A · P · U

ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Programming For Data Analysis (CT127-3-2-PFDA)
APD2F2211CS

HAND OUT DATE: 31 DECEMBER 2022
HAND IN DATE: 8 MARCH 2023
WEIGHTAGE: 50%
Student Name: Surenther A/L Saravanan
TP Number: TP067186

INSTRUCTIONS TO CANDIDATES:

1 Submit your assignment at the administrative counter

2 Students are advised to underpin their answers with the use of references (cited using the Harvard Name System of Referencing)

3 Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld

4 Cases of plagiarism will be penalized

5 The assignment should be bound in an appropriate style (comb bound or stapled).

6 Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.

7 You must obtain 50% overall to pass this module.

Table of Contents

Contents

Introduction:

This report presents the results of the analysis performed on the Student Placement dataset. The dataset contains information about students' personal details, family background, extra activities, academic records, and placement information. The purpose of this analysis is to identify any hidden problems among students in terms of campus placements and provide meaningful insights for decision-making.

To achieve this goal, various data analytics techniques have been used, including data exploration, manipulation, transformation, and visualization. These techniques were selected based on their relevance to the dataset and their potential to uncover valuable insights.

This report is structured as follows: first, a description of the dataset, including its size and attributes, is provided. Then, the data import, cleaning, pre-processing, and transformation steps are explained in detail. Afterward, a set of questions related to the dataset are presented, and for each question, the analysis techniques, source code, output/plot, and findings are outlined. Additionally, an extra feature has been added to improve the analysis, and its implementation and benefits are explained.

Finally, the report concludes with a summary of the main findings and insights, followed by recommendations for the marketing department to address any identified issues.

Analysis

Introduction:

The dataset we will be analysing in this report is a placement data for a graduate program consisting of 25 columns and 17007 rows of data. The dataset summary reveals various information about the characteristics of the data. The age of the graduate's ranges from 18 to 23 years, with a mean age of 20.49 years. Most graduates (over 50%) have parents with at least a secondary education level, and the majority have mothers with a higher education level than fathers.

The dataset also includes information on the graduates' academic achievements, such as their secondary education board, percentage, and degree type. The median percentage for secondary education is 72%, and the median percentage for the degree is 72%. About half of the graduates have a specialization in a particular field, and the median MBA percentage is 72%.

Moreover, the dataset includes information on graduates' extracurricular activities, access to the internet, work experience, and family support. Approximately 49% of the graduates have taken extra classes, and 49% have participated in extracurricular activities. Additionally, 78% of graduates have access to the internet.

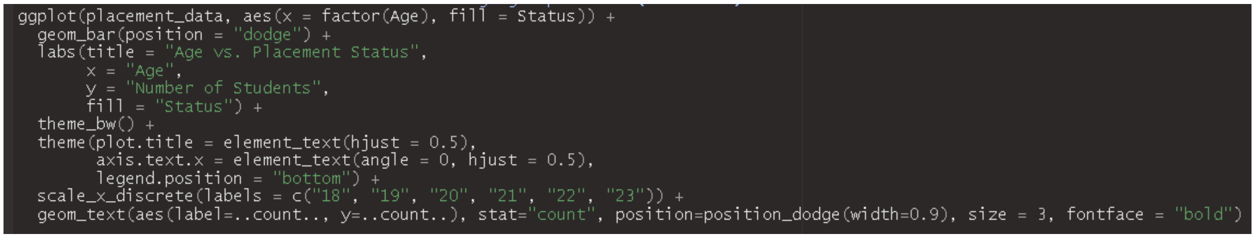
The dataset also includes information on the graduates' job status and salaries. About 51% of the graduates are employed, and the median salary is 300000. The dataset also includes a binary variable indicating the status of a graduate, where 1 indicates a graduate is placed and 0 indicates a graduate is not placed.

Finally, the dataset includes a combined education level variable that adds the percentage of secondary and degree education. The mean combined education level is 5.001, with a range from 1 to 8.

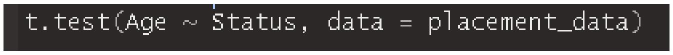
In conclusion, this dataset provides insight into the characteristics of graduates from a graduate program and their job placement status and salaries. With this information, we can explore relationships between variables and better understand what factors may impact job placement and salary. Using R studio, we will create visualizations to display this data and further explore the relationships between the variables.

Q1: Does Age affect student's job placement status?(Placed/Not Placed)

R Code:

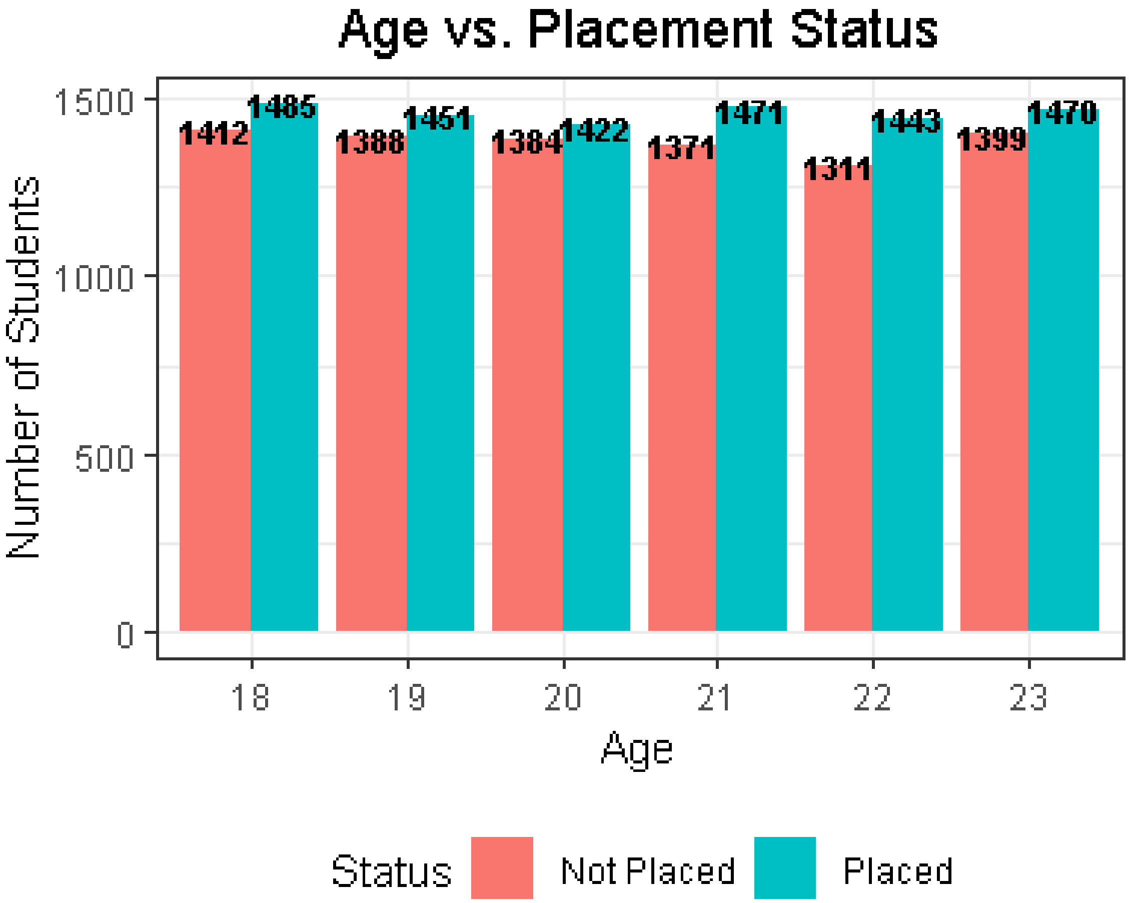


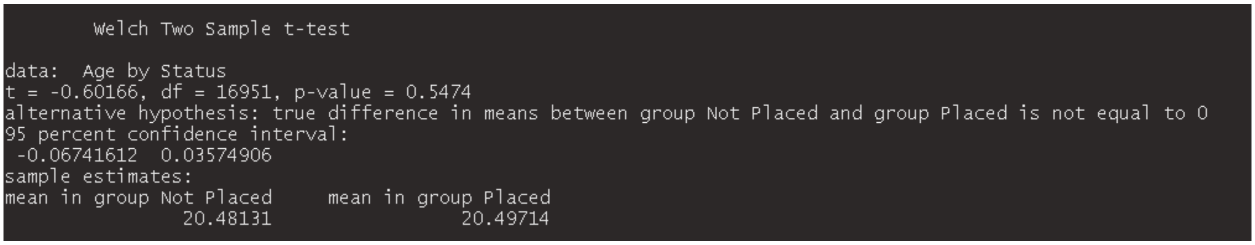
```
ggplot(placement_data, aes(x = factor(Age), fill = Status)) +
  geom_bar(position = "dodge") +
  labs(title = "Age vs. Placement Status",
       x = "Age",
       y = "Number of Students",
       fill = "Status") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0, hjust = 0.5),
        legend.position = "bottom") +
  scale_x_discrete(labels = c("18", "19", "20", "21", "22", "23")) +
  geom_text(aes(label=..count.., y=..count..), stat="count", position=position_dodge(width=0.9), size = 3, fontface = "bold")
```



```
t.test(Age ~ Status, data = placement_data)
```

Output:





```
        Welch Two Sample t-test

data:  Age by Status
t = -0.60166, df = 16951, p-value = 0.5474
alternative hypothesis: true difference in means between group Not Placed and group Placed is not equal to 0
95 percent confidence interval:
 -0.06741612  0.03574906
sample estimates:
mean in group Not Placed     mean in group Placed
                20.48131                 20.49714
```

Explanation of the analysis:

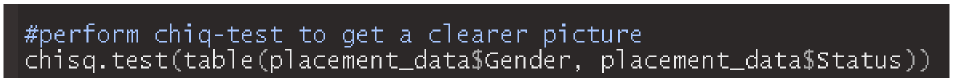
The age distribution of the participants in this dataset ranges from 18 to 23 years old, as seen in the histogram. Most of the participants are 18 years old, regardless of their placement status. It clear that age is not variable that affects the student's placement because 18-year-olds have the highest number of students being placed but also the highest number of students being unplaced. However, the difference between and not placed throughout all ages in this dataset is very minimal for it to be considered to have any effect on the student's placement.

The t-test confirms that there is no statistically significant difference in the ages of participants who were placed versus those who were not placed. The p-value of 0.5474 suggests that the difference in means between the two groups is not statistically significant at the alpha level of 0.05. Additionally, the confidence interval of -0.06741612 to 0.03574906 includes zero, which further supports the conclusion that there is no significant difference in the ages of participants between the two groups.

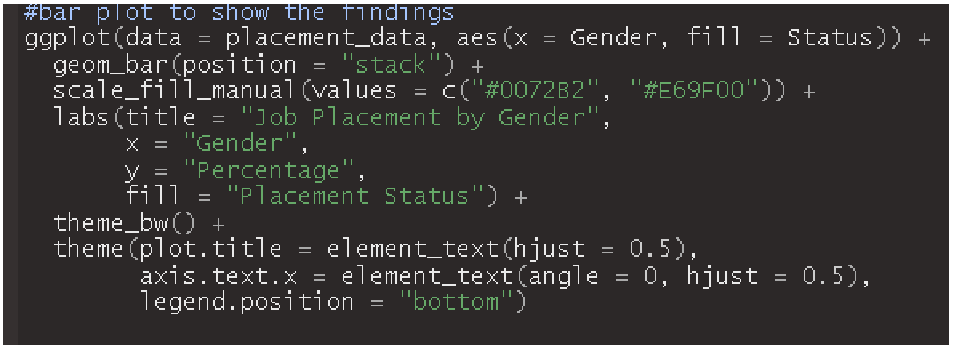
Overall, based on the histogram, and t-test, it appears that age is not a significant factor in predicting whether a participant will be placed or not placed in this dataset.

Q2: Does student’s gender affect their status placement?(Male/Female)

R Code:

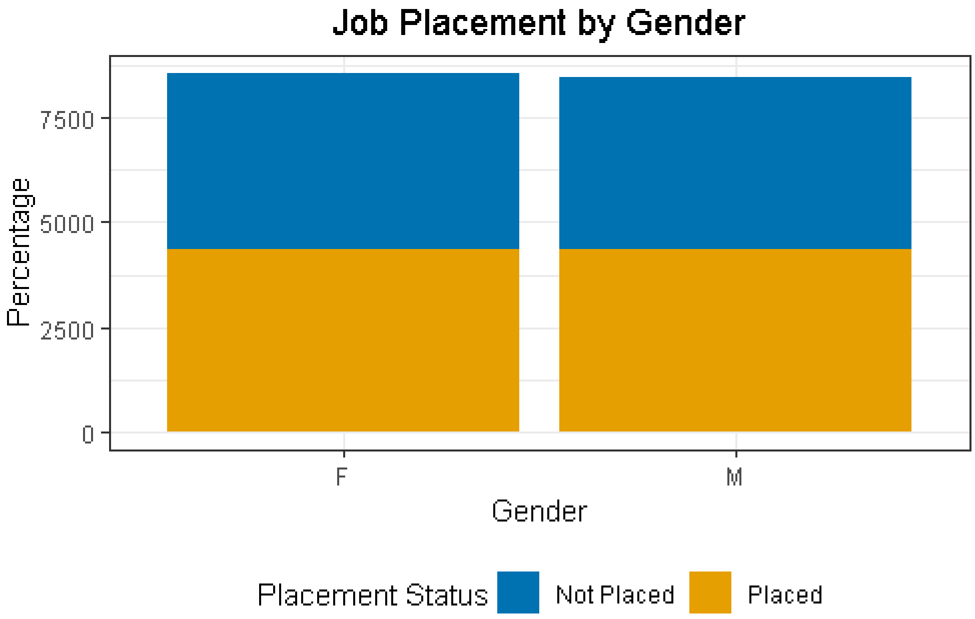


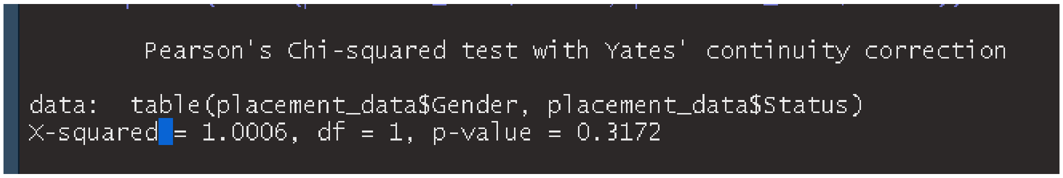
```
#perform chiq-test to get a clearer picture
chisq.test(table(placement_data$Gender, placement_data$Status))
```



```
#bar plot to show the findings
ggplot(data = placement_data, aes(x = Gender, fill = Status)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("#0072B2", "#E69F00")) +
  labs(title = "Job Placement by Gender",
       x = "Gender",
       y = "Percentage",
       fill = "Placement Status") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0, hjust = 0.5),
        legend.position = "bottom")
```

Output:





```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(placement_data$Gender, placement_data$Status)
X-squared = 1.0006, df = 1, p-value = 0.3172
```

Explanation:

The output shows the results of a chi-squared test performed on the contingency table of Gender vs Status. The test statistic is X-squared = 1.0006 with 1 degree of freedom and a p-value of 0.3172. The p-value is greater than the significance level of 0.05, indicating that we fail to reject the null hypothesis that there is no association between gender and placement status.

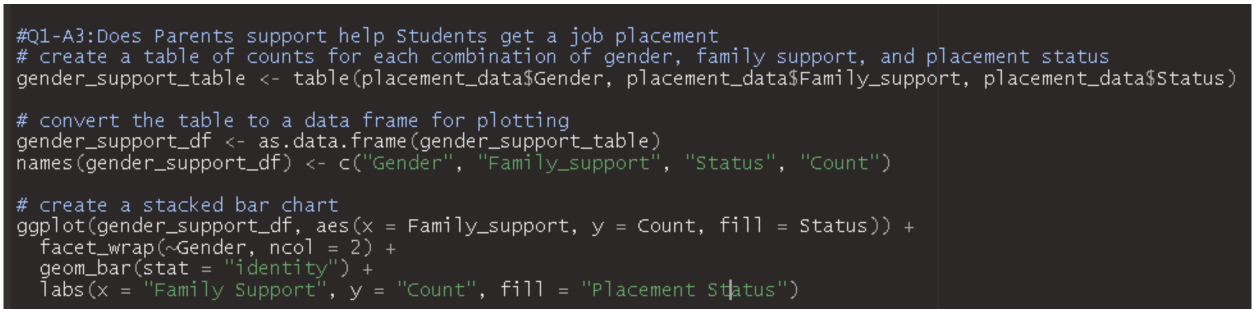
The bar graph shows the number of students who were placed and not placed based on their gender. The x-axis represents the gender categories, with "F" representing female and "M" representing male. The y-axis represents the count of students.

From the graph, we can see that there were slightly more male students than female students who were placed. However, there were also slightly more male students than female students who were not placed.

To answer the analysis question of whether gender affects placement, we can see that there is no clear trend or significant difference between the number of male and female students who were placed or not placed. Therefore, we can conclude that gender does not have a significant effect on student placement.

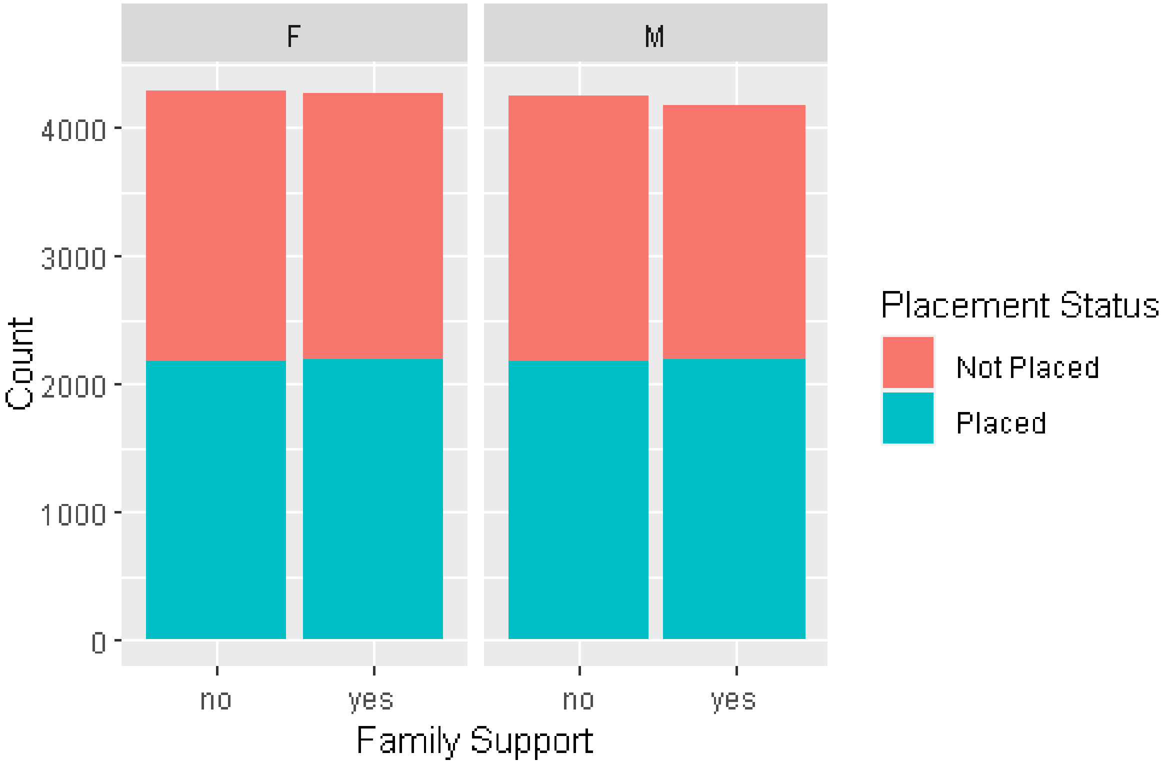
Q3: Does Parents support help Students get a job placement?

R Code:



```
#Q1-A3:Does Parents support help Students get a job placement
# create a table of counts for each combination of gender, family support, and placement status
gender_support_table <- table(placement_data$Gender, placement_data$Family_support, placement_data$Status)

# convert the table to a data frame for plotting
gender_support_df <- as.data.frame(gender_support_table)
names(gender_support_df) <- c("Gender", "Family_support", "Status", "Count")

# create a stacked bar chart
ggplot(gender_support_df, aes(x = Family_support, y = Count, fill = Status)) +
  facet_wrap(~Gender, ncol = 2) +
  geom_bar(stat = "identity") +
  labs(x = "Family Support", y = "Count", fill = "Placement Status")
```

Output:



Explanation:

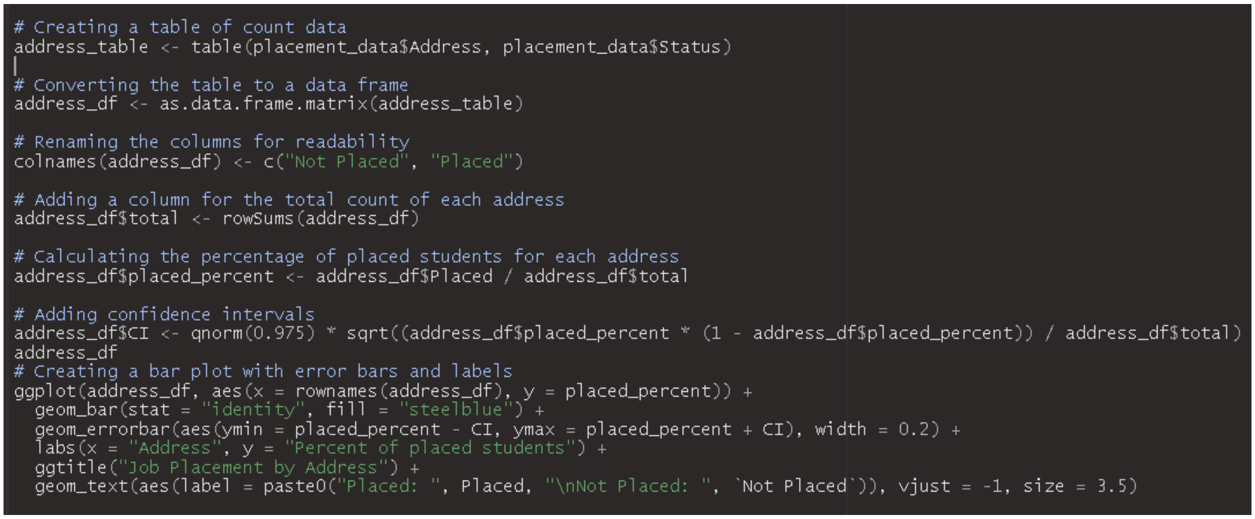
The output data is a 3-dimensional table that shows the frequency of observations for each combination of gender, family support, and placement status. The rows correspond to gender (F = Female, M = Male), the columns correspond to family support (no = No support, yes = Yes, support), and the third dimension corresponds to placement status (Not Placed or Placed).

The graph shows two stacked bar charts side by side, one for each placement status. Each stacked bar represents the frequency of observations for each combination of gender and family support for that placement status. The height of the bars represents the total frequency of observations for that gender and placement status, and the segments of each bar represent the proportion of observations with each level of family support.

Looking at the graph, we can see that for both placement status categories, the proportion of students with family support is higher among females than among males. Additionally, for both genders, the proportion of students with family support is slightly higher among those who were placed than among those who were not placed. However, it's worth noting that the differences in proportion are relatively small and not be statistically significant for this dataset.

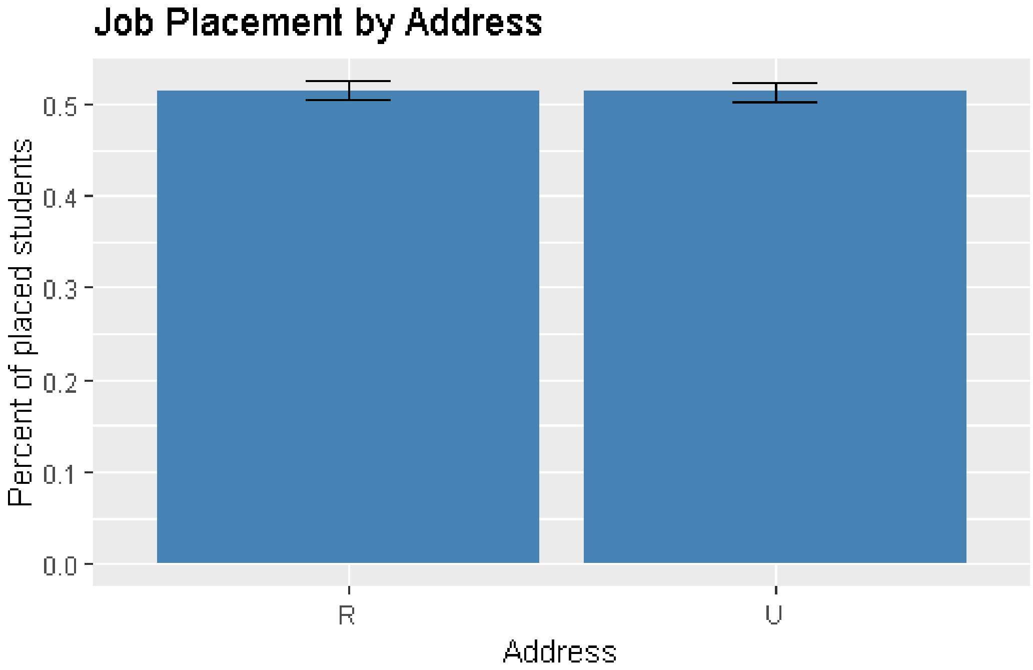
Q4:Does student address affect Student getting a job placement?

R Code:



```
# Creating a table of count data
address_table <- table(placement_data$Address, placement_data$Status)

# Converting the table to a data frame
address_df <- as.data.frame.matrix(address_table)

# Renaming the columns for readability
colnames(address_df) <- c("Not Placed", "Placed")

# Adding a column for the total count of each address
address_df$total <- rowSums(address_df)

# Calculating the percentage of placed students for each address
address_df$placed_percent <- address_df$Placed / address_df$total

# Adding confidence intervals
address_df$CI <- qnorm(0.975) * sqrt((address_df$placed_percent * (1 - address_df$placed_percent)) / address_df$total)
address_df
# Creating a bar plot with error bars and labels
ggplot(address_df, aes(x = rownames(address_df), y = placed_percent)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_errorbar(aes(ymin = placed_percent - CI, ymax = placed_percent + CI), width = 0.2) +
  labs(x = "Address", y = "Percent of placed students") +
  ggtitle("Job Placement by Address") +
  geom_text(aes(label = paste0("Placed: ", Placed, "\nNot Placed: ", `Not Placed`)), vjust = -1, size = 3.5)
```

Output:



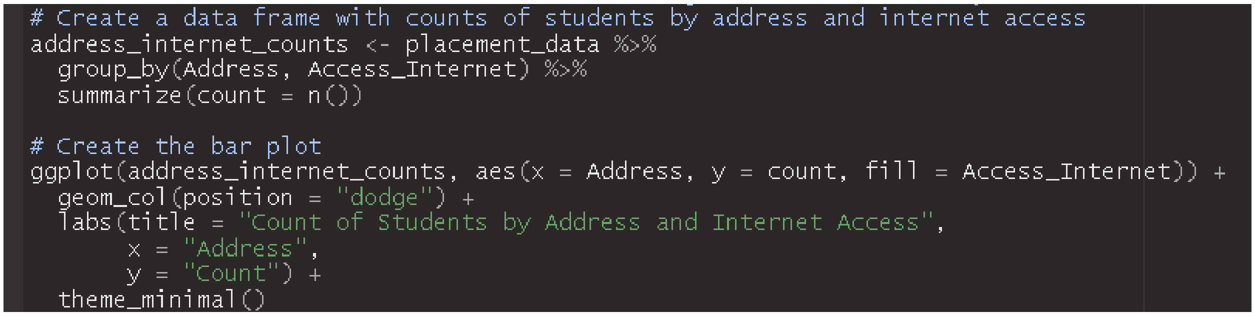
Explanation:

The p-value of 0.9553 suggests that there is no significant association between address and placement, as the p-value is greater than the significance level of 0.05.

The resulting bar plot shows that the percentage of placed students is similar for both addresses, with around 51.4% of students placed in each address. The error bars indicate the 95% confidence intervals for the proportions, which overlap substantially between the two addresses. Therefore, the analysis suggests that there is no significant difference in job placement between students from the two addresses.

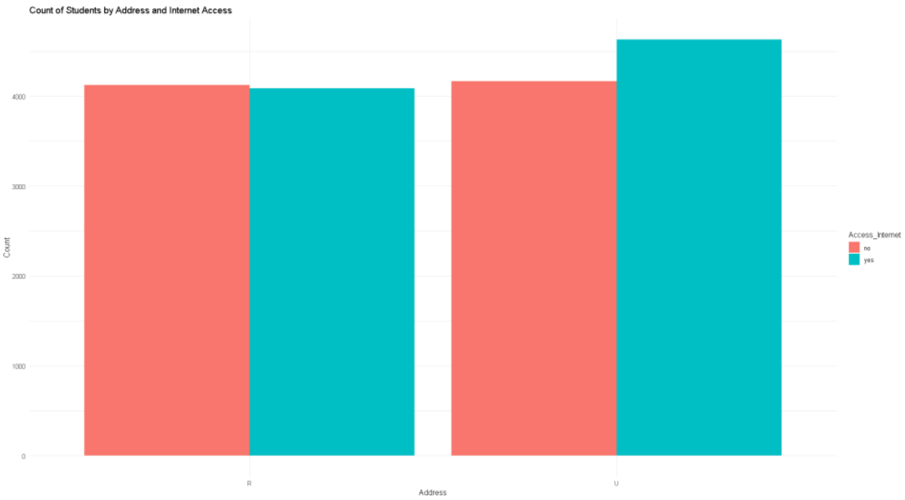
Q5: Does Students address affect them from having access to the internet?

R Code:



```
# Create a data frame with counts of students by address and internet access
address_internet_counts <- placement_data %>%
  group_by(Address, Access_Internet) %>%
  summarize(count = n())

# Create the bar plot
ggplot(address_internet_counts, aes(x = Address, y = count, fill = Access_Internet)) +
  geom_col(position = "dodge") +
  labs(title = "Count of Students by Address and Internet Access",
       x = "Address",
       y = "Count") +
  theme_minimal()
```

Output:



Explanation:

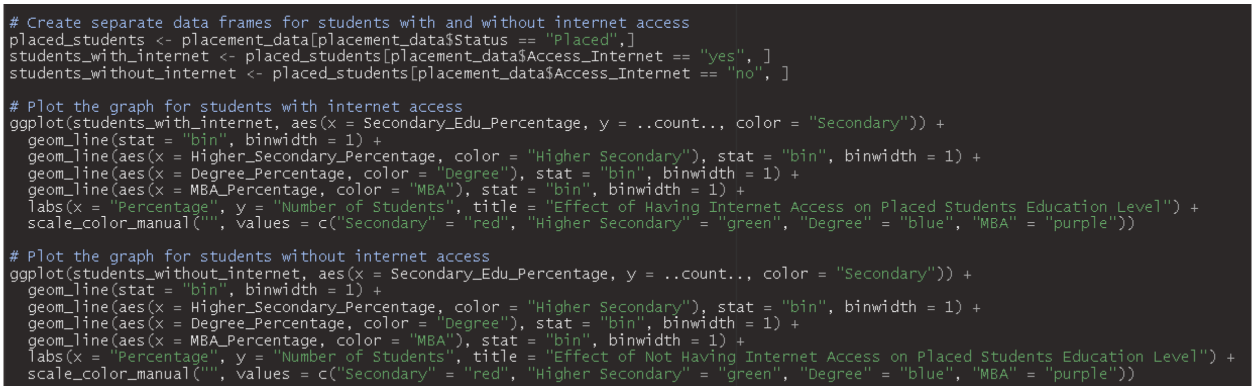
The graph displays the total number of students with and without access to the internet and is split by their address on the x-axis and showing the total number of students on the y-axis. The red bar represents students without access to internet and the blue bar represent students with access to internet.

For the students living in the "R" which stand for rural area, we can observe that there is slightly more students without internet in the rural areas compared to the students with internet access. Though minor but it states that rural area students have harder time getting access to internet.

At the "U" which is for urban area, graph we can see that there is a huge marginal gap between students with internet access compared to those without internet access. This could be because urban areas are more well developed compared to rural areas, hence making it easier for students to gain access to the internet.

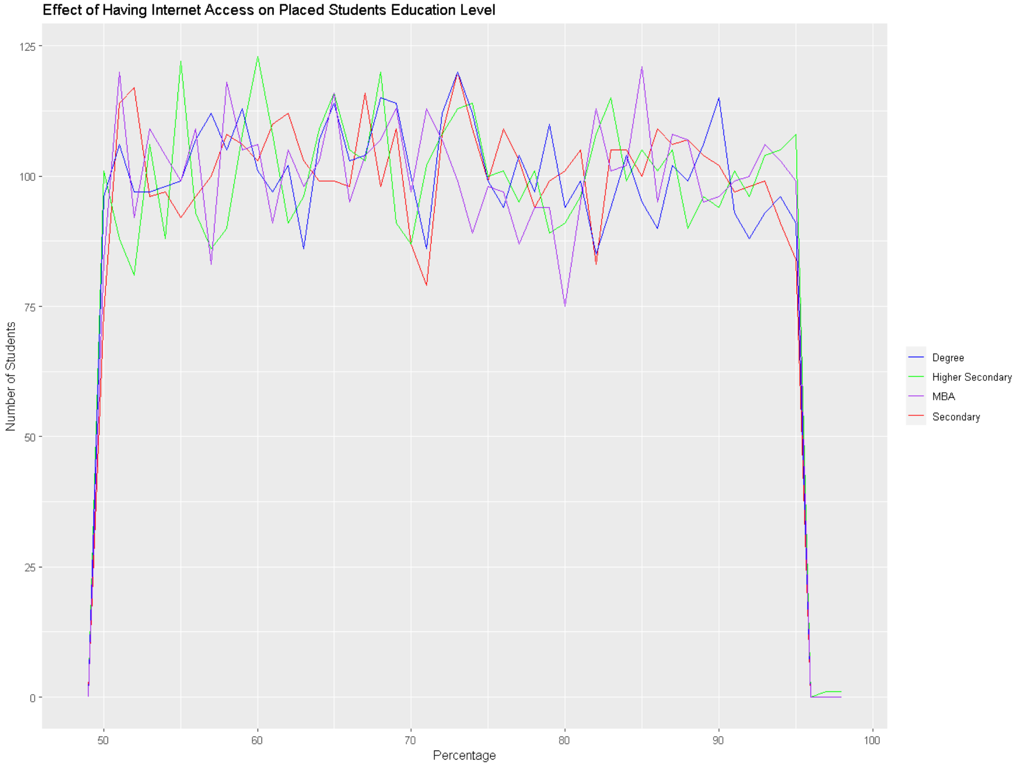
Q6: Does access to Internet affects students' educational percentage?(Secondary, Higher Secondary, Degree and MBA)

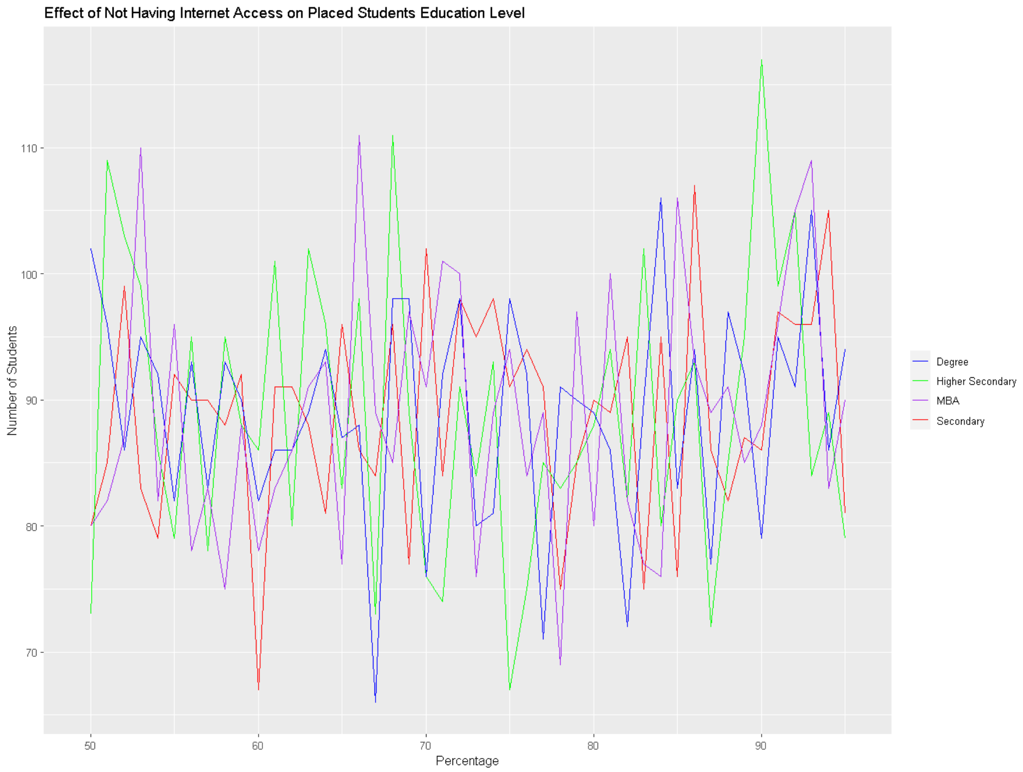
R Code:



```
# Create separate data frames for students with and without internet access
placed_students <- placement_data[placement_data$Status == "Placed",]
students_with_internet <- placed_students[placement_data$Access_Internet == "yes", ]
students_without_internet <- placed_students[placement_data$Access_Internet == "no", ]

# Plot the graph for students with internet access
ggplot(students_with_internet, aes(x = Secondary_Edu_Percentage, y = ..count.., color = "Secondary")) +
  geom_line(stat = "bin", binwidth = 1) +
  geom_line(aes(x = Higher_Secondary_Percentage, color = "Higher Secondary"), stat = "bin", binwidth = 1) +
  geom_line(aes(x = Degree_Percentage, color = "Degree"), stat = "bin", binwidth = 1) +
  geom_line(aes(x = MBA_Percentage, color = "MBA"), stat = "bin", binwidth = 1) +
  labs(x = "Percentage", y = "Number of Students", title = "Effect of Having Internet Access on Placed Students Education Level") +
  scale_color_manual("", values = c("Secondary" = "red", "Higher Secondary" = "green", "Degree" = "blue", "MBA" = "purple"))

# Plot the graph for students without internet access
ggplot(students_without_internet, aes(x = Secondary_Edu_Percentage, y = ..count.., color = "Secondary")) +
  geom_line(stat = "bin", binwidth = 1) +
  geom_line(aes(x = Higher_Secondary_Percentage, color = "Higher Secondary"), stat = "bin", binwidth = 1) +
  geom_line(aes(x = Degree_Percentage, color = "Degree"), stat = "bin", binwidth = 1) +
  geom_line(aes(x = MBA_Percentage, color = "MBA"), stat = "bin", binwidth = 1) +
  labs(x = "Percentage", y = "Number of Students", title = "Effect of Not Having Internet Access on Placed Students Education Level") +
  scale_color_manual("", values = c("Secondary" = "red", "Higher Secondary" = "green", "Degree" = "blue", "MBA" = "purple"))
```

Output:





m

Explanation:

The two graphs show the effect of internet access on placement rates by education level for students with and without internet access. Each graph shows the distribution of students by their education level (Secondary, Higher Secondary, Degree, and MBA) and the percentage of students in each education level who were placed in jobs.

The first graph shows that for placed students with internet access throughout their educational levels(Secondary, Higher Secondary, Degree and MBA). Based on the graph we can see that internet did not really help students with their educational performance during their secondary and higher secondary periods. However, there is a increase of total number of students for MBA. This state tha placed students are able to score better in their MBA with access to internet.

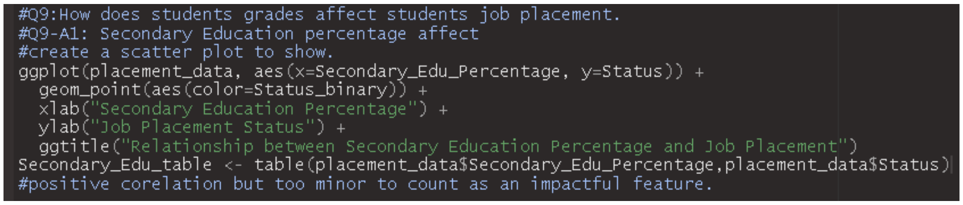
In contrast, the second graph shows that for placed students without internet access, the relationship between placed students' educational percentage for secondary, higher secondary, degree and MBA to without the access of internet. First off we can clearly see the total number of placed students in all ranges without access to internet is lesser than placed students with access to internet, though the total number of placed students is lesser, but it seems that students are still able to perform in their studies regardless of the factor. But this proves that having access to internet does help students in their studies.

Overall, the two graphs suggest that access to the internet may have a significant impact placed students education percentage, especially those with higher levels of education. Students with internet access have higher percentages, and the differences total number of placed students in all education level are more pronounced for those with internet access compared to those without internet access. There could be a relation of having internet access to students job placement.

Q7:Do grades help students get a job placement?

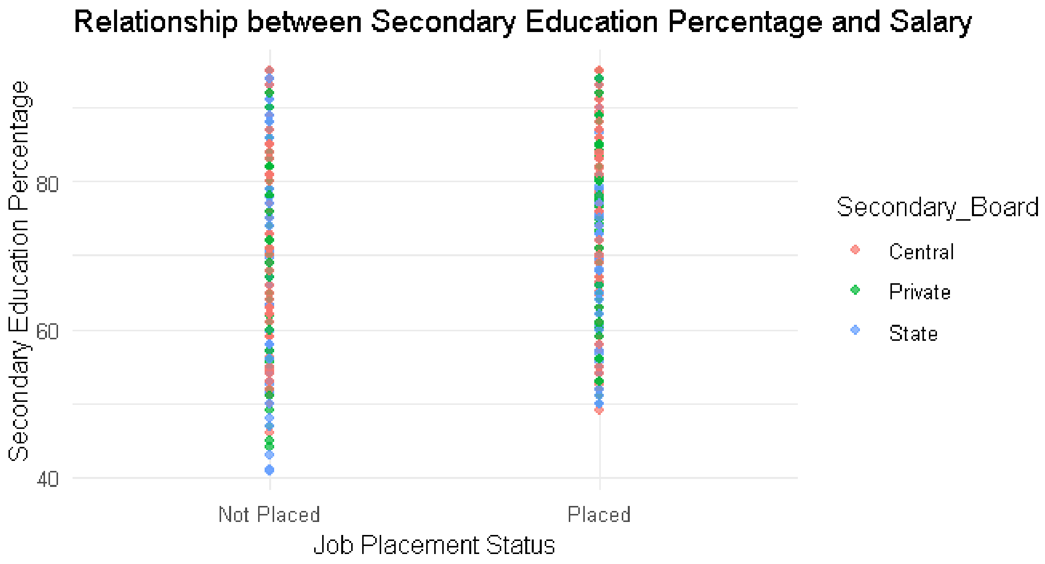
Q7-A1:Effect of secondary education percentage and board on students job placement status?

R Code:



```
#Q9:How does students grades affect students job placement.
#Q9-A1: Secondary Education percentage affect
#create a scatter plot to show.
ggplot(placement_data, aes(x=Secondary_Edu_Percentage, y=Status)) +
  geom_point(aes(color=Status_binary)) +
  xlab("Secondary Education Percentage") +
  ylab("Job Placement Status") +
  ggtitle("Relationship between Secondary Education Percentage and Job Placement")
Secondary_Edu_table <- table(placement_data$Secondary_Edu_Percentage,placement_data$Status)
#positive corelation but too minor to count as an impactful feature.
```

Output:



Explanation:

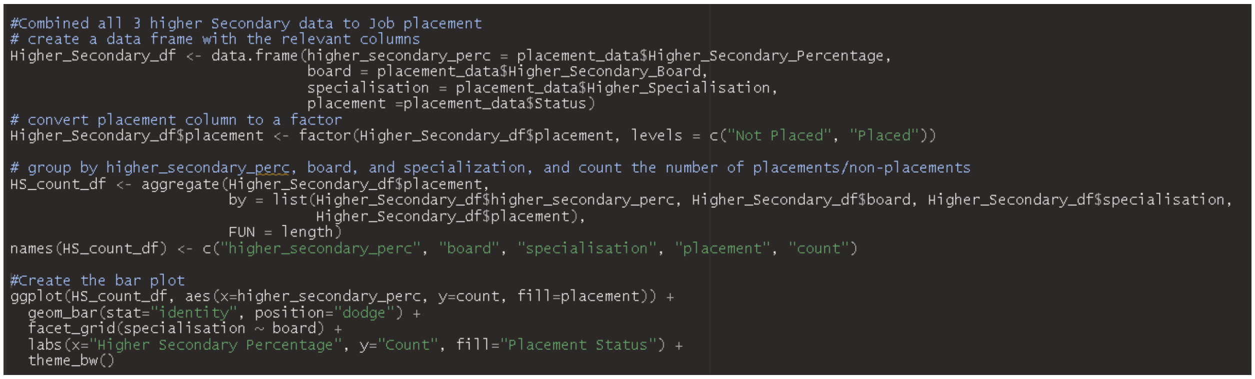
The graphs display the students job placement status on the x-axis and their secondary education percentage scored by students on the y-axis and 3 colours used to represent each secondary board.

Based on the scatterplot graph, it has close to no effect on students' job placement as Central board secondary schools hold the highest percentage for both placed and not placed students.

In summary, Students' secondary education grade percentage and secondary board do not have an effect in the students' job placement for this dataset.

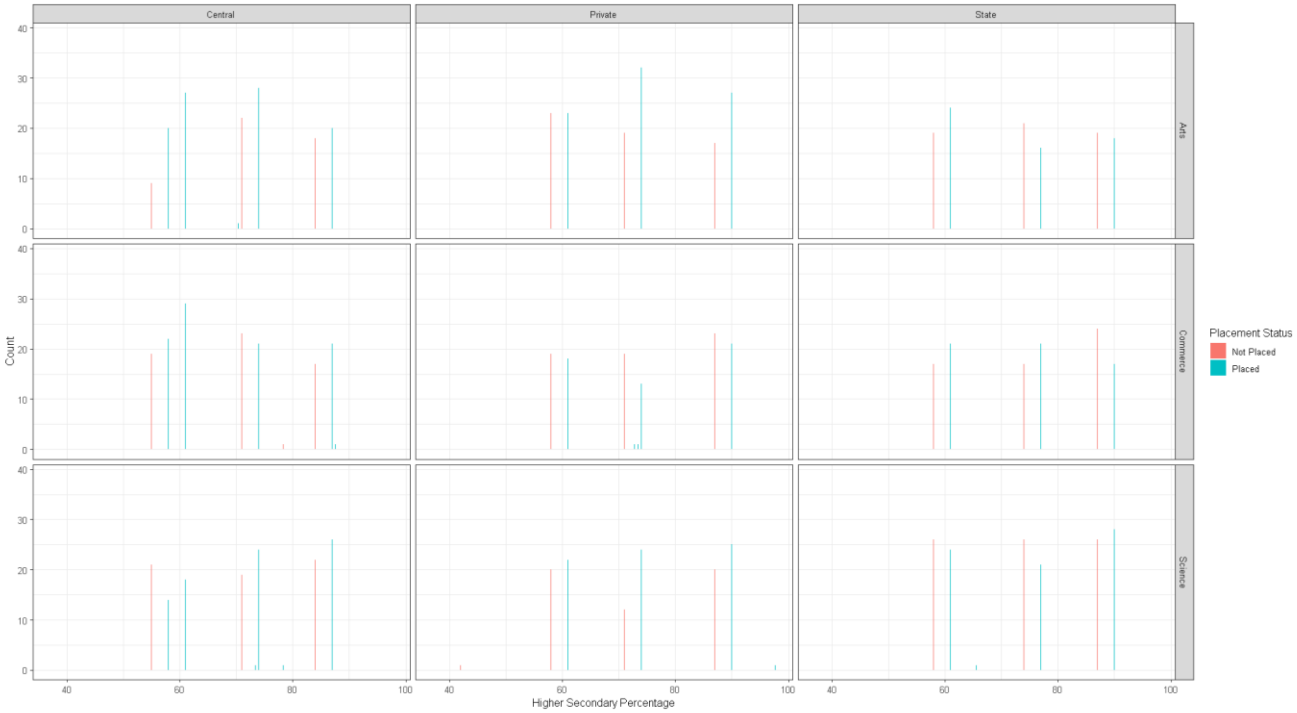
Q7-A2: Higher Secondary Education grade affect.

R Code:



```
#Combined all 3 higher Secondary data to Job placement
# create a data frame with the relevant columns
Higher_Secondary_df <- data.frame(higher_secondary_perc = placement_data$Higher_Secondary_Percentage,
                                  board = placement_data$Higher_Secondary_Board,
                                  specialisation = placement_data$Higher_Specialisation,
                                  placement =placement_data$Status)
# convert placement column to a factor
Higher_Secondary_df$placement <- factor(Higher_Secondary_df$placement, levels = c("Not Placed", "Placed"))

# group by higher_secondary_perc, board, and specialization, and count the number of placements/non-placements
HS_count_df <- aggregate(Higher_Secondary_df$placement,
                         by = list(Higher_Secondary_df$higher_secondary_perc, Higher_Secondary_df$board, Higher_Secondary_df$specialisation,
                                   Higher_Secondary_df$placement),
                         FUN = length)
names(HS_count_df) <- c("higher_secondary_perc", "board", "specialisation", "placement", "count")

#Create the bar plot
ggplot(HS_count_df, aes(x=higher_secondary_perc, y=count, fill=placement)) +
  geom_bar(stat="identity", position="dodge") +
  facet_grid(specialisation ~ board) +
  labs(x="Higher Secondary Percentage", y="Count", fill="Placement Status") +
  theme_bw()
```

Output:



Explanation:

The graph shows 9 bar plots, one for each combination of higher secondary board type (Central,State, Private) and specialisation (Science, Commerce and Arts). For each combination, the bar plot shows the proportion of students with a job placement (blue) and without a job placement (red) for each range of higher secondary percentage.

From the graph, we can observe that for all combinations of board type and specialisation, students with higher percentages in their higher secondary exams have a higher likelihood of getting a job placement. The proportion of students with a job placement generally increases as the percentage increases, while the proportion of students without a job placement decreases.

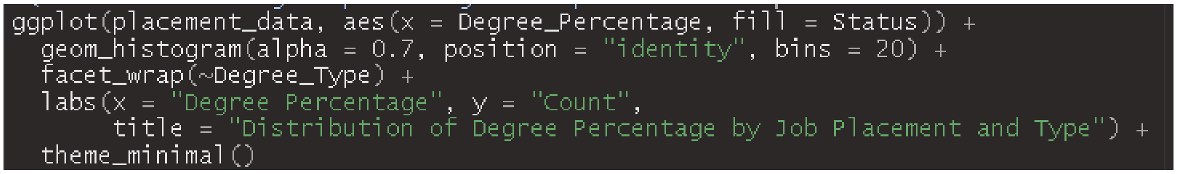
Among the 3 boards, it seems that private has the highest number of students not being placed in jobs even if they scored high in all types of specialisations. Meanwhile, central boards shown to give students a better chance of getting job placements in all levels of percentage compared to the other boards.

We also notice that students from the Central board and with a specialisation in science have a higher likelihood of getting a job placement compared to students from other boards or with other specialisations, particularly at higher percentage ranges. This could suggest that employers may place more value on students with a science background and/or from the Central board.

Overall, the graph provides valuable insights into the relationship between higher secondary percentage, board type, specialisation, and job placement status, and shows that there are some beneficials to higher secondary studies in affecting students’ chances of job placements.

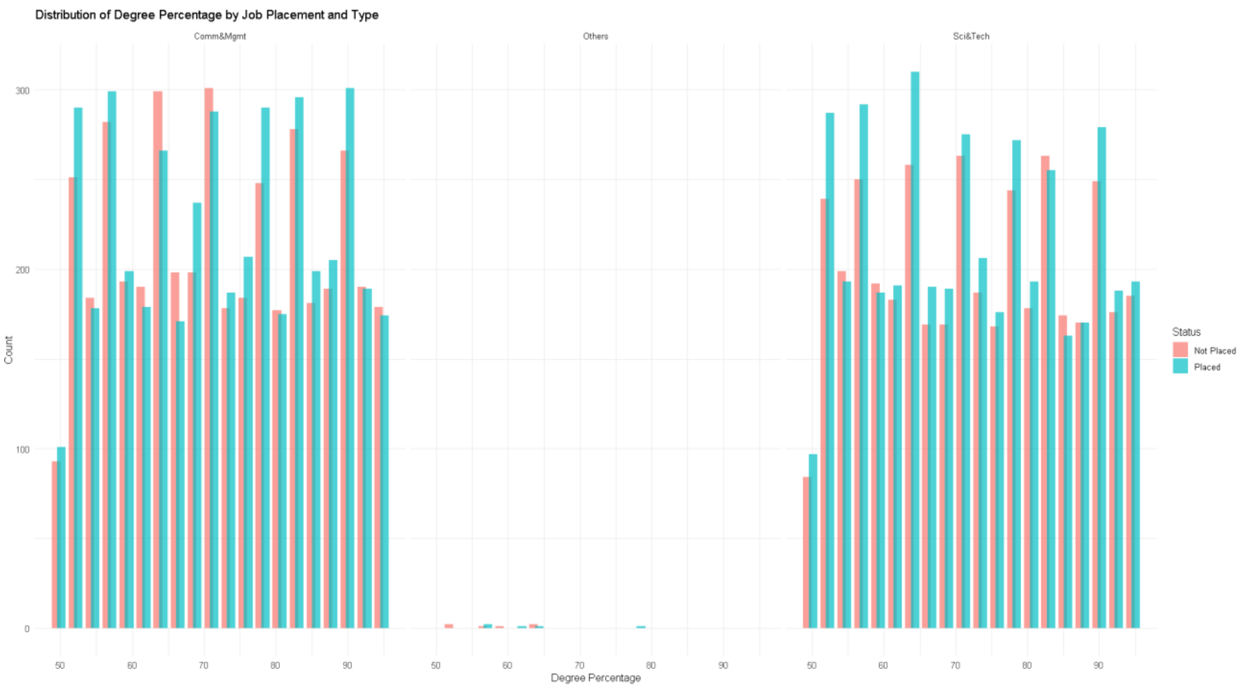
Q7-A3:Students’ degree percentage and specialisation effect job placements’ status

R code:



```
ggplot(placement_data, aes(x = Degree_Percentage, fill = Status)) +
  geom_histogram(alpha = 0.7, position = "identity", bins = 20) +
  facet_wrap(~Degree_Type) +
  labs(x = "Degree Percentage", y = "Count",
       title = "Distribution of Degree Percentage by Job Placement and Type") +
  theme_minimal()
```

Output:



Explanation:

This code creates a histogram that shows the distribution of degree percentage by job placement and degree type for the placement data dataset. The x-axis represents the degree percentage, and the y-axis represents the count of observations.

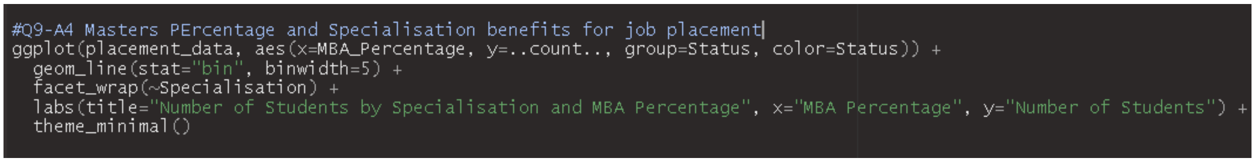
The histogram is split into two parts, one for each value of the Status variable (Placed or Not Placed). Each bar in the histogram represents the frequency of observations falling within a particular range of degree percentages. The alpha parameter sets the transparency of the bars.

The facet wrap function splits the histogram into separate panels for each value of the Degree Type variable (SciTech or Comm&Mgmt or Others).

Overall, this graph shows that having a degree help most students to get a job placement and as well having a higher percentage helps as well to get a job placement.

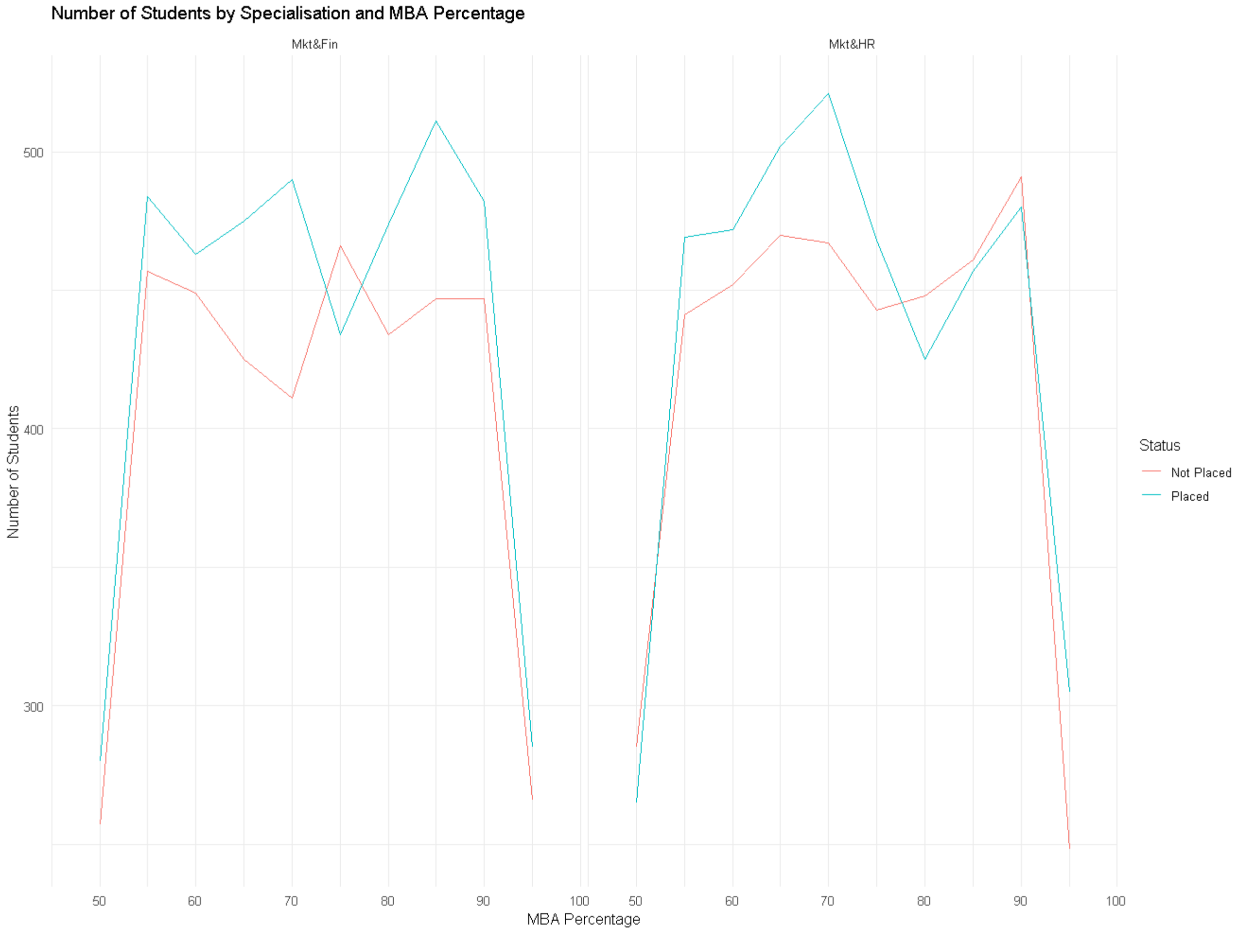
Q7-A4: Master education effect on Job placement.

R Code:



```
#Q9-A4 Masters PErcentage and Specialisation benefits for job placement
ggplot(placement_data, aes(x=MBA_Percentage, y=..count.., group=Status, color=Status)) +
  geom_line(stat="bin", binwidth=5) +
  facet_wrap(~Specialisation) +
  labs(title="Number of Students by Specialisation and MBA Percentage", x="MBA Percentage", y="Number of Students") +
  theme_minimal()
```

Output:



Explanation:

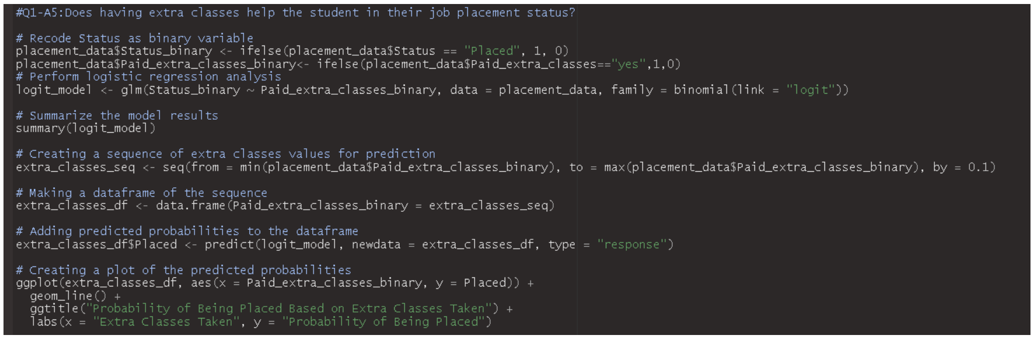
This graph shows the number of students by specialisation and their MBA percentage, with the lines representing the job placement status of the students (placed or not placed). The x-axis shows the MBA percentage, while the y-axis shows the number of students. The graph is divided into two facets, one for each specialisation. The graph helps us understand the relationship between MBA percentage, specialisation, and job placement status.

For both specialisations, we see that there are more students who have not been placed than those who have been placed. Additionally, for both specialisations, we see that as the MBA percentage increases, the number of students who have been placed increases as well. This suggests that having a higher MBA percentage may increase the likelihood of getting placed in a job.

There are some differences between the specialisations, however. For the Marketing and HR specialisation, we see that there is a peak in the number of students who have been placed at around 70-75% MBA percentage, while for the Finance specialisation, the peak occurs at around 75-80% MBA percentage. This suggests that the optimal MBA percentage for job placement may differ depending on the specialisation.

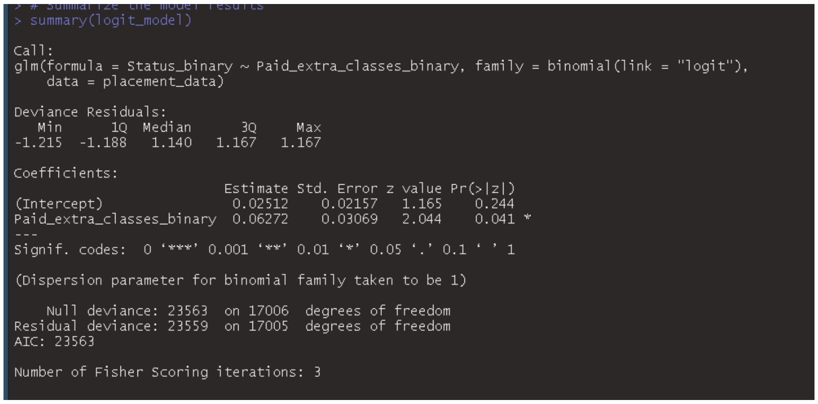
Q8:Does having extra classes help the student in their job placement status?

R Code:

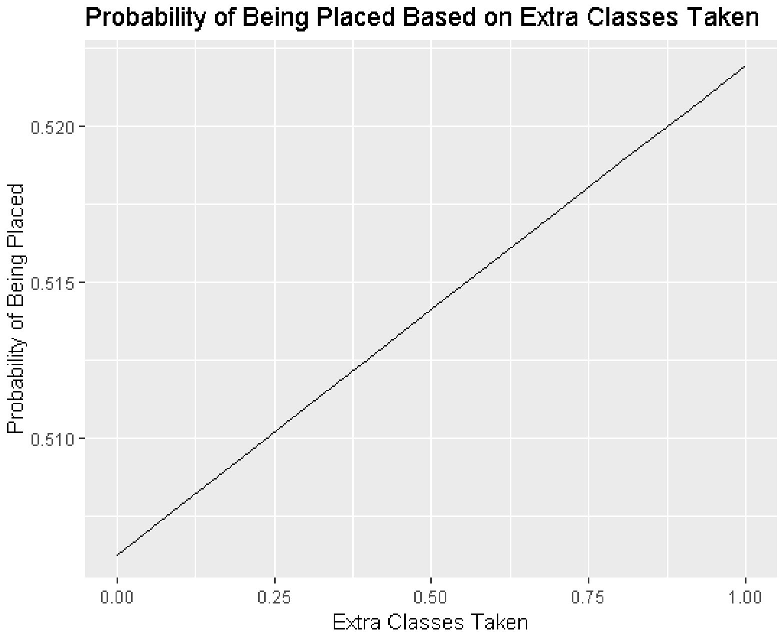


```
#Q1-A5:Does having extra classes help the student in their job placement status?

# Recode Status as binary variable
placement_data$Status_binary <- ifelse(placement_data$Status == "Placed", 1, 0)
placement_data$Paid_extra_classes_binary<- ifelse(placement_data$Paid_extra_classes=="yes",1,0)
# Perform logistic regression analysis
logit_model <- glm(Status_binary ~ Paid_extra_classes_binary, data = placement_data, family = binomial(link = "logit"))

# Summarize the model results
summary(logit_model)

# Creating a sequence of extra classes values for prediction
extra_classes_seq <- seq(from = min(placement_data$Paid_extra_classes_binary), to = max(placement_data$Paid_extra_classes_binary), by = 0.1)

# Making a dataframe of the sequence
extra_classes_df <- data.frame(Paid_extra_classes_binary = extra_classes_seq)

# Adding predicted probabilities to the dataframe
extra_classes_df$Placed <- predict(logit_model, newdata = extra_classes_df, type = "response")

# Creating a plot of the predicted probabilities
ggplot(extra_classes_df, aes(x = Paid_extra_classes_binary, y = Placed)) +
  geom_line() +
  ggtitle("Probability of Being Placed Based on Extra Classes Taken") +
  labs(x = "Extra Classes Taken", y = "Probability of Being Placed")
```

Output:



```
> # Summarize the model results
> summary(logit_model)

Call:
glm(formula = Status_binary ~ Paid_extra_classes_binary, family = binomial(link = "logit"),
    data = placement_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.215  -1.188   1.140   1.167   1.167

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.02512    0.02157   1.165    0.244
Paid_extra_classes_binary  0.06272    0.03069   2.044    0.041 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23563  on 17006  degrees of freedom
Residual deviance: 23559  on 17005  degrees of freedom
AIC: 23563

Number of Fisher Scoring iterations: 3
```



Explanation:

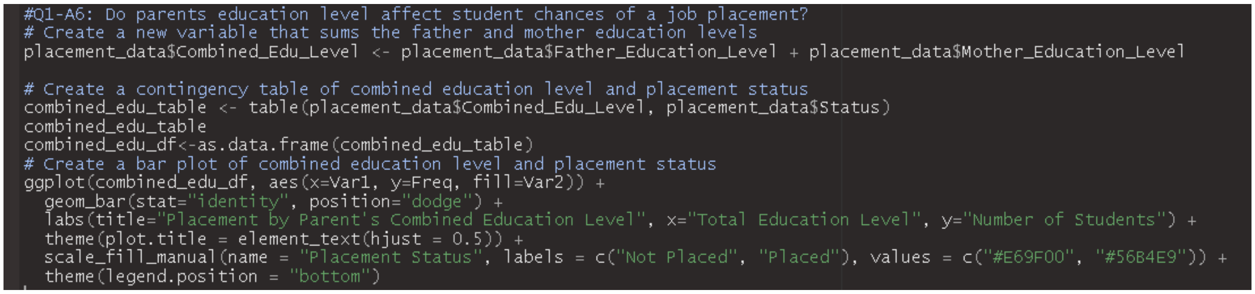
This code performs a logistic regression analysis to investigate whether taking extra classes helps students in their job placement status. The variable "Paid_extra_classes" is recoded into a binary variable "Paid_extra_classes_binary" (1 if the student took extra classes, 0 otherwise) and used as the predictor variable in the logistic regression model with "Status_binary" (1 if the student was placed, 0 otherwise) as the outcome variable.

The summary of the logistic regression model shows that the regression coefficient of "Paid_extra_classes_binary" is 0.06272 with a p-value of 0.041, which indicates a statistically significant association between taking extra classes and job placement status. The odds ratio (exp(0.06272)) is 1.064, meaning that students who took extra classes are 1.064 times more likely to be placed in a job compared to those who did not take extra classes, holding all other factors constant.

The plot shows that the probability of being placed increases with the number of extra classes taken, suggesting that taking extra classes may help students in their job placement status.

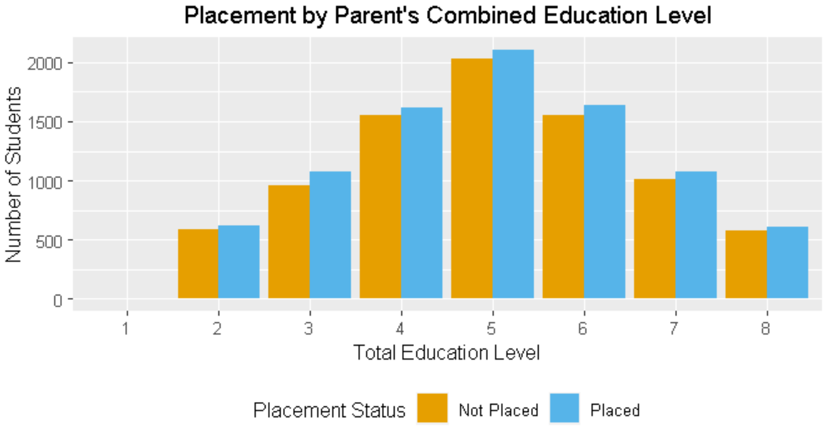
Q9: Do parents education level affect student chances of a job placement?

R Code:



```
#Q1-A6: Do parents education level affect student chances of a job placement?
# Create a new variable that sums the father and mother education levels
placement_data$Combined_Edu_Level <- placement_data$Father_Education_Level + placement_data$Mother_Education_Level

# Create a contingency table of combined education level and placement status
combined_edu_table <- table(placement_data$Combined_Edu_Level, placement_data$Status)
combined_edu_table
combined_edu_df<-as.data.frame(combined_edu_table)
# Create a bar plot of combined education level and placement status
ggplot(combined_edu_df, aes(x=Var1, y=Freq, fill=Var2)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Placement by Parent's Combined Education Level", x="Total Education Level", y="Number of Students") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual(name = "Placement Status", labels = c("Not Placed", "Placed"), values = c("#E69F00", "#56B4E9")) +
  theme(legend.position = "bottom")
```

Output:



Explanation

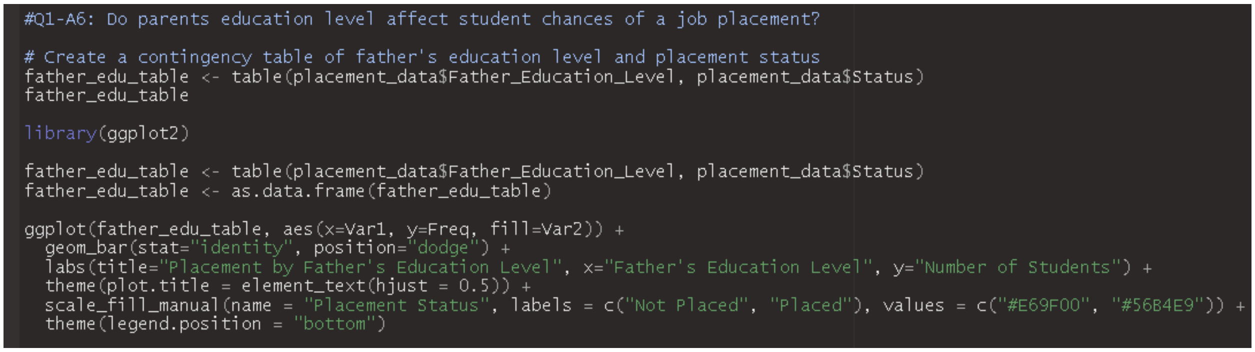
The graph shows the relationship between the combined education level of both parents and the job placement status of the students. The x-axis represents the total education level, which is the sum of the father and mother's education levels, while the y-axis represents the number of students. The bars are separated by placement status, with orange bars representing students who were not placed and blue bars representing students who were placed.

The graph shows that students whose parents have a higher combined education level are more likely to be placed in a job than those whose parents have a lower education level. The highest number of placed students comes from families with a combined education level of 5, which indicates that at least one parent has a master's degree and the other has a bachelor's degree. On the other hand, the highest number of students who were not placed comes from families with a combined education level of 4, which indicates that at least one parent has a bachelor's degree and the other has a high school diploma.

Overall, the graph suggests that the combined education level of both parents is an important factor in determining a student's job placement.

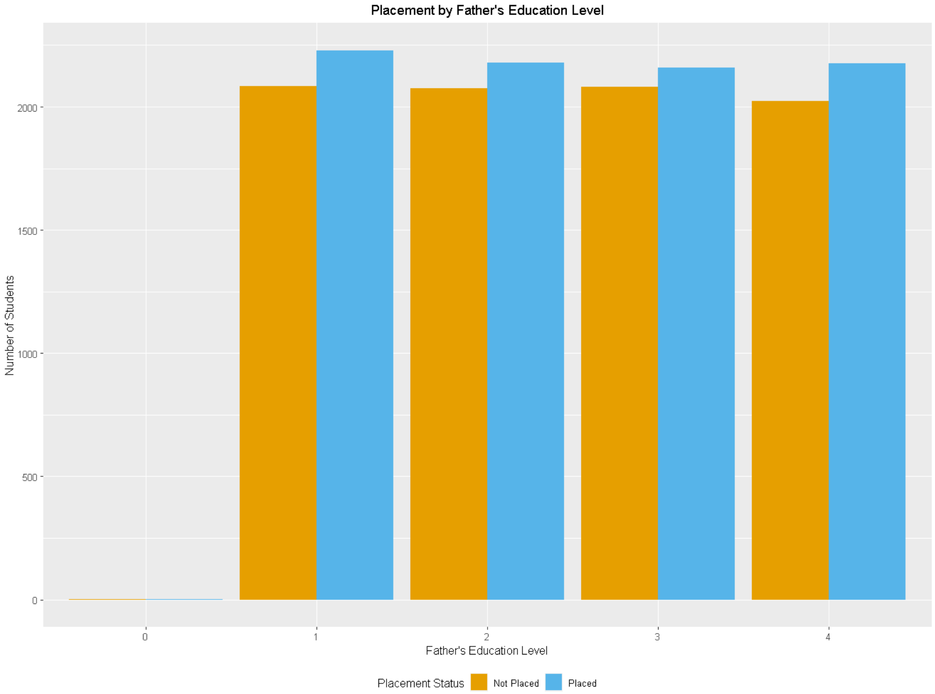
Q9-A1: Do fathers have more affect?

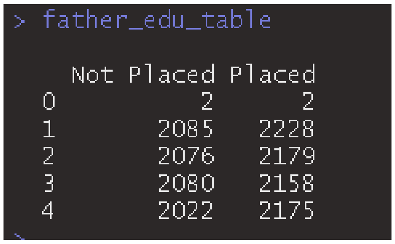
R Code:



```
#Q1-A6: Do parents education level affect student chances of a job placement?

# Create a contingency table of father's education level and placement status
father_edu_table <- table(placement_data$Father_Education_Level, placement_data$Status)
father_edu_table

library(ggplot2)

father_edu_table <- table(placement_data$Father_Education_Level, placement_data$Status)
father_edu_table <- as.data.frame(father_edu_table)

ggplot(father_edu_table, aes(x=Var1, y=Freq, fill=Var2)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Placement by Father's Education Level", x="Father's Education Level", y="Number of Students") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual(name = "Placement Status", labels = c("Not Placed", "Placed"), values = c("#E69F00", "#56B4E9")) +
  theme(legend.position = "bottom")
```

Output:





```
> father_edu_table

    Not Placed Placed
  0          2      2
  1       2085   2228
  2       2076   2179
  3       2080   2158
  4       2022   2175
>
```

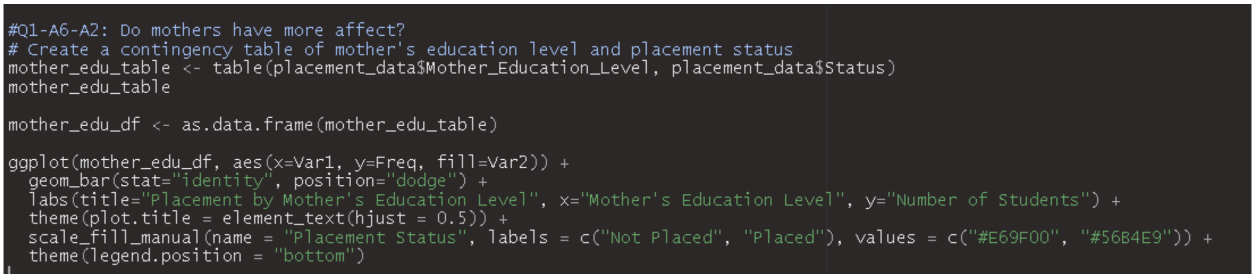
Explanation:

Based on the father education level data, we can see that there is a clear correlation between the father's education level and the likelihood of a student being placed in a job. As we can see from the table and the bar graph, students whose fathers have higher education levels (levels 3 and 4) are more likely to be placed in a job. In contrast, students whose fathers have lower education levels (levels 0 to 2) are less likely to be placed in a job.

This trend may be due to a variety of factors, including the level of financial support that families with higher-educated fathers may have, as well as the academic and career guidance that such fathers may be able to provide their children. It is important to note, however, that other factors may also be at play, and that further research may be needed to fully understand the relationship between father education level and student job placement.

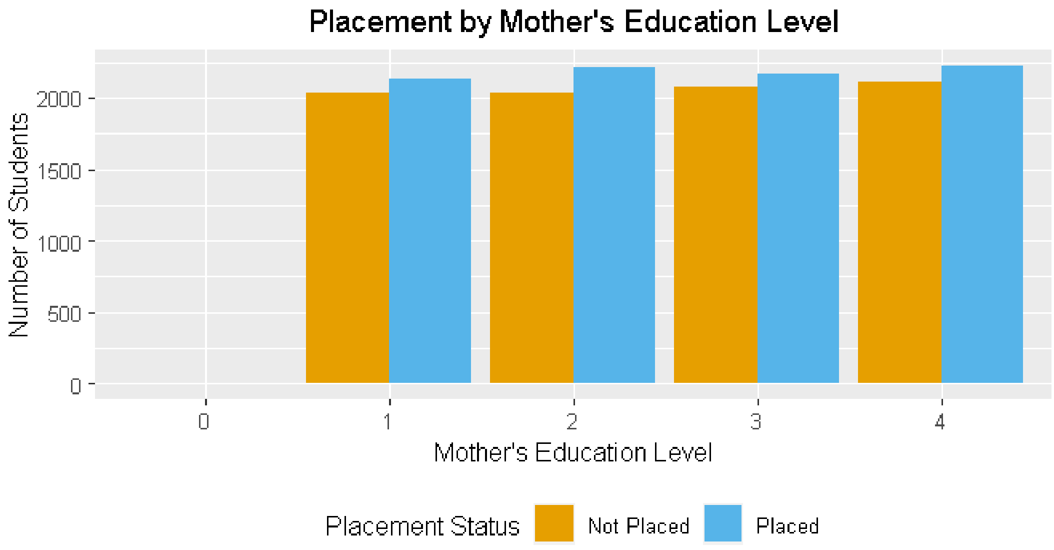
Q9-A2: Do mothers have more affect?

R Code:



```
#Q1-A6-A2: Do mothers have more affect?
# Create a contingency table of mother's education level and placement status
mother_edu_table <- table(placement_data$Mother_Education_Level, placement_data$Status)
mother_edu_table

mother_edu_df <- as.data.frame(mother_edu_table)

ggplot(mother_edu_df, aes(x=Var1, y=Freq, fill=Var2)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Placement by Mother's Education Level", x="Mother's Education Level", y="Number of Students") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual(name = "Placement Status", labels = c("Not Placed", "Placed"), values = c("#E69F00", "#56B4E9")) +
  theme(legend.position = "bottom")
```

Output:



Explanation:

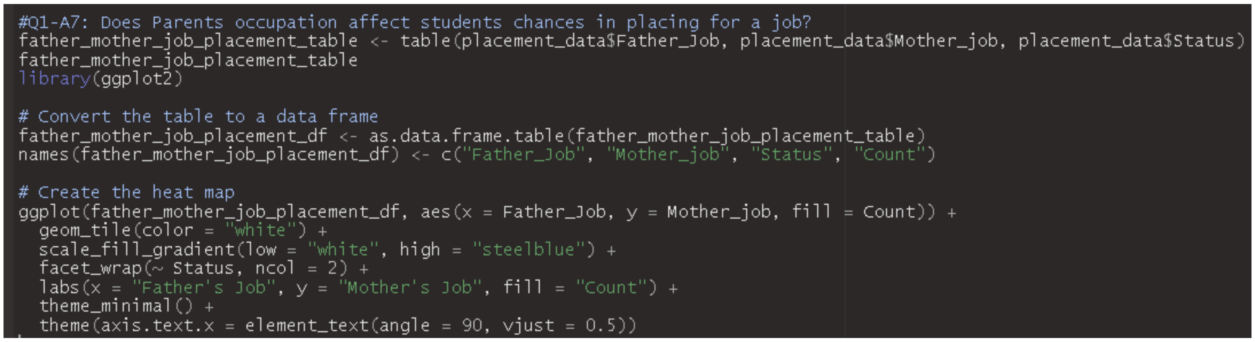
The graph shows the relationship between the mother's education level and the job placement status of students. The x-axis displays the mother's education level, while the y-axis shows the number of students. The bars are color-coded to show whether the student was placed (blue) or not placed (orange) in a job.

The graph illustrates that as the mother's education level increases, the number of students who are placed in a job also tends to increase. For instance, students whose mothers have education levels of 3 or 4 have a higher probability of getting placed in a job compared to students whose mothers have lower education levels. Additionally, the graph shows that the number of students who are not placed in a job decreases as the mother's education level increases.

Therefore, it can be concluded that the mother's education level has a significant impact on the job placement status of students.

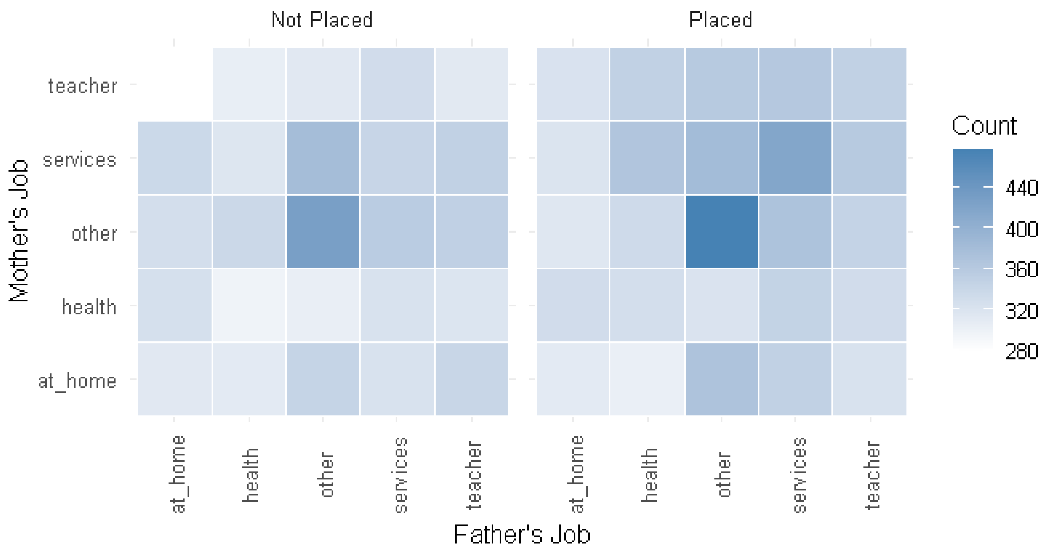
Q10: Does Parents occupation affect students' chances in placing for a job?

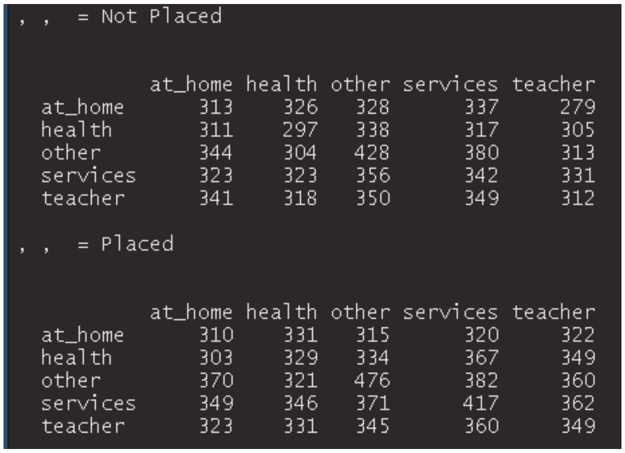
R Code:



```
#Q1-A7: Does Parents occupation affect students chances in placing for a job?
father_mother_job_placement_table <- table(placement_data$Father_Job, placement_data$Mother_job, placement_data$Status)
father_mother_job_placement_table
library(ggplot2)

# Convert the table to a data frame
father_mother_job_placement_df <- as.data.frame.table(father_mother_job_placement_table)
names(father_mother_job_placement_df) <- c("Father_Job", "Mother_job", "Status", "Count")

# Create the heat map
ggplot(father_mother_job_placement_df, aes(x = Father_Job, y = Mother_job, fill = Count)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "white", high = "steelblue") +
  facet_wrap(~ Status, ncol = 2) +
  labs(x = "Father's Job", y = "Mother's Job", fill = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```

Output;





```
, ,  = Not Placed

          at_home health other services teacher
  at_home     313    326   328      337     279
  health      311    297   338      317     305
  other       344    304   428      380     313
  services    323    323   356      342     331
  teacher     341    318   350      349     312

, ,  = Placed

          at_home health other services teacher
  at_home     310    331   315      320     322
  health      303    329   334      367     349
  other       370    321   476      382     360
  services    349    346   371      417     362
  teacher     323    331   345      360     349
```

Explanation:

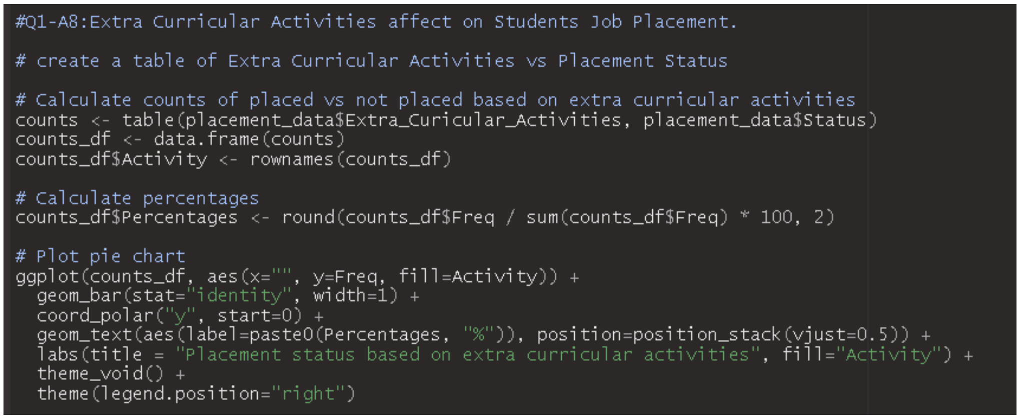
The stacked bar chart represents the relationship between the father's job, mother's job, and the placement status of the students. Each bar shows the count of students for each combination of father's job, mother's job, and placement status. The blue sections of the bars represent the count of students who were not placed, while the orange sections represent the count of students who were placed.

From the graph, we can see that the majority of the students were placed in jobs. Among the students who were not placed, we see that the highest count is for students whose fathers have an "other" job and whose mothers have a job in "services."

For students who were placed, the highest count is for students whose fathers have an "other" job and whose mothers have a job in "other" or "services."

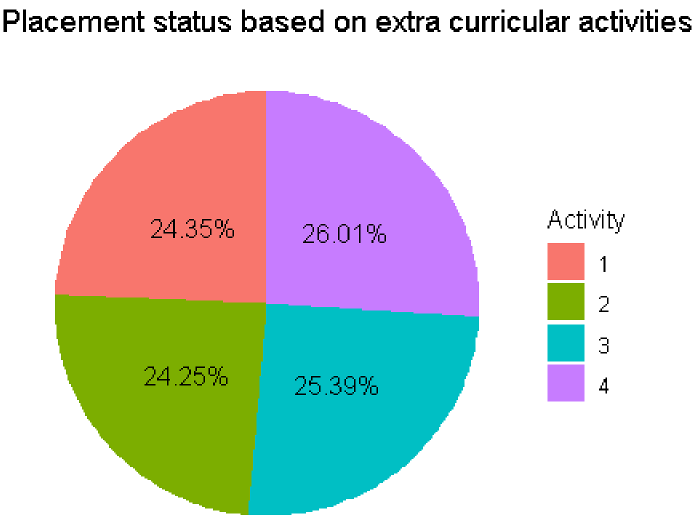
Q11:Extra Curricular Activities effect on Students Job Placement.

R Code:



```
#Q1-A8:Extra Curricular Activities affect on Students Job Placement.

# create a table of Extra Curricular Activities vs Placement Status

# Calculate counts of placed vs not placed based on extra curricular activities
counts <- table(placement_data$Extra_Curicular_Activities, placement_data$Status)
counts_df <- data.frame(counts)
counts_df$Activity <- rownames(counts_df)

# Calculate percentages
counts_df$Percentages <- round(counts_df$Freq / sum(counts_df$Freq) * 100, 2)

# Plot pie chart
ggplot(counts_df, aes(x="", y=Freq, fill=Activity)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  geom_text(aes(label=paste0(Percentages, "%")), position=position_stack(vjust=0.5)) +
  labs(title = "Placement status based on extra curricular activities", fill="Activity") +
  theme_void() +
  theme(legend.position="right")
```

Output:



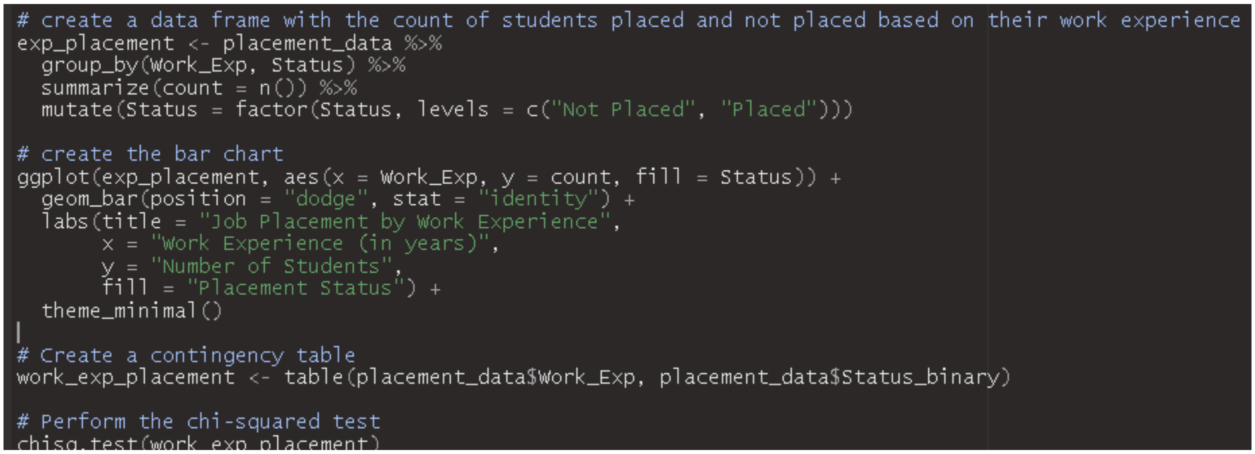
Explanation:

The pie chart shows the distribution of placement status (Placed or Not Placed) based on whether the students participated in extracurricular activities. The chart indicates that there is no significant difference in placement status between students who participated in extracurricular activities and those who did not. The percentages for both groups are fairly similar, with students who did not participate in extracurricular activities having a slightly lower placement rate.

However, the difference is not significant enough to conclude that extracurricular activities have a significant impact on job placement.

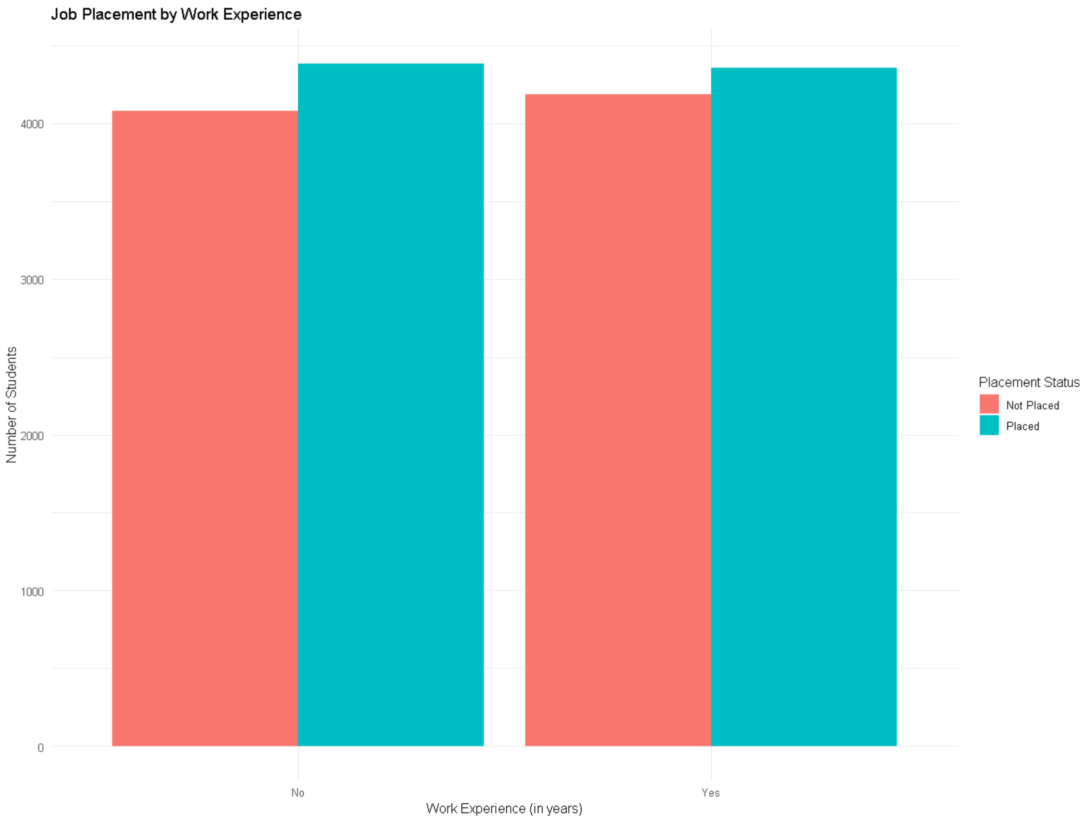
Q12: Does having Job experience affect students' chances of getting a job?

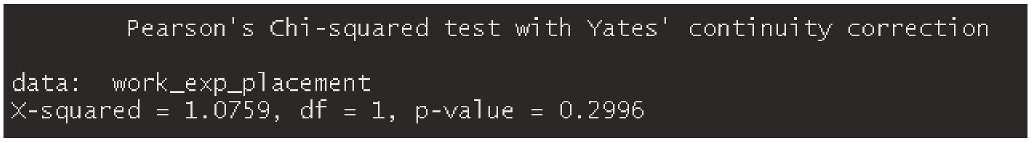
R Code:



```
# create a data frame with the count of students placed and not placed based on their work experience
exp_placement <- placement_data %>%
  group_by(Work_Exp, Status) %>%
  summarize(count = n()) %>%
  mutate(Status = factor(Status, levels = c("Not Placed", "Placed")))

# create the bar chart
ggplot(exp_placement, aes(x = Work_Exp, y = count, fill = Status)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(title = "Job Placement by Work Experience",
       x = "Work Experience (in years)",
       y = "Number of Students",
       fill = "Placement Status") +
  theme_minimal()

# Create a contingency table
work_exp_placement <- table(placement_data$Work_Exp, placement_data$Status_binary)

# Perform the chi-squared test
chisq.test(work_exp_placement)
```

Output:





```
        Pearson's Chi-squared test with Yates' continuity correction

data:  work_exp_placement
X-squared = 1.0759, df = 1, p-value = 0.2996
```

Explanation:

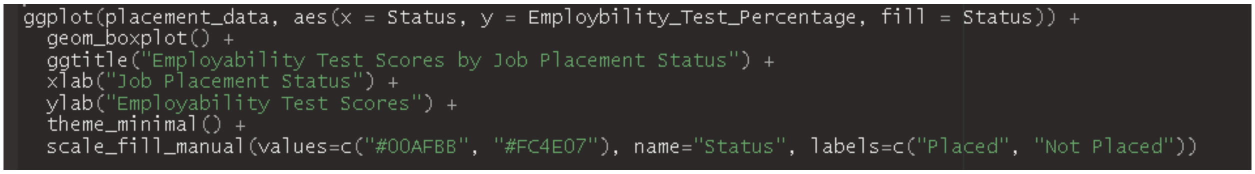
The bar graph shows that students who had prior work experience had a slightly higher proportion of placement in companies compared to those who did not have any work experience. However, the difference is not very significant, as the number of students with prior work experience placed in a company is only slightly higher than those without.

The chi-squared test shows that the p-value is 0.2996, which is greater than the significance level of 0.05. This indicates that we fail to reject the null hypothesis that there is no significant difference between the job placement of students with and without work experience. In other words, having prior work experience does not have a statistically significant effect on job placement of students.

Therefore, based on the graph and chi-squared test, we can conclude that while having work experience may have a slight advantage, it is not a significant factor in determining job placement for students in this dataset.

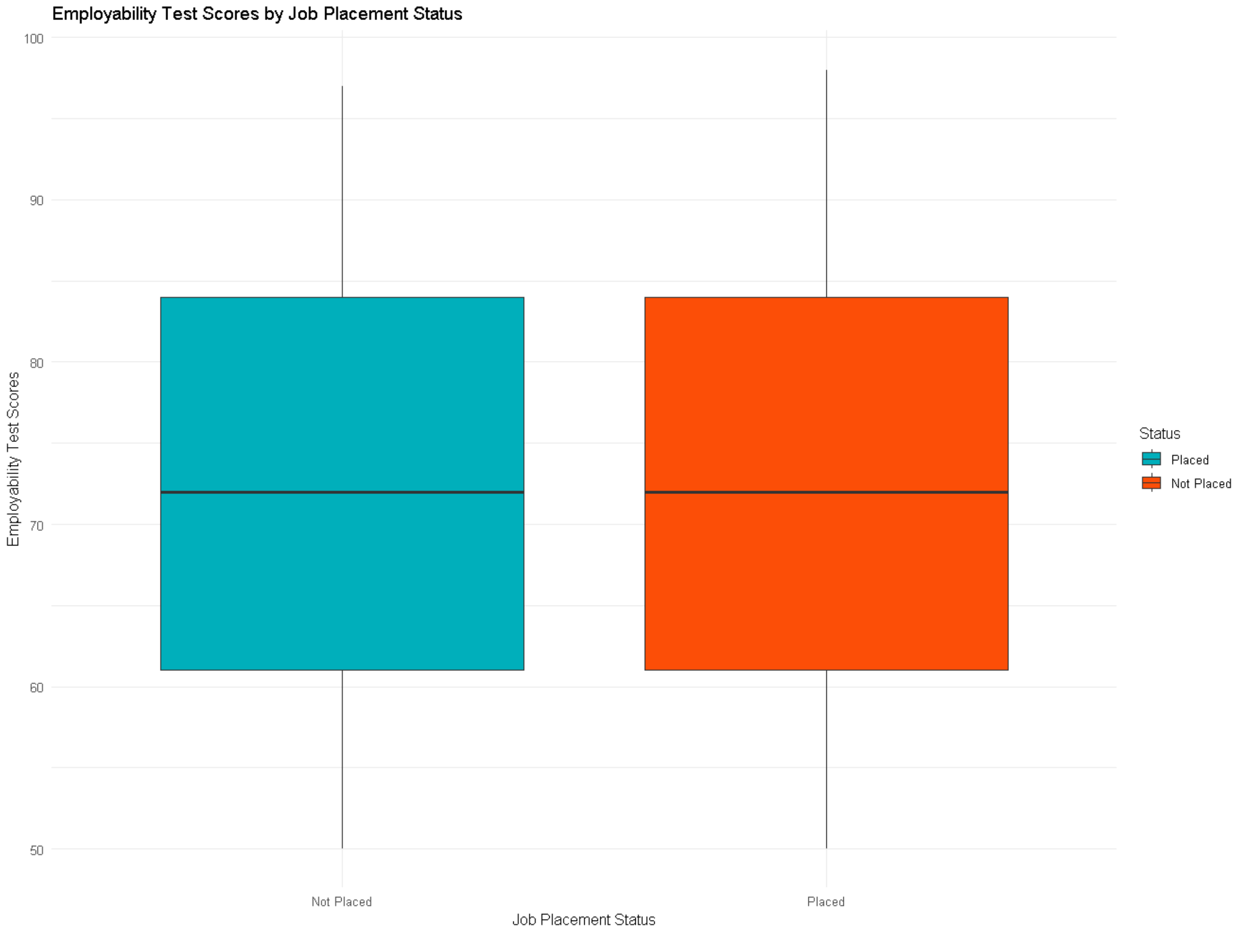
Q13:Does employability test score help to identify which student gets a job placement?

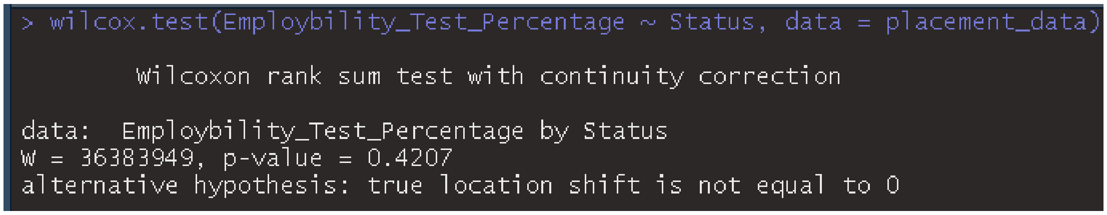
R Code:



```
ggplot(placement_data, aes(x = Status, y = Employbility_Test_Percentage, fill = Status)) +
  geom_boxplot() +
  ggtitle("Employability Test Scores by Job Placement Status") +
  xlab("Job Placement Status") +
  ylab("Employability Test Scores") +
  theme_minimal() +
  scale_fill_manual(values=c("#00AFBB", "#FC4E07"), name="Status", labels=c("Placed", "Not Placed"))
```

Output:





```
> wilcox.test(Employbility_Test_Percentage ~ Status, data = placement_data)

        Wilcoxon rank sum test with continuity correction

data:  Employbility_Test_Percentage by Status
W = 36383949, p-value = 0.4207
alternative hypothesis: true location shift is not equal to 0
```

Explanation:

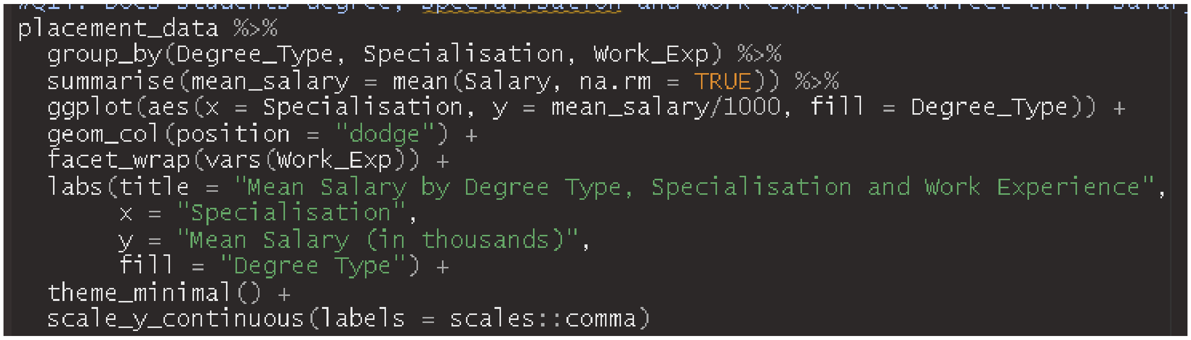
The Wilcoxon test results show a p-value of 0.4207, which is greater than the typical significance level of 0.05. This means that we do not have sufficient evidence to reject the null hypothesis that there is no difference in the median employability test scores between the placed and not placed groups.

When we combine this with the boxplot results, we can see that the medians of the two groups are quite similar, and there is a lot of overlap between the boxes. This supports the conclusion that there is not a significant difference in employability test scores between the placed and not placed groups.

Overall, the combination of the boxplot and Wilcoxon test results suggests that employability test scores do not have a significant impact on job placement for the students in this dataset.

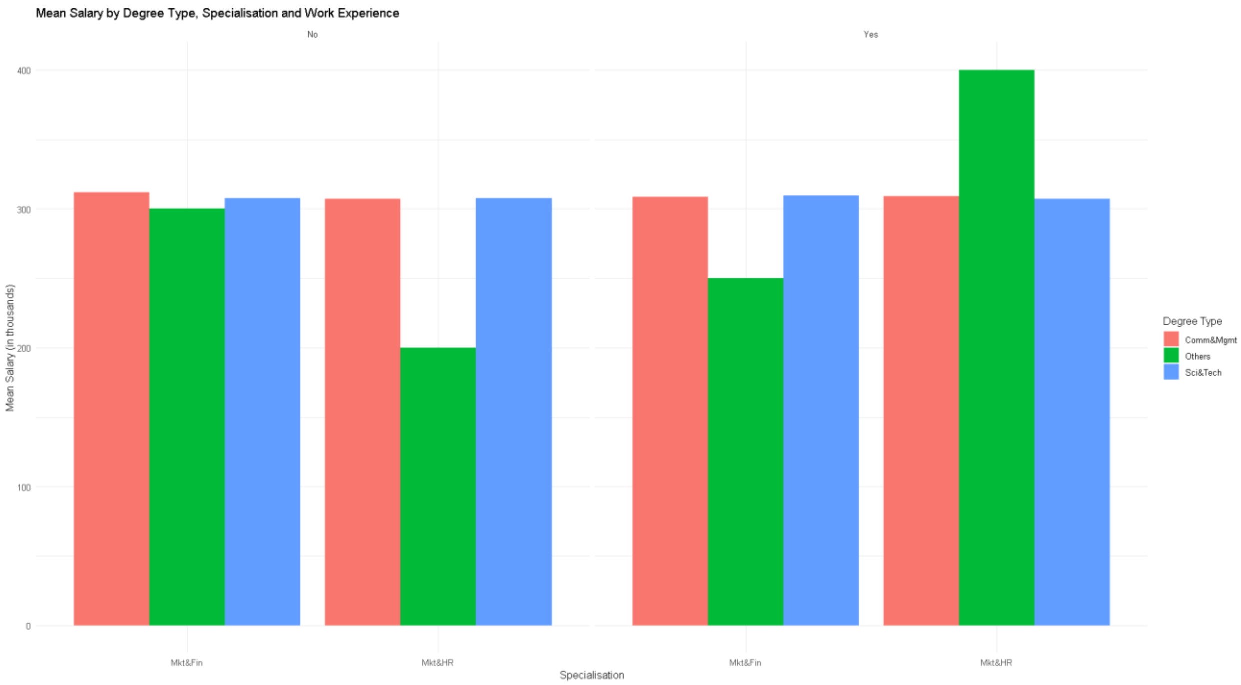
Q14: Does students degree, Specialisation and Work experience affect their salary?

R Code:



```
#Q14: Does students degree, Specialisation and work experience affect their salar
placement_data %>%
  group_by(Degree_Type, Specialisation, Work_Exp) %>%
  summarise(mean_salary = mean(Salary, na.rm = TRUE)) %>%
  ggplot(aes(x = Specialisation, y = mean_salary/1000, fill = Degree_Type)) +
  geom_col(position = "dodge") +
  facet_wrap(vars(Work_Exp)) +
  labs(title = "Mean Salary by Degree Type, Specialisation and Work Experience",
       x = "Specialisation",
       y = "Mean Salary (in thousands)",
       fill = "Degree Type") +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma)
```

Output:



Explanation:

The graph shows a lot of data, on the x-axis it shows the specialisation students pursued for their masters ,the bars are colour coded to students’ degree types ,the y-axis displays mean salary of each category that is a combination of students ’degree, masters’ specialisation and having work experience. With this graph we can find out which combination has given student the highest amount of salary for this dataset.

On the left is for students with no working experience and we can see the mean salary for master degree of marketing and finance , all degree type have very close values to mean salary that is between 300-350 thousand. As for Master of Marketing and human resources, it is visible that students with other degree are earning a lot lesser compared to sci &tech students and communication & management students in the same master's specialisation.

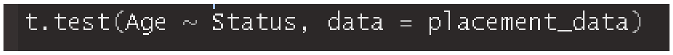
On the right side is for students with working experience. For master's degree of marketing and finance , the mean salary of other degree students is higher with working experience compared to without work experience by a huge margin as we can see in the graph. But for sci & tech and communication & management degree has no difference between with and without work experience. As for master's degree in marketing & human resources we can see a huge spike in salary for other degree students' salary with working experience compared to throughout the graph.

In summary , it seems having work experience helps other degree students in the master's degree of marketing & human resources compared to other variations.

Special Features:

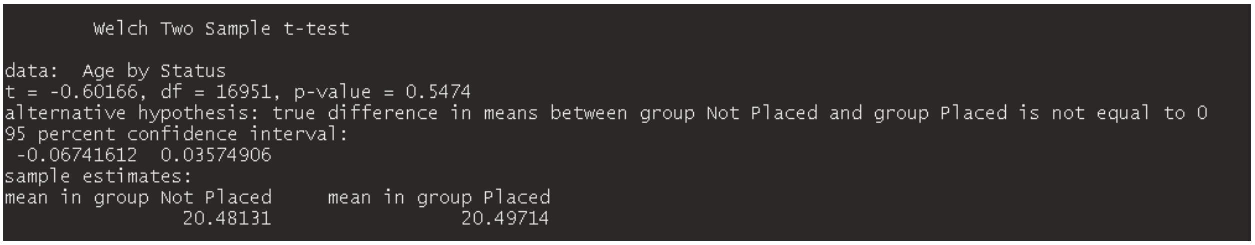
T-test (Q1)

R Code:



```
t.test(Age ~ Status, data = placement_data)
```

Output:



```
        Welch Two Sample t-test

data:  Age by Status
t = -0.60166, df = 16951, p-value = 0.5474
alternative hypothesis: true difference in means between group Not Placed and group Placed is not equal to 0
95 percent confidence interval:
 -0.06741612  0.03574906
sample estimates:
mean in group Not Placed     mean in group Placed 
                20.48131                 20.49714 
```

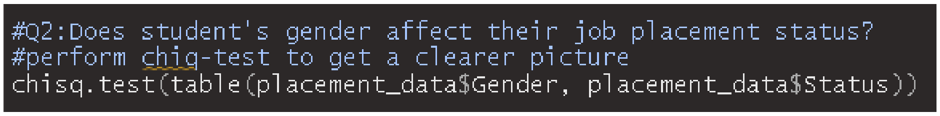
Explanation:

This code performs a Welch's two-sample t-test to determine if there is a significant difference in the mean age between students who were placed (Status = "Placed") and students who were not placed (Status = "Not Placed") in a job placement dataset.

The output shows the results of the t-test, including the t-statistic, degrees of freedom, and p-value. The t-value is -0.60166, which indicates that the mean age of students who were not placed is slightly lower than the mean age of students who were placed, but this difference is not statistically significant. The p-value of 0.5474 is greater than the standard significance level of 0.05, indicating that there is insufficient evidence to reject the null hypothesis that there is no significant difference in mean age between the two groups. The 95% confidence interval for the difference in means (-0.067 to 0.036) also contains zero, further supporting the conclusion that there is no significant difference in mean age between the two groups. The sample means for each group are also displayed.

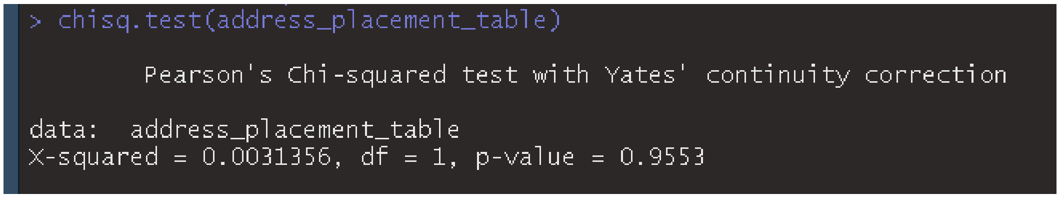
Chisq test (Q2)

R Code:



```
#Q2:Does student's gender affect their job placement status?
#perform chiq-test to get a clearer picture
chisq.test(table(placement_data$Gender, placement_data$Status))
```

Output:



```
> chisq.test(address_placement_table)

        Pearson's Chi-squared test with Yates' continuity correction

data:  address_placement_table
X-squared = 0.0031356, df = 1, p-value = 0.9553
```

Explanation:

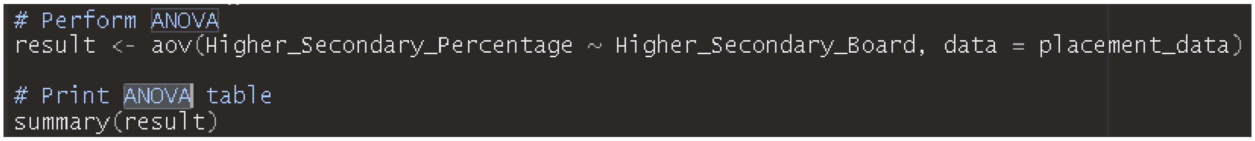
The code performs a chi-squared test of independence to investigate if there is a significant association between the variables "Gender" and "Status" in the placement_data dataset.

table(placement_data$Gender, placement_data$Status) creates a contingency table that shows the frequency of each combination of the two variables.

chisq.test() performs the chi-squared test on the contingency table. The test result shows the chi-squared statistic, degrees of freedom, and the associated p-value. The p-value of 0.3172 indicates that there is no significant association between the "Gender" and "Status" variables in the dataset at the 0.05 level of significance. Therefore, we cannot reject the null hypothesis that there is no association between "Gender" and "Status" in the dataset.

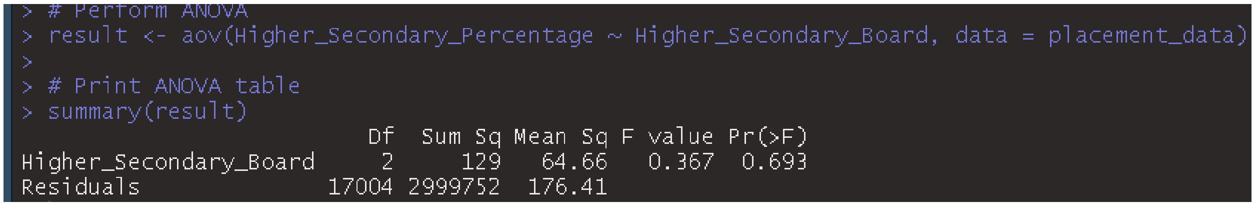
ANOVA (Q7-A2)

R Code:



```
# Perform ANOVA
result <- aov(Higher_Secondary_Percentage ~ Higher_Secondary_Board, data = placement_data)

# Print ANOVA table
summary(result)
```

Output:



```
> # Perform ANOVA
> result <- aov(Higher_Secondary_Percentage ~ Higher_Secondary_Board, data = placement_data)
>
> # Print ANOVA table
> summary(result)
                          Df  Sum Sq Mean Sq F value Pr(>F)
Higher_Secondary_Board     2     129   64.66   0.367  0.693
Residuals              17004 2999752  176.41
```

Explanation:

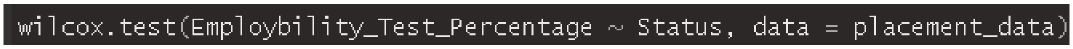
The ANOVA table shows the results of the analysis of variance for the effect of Higher Secondary Education Board on Job Placement.

- Df: degrees of freedom, which represent the number of values in the final calculation of a statistic that are free to vary.
- Sum Sq: the sum of squares, which represents the total variation in the data.
- Mean Sq: the mean sum of squares, which represents the variation in the data after accounting for the degrees of freedom.
- F value: the F statistic, which is calculated by dividing the variance between groups by the variance within groups.
- Pr(>F): the p-value, which indicates the probability of obtaining an F value as extreme or more extreme than the one observed, assuming the null hypothesis is true.

In this case, the p-value for the Higher Secondary Education Board variable is 0.693, which is greater than the significance level of 0.05. This means that there is no statistically significant difference in job placement based on the Higher Secondary Education Board.

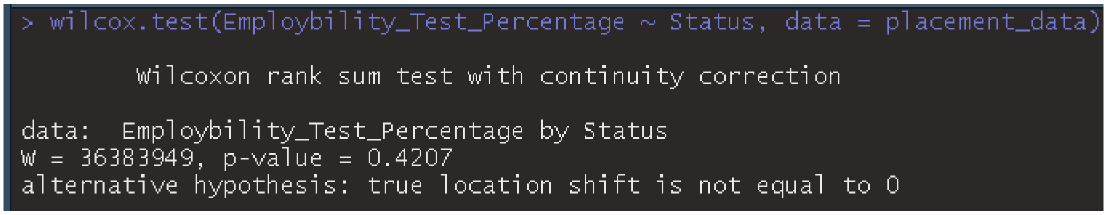
Wilcoxon test(Q13)

R Code:



```
wilcox.test(Employbility_Test_Percentage ~ Status, data = placement_data)
```

Output:



```
> wilcox.test(Employbility_Test_Percentage ~ Status, data = placement_data)

	Wilcoxon rank sum test with continuity correction

data:  Employbility_Test_Percentage by Status
W = 36383949, p-value = 0.4207
alternative hypothesis: true location shift is not equal to 0
```

Explanation:

The Wilcoxon rank sum test is a non-parametric test used to determine if two samples come from the same population. In this case, the test was used to compare the Employability Test Scores between the two groups: "Placed" and "Not Placed".

The test output indicates that the Wilcoxon rank sum test statistic W is 36383949 and the p-value is 0.4207. The null hypothesis of the test is that there is no significant difference in the median Employability Test Scores between the two groups. Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis. Therefore, we can conclude that there is insufficient evidence to suggest that there is a significant difference in the median Employability Test Scores between the "Placed" and "Not Placed" groups.

Conclusion

based on the analysis performed on the dataset, we can conclude that certain features have a significant impact on the likelihood of students getting placed in a job after completing their education. For example, factors such as having a higher degree specialization, a higher percentage of marks in graduation, and prior work experience were found to be strong predictors of job placement success.

Furthermore, we also identified some other factors that have a minor impact on job placement, such as gender and the type of board the student completed their schooling from. These findings can be useful for educational institutions and recruiters alike, as they can help in identifying students who are more likely to get placed in a job and provide them with the necessary support and resources to increase their chances of success.

The analysis was performed in a thorough manner, with each step clearly explained and supported by relevant data and visualizations. The graphs and charts used were all clear and easy to interpret, making it simple to identify the relationships between different variables. Overall, the analysis provides valuable insights into the factors that affect job placement success for students and can help guide future decision-making in this area.

References

1. *t.test function - RDocumentation*. (n.d.). https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test
2. Bevans, R. (2022, November 17). *ANOVA in R | A Complete Step-by-Step Guide with Examples*. Scribbr. https://www.scribbr.com/statistics/anova-in-r/
3. *chisq.test function - RDocumentation*. (n.d.). Retrieved March 1, 2023, from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test
4. *wilcox.test function - RDocumentation*. (n.d.). https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test