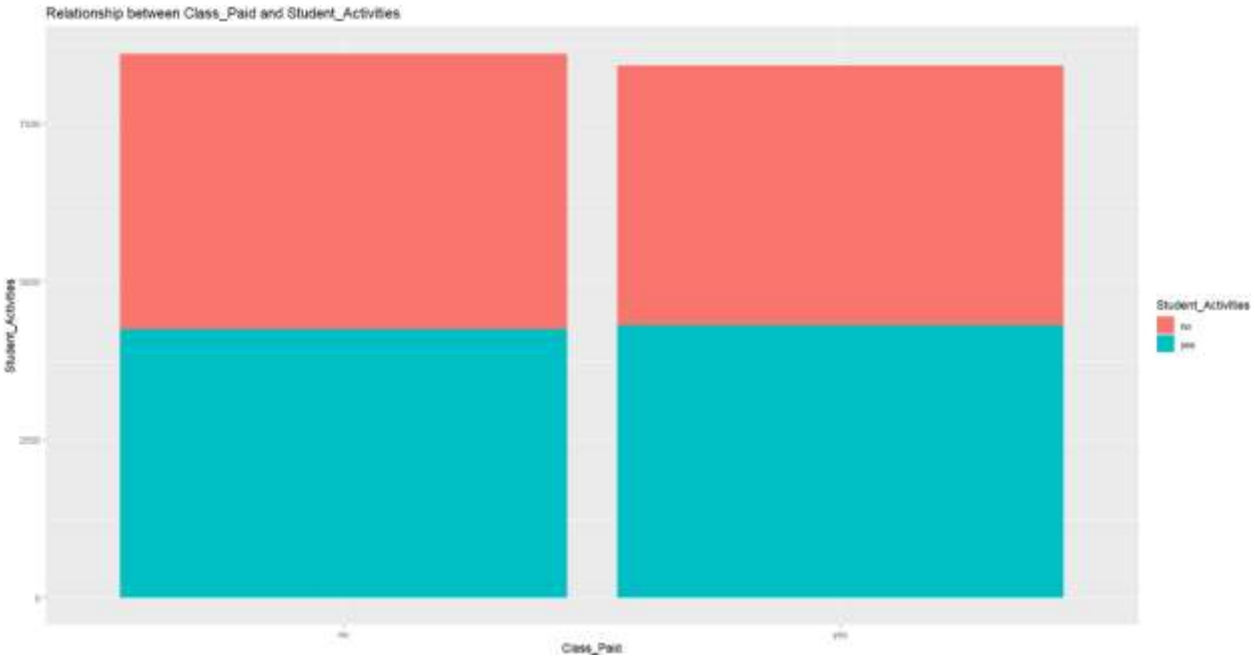


From the graph, it can be understood that it is hardly to differentiate the differences as both family support and placement status data are same.

Analysis 5.2: What is the relationship between Class Paid and Student Activities?

```
ggplot(importData, aes(x=Class_Paid, fill=Student_Activities)) +
  geom_bar() +
  ggtitle("Relationship between Class_Paid and Student_Activities") +
  labs(x="Class_Paid", y="Student_Activities")
```

The graph is created with geom_bar() and combine two data to compare the data.

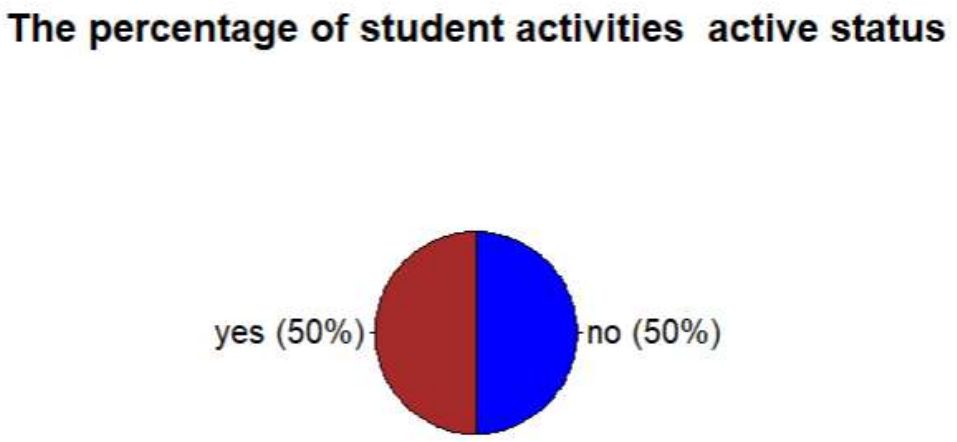


From the graph, it can be understood that it is hardly to differentiate the differences as both family support and placement status data are same.

Analysis 5.3: Does active in participant student activities will get highest salary paid?

```
Student_Activities = c("yes","no")
prop <- prop.table(table(Student_Activities)) * 100
pie(prop, labels = paste0(names(prop), " (", round(prop, 1), "%)"),
    main = "The percentage of student activities  active status",
    col = c("blue", "brown"), clockwise = TRUE)
```

The chart is created with piechart with function pie(), add data label, color, and clockwise.

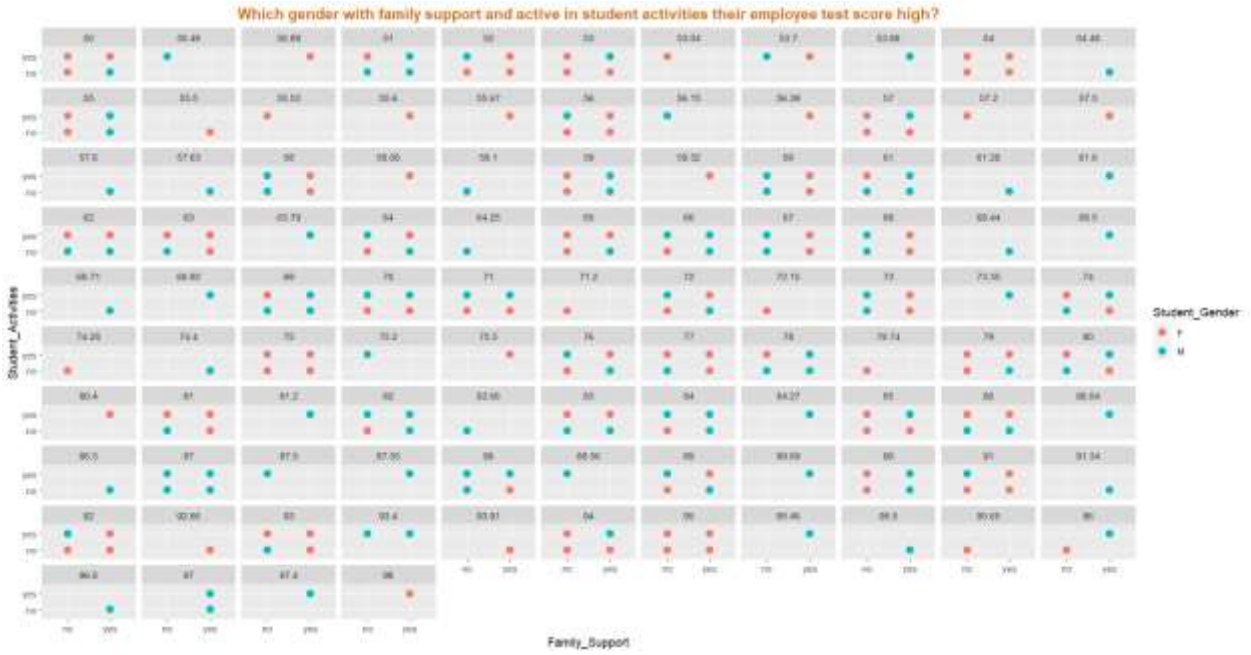


The students who active in participants the student activities are equal.

Analysis 5.4: Which gender with family support and active in student activities their employee test score high?

```
ggplot(importData, aes(Family_Support, Student_Activities, colour=Student_Gender))+
  geom_point(size=3)+
  ggtitle("Which gender with family support and active in student activities their employee test score high?")+
  facet_wrap(~Employee_Test)+
  theme(plot.title = element_text(hjust = 0.5, size=14,
                                  face='bold', color='#CC6600'))
```

The graph is created with geom_point() function size 3, color display the student gender and facet employee test marks into multiple graph.

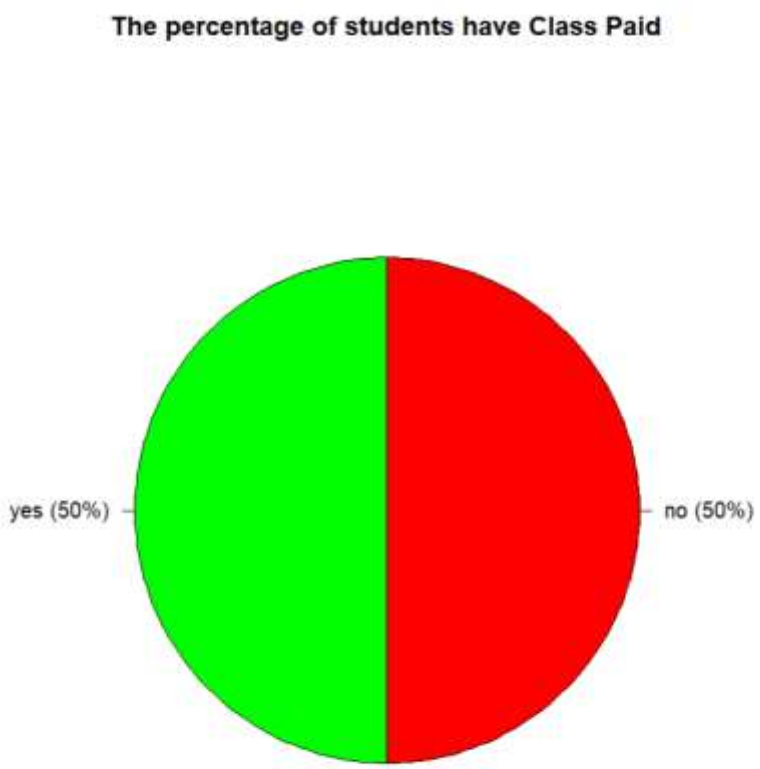


The axis-x is indicating family support, and axis-y is indicating student activities. Based on the graph, we can understand that female student with employee test score 98% have family support and active in participant activities.

Analysis 5.5: How many students have class paid?

```
Class_Paid = c("yes","no")
prop <- prop.table(table(Class_Paid)) * 100
pie(prop, labels = paste0(names(prop), " (", round(prop, 1), "%)"),
    main = "The percentage of students have Class Paid",
    col = c("red", "green"), clockwise = TRUE)
```

The chart is created with pie chart with color red and green.

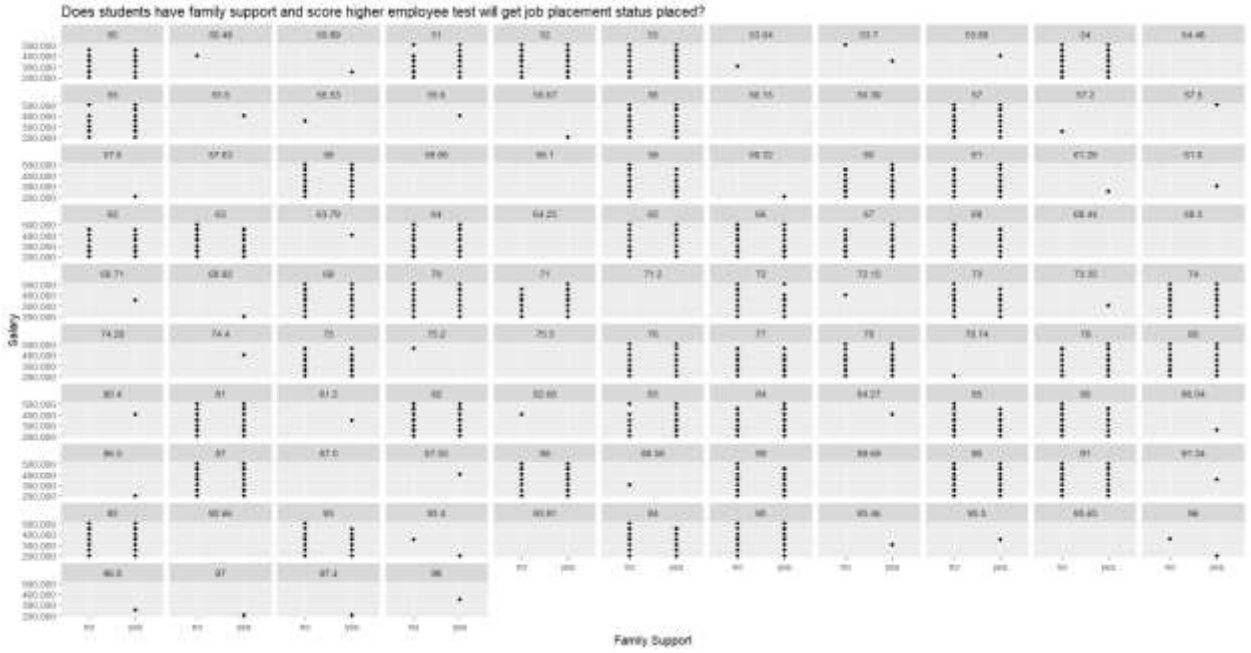


The students who have class paid are equal to students who do not have class paid.

Analysis 5.6: Does students have family support and score the highest employee test will get HIGHEST salary paid?

```
#analysis 5.6 : family support and employee test

ggplot(importData, aes(Family_Support, Salary)) +
  geom_point(stat = "identity") +
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  facet_wrap(~Employee_Test) +
  labs(x = "Family Support", y = "Salary") +
  ggtitle("Does students have family support and score higher employee test will get job placement status placed?")
```

The graph is created with geom_point() function stat identity, and facet employee test marks into multiple graph.

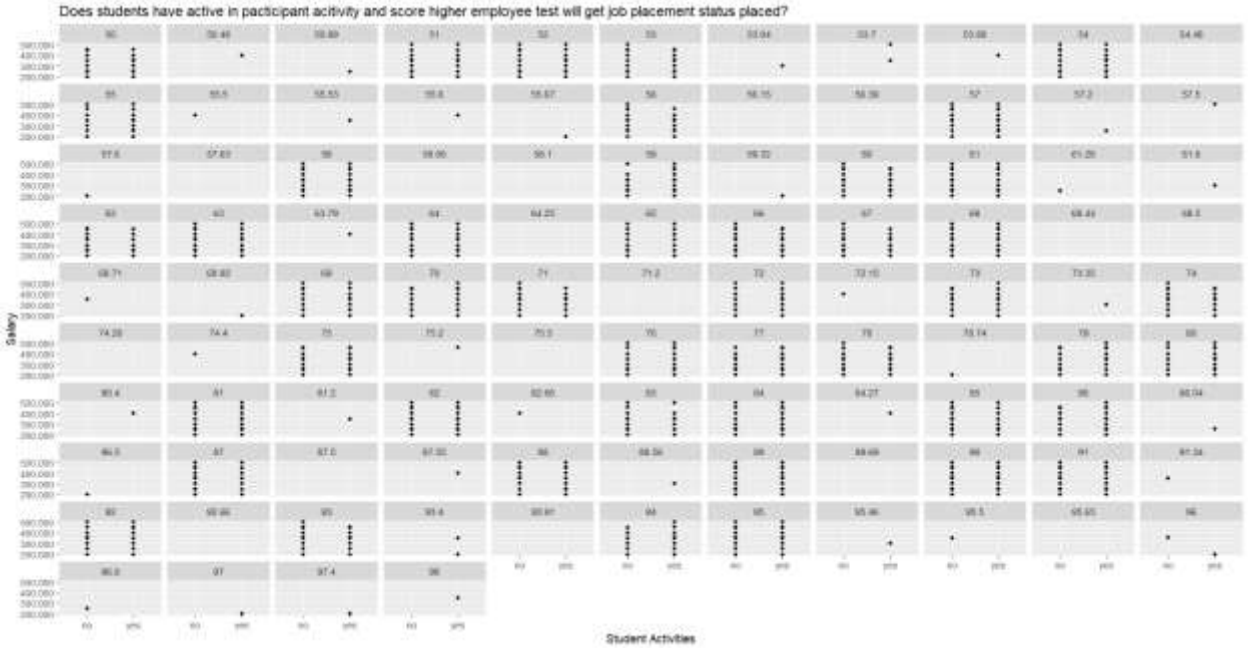


Based on the graph, we can understand students who score their highest employee test mark with 98%, whose salary amount is $350,000, and with family support. It can be assumed that if they score the highest employee test marks, they are impossible to get the highest salary paid.

Analysis 5.7: Does students have active in participant activity and score higher employee test will get HIGHEST salary paid?

```
#analysis 5.7 : student activity by employee test

ggplot(importData, aes(Student_Activities, Salary)) +
  geom_point(stat = "identity") +
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  facet_wrap(~Employee_Test) +
  labs(x = "Student Activities", y = "Salary") +
  ggtitle("Does students have active in pacticipant acitivity and score higher employee test will get job placement status placed?")
```

The graph is created with geom_point() function stat identity, and facet employee test marks into multiple graph.

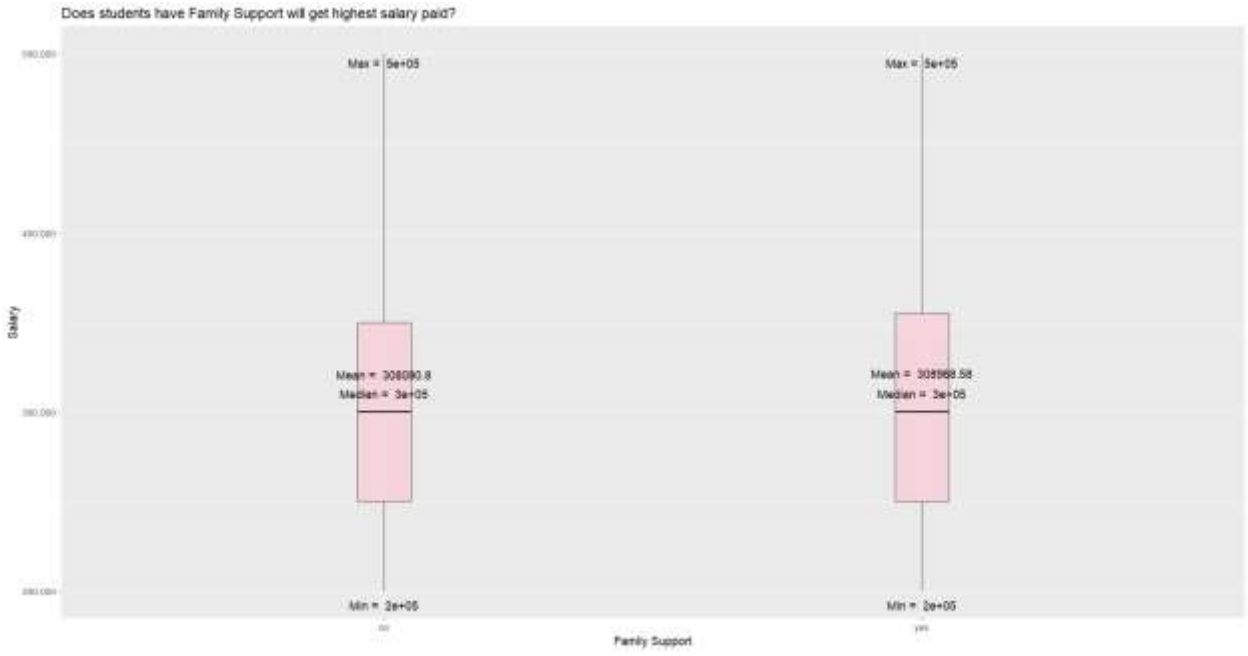


Based on the graph, we can understand students who score their highest employee test mark with 98%, whose salary amount is $350,000, and with active in participants student activities. It can be assumed that if they score the highest employee test marks, they are impossible to get the highest salary paid.

Analysis 5.8: Does students have Family Support will get highest salary paid?

```
#analysis 5.8 : family support and salary
ggplot(importData, aes(x = Family_Support, y = Salary)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Family Support", y = "Salary") +
  ggtitle("Salary by Family Support")
#label and color
ggplot(importData, aes(x = Family_Support, y = Salary)) +
  geom_boxplot(width=0.1, fill="pink", alpha=0.5, outlier.color="brown") +
  stat_summary(fun = median, geom = "text", aes(label = paste("Median = ", round(..y..,2))),
               vjust = -1.5, show.legend = FALSE) +
  stat_summary(fun = mean, geom = "text", aes(label = paste("Mean = ", round(..y..,2))),
               vjust = -2, show.legend = FALSE) +
  stat_summary(fun = min, geom = "text", aes(label = paste("Min = ", round(..y..,2))),
               vjust = 2, show.legend = FALSE) +
  stat_summary(fun = max, geom = "text", aes(label = paste("Max = ", round(..y..,2))),
               vjust = 1.5, show.legend = FALSE) +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Family Support", y = "Salary") +
  ggtitle("Salary by Family Support")
```

The graph is created with boxplot and stat_summary() function to show the data label with minimum, maximum, median and mean.

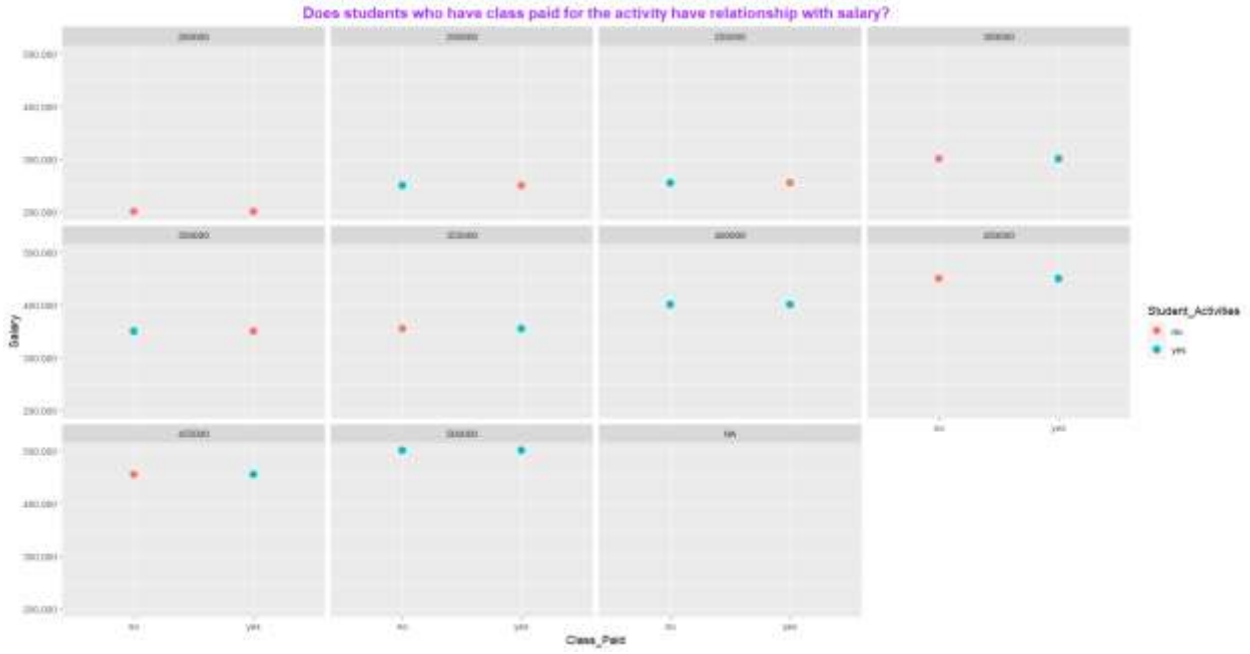


Based on the graph, with family support or without family support students are possible to get highest salary paid. However, their mean value is different for no family support is 308090.8 and have family support is 308968.58.

Analysis 5.9: Does class paid have relationship with salary?

```
ggplot(importData, aes(Class_Paid, Salary, colour=Student_Activities))+
  geom_point(size=3)+
  ggtitle("Does students who have class paid for the activity have relationship with salary?")+
  facet_wrap(~Salary)+
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  theme(plot.title = element_text(hjust = 0.5, size=14,
                                  face='bold', color='purple'))
```

The graph is created by geom_point() function with size 3 and facet wrap the test marks into multiple salary value.



Students who did paid for the class paid and active participant for the activities their salary have occupied with the amount of $500,000. While students who did not paid for the class paid and not active in participant the activities their salary amount are $200,000.

Conclusion for question 5:

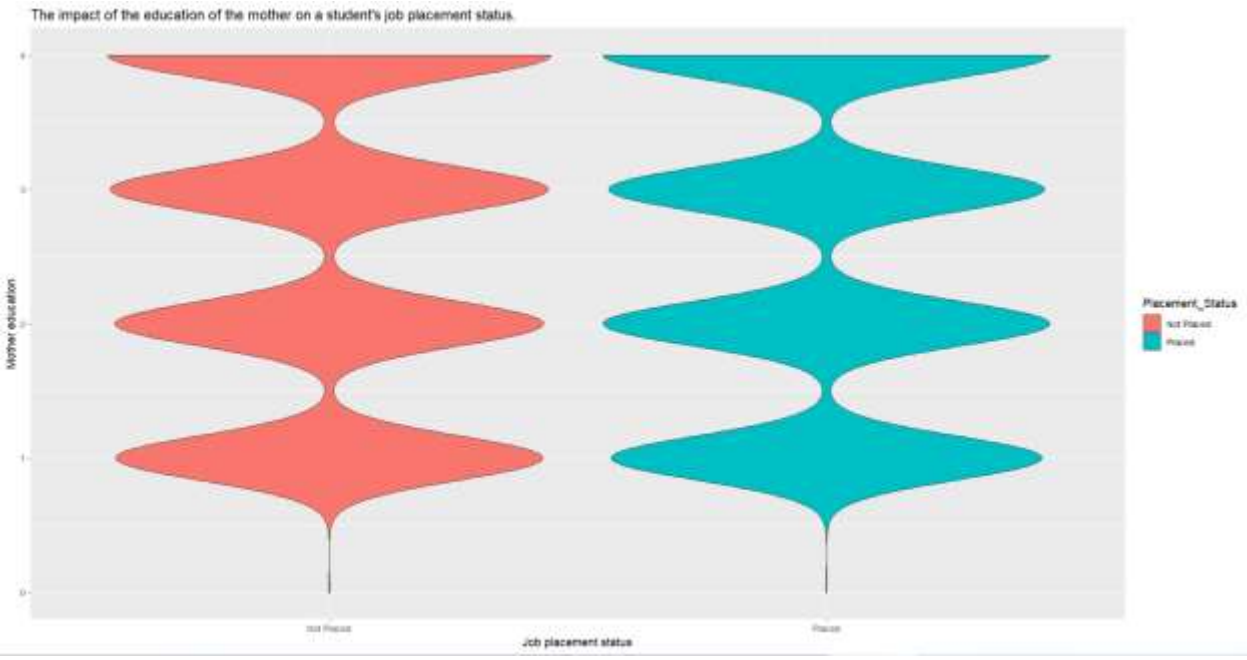
In conclusion, the graph provides insights into the relationship between various factors such as class payment, family support, student activities, and employee test scores with the salary of students. The data shows that students who paid for the class and participated actively in activities earn a higher salary compared to those who did not. Family support does not seem to have a significant impact on the salary earned by students, although there may be slight variations in the mean value. The highest earning students seem to have scored the highest marks in the employee test and participated actively in student activities. However, it is challenging to compare the differences between family support and placement status data as their data are mostly equal. Overall, the graph provides useful information for understanding the factors that contribute to the salary earned by students.

Question 6: Does a student's job placement status will be affected by their parents' education level and job type?

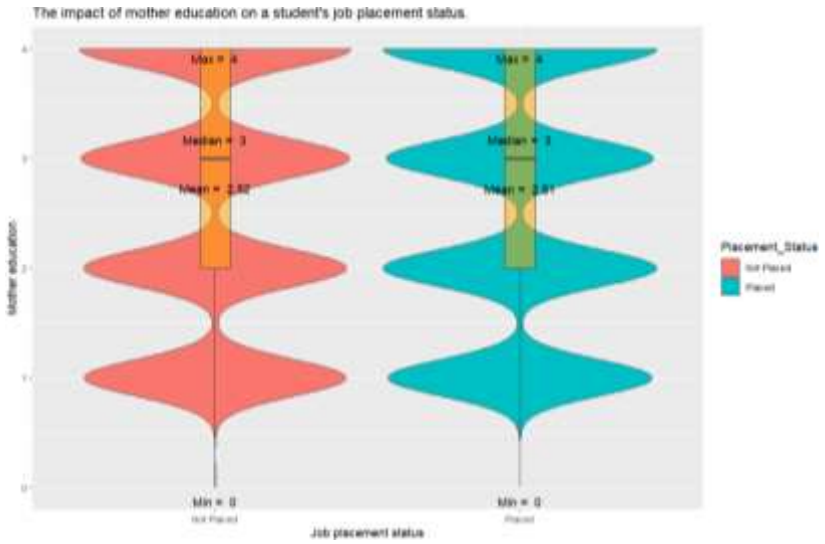
Analysis 6.1: The impact of the mother education level on a student's job placement status.

```
#Analysis 6.1: The impact of mother education on a student's job placement status.

ggplot(importData, aes(x = Placement_Status, y = Student_Medu, fill= Placement_Status)) +
  geom_violin() +
  labs(x = "Job placement status", y = "Mother education",
       title = "The impact of the education of the mother on a student's job placement status.")

#add the box plot in violist
ggplot(importData, aes(x = Placement_Status, y = Student_Medu, fill= Placement_Status)) +
  geom_violin() +
  geom_boxplot(width=0.1, fill="orange", alpha=0.5, outlier.color="red") +
  stat_summary(fun = median, geom = "text", aes(label = paste("Median = ", round(..y..,2))),
               vjust = -1.5, show.legend = FALSE) +
  stat_summary(fun = mean, geom = "text", aes(label = paste("Mean = ", round(..y..,2))),
               vjust = -2, show.legend = FALSE) +
  stat_summary(fun = min, geom = "text", aes(label = paste("Min = ", round(..y..,2))),
               vjust = 2, show.legend = FALSE) +
  stat_summary(fun = max, geom = "text", aes(label = paste("Max = ", round(..y..,2))),
               vjust = 1.5, show.legend = FALSE) +
  labs(x = "Job placement status", y = "Mother education",
       title = "The impact of mother education on a student's job placement status.")
```

The graph is created in geom_violion() and geom_boxplot() function with fill color orange and outlier color red. The graph also added stat_summary() to show the data label minimum, maximum, mean and median.



The axis-x is indicating job placement status and axis-y is indicating mother education level. The graph did not have added with boxplot yet.

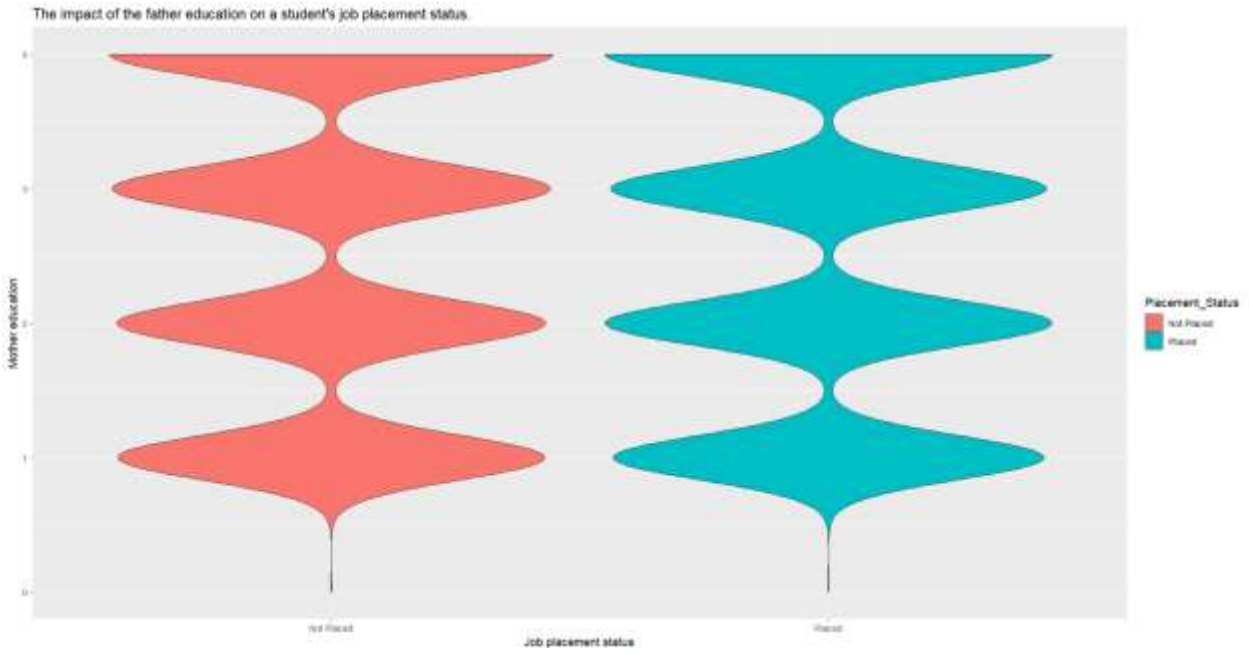


The plot indicates that there is a high degree of similarity between the data for mother's education level in terms of job placement status, whether placed or not placed. The minimum value is 0, the median is 3, and the maximum is 4 for both categories. The only difference between the two is that the mean for the unplaced status is 2.52, while the mean for the placed status is 2.51.

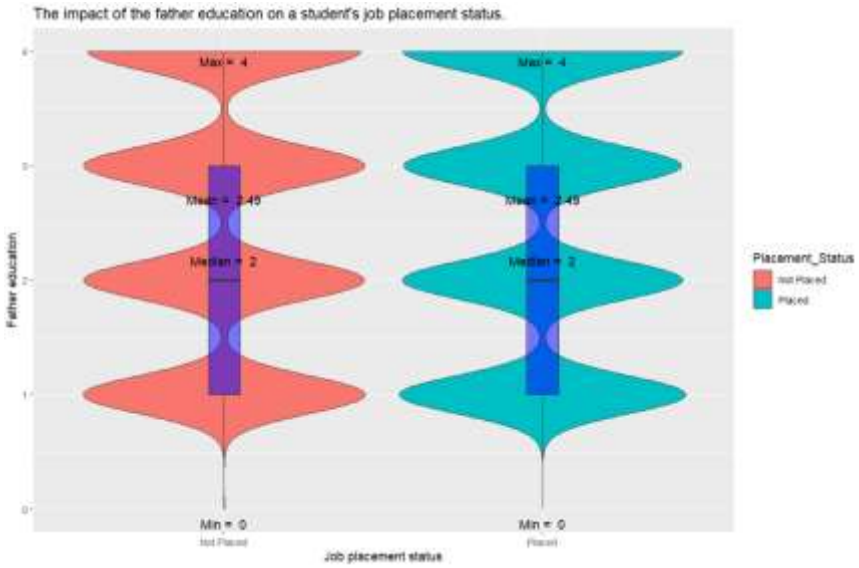
Analysis 6.2: The impact of the father education level on a student's job placement status.

```
#Analysis 6.2: The impact of the father education on a student's job placement status.

ggplot(importData, aes(x = Placement_Status, y = Student_Medu, fill= Placement_Status)) +
  geom_violin() +
  labs(x = "Job placement status", y = "Mother education",
       title = "The impact of the father education on a student's job placement status.")

#add the box plot in violist
ggplot(importData, aes(x = Placement_Status, y = Student_Fedu, fill= Placement_Status)) +
  geom_violin() +
  geom_boxplot(width=0.1, fill="blue", alpha=0.5, outlier.color="white") +
  stat_summary(fun = median, geom = "text", aes(label = paste("Median = ", round(..y..,2))),
               vjust = -1.5, show.legend = FALSE) +
  stat_summary(fun = mean, geom = "text", aes(label = paste("Mean = ", round(..y..,2))),
               vjust = -2, show.legend = FALSE) +
  stat_summary(fun = min, geom = "text", aes(label = paste("Min = ", round(..y..,2))),
               vjust = 2, show.legend = FALSE) +
  stat_summary(fun = max, geom = "text", aes(label = paste("Max = ", round(..y..,2))),
               vjust = 1.5, show.legend = FALSE) +
  labs(x = "Job placement status", y = "Father education",
       title = "The impact of the father education on a student's job placement status.")
```

The graph is created in geom_violion() and geom_boxplot() function with fill color blue and outlier color white. The graph also added stat_summary() to show the data label minimum, maximum, mean and median.



The axis-x is indicating job placement status and axis-y is indicating father education level. The graph did not have added with boxplot yet.

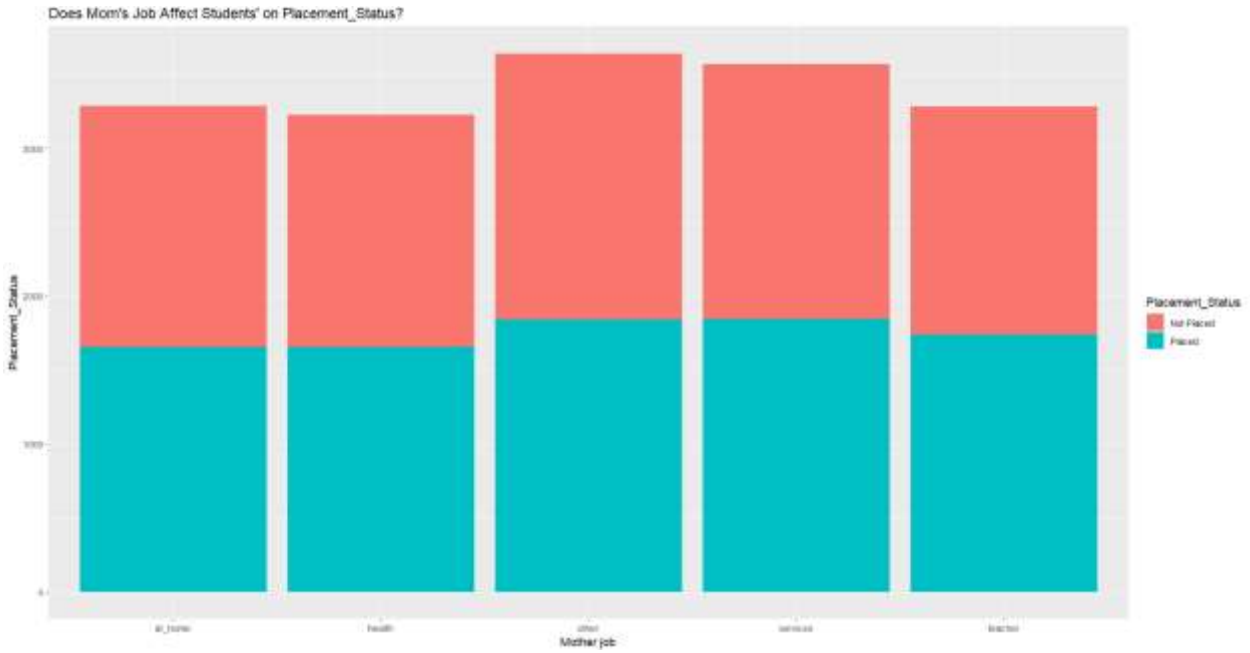


The plot shows that there is similarity between the data for father's education level in terms of job placement status, whether placed or not placed. The data has a minimum value of 0, a median value of 2, a mean value of 2.49, and a maximum value of 4.

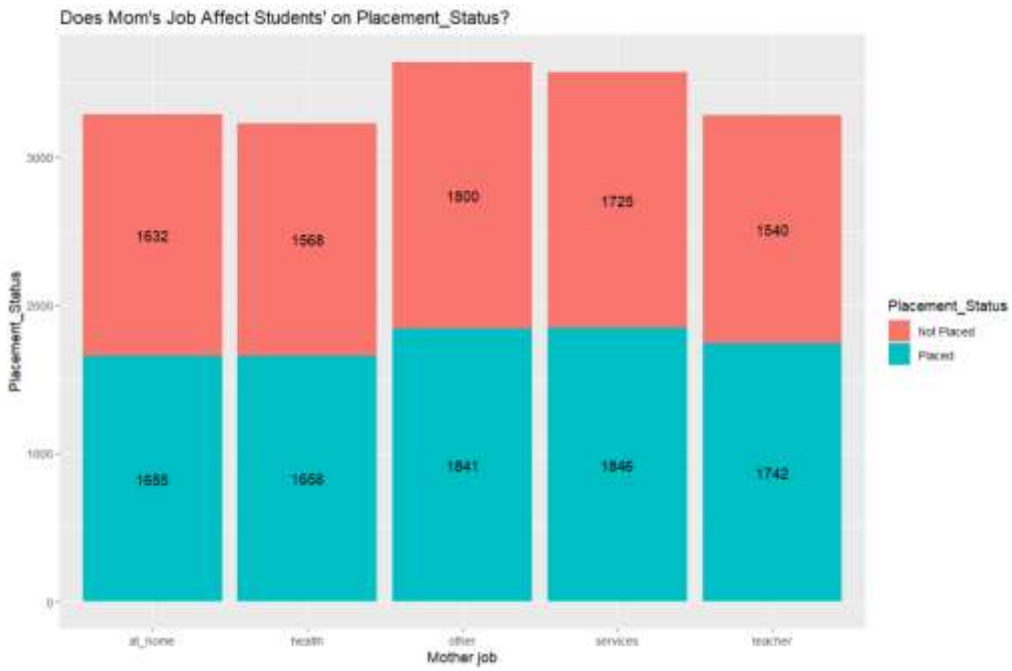
Analysis 6.3: Does Mom's Job Affect Students on Placement Status?

```
#Analysis 6.3: The influence of the mother's occupation on a student's secondary education gred.
ggplot(importData, aes(x=Student_Mejob, fill=Placement_Status)) +
  geom_bar(stat="count") +
  ggtitle("Does Mom's Job Affect Students' on Placement_Status?") +
  labs(x="Mother job", y="Placement_Status")
#add label - show the amount of placed or not placed for mother job
ggplot(importData, aes(x=Student_Mejob, fill=Placement_Status)) +
  geom_bar(stat="count") +
  geom_text(stat="count", aes(label=..count..), position=position_stack(vjust=0.5)) +
  ggtitle("Does Mom's Job Affect Students' on Placement_Status?") +
  labs(x="Mother job", y="Placement_Status")
```

The graph is created in function geom_bar() with stat count and label with function geom_text().



Axis-x is indicating the mother job type, and axis-y is indicating the placement status. The fill will show two color to indicate not placed and placed status.

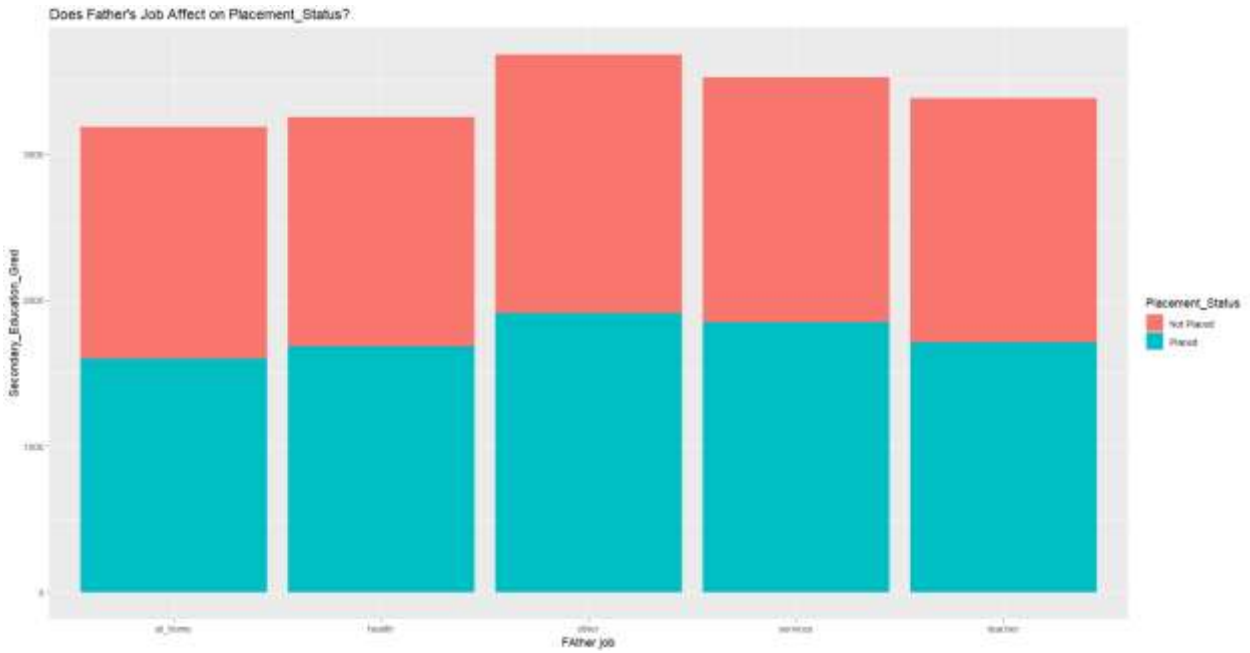


The graph indicates that individuals with mothers' job in the "others" occupation have the highest placement status, with 1841 individuals placed and 1800 individuals not placed. On the other hand, those with mother's job in the "health" occupation have the lowest placement status, with 1658 individuals placed and 1568 individuals not placed.

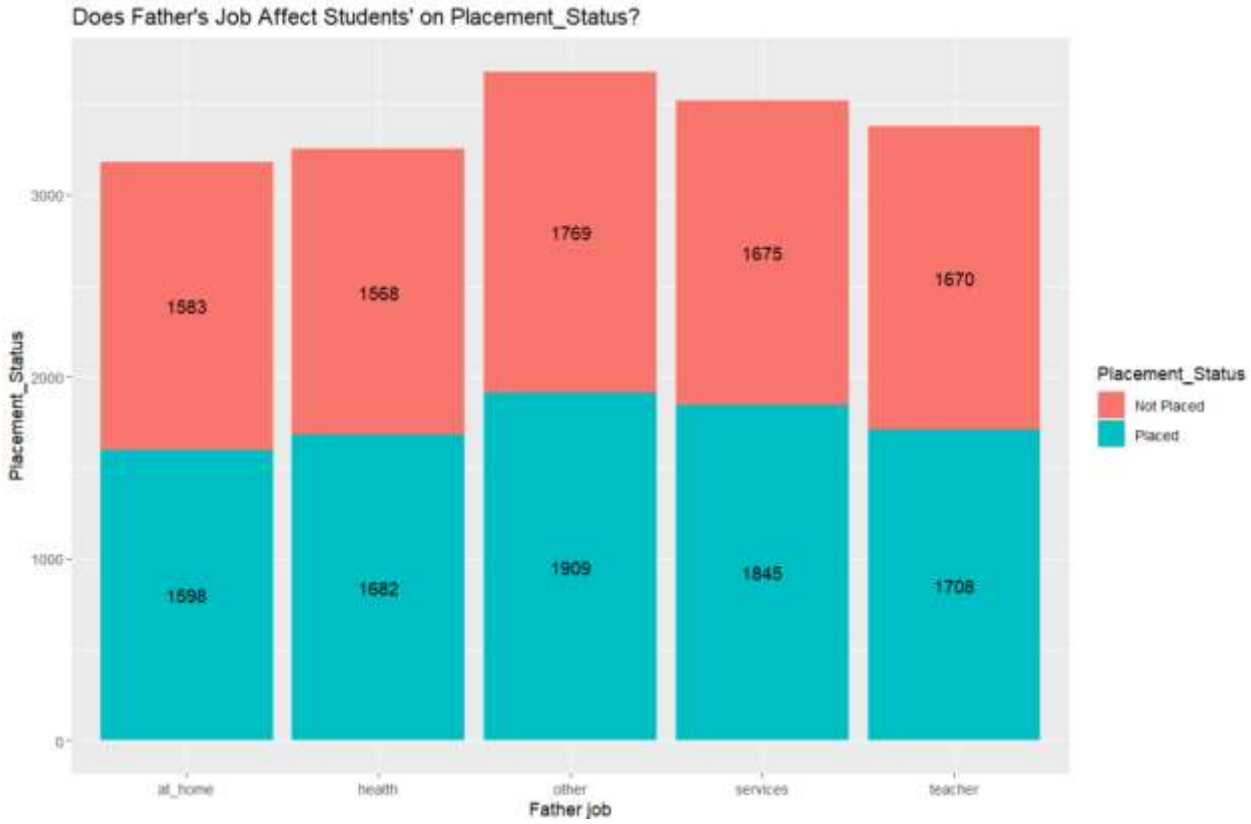
Analysis 6.4: Does Father's Job Affect Students on Placement Status?

```
#Analysis 6.4: The influence of the father's occupation on a student's job placement.
ggplot(importData, aes(x=Student_Fejob, fill=Placement_Status)) +
  geom_bar(stat="count") +
  ggtitle("Does Father's Job Affect on Placement_Status?") +
  labs(x="FAther job", y="Secondary_Education_Gred")

#add label - show the amount of placed or not placed for father job
ggplot(importData, aes(x=Student_Fejob, fill=Placement_Status)) +
  geom_bar(stat="count") +
  geom_text(stat="count", aes(label=..count..), position=position_stack(vjust=0.5)) +
  ggtitle("Does Father's Job Affect Students' on Placement_Status?") +
  labs(x="Father job", y="Placement_Status")
```

The graph type is geom_bar() with position stacks to represent the placement status.



The axis-x is indicating father job type, and the axis-y is indicating student's secondary education test marks.



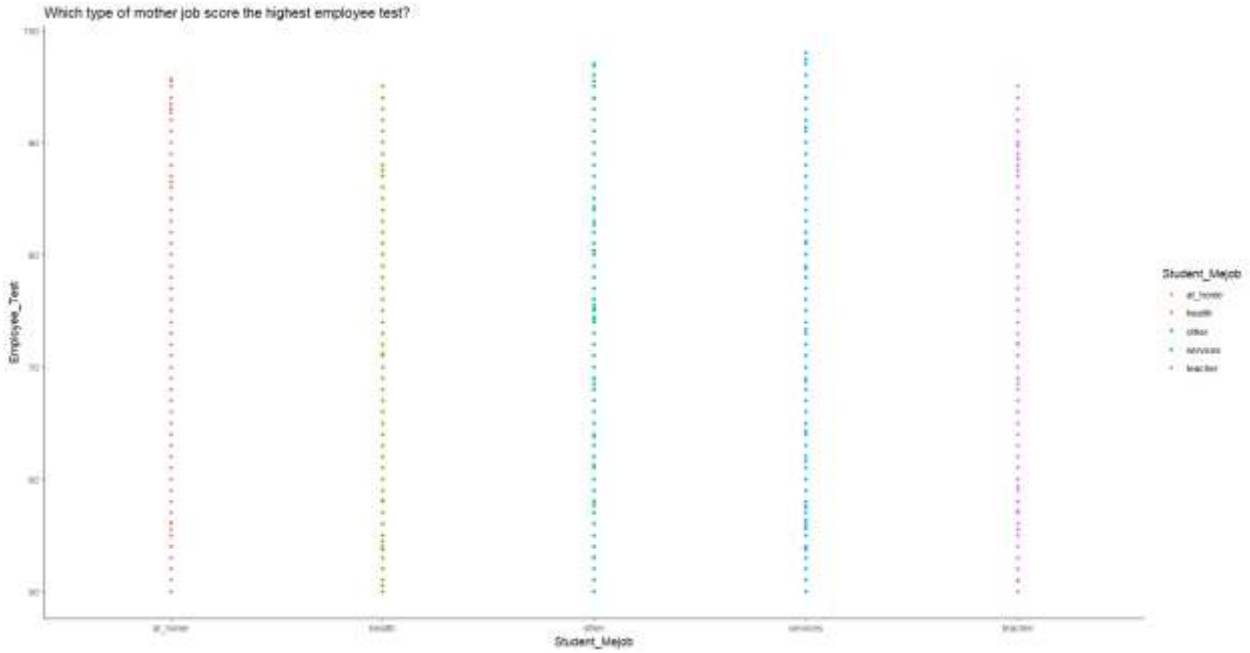
By examining the labeled graph, we can deduce that individuals with father's job in the "others" occupation have the highest placement status, with 1909 individuals placed and 1769 individuals not placed. Conversely, those with father's job in the "at home" occupation have the lowest placement status, with 1598 individuals placed and 1583 individuals not placed.

Analysis 6.5: Which type of mother job score the highest employee test?

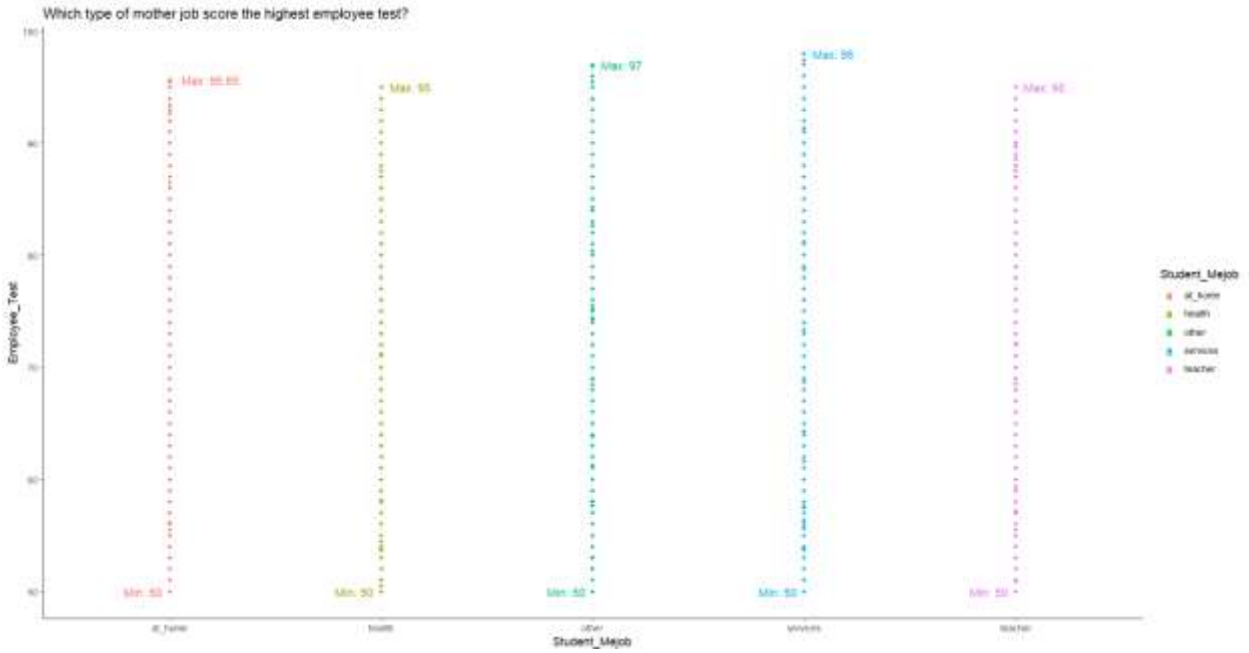
```
ggplot(importData, aes(x=Student_Mejob, y=Employee_Test, color=Student_Mejob)) +
  geom_point() +
  labs(x = "Student_Mejob", y = "Employee_Test") +
  ggtitle("Which type of mother job score the highest employee test?") +
  theme_classic()
```

```
ggplot(importData, aes(x=Student_Mejob, y=Employee_Test, color=Student_Mejob)) +
  geom_point() +
  labs(x = "Student_Mejob", y = "Employee_Test") +
  ggtitle("Which type of mother job score the highest employee test?") +
  theme_classic()+
  stat_summary(fun.y = max, aes(label = paste0("Max: ", round(..y.., 2))), geom = "text", hjust = -0.2) +
  stat_summary(fun.y = min, aes(label = paste0("Min: ", round(..y.., 2))), geom = "text", hjust = 1.2)
```

The graph is create in function geom_point() and with show data label function stat_summary().



The graph show that axis-x is student's mother job type and employee test score on axis-y.



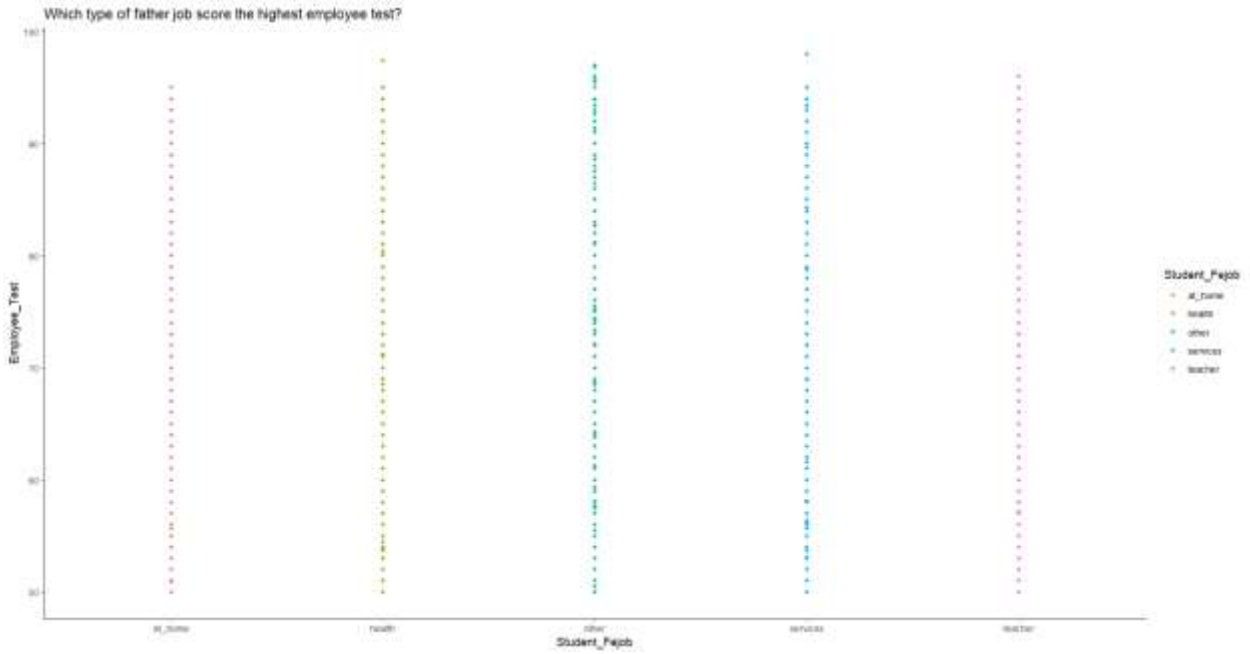
The graph with show data label, we can understand that student's mother job on "services" have the highest employee test marks, which is 98% While student's lowest employee test marks are on "health" and "teacher" mother job type, which is 95%. The minimum marks are same for all mother job, which is 50%.

Analysis 6.6: Which type of father job score the highest employee test?

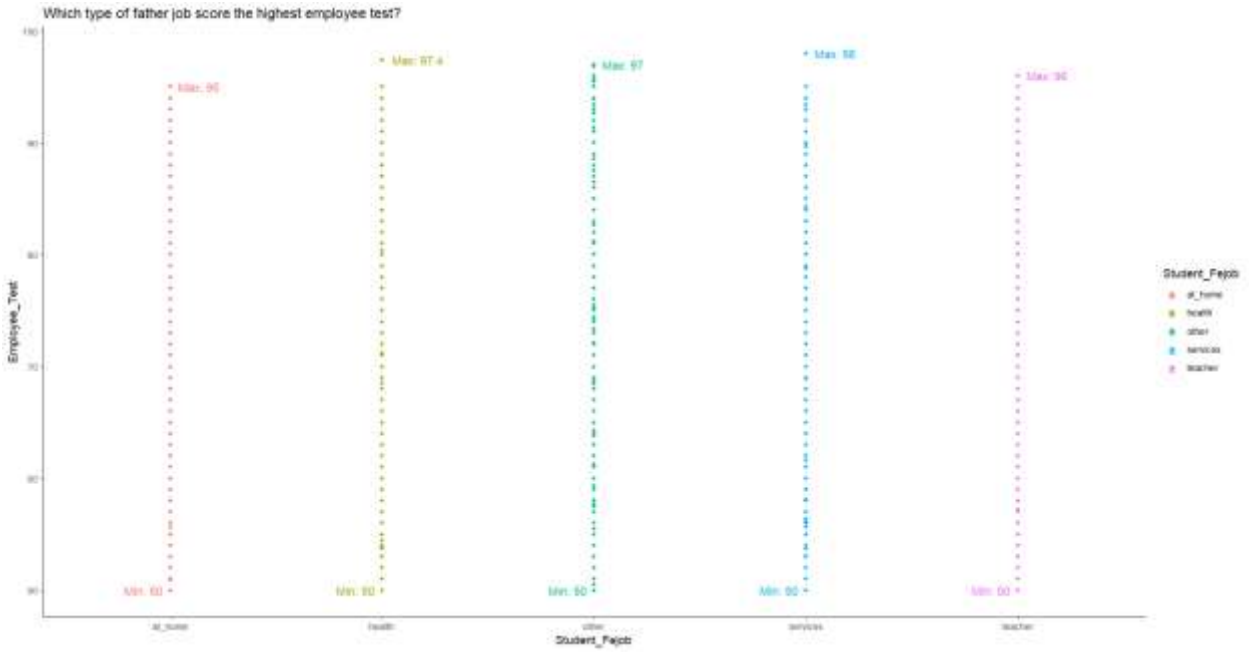
```
ggplot(importData, aes(x=Student_Fejob, y=Employee_Test, color=Student_Fejob)) +
  geom_point() +
  labs(x = "Student_Fejob", y = "Employee_Test") +
  ggtitle("Which type of father job score the highest employee test?") +
  theme_classic()
```

```
ggplot(importData, aes(x=Student_Fejob, y=Employee_Test, color=Student_Fejob)) +
  geom_point() +
  labs(x = "Student_Fejob", y = "Employee_Test") +
  ggtitle("Which type of father job score the highest employee test?") +
  theme_classic()+
  stat_summary(fun.y = max, aes(label = paste0("Max: ", round(..y.., 2))), geom = "text", hjust = -0.2) +
  stat_summary(fun.y = min, aes(label = paste0("Min: ", round(..y.., 2))), geom = "text", hjust = 1.2)
```

The graph is created in geom_point() function and stat_summary to show the data label.



The axis-x is indicating student father job type and employee test marks is indicating employee test marks.

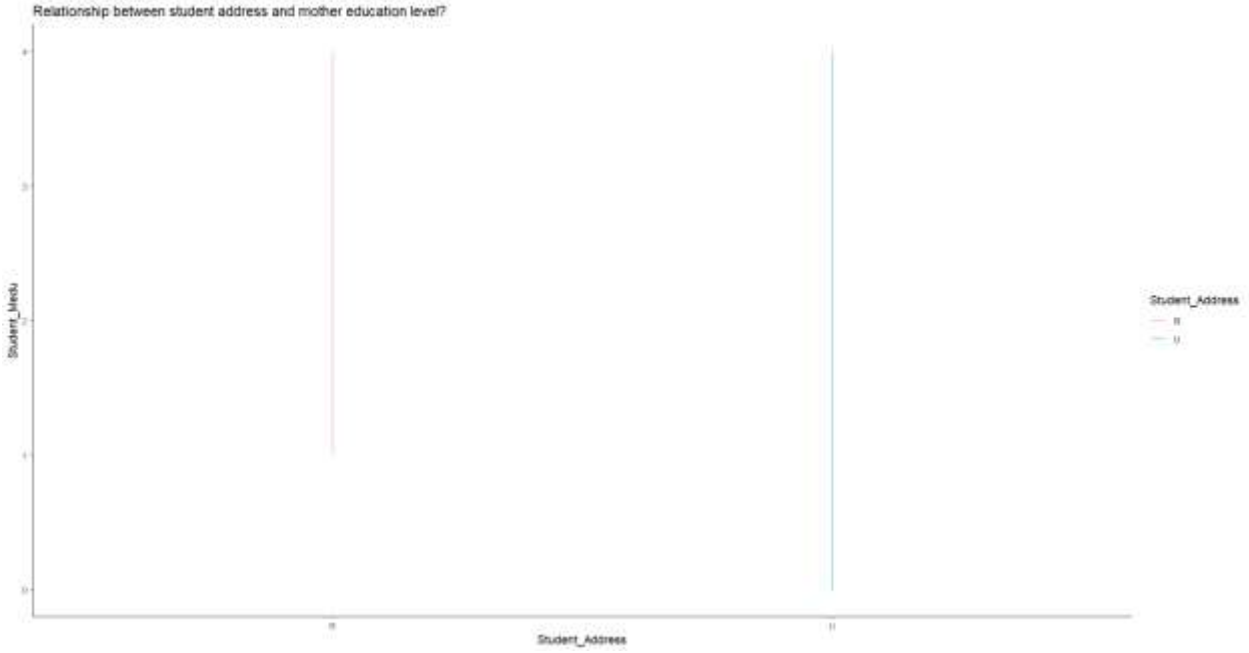


With the show data label graph, we can understand that “services” father job has the highest employee test marks, which is 98%. The minimum marks are same for all father job, which is 50%.

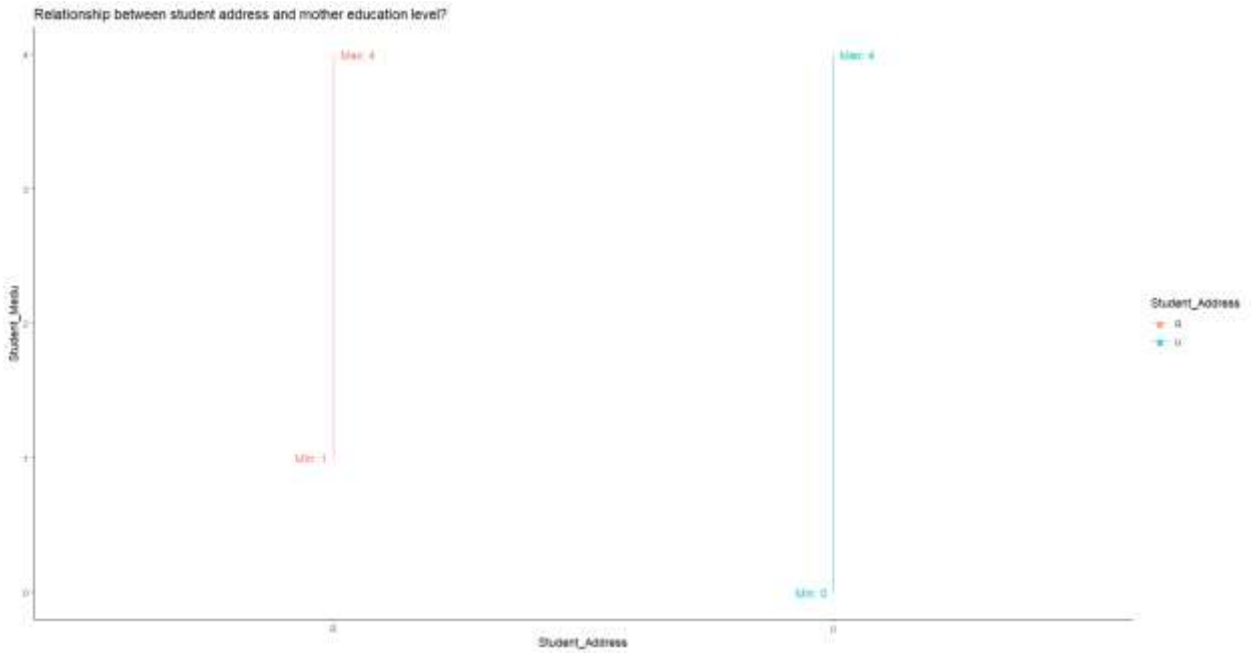
Analysis 6.7: Relationship between student address and mother education level?

```
#analysis 6.7: student address and mother education level
ggplot(importData, aes(x=Student_Address, y=Student_Medu, color=Student_Address)) +
  geom_line() +
  labs(x = "Student_Address", y = "Student_Medu") +
  ggtitle("Relationship betwen student address and mother education level?") +
  theme_classic()
#show data label
ggplot(importData, aes(x=Student_Address, y=Student_Medu, color=Student_Address)) +
  geom_line() +
  stat_summary(fun.y = max, aes(label = paste0("Max: ", round(..y.., 2))), geom = "text", hjust = -0.2) +
  stat_summary(fun.y = min, aes(label = paste0("Min: ", round(..y.., 2))), geom = "text", hjust = 1.2) +
  labs(x = "Student_Address", y = "Student_Medu") +
  ggtitle("Relationship between student address and mother education level?") +
  theme_classic()
```

The graph is created with geom_line() function and with stat_summary to show data label.



The axis-x is represented student address area and they axis-y is represent student mother education level.

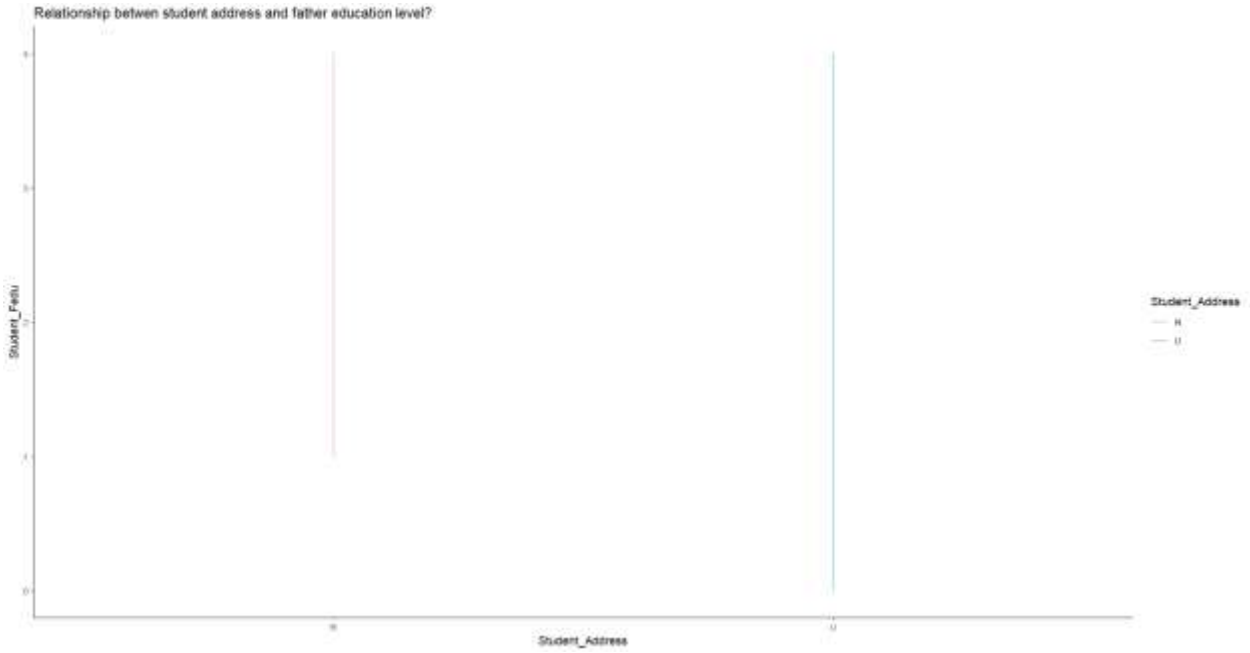


The graph show that in rural area mother education level minimum are 1 and maximum are 4. Whereas in urban area mother education level minimum are 0 and maximum are 4.

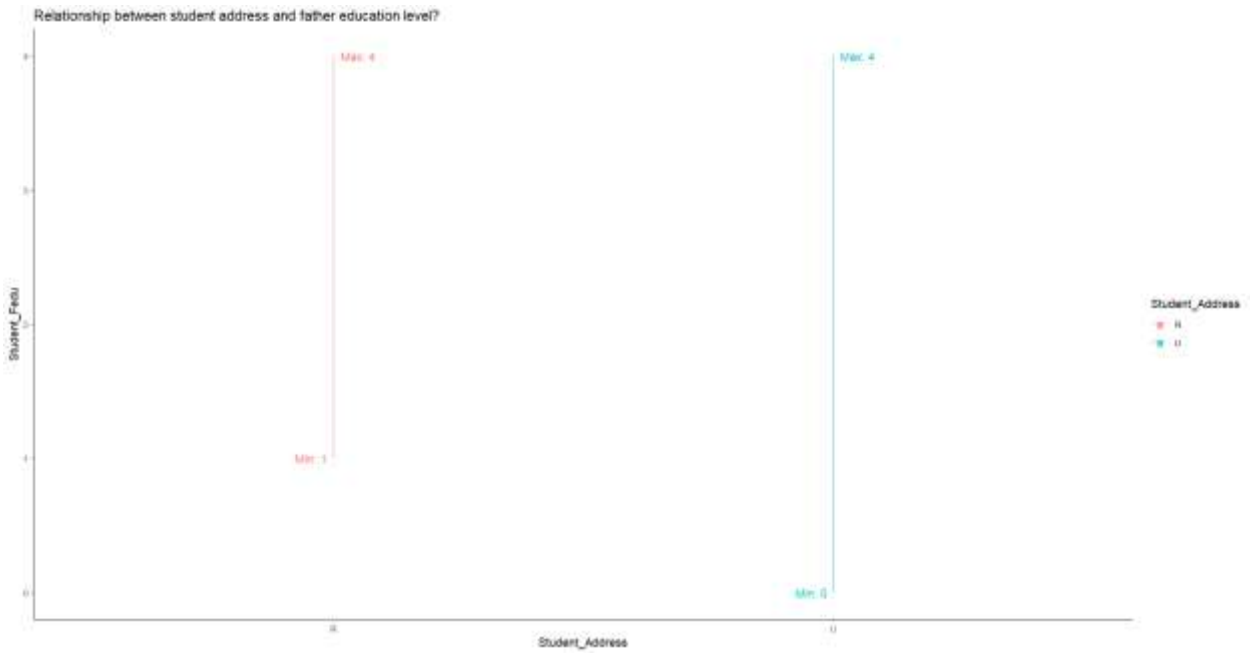
Analysis 6.8: Relationship between student address and father education level?

```
#analysis 6.8 : student address and father education level
ggplot(importData, aes(x=Student_Address, y=Student_Fedu, color=Student_Address)) +
  geom_line() +
  labs(x = "Student_Address", y = "Student_Fedu") +
  ggtitle("Relationship betwen student address and father education level?") +
  theme_classic()

#show data label
ggplot(importData, aes(x=Student_Address, y=Student_Fedu, color=Student_Address)) +
  geom_line() +
  stat_summary(fun.y = max, aes(label = paste0("Max: ", round(..y.., 2))), geom = "text", hjust = -0.2) +
  stat_summary(fun.y = min, aes(label = paste0("Min: ", round(..y.., 2))), geom = "text", hjust = 1.2) +
  labs(x = "Student_Address", y = "Student_Fedu") +
  ggtitle("Relationship between student address and father education level?") +
  theme_classic()
```

The graph is created with geom_line() function and with stat_summary to show data label.



The axis-x is represented student address area and they axis-y is represent student father education level.

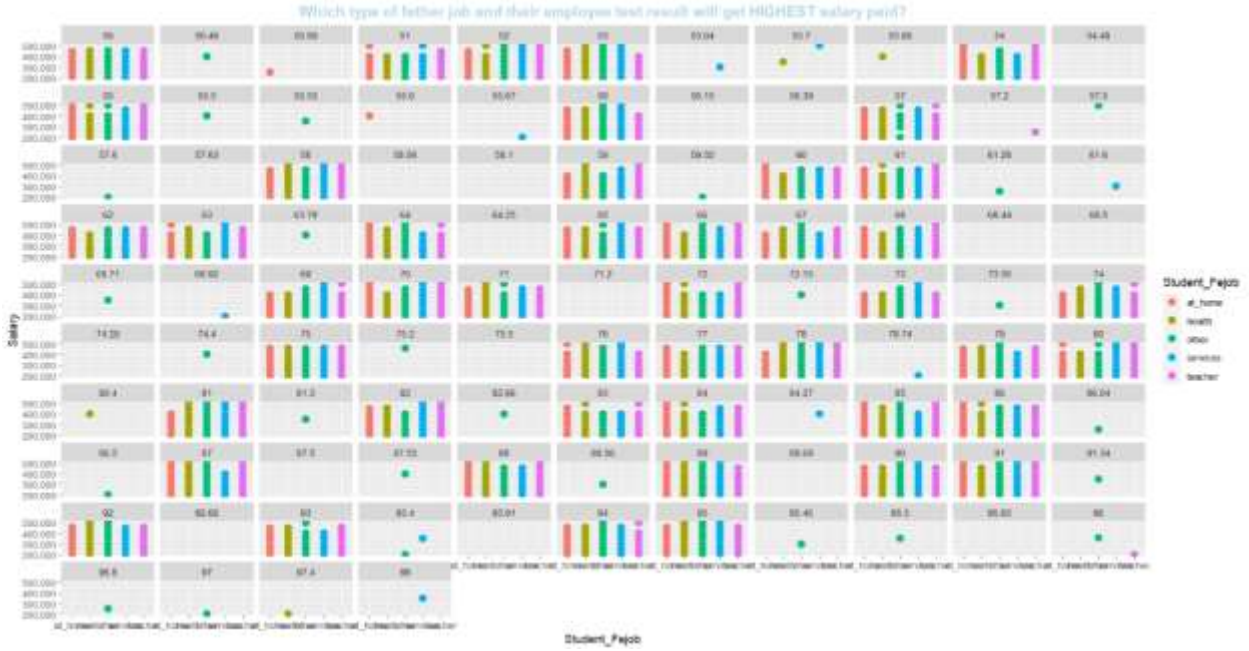


The graph show that in rural area father education level minimum are 1 and maximum are 4. Whereas in urban area father education level minimum are 0 and maximum are 4.

Analysis 6.9: Which type of father jobs and the test results of their employees will result in the highest salaries being paid?

```
ggplot(importData, aes(Student_Fejob, Salary, colour=Student_Fejob))+
  geom_point(size=3)+
  ggtitle("Which type of father job and their employee test result will get HIGHEST salary paid?")+
  facet_wrap(~Employee_Test)+
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  theme(plot.title = element_text(hjust = 0.5, size=14,
                                  face='bold', color='lightblue'))
```

The graph is created by geom_point() function with size 3 and facet wrap the test marks into multiple graphs.



Students who score their employee test with 98%, salary amount is paid with $350,000 and the father job type is services. While students for others, services and health father's job type have the highest salary paid which is $500,000 when their employee test is 95%.

Analysis 6.10: Which type of mother jobs and the test results of their employees will result in the highest salaries being paid?

```
ggplot(importData, aes(Student_Mejob, Salary, colour=Student_Mejob))+
  geom_point(size=3)+
  ggtitle("Which type of MOTHER job and their employee test result get HIGHEST salary paid?")+
  facet_wrap(~Employee_Test)+
  scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))+
  theme(plot.title = element_text(hjust = 0.5, size=14,
                                  face='bold', color='red'))
```

The graph is created by geom_point() function with size 3 and facet wrap the test marks into multiple graphs.

Students who score their employee test with 98%, salary amount is paid with $200,000 and the mother job type is also services. While students for others, services and teacher mother's job type have the highest salary paid which is $500,000 when their employee test is 95%.

Conclusion for Question 6:

In conclusion, students whose fathers and mothers work in service jobs tend to score higher on tests, but their salaries are typically between $200,000 to $350,000. To get the highest salaries paid with amount $500,000 when their employee test result marks is 95%. Additionally, students residing in urban areas whose fathers and mothers have no formal education and students residing in rural areas whose fathers and mothers have at least one year of education tend to have lower test scores. Interestingly, the placement status of students whose fathers and mothers work in other professions has the highest rate of being placed. These findings suggest that job type and parental education may play a role in a student's academic performance and job placement status.

Extra features

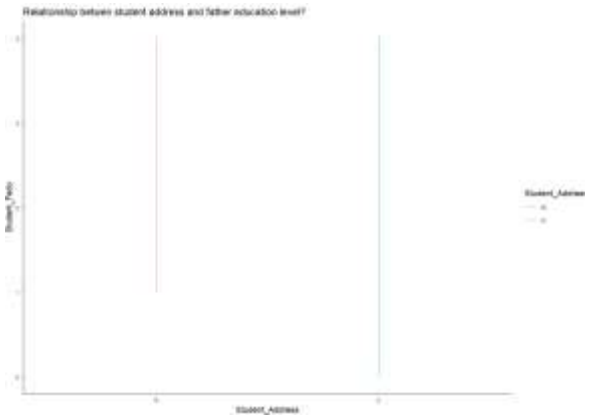
1. scale_y_continuous(labels = scales::comma, limits = c(2e5, 5e5))

This code can be used for scales on plot axes, marked by formatting values with commas and setting min and max values from 200 000 to 500 000. The reason I set the maximum value to 500 000 is that I assume that from the data, the maximum value of the salary will be as high as 400 000.

2. theme(plot.title = element_text(hjust = 0.5, size=14, face='bold', color='red'))

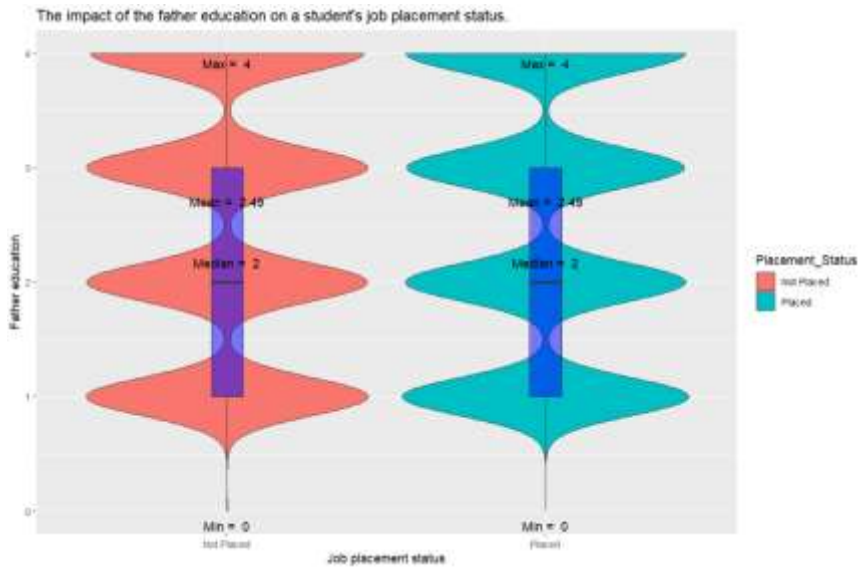
The code can used for change color the specific color for the text size, font type and the main title.

3. theme_classic()



The function theme_classic() applies a traditional aesthetic to the plot by rendering x and y axis lines without any gridlines (ggplot2.tidyverse.org, n.d.).

4. stat_summary(fun = median, geom = "text", aes(label = paste("Median = ", round(..y..,2))), vjust = -1.5, show.legend = FALSE)



With the stat_summary() function by adding the label, the graph can clearly display the data label and value for showing their maximum, minimum, mean and median.

Conclusion

After analyzing the data provided by the marketing department, it was found that the data was clean, and no data cleaning was required to carry out the analysis. However, one issue was identified: most of the students who were not placed in a job had a blank salary value in the imported data. To avoid this issue from affecting the median, mean, max, and min values, it was decided to skip. Instead of replacing the blank columns with zero values. This helped to minimize the impact of the issue on doing further analysis. Based on all the questions and analysis, one can understand that all the data are almost equal except their median or mean value. In order to carry out useful analysis, it is suggested can combine all these data together into 2 to 3 to compare their relationship. Therefore, we will get a better result of their differences to find out the hidden issues.

References

(n.d.). Retrieved from ggplot2.tidyverse.org: https://ggplot2.tidyverse.org/reference/ggtheme.html

Ledolter, J. (2013). Data mining and business analytics with R. John Wiley & Sons.