A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APD2F2111CS(CYB)

HAND OUT DATE: 6 DECEMBER 2021

HAND IN DATE: 31 JANUARY 2022

WEIGHTAGE: 50%

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 American Psychological Association (APA) 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.**

Name: Lee Jun Lam TP Number: TP055697 Intake Code: APD2F2111CS(CYB)

Table Of Contents

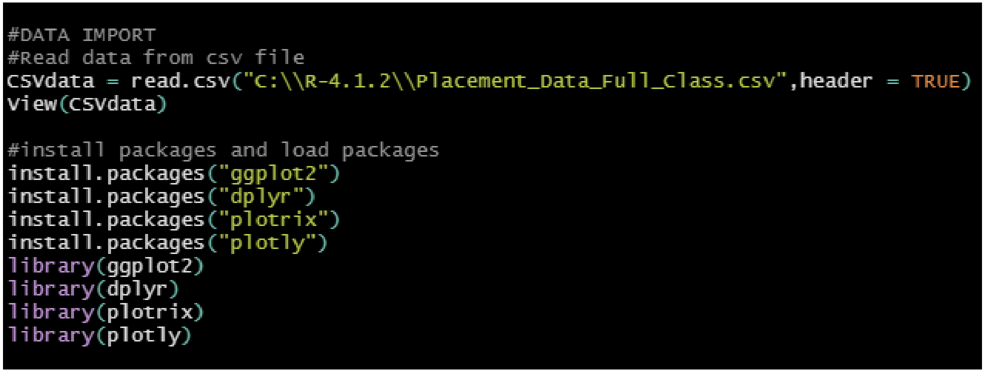
1.0 Introduction and Assumption

Introduction

In the educational system, the significance of overall student education performance (scores/grades) and working experience in obtaining student placements roles is extensively emphasized. MBA studies are commonly pursued after completing undergraduate education (Degree) in order to improve employability chances for students to enroll in the job after completing postsecondary education. Employability test is also conducted in college to evaluate students' own perceptions of their attitude, behavior, skills, and abilities in light of potential perceptions which could be made by interviewers. Therefore, Placement_Data_Full_Class.csv Dataset was assigned for the students to analyses and identify the hidden requirement and condition for students to receive placement after completing their studies. The dataset comprises personal information of the students such as personal data, education records, family background records, placement status, salary amount and so forth. From the dataset, various question analysis needs to be conduct and executed by applying the concept of R programming language and various techniques such as data exploration, data visualization, data manipulation and data transformation. As the dataset given is clean dataset, thus no data cleaning is executed.

2.0 Data Import/ Pre-Processing/ Exploration

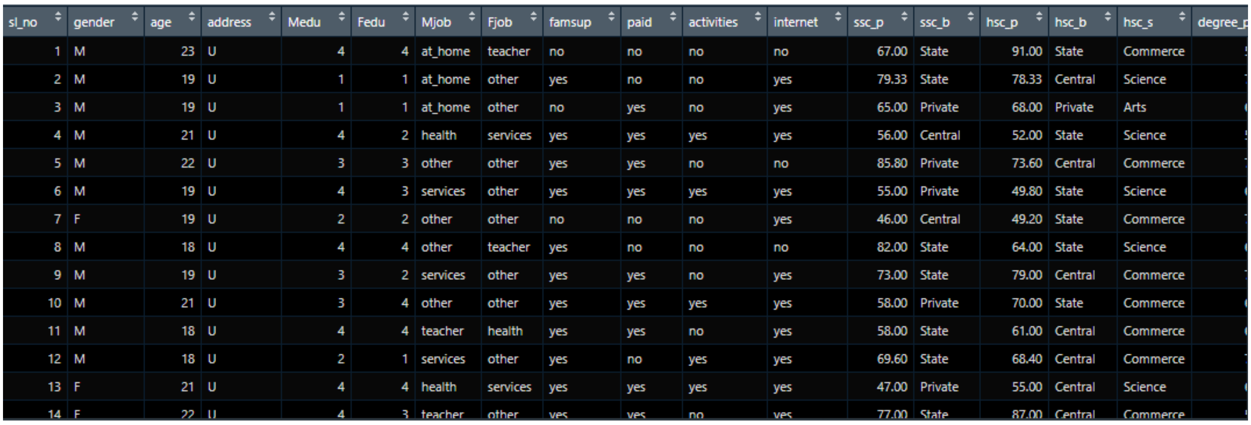
2.1 Data Import



```
#DATA IMPORT
#Read data from csv file
CSVdata = read.csv("C:\\R-4.1.2\\Placement_Data_Full_Class.csv",header = TRUE)
View(CSVdata)

#install packages and load packages
install.packages("ggplot2")
install.packages("dplyr")
install.packages("plotrix")
install.packages("plotly")
library(ggplot2)
library(dplyr)
library(plotrix)
library(plotly)
```

Figure 1 – Data Importing to RStudio

Data import is the first step needs to be executed to allow analyst to join the data generated in the R Studio with the data collected. Analyst can organize, analyze and act upon this unified data view in ways. The dataset given is in .csv format, thus applying read.csv () function reads a csv format file into the memory. In this scenario, 2 parameters were used in read.csv () function: The first parameter reads the csv file located in C:\\R-4.1.2\\Placement_Data_Full_Class.csv and assign it to the variable named CSVdata. The header line containing the names of the columns in the CSV file. This is specified by the second parameter header=TRUE. After that, using the View () to view the data import as new tab will display the content of data import as shown in Figure 2. It is also necessary to install packages to run certain function such as ggplot2, dplyr, plotrix and plotly. To load the packages, typing “library()” and insert the package name inside the bracket and run.

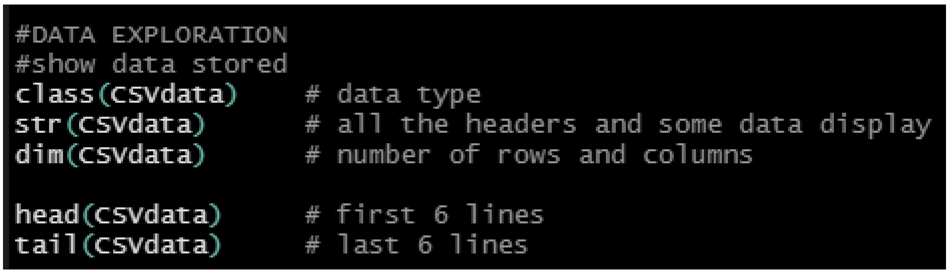


sl_no	gender	age	address	Medu	Fedu	Mjob	Fjob	famsup	paid	activities	internet	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_
1	M	23	U	4	4	at_home	teacher	no	no	no	no	67.00	State	91.00	State	Commerce	
2	M	19	U	1	1	at_home	other	yes	no	no	yes	79.33	State	78.33	Central	Science	
3	M	19	U	1	1	at_home	other	no	yes	no	yes	65.00	Private	68.00	Private	Arts	
4	M	21	U	4	2	health	services	yes	yes	yes	yes	56.00	Central	52.00	State	Science	
5	M	22	U	3	3	other	other	yes	yes	no	no	85.80	Private	73.60	Central	Commerce	
6	M	19	U	4	3	services	other	yes	yes	yes	yes	55.00	Private	49.80	State	Science	
7	F	19	U	2	2	other	other	no	no	no	yes	46.00	Central	49.20	State	Commerce	
8	M	18	U	4	4	other	teacher	yes	no	no	no	82.00	State	64.00	State	Science	
9	M	19	U	3	2	services	other	yes	yes	no	yes	73.00	State	79.00	Central	Commerce	
10	M	21	U	3	4	other	other	yes	yes	yes	yes	58.00	Private	70.00	State	Commerce	
11	M	18	U	4	4	teacher	health	yes	yes	no	yes	58.00	State	61.00	Central	Commerce	
12	M	18	U	2	1	services	other	yes	no	yes	yes	69.60	State	68.40	Central	Commerce	
13	F	21	U	4	4	health	services	yes	yes	yes	yes	47.00	Private	55.00	Central	Science	
14	F	22	U	4	3	teacher	other	yes	yes	no	yes	77.00	State	87.00	Central	Commerce	

Figure 2: The dataset has been imported successful and open in RStudio

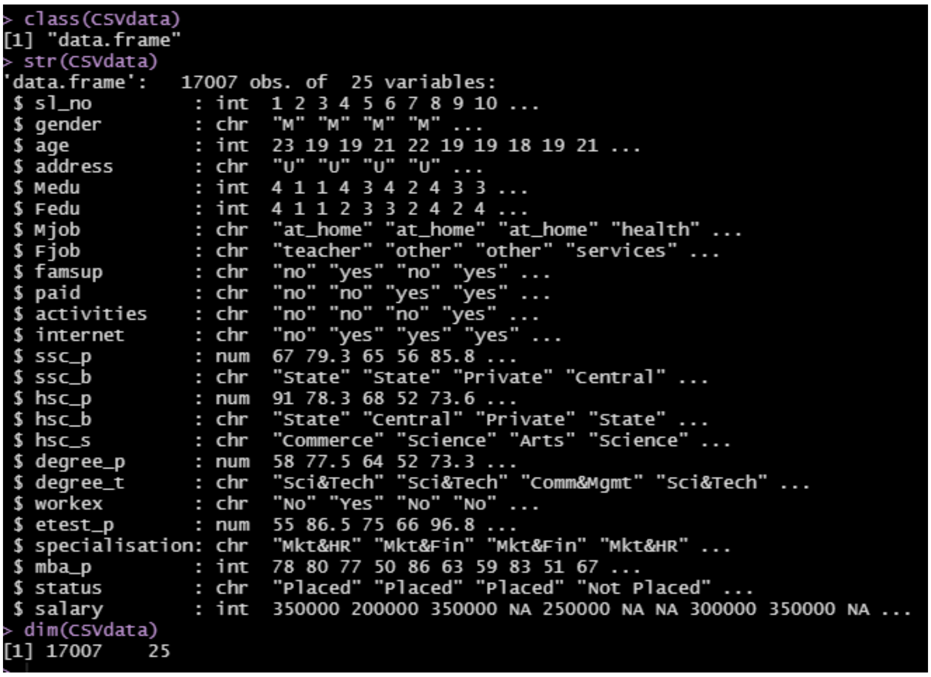
2.2 Data Exploration

Once the data has been successfully imported and open in RStudio, the second step is to conduct data exploration to investigate about general characteristics and potential problems of a data set without the need to formulate assumptions about the data beforehand. it is necessary to start exploring the attribute and understanding the content of the dataset. First of all, analyst should start exploring number of rows and columns and review first 6 line and last 6 line of data to ensure the data is importing accordingly. Then, analyst should understand class data type had been stored, structures, and column names.



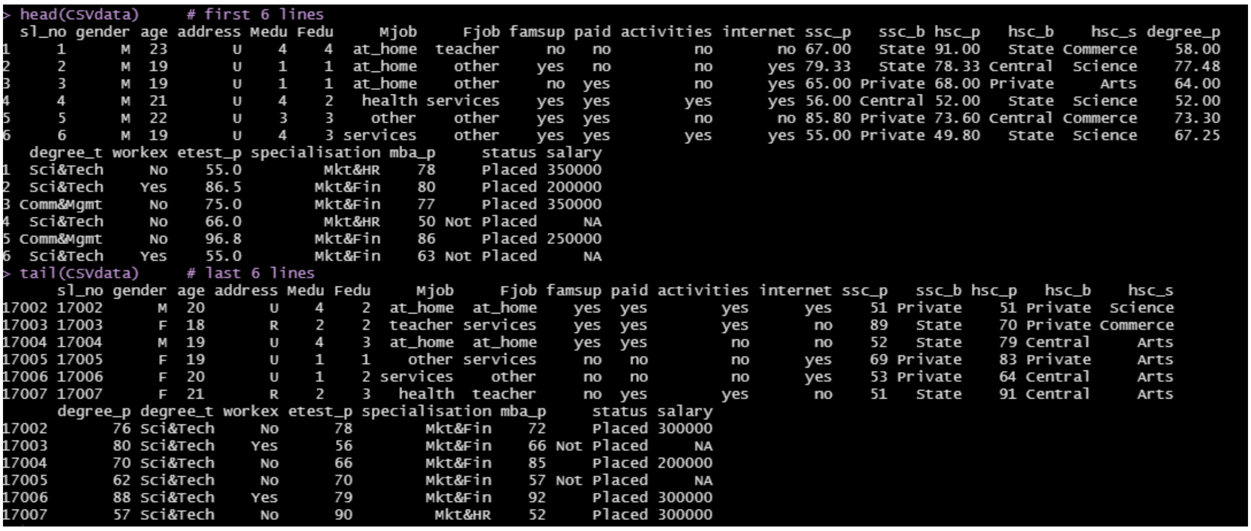
```
#DATA EXPLORATION
#show data stored
class(CSVdata)     # data type
str(CSVdata)       # all the headers and some data display
dim(CSVdata)       # number of rows and columns

head(CSVdata)      # first 6 lines
tail(CSVdata)      # last 6 lines
```

Figure 3: Example of Data Exploration applied



```
> class(CSVdata)
[1] "data.frame"
> str(CSVdata)
'data.frame':   17007 obs. of  25 variables:
 $ sl_no         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ gender        : chr  "M" "M" "M" "M" ...
 $ age           : int  23 19 19 21 22 19 19 18 19 21 ...
 $ address       : chr  "U" "U" "U" "U" ...
 $ Medu          : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu          : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob          : chr  "at_home" "at_home" "at_home" "health" ...
 $ Fjob          : chr  "teacher" "other" "other" "services" ...
 $ famsup        : chr  "no" "yes" "no" "yes" ...
 $ paid          : chr  "no" "no" "yes" "yes" ...
 $ activities    : chr  "no" "no" "no" "yes" ...
 $ internet      : chr  "no" "yes" "yes" "yes" ...
 $ ssc_p         : num  67 79.3 65 56 85.8 ...
 $ ssc_b         : chr  "State" "State" "Private" "Central" ...
 $ hsc_p         : num  91 78.3 68 52 73.6 ...
 $ hsc_b         : chr  "State" "Central" "Private" "State" ...
 $ hsc_s         : chr  "Commerce" "Science" "Arts" "Science" ...
 $ degree_p      : num  58 77.5 64 52 73.3 ...
 $ degree_t      : chr  "Sci&Tech" "Sci&Tech" "Comm&Mgmt" "Sci&Tech" ...
 $ workex        : chr  "No" "Yes" "No" "No" ...
 $ etest_p       : num  55 86.5 75 66 96.8 ...
 $ specialisation: chr  "Mkt&HR" "Mkt&Fin" "Mkt&Fin" "Mkt&HR" ...
 $ mba_p         : int  78 80 77 50 86 63 59 83 51 67 ...
 $ status        : chr  "Placed" "Placed" "Placed" "Not Placed" ...
 $ salary        : int  350000 200000 350000 NA 250000 NA NA 300000 350000 NA ...
> dim(CSVdata)
[1] 17007    25
```

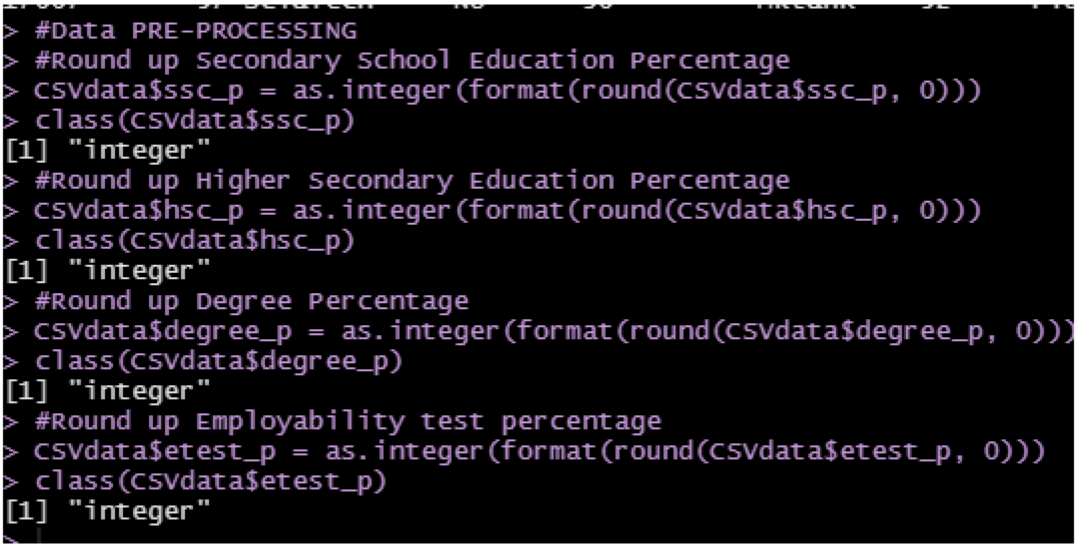
Figure 4: Output of class(CSVdata), str(CSVdata), dim(CSVdata)



```
> head(CSVdata)      # first 6 lines
  sl_no gender age address Medu Fedu     Mjob     Fjob famsup paid activities internet ssc_p   ssc_b hsc_p   hsc_b    hsc_s degree_p
1     1      M  23       U    4    4  at_home  teacher     no   no         no       no 67.00   State 91.00   State Commerce    58.00
2     2      M  19       U    1    1  at_home    other    yes   no         no      yes 79.33   State 78.33 Central  Science    77.48
3     3      M  19       U    1    1  at_home    other     no  yes         no      yes 65.00 Private 68.00 Private     Arts    64.00
4     4      M  21       U    4    2   health services    yes  yes        yes      yes 56.00 Central 52.00   State  Science    52.00
5     5      M  22       U    3    3    other    other    yes  yes         no       no 85.80 Private 73.60 Central Commerce    73.30
6     6      M  19       U    4    3 services    other    yes  yes        yes      yes 55.00 Private 49.80   State  Science    67.25
   degree_t workex etest_p specialisation mba_p     status salary
1  Sci&Tech     No    55.0         Mkt&HR    78     Placed 350000
2  Sci&Tech    Yes    86.5        Mkt&Fin    80     Placed 200000
3 Comm&Mgmt     No    75.0        Mkt&Fin    77     Placed 350000
4  Sci&Tech     No    66.0         Mkt&HR    50 Not Placed     NA
5 Comm&Mgmt     No    96.8        Mkt&Fin    86     Placed 250000
6  Sci&Tech    Yes    55.0        Mkt&Fin    63 Not Placed     NA
> tail(CSVdata)      # last 6 lines
      sl_no gender age address Medu Fedu     Mjob     Fjob famsup paid activities internet ssc_p   ssc_b hsc_p   hsc_b    hsc_s
17002 17002      M  20       U    4    2  at_home  at_home    yes  yes        yes      yes    51 Private    51 Private  Science
17003 17003      F  18       R    2    2  teacher services    yes  yes        yes       no    89   State    70 Private Commerce
17004 17004      M  19       U    4    3  at_home  at_home    yes  yes         no       no    52   State    79 Central     Arts
17005 17005      F  19       U    1    1    other services     no   no         no      yes    69 Private    83 Private     Arts
17006 17006      F  20       U    1    2 services    other     no   no         no      yes    53 Private    64 Central     Arts
17007 17007      F  21       R    2    3   health  teacher     no  yes        yes       no    51   State    91 Central     Arts
      degree_p degree_t workex etest_p specialisation mba_p     status salary
17002       76 Sci&Tech     No      78        Mkt&Fin    72     Placed 300000
17003       80 Sci&Tech    Yes      56        Mkt&Fin    66 Not Placed     NA
17004       70 Sci&Tech     No      66        Mkt&Fin    85     Placed 200000
17005       62 Sci&Tech     No      70        Mkt&Fin    57 Not Placed     NA
17006       88 Sci&Tech    Yes      79        Mkt&Fin    92     Placed 300000
17007       57 Sci&Tech     No      90         Mkt&HR    52     Placed 300000
```

Figure 5: Output of head(CSVdata) and tail(CSVdata)

2.3 Data Pre-Processing



```
> #Data PRE-PROCESSING
> #Round up Secondary School Education Percentage
> csvdata$ssc_p = as.integer(format(round(csvdata$ssc_p, 0)))
> class(csvdata$ssc_p)
[1] "integer"
> #Round up Higher Secondary Education Percentage
> csvdata$hsc_p = as.integer(format(round(csvdata$hsc_p, 0)))
> class(csvdata$hsc_p)
[1] "integer"
> #Round up Degree Percentage
> csvdata$degree_p = as.integer(format(round(csvdata$degree_p, 0)))
> class(csvdata$degree_p)
[1] "integer"
> #Round up Employability test percentage
> csvdata$etest_p = as.integer(format(round(csvdata$etest_p, 0)))
> class(csvdata$etest_p)
[1] "integer"
```

Figure 6: Data Preprocessing

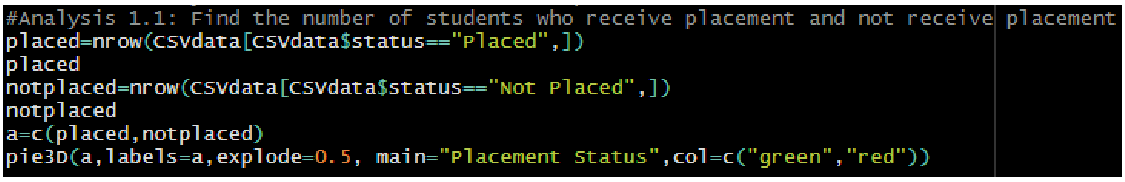
After review the dataset, certain data contains numeric data type (due to decimal), thus it is difficult to do analysis especially if bar chart is applied. Thus, data preprocessing is applied to round off the data and convert the data type to integer (Cottman, 2021).

3.0 Question and Analysis

Question 1: Why certain students can receive placement while certain students did not?

Analysis 1.1: Find the number of students who receive placement and not receive placement.

Source Code

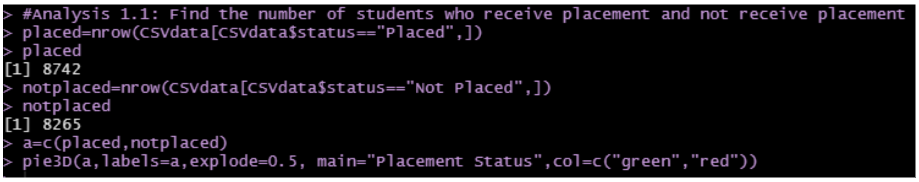


```
#Analysis 1.1: Find the number of students who receive placement and not receive placement
placed=nrow(CSVdata[CSVdata$status=="Placed",])
placed
notplaced=nrow(CSVdata[CSVdata$status=="Not Placed",])
notplaced
a=c(placed,notplaced)
pie3D(a,labels=a,explode=0.5, main="Placement Status",col=c("green","red"))
```

Figure 1.1.1: Creating Pie Chart to find the number of students who received placement and not receive placement.

The source code above uses data exploration to calculate number of students whose placement status = "Placed" and "Not Placed" and assign into "placed" and "notplaced" variables respectively. Then, the variables are combined in another variable named "a" to represent input values for the pie chart syntax. Lastly, a 3D pie chart is created by using pie3D syntax which is pie3D(values, labels, explodes, title, color).

Data Visualization



```
> #Analysis 1.1: Find the number of students who receive placement and not receive placement
> placed=nrow(CSVdata[CSVdata$status=="Placed",])
> placed
[1] 8742
> notplaced=nrow(CSVdata[CSVdata$status=="Not Placed",])
> notplaced
[1] 8265
> a=c(placed,notplaced)
> pie3D(a,labels=a,explode=0.5, main="Placement Status",col=c("green","red"))
```

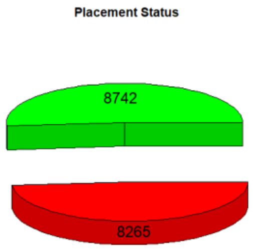


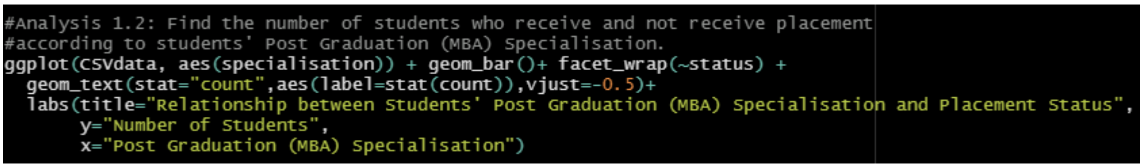
Figure 1.1.2: Number of students who received placement and not receive placement +

Figure 1.1.3: 3D Pie Chart

Based on Figure 1.1.2, the number of students who received placement is 8742 while the number of students who does not receive placement is 8265 and the result is also shown in the pie chart which is Figure 1.1.3. Thus, **51.40%** of the students received placement (8742/17007 * 100 = 51.40%).

Analysis 1.2: Find the number of students who receive and not receive placement according to students' Post Graduation (MBA) Specialisation.

Source Code



```
#Analysis 1.2: Find the number of students who receive and not receive placement
#according to students' Post Graduation (MBA) Specialisation.
ggplot(CSVdata, aes(specialisation)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Post Graduation (MBA) Specialisation and Placement Status",
       y="Number of Students",
       x="Post Graduation (MBA) Specialisation")
```

Figure 1.2.1: Using ggplot to create bar chart graph

The source code above is applied with uses ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between students' post-graduation (MBA) specialisation and placement status. The x-axis displays the post-graduation (MBA) specialisation while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart. By using bar chart, it is easier to identify which specialisation has more students received placement.

Data Visualization

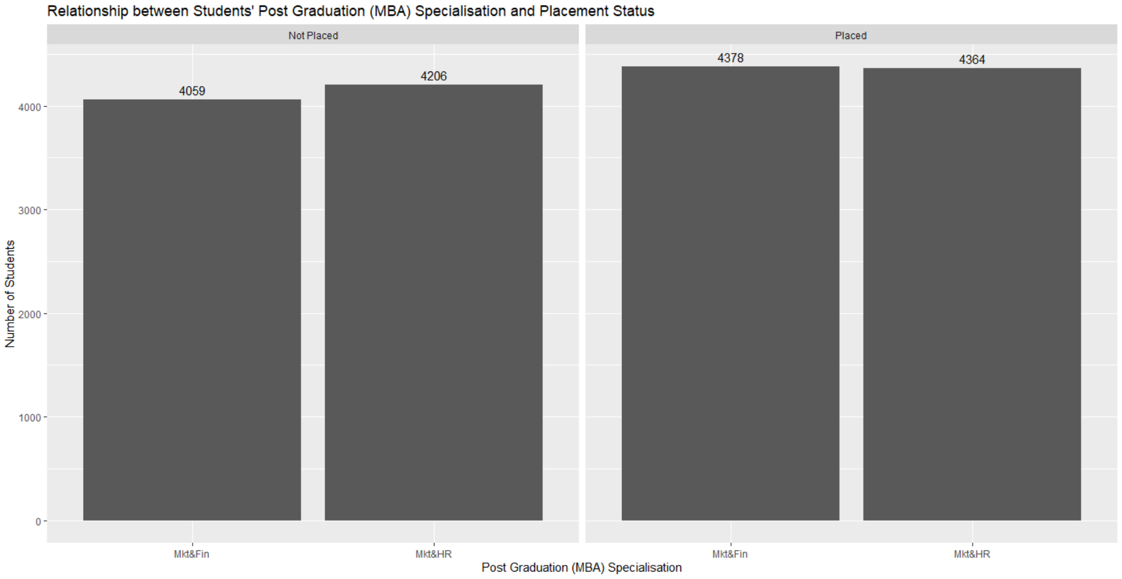


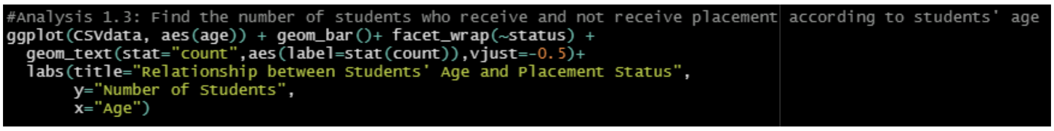
Figure 1.2.2: Relationship between Students' Post Graduation (MBA) Specialisation and Placement Status

Based on the bar chart, the number of students who studies in Mkt&Fin are (4378 + 4059 = **8437**) and 4378 students receive placement (4378/8437 * 100 = **51.89%**), while number of students who studies in Mkt&HR are (4364 + 4206 = **8570**) and 4364 students receive placement (4364/8570 * 100 = **50.92%**).

In conclusion, student who studied Mkt&Fin has a slightly higher percentage receive placement than student who studies Mkt&HR. However, more analysis needs to be conducted first before conclude the relationship beween MBA specialization and placement status.

Analysis 1.3: Find the number of students who receive and not receive placement according to students' age.

Source Code



```
#Analysis 1.3: Find the number of students who receive and not receive placement according to students' age
ggplot(CSVdata, aes(age)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Age and Placement Status",
       y="Number of Students",
       x="Age")
```

Figure 1.3.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between students' age and placement status. The x-axis displays the age while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart. By using bar chart, it is easier to identify which age has more students received placement.

Data Visualization

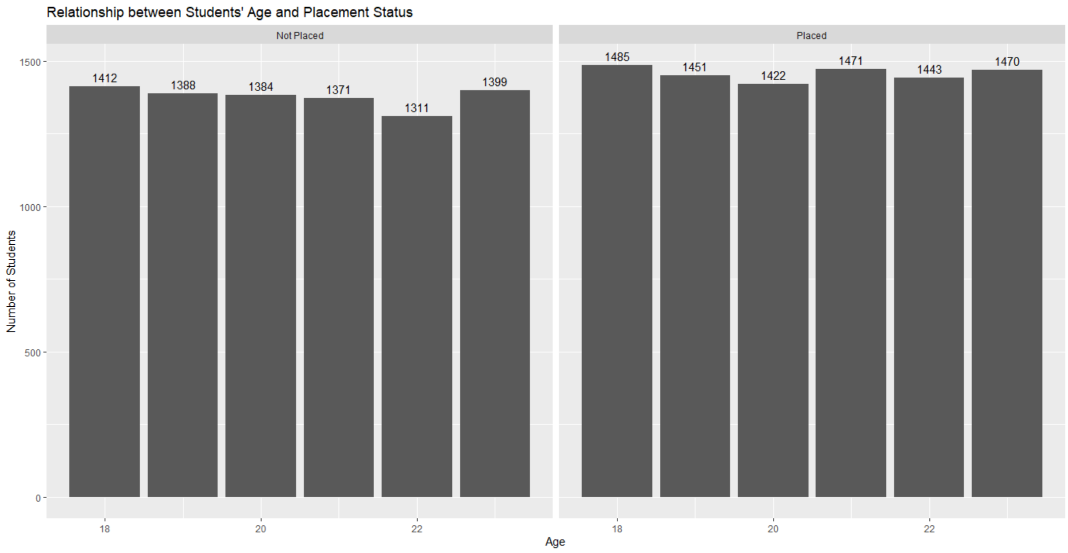


Figure 1.3.2: Relationship between Students' Age and Placement Status

Based on the bar chart, the total number of students according to the age which are: ~

age 18 (1485 + 1412 = **2897**), age 19 (1451 + 1388 = **2839**), age 20 (1422 + 1384 = **2806**), age 21 (1471 + 1371 = **2842**), age 22 (1443 + 1311 = **2754**), age 23 (1470 + 1399 = **2869**)

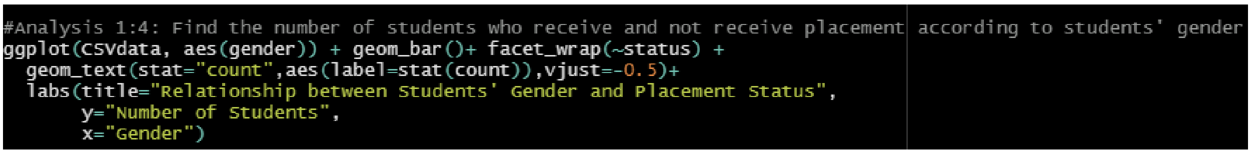
The percentage of students who receive placement according to the age which are: ~

age 18 (1485/ 2897 * 100 = **51.26%**), age 19 (1451/ 2839 * 100 = **51.11%**), age 20 (1422/ 2806 * 100 = **50.68%**), age 21 (1471/ 2842 * 100 = **51.76%**), age 22 (1443/ 2754 * 100 = **52.40%**), age 23 (1470/ 2869 * 100 = **51.24%**)

In conclusion, the percentage of receiving placement is relatively close, thus any of student with any age >= 18 and <= 23 has a fair chance to receive placement.

Analysis 1.4: Find the number of students who receive and not receive placement according to students' gender.

Source Code



```
#Analysis 1:4: Find the number of students who receive and not receive placement according to students' gender
ggplot(CSVdata, aes(gender)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Gender and Placement Status",
       y="Number of Students",
       x="Gender")
```

Figure 1.4.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between students' gender and placement status. The x-axis displays the gender while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart. By using bar chart, it is easier to identify which gender has more students received placement.

Data Visualization

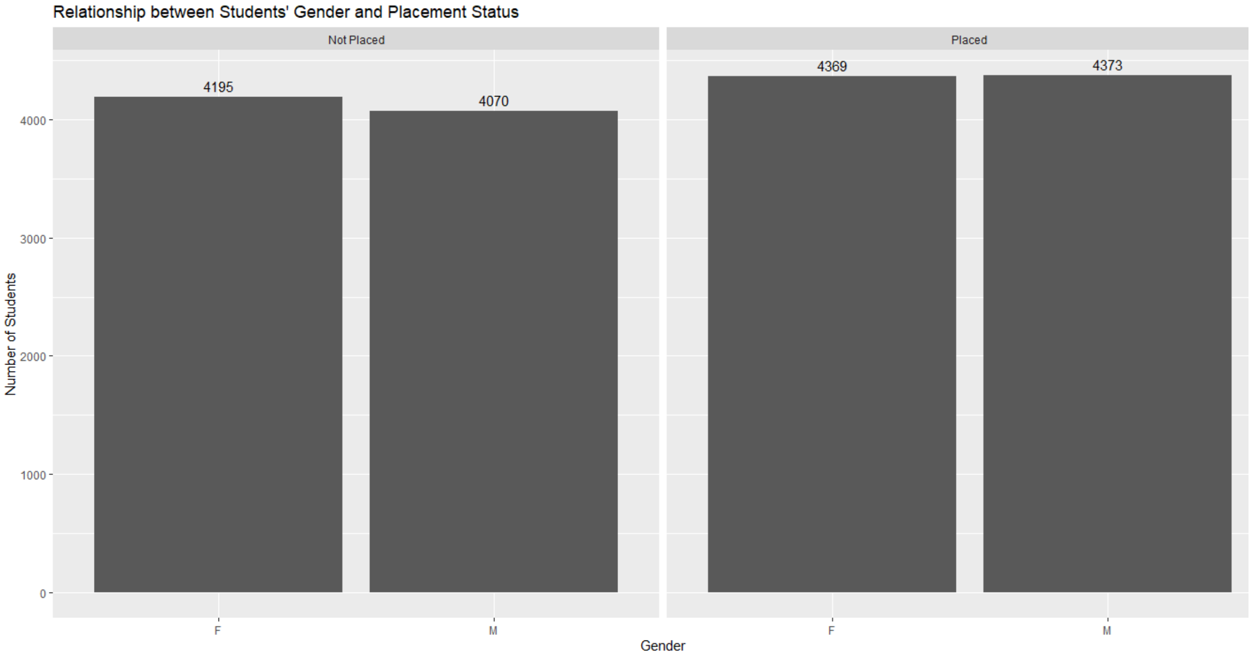


Figure 1.4.2: Relationship between Students' Gender and Placement Status

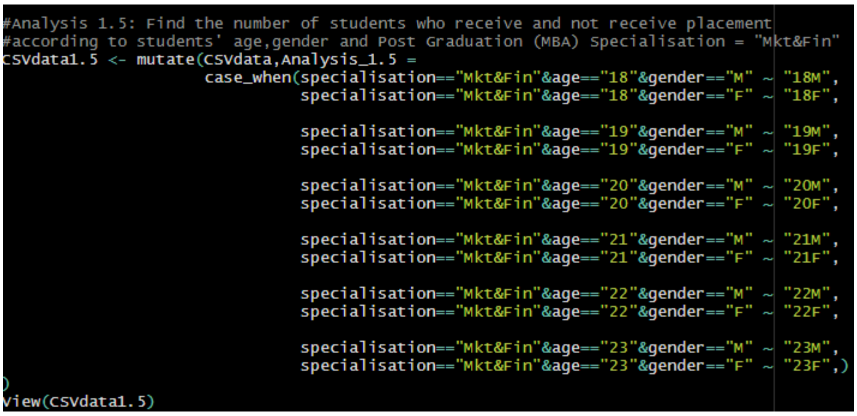
Based on the bar chart, the total number of students who are gender "F" are (4369 + 4195 = **8564**) and 4369 students receive placement (4369/8564 * 100 = **51.02%**), while number of students who are gender "M" are (4373 + 4070 = **8443**) and 4373 students receive placement (4373/8443 * 100 = **51.79%**).

In conclusion, the percentage of receiving placement according to two different gender is relatively close to each other, thus any gender of student has a fair chance to receive placement.

Analysis 1.5: Find the number of students who receive and not receive placement according to students' age, gender and Post Graduation (MBA) Specialisation = "Mkt&Fin".

Source Code

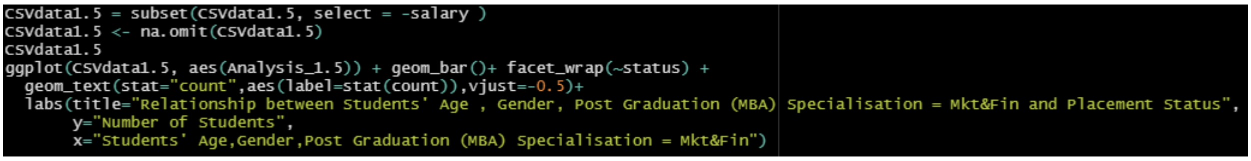
Data Manipulation and Transformation



```
#Analysis 1.5: Find the number of students who receive and not receive placement
#according to students' age,gender and Post Graduation (MBA) Specialisation = "Mkt&Fin"
CSVdata1.5 <- mutate(CSVdata,Analysis_1.5 =
                     case_when(specialisation=="Mkt&Fin"&age=="18"&gender=="M" ~ "18M",
                               specialisation=="Mkt&Fin"&age=="18"&gender=="F" ~ "18F",

                               specialisation=="Mkt&Fin"&age=="19"&gender=="M" ~ "19M",
                               specialisation=="Mkt&Fin"&age=="19"&gender=="F" ~ "19F",

                               specialisation=="Mkt&Fin"&age=="20"&gender=="M" ~ "20M",
                               specialisation=="Mkt&Fin"&age=="20"&gender=="F" ~ "20F",

                               specialisation=="Mkt&Fin"&age=="21"&gender=="M" ~ "21M",
                               specialisation=="Mkt&Fin"&age=="21"&gender=="F" ~ "21F",

                               specialisation=="Mkt&Fin"&age=="22"&gender=="M" ~ "22M",
                               specialisation=="Mkt&Fin"&age=="22"&gender=="F" ~ "22F",

                               specialisation=="Mkt&Fin"&age=="23"&gender=="M" ~ "23M",
                               specialisation=="Mkt&Fin"&age=="23"&gender=="F" ~ "23F",)
)
View(CSVdata1.5)
```

Figure 1.5.1: Step 1 – Duplicate data frame and create new column by using mutate and case_when function

Due to all students take MBA studies regardless of ages and gender, thus this analysis is conducted to investigate which age and gender that studies MBA Specialisation = “Mkt&Fin” has a higher chance receive placement. Thus, the first step is applying the concept of data manipulation and transformation to duplicate a new data frame (CSVdata1.5) to avoid alter the original data frame and create new column (Analysis_1.5) by using mutate function which combine students’ age and gender together who studies MBA Specialisation = “Mkt&Fin”. The case_when function has been used inside the mutate function to implement conditional logic like if/else and if/else if/else.



```
CSVdata1.5 = subset(CSVdata1.5, select = -salary )
CSVdata1.5 <- na.omit(CSVdata1.5)
CSVdata1.5
ggplot(CSVdata1.5, aes(Analysis_1.5)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Age , Gender, Post Graduation (MBA) Specialisation = Mkt&Fin and Placement Status",
       y="Number of Students",
       x="Students' Age,Gender,Post Graduation (MBA) Specialisation = Mkt&Fin")
```

Figure 1.5.2: Step 2 – Remove salary column, purge any row with N/A data and create bar chart by using ggplot function

After new column has been create in the duplicate data frame, the salary column has been remove through subset function before purge any row that contains N/A value by using na.omit function as all the rows which contains “Mkt&Fin” needs to be reserve to create bar chart. Then ggplot function is applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_1.5) and placement status. The x-axis displays the data in the new column (combination of student age and gender where Post Graduation (MBA) specialization = “Mkt&Fin” while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

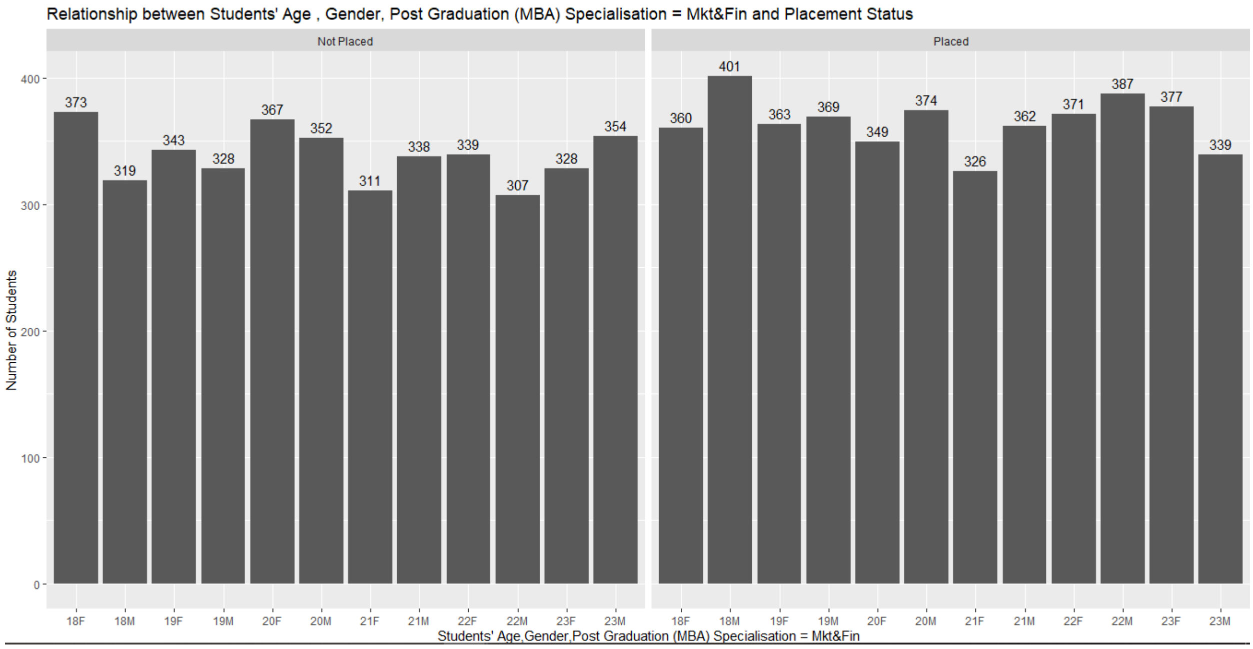


Figure 1.5.3: Relationship between Students' Age, Gender, Post-Graduation (MBA) Specialisation = "Mkt&Fin" and Placement Status

Based on the bar chart, the total number of students according to the age and gender which are: ~

18F (360 + 373 = **733**), 18M (401 + 319 = **720**)

19F (363 + 343 = **706**), 19M (369 + 328 = **697**)

20F (349 + 367 = **716**), 20M (374 + 352 = **726**)

21F (326 + 311 = **637**), 21M (362 + 338 = **700**)

22F (371 + 339 = **710**), 22M (387 + 307 = **694**)

23F (377 + 328 = **705**), 23M (339 + 354 = **693**)

Total F: **4207** Total M: **4230** Total Student: **8437**

The percentage of students who receive placement according to the age and gender which are: ~

18F (360/733 * 100 = **49.11%**), 18M (401/720 * 100 = **55.69%**)

19F (363/706 * 100 = **51.42%**), 19M (369/697 * 100 = **52.94%**)

20F (349/716 * 100 = **48.74%**), 20M (374/726 * 100 = **51.52%**)

21F (326/637 * 100 = **51.18%**), 21M (362/700 * 100 = **51.71%**)

22F (371/710 * 100 = **52.25%**), 22M (387/694 * 100 = **55.76%**)

23F (377/705 * 100 = **53.48%**), 23M (339/693 * 100 = **48.92%**)

Average percentage F: **51.03%** Average percentage M: **52.76%**

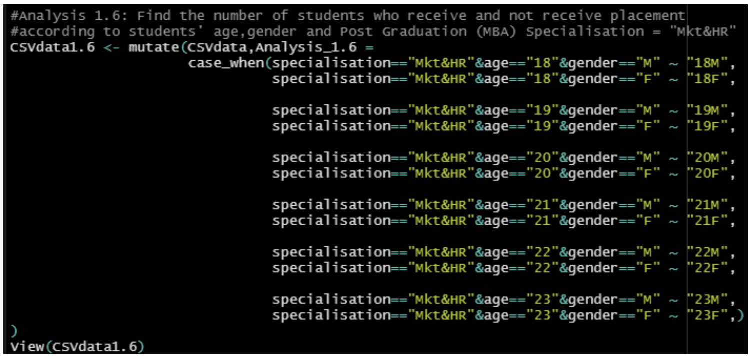
In conclusion, the average percentage between female and male student who receive placement are quite close, each student has rather fair chance in receiving placement if compare with their own gender. Meanwhile for age, certain age receive slightly lower percent is perhaps due to other factor causes, thus more analysis is required.

If compare male and female, male student has slightly higher chance receiving placement due to total number of male students studied in specialisation = “Mkt&Fin” is slightly higher.

Analysis 1.6: Find the number of students who receive and not receive placement according to students' age, gender and Post Graduation (MBA) Specialisation = "Mkt&HR".

Source Code

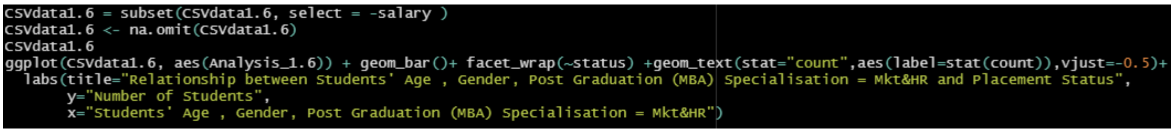
Data Manipulation and Transformation



```
#Analysis 1.6: Find the number of students who receive and not receive placement
#according to students' age,gender and Post Graduation (MBA) Specialisation = "Mkt&HR"
CSVdata1.6 <- mutate(CSVdata,Analysis_1.6 =
                       case_when(specialisation=="Mkt&HR"&age=="18"&gender=="M" ~ "18M",
                                 specialisation=="Mkt&HR"&age=="18"&gender=="F" ~ "18F",

                                 specialisation=="Mkt&HR"&age=="19"&gender=="M" ~ "19M",
                                 specialisation=="Mkt&HR"&age=="19"&gender=="F" ~ "19F",

                                 specialisation=="Mkt&HR"&age=="20"&gender=="M" ~ "20M",
                                 specialisation=="Mkt&HR"&age=="20"&gender=="F" ~ "20F",

                                 specialisation=="Mkt&HR"&age=="21"&gender=="M" ~ "21M",
                                 specialisation=="Mkt&HR"&age=="21"&gender=="F" ~ "21F",

                                 specialisation=="Mkt&HR"&age=="22"&gender=="M" ~ "22M",
                                 specialisation=="Mkt&HR"&age=="22"&gender=="F" ~ "22F",

                                 specialisation=="Mkt&HR"&age=="23"&gender=="M" ~ "23M",
                                 specialisation=="Mkt&HR"&age=="23"&gender=="F" ~ "23F",)
)
View(CSVdata1.6)
```

Figure 1.6.1: Step 1 – Duplicate data frame and create new column by using mutate and case_when function

Similar to Analysis 1.5, due to all students take MBA studies regardless of ages and gender, thus this analysis is conducted to investigate which age and gender that studies MBA Specialisation = “Mkt&HR” has a higher chance receive placement. Thus, the first step is applying the concept of data manipulation and transformation to duplicate a new data frame (CSVdata1.6) to avoid alter the original data frame and create new column (Analysis_1.6) by using mutate function which combine students’ age and gender together who studies MBA Specialisation = “Mkt&HR”. The case_when function has been used inside the mutate function to implement conditional logic like if/else and if/else if/else.



```
CSVdata1.6 = subset(CSVdata1.6, select = -salary )
CSVdata1.6 <- na.omit(CSVdata1.6)
CSVdata1.6
ggplot(CSVdata1.6, aes(Analysis_1.6)) + geom_bar()+ facet_wrap(~status) +geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Age , Gender, Post Graduation (MBA) Specialisation = Mkt&HR and Placement Status",
       y="Number of Students",
       x="Students' Age , Gender, Post Graduation (MBA) Specialisation = Mkt&HR")
```

Figure 1.6.2: Step 2 – Remove salary column, purge any row with N/A data and create bar chart by using ggplot function

After new column has been create in the duplicate data frame, the salary column has been remove through subset function before purge any row that contains N/A value by using na.omit function as all the rows which contains “Mkt&HR” needs to be reserve to create bar chart. Then ggplot function is applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_1.6) and placement status. The x-axis displays the data in the new column (combination of student age and gender where Post Graduation (MBA) specialization = “Mkt&HR” while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

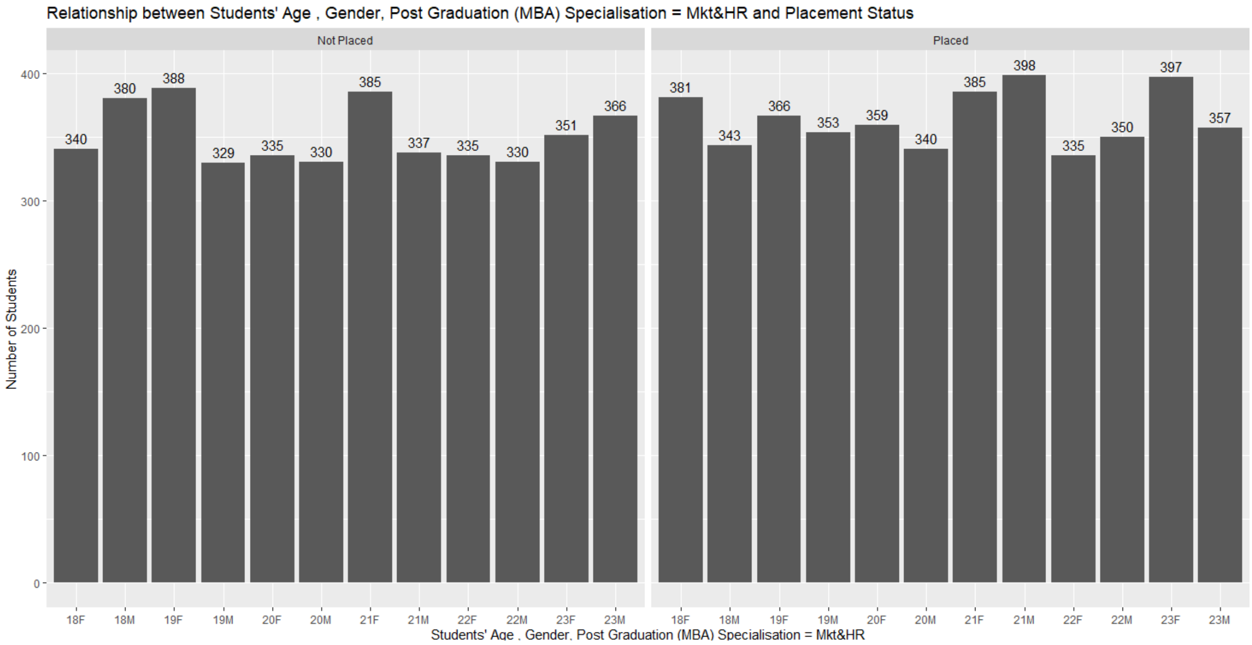


Figure 1.6.3: Relationship between Students' Age, Gender, Post-Graduation (MBA) Specialisation = "Mkt&HR" and Placement Status

Based on the bar chart, the number of students according to the age and gender which are: ~

18F (381 + 340 = **721**), 18M (343 + 380 = **723**)

19F (366 + 388 = **754**), 19M (353 + 329 = **682**)

20F (359 + 335 = **694**), 20M (340 + 330 = **670**)

21F (385 + 385 = **770**), 21M (398 + 337 = **735**)

22F (335 + 335 = **670**), 22M (350 + 330 = **680**)

23F (397 + 351 = **748**), 23M (357 + 366 = **723**)

Total F: **4357** Total M: **4213** Total Student: **8570**

The percentage of students who receive placement according to the age and gender which are: ~

18F (381/721 * 100 = **52.84%**), 18M (343/723 * 100 = **47.44%**)

19F (366/754 * 100 = **48.54%**), 19M (353/682 * 100 = **51.76%**)

20F (359/694 * 100 = **51.73%**), 20M (340/670 * 100 = **50.75%**)

21F (385/770 * 100 = **50.00%**), 21M (398/735 * 100 = **54.15%**)

22F (335/670 * 100 = **50.00%**), 22M (350/680 * 100 = **51.47%**)

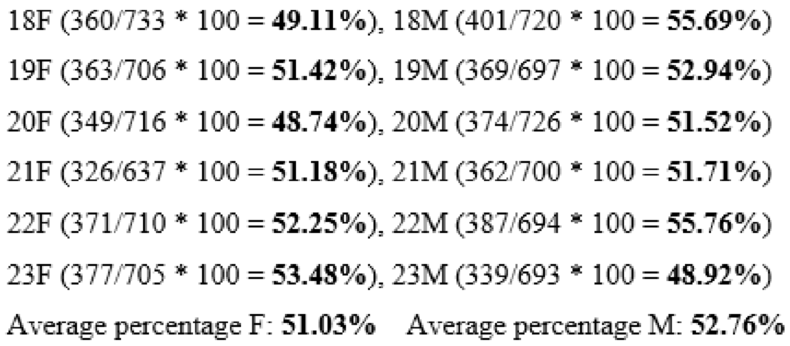
23F (397/748 * 100 = **53.07%**), 23M (357/723 * 100 = **49.38%**)

Average percentage F: **51.03%** Average percentage M: **50.83%**

In conclusion, the average percentage between female and male student who receive placement are quite close, each student has rather fair chance in receiving placement if compare with their own gender. Meanwhile for age, certain age receive slightly lower percent is perhaps due to other factor causes, thus more analysis is required.

Meanwhile in age, female student has slightly higher chance receiving placement due to total number of female students studied in specialisation = “Mkt&HR” is slightly higher.

Conclusion for Question 1



18F (360/733 * 100 = **49.11%**), 18M (401/720 * 100 = **55.69%**)

19F (363/706 * 100 = **51.42%**), 19M (369/697 * 100 = **52.94%**)

20F (349/716 * 100 = **48.74%**), 20M (374/726 * 100 = **51.52%**)

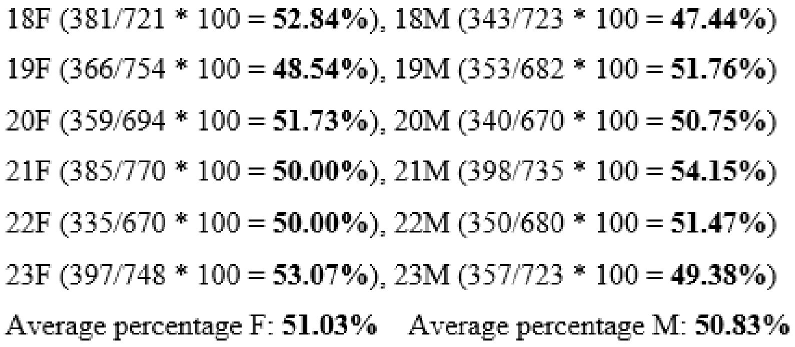
21F (326/637 * 100 = **51.18%**), 21M (362/700 * 100 = **51.71%**)

22F (371/710 * 100 = **52.25%**), 22M (387/694 * 100 = **55.76%**)

23F (377/705 * 100 = **53.48%**), 23M (339/693 * 100 = **48.92%**)

Average percentage F: **51.03%** Average percentage M: **52.76%**

Figure 1.7.1: Screenshot of Percentage calculation in Analysis 1.5



18F (381/721 * 100 = **52.84%**), 18M (343/723 * 100 = **47.44%**)

19F (366/754 * 100 = **48.54%**), 19M (353/682 * 100 = **51.76%**)

20F (359/694 * 100 = **51.73%**), 20M (340/670 * 100 = **50.75%**)

21F (385/770 * 100 = **50.00%**), 21M (398/735 * 100 = **54.15%**)

22F (335/670 * 100 = **50.00%**), 22M (350/680 * 100 = **51.47%**)

23F (397/748 * 100 = **53.07%**), 23M (357/723 * 100 = **49.38%**)

Average percentage F: **51.03%** Average percentage M: **50.83%**

Figure 1.7.2: Screenshot of Percentage calculation in Analysis 1.6

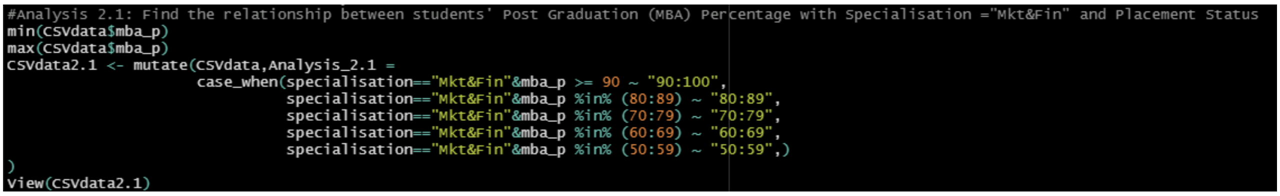
Based on the figure 1.7.1 and 1.7.2 as shown above, the average percentage of female student and male student that receive placement is rather close to each other regardless any MBA specialization is being chosen and age. In certain age male student has slightly higher chance receive placement than female student and vice versa due to other factors causes.

After conducting various analysis in Question 1, the result has shown that each student has fair chance to receive placement regardless of MBA specialization that chosen, gender and age as the percentage of receiving placement is very close to each other as shown in analysis 1.2, analysis 1.3 and analysis 1.4. The factor is also combined in analysis 1.5 and analysis 1.6 to conduct further analysis and the percentage of receiving placement is very close to each other. Thus, more question and analysis need to be conducted first before conclude the factors in Question 1.

Question 2: Does Student's University Education Records Affect Student's Placement Status?

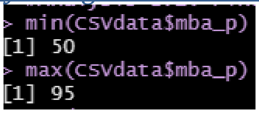
Analysis 2.1: Find the relationship between students' Post Graduation (MBA) Percentage with Specialisation ="Mkt&Fin" and Placement Status.

Source Code



```
#Analysis 2.1: Find the relationship between students' Post Graduation (MBA) Percentage with Specialisation ="Mkt&Fin" and Placement Status
min(CSVdata$mba_p)
max(CSVdata$mba_p)
CSVdata2.1 <- mutate(CSVdata,Analysis_2.1 =
                     case_when(specialisation=="Mkt&Fin"&mba_p >= 90 ~ "90:100",
                               specialisation=="Mkt&Fin"&mba_p %in% (80:89) ~ "80:89",
                               specialisation=="Mkt&Fin"&mba_p %in% (70:79) ~ "70:79",
                               specialisation=="Mkt&Fin"&mba_p %in% (60:69) ~ "60:69",
                               specialisation=="Mkt&Fin"&mba_p %in% (50:59) ~ "50:59",)
)
View(CSVdata2.1)
```

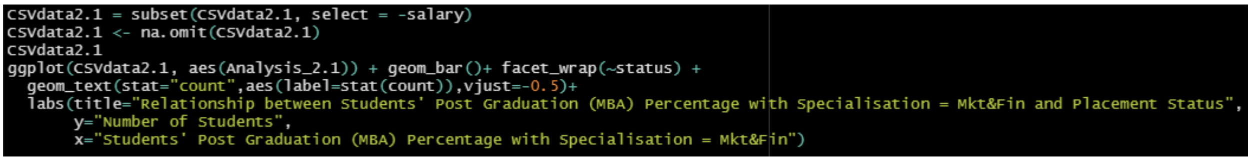
Figure 2.1.1: Step 1: Identify lowest and highest MBA percentage, duplicate data frame and create new column by using mutate and case_when function



```
> min(CSVdata$mba_p)
[1] 50
> max(CSVdata$mba_p)
[1] 95
```

Figure 2.1.2: The lowest and highest MBA percentage

As all the student has studied in MBA and the chances of receive placement is approximately 50 percent more or less regardless of age and gender, thus this analysis is conducted to investigate which minimum range of MBA percentage student should score in order to secure higher chance to receive placement. The first 2 line in the source code above uses data exploration concept to identify the lowest and highest MBA percentage through the usage min and max function. Then, data manipulation and transformation concept are applied to duplicate a new data frame (CSVdata2.1) to avoid alter the original data frame and create new column (Analysis_2.1) by using mutate function and case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the Specialization = “Mkt&Fin”.



```
CSVdata2.1 = subset(CSVdata2.1, select = -salary)
CSVdata2.1 <- na.omit(CSVdata2.1)
CSVdata2.1
ggplot(CSVdata2.1, aes(Analysis_2.1)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&Fin and Placement Status",
       y="Number of Students",
       x="Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&Fin")
```

Figure 2.1.3: Step 2 – Remove salary column, purge any row with N/A data and create bar chart by using ggplot function

After new column has been create in the duplicate data frame, the salary column has been remove through subset function before purge any row that contains N/A value by using na.omit function as all the rows which contains “Mkt&Fin” needs to be reserve to create bar chart. Then ggplot function is applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_2.1) and placement status. The x-axis displays the range of Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&Fin while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

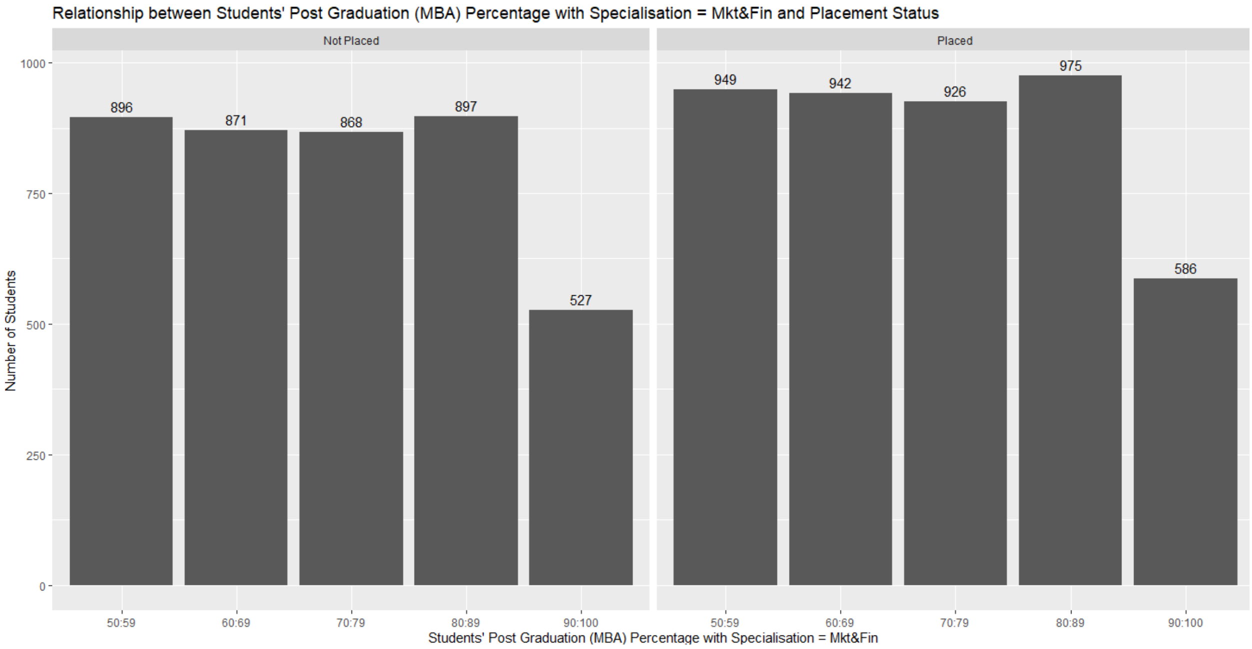


Figure 2.1.4: Relationship between Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&Fin and Placement Status

Based on the bar chart, the total number of students according to the percentage range which are: ~

50:59 (949 + 896 = **1845**), 60:69 (942 + 871 = **1813**) ,70:79 (926 + 868 = **1794**)

80:89 (975 + 897 = **1872**), 90:100 (586 + 527 = **1113**)

The percentage of students who receive placement according to the range of Post Graduation (MBA) Percentage with Specialisation = Mkt&Fin which are: ~

50:59 (949/1845 * 100 = **51.44%**), 60:69 (942/1813 * 100 = **51.96%**)

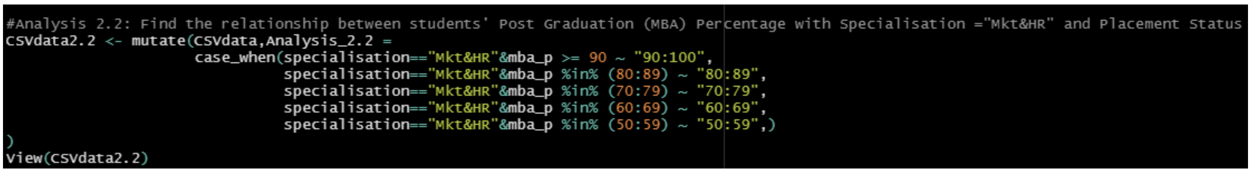
70:79 (926/1794 * 100 = **51.62%**), 80:89 (975/1872 * 100 = **52.08%**)

90:100 (586/1113 * 100 = **52.65%**)

In conclusion, the percentage of students who receive placement according to the range of Post Graduation (MBA) Percentage with Specialisation = “Mkt&Fin” are quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless any MBA percentage that had achieved in the Specialisation = “Mkt&Fin”. The bar chart also suggested that more analysis should be conducted.

Analysis 2.2: Find the relationship between students' Post Graduation (MBA) Percentage with Specialisation ="Mkt&HR" and Placement Status.

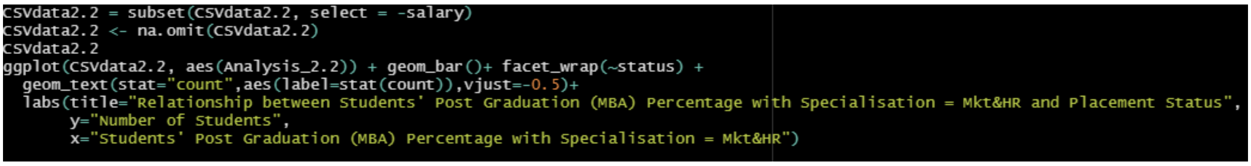
Source Code



```
#Analysis 2.2: Find the relationship between students' Post Graduation (MBA) Percentage with Specialisation ="Mkt&HR" and Placement Status
CSVdata2.2 <- mutate(CSVdata,Analysis_2.2 =
                     case_when(specialisation=="Mkt&HR"&mba_p >= 90 ~ "90:100",
                               specialisation=="Mkt&HR"&mba_p %in% (80:89) ~ "80:89",
                               specialisation=="Mkt&HR"&mba_p %in% (70:79) ~ "70:79",
                               specialisation=="Mkt&HR"&mba_p %in% (60:69) ~ "60:69",
                               specialisation=="Mkt&HR"&mba_p %in% (50:59) ~ "50:59",)
)
View(CSVdata2.2)
```

Figure 2.2.1: Step 1: Duplicate data frame and create new column by using mutate and case_when function

Similar to Analysis 2.1, this analysis is conducted to investigate which minimum range of MBA percentage student should score in order to secure higher chance to receive placement. According to the source code above, data manipulation and transformation concept are applied to duplicate a new data frame (CSVdata2.2) to avoid alter the original data frame and create new column (Analysis_2.2) by using mutate function and case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the Specialization = “Mkt&HR”.



```
CSVdata2.2 = subset(CSVdata2.2, select = -salary)
CSVdata2.2 <- na.omit(CSVdata2.2)
CSVdata2.2
ggplot(CSVdata2.2, aes(Analysis_2.2)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&HR and Placement Status",
       y="Number of Students",
       x="Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&HR")
```

Figure 2.2.2: Step 2 – Remove salary column, purge any row with N/A data and create bar chart by using ggplot function

After new column has been create in the duplicate data frame, the salary column has been remove through subset function before purge any row that contains N/A value by using na.omit function as all the rows which contains “Mkt&HR” needs to be reserve to create bar chart. Then ggplot function is applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_2.2) and placement status. The x-axis displays the range of Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&HR while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

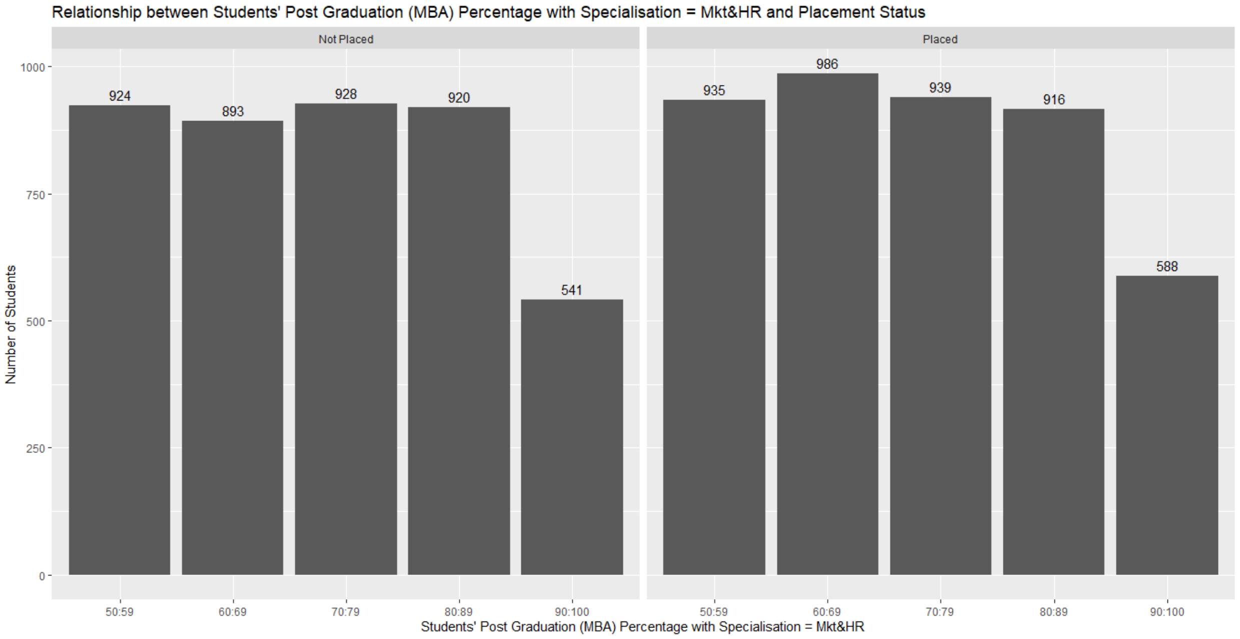


Figure 2.2.3: Relationship between Students' Post Graduation (MBA) Percentage with Specialisation = Mkt&HR and Placement Status

Based on the bar chart, the total number of students according to the percentage range which are: ~

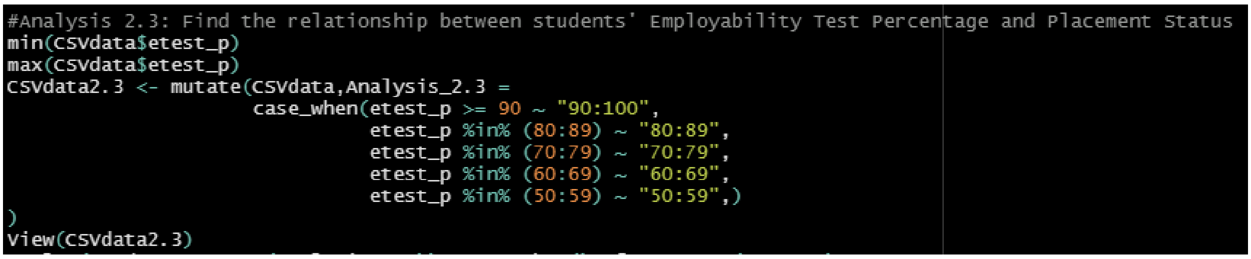
50:59 (935 + 924 = **1859**), 60:69 (986 + 893 = **1879**) ,70:79 (939+ 928 = **1867**)
80:89 (916 + 920 = **1836**), 90:100 (588 + 541 = **1129**)

The percentage of students who receive placement according to the range of Post Graduation (MBA) Percentage with Specialisation = Mkt&HR which are: ~
50:59 (935/1859 * 100 = **50.30%**), 60:69 (986/1879 * 100 = **52.47%**)
70:79 (939/1867 * 100 = **50.29%**), 80:89 (916/1836 * 100 = **49.89%**)
90:100 (588/1129 * 100 = **52.08%**)

In conclusion, the percentage of students who receive placement according to the range of Post Graduation (MBA) Percentage with Specialisation = "Mkt&HR" are quite close to each other. Student who scores between 80:89 has lowest chance to receive placement due to other reason. Overall, the bar chart shows each student has a rather fair chance to receive placement regardless any MBA percentage that had achieved in the Specialisation = "Mkt&HR". The bar chart also suggested that more analysis should be conducted.

Analysis 2.3: Find the relationship between students' Employability Test Percentage and Placement Status

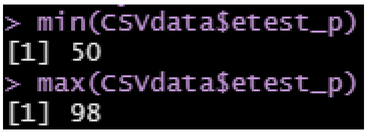
Source Code



```
#Analysis 2.3: Find the relationship between students' Employability Test Percentage and Placement Status
min(CSVdata$etest_p)
max(CSVdata$etest_p)
CSVdata2.3 <- mutate(CSVdata,Analysis_2.3 =
                       case_when(etest_p >= 90 ~ "90:100",
                                 etest_p %in% (80:89) ~ "80:89",
                                 etest_p %in% (70:79) ~ "70:79",
                                 etest_p %in% (60:69) ~ "60:69",
                                 etest_p %in% (50:59) ~ "50:59",)
)
View(CSVdata2.3)
```

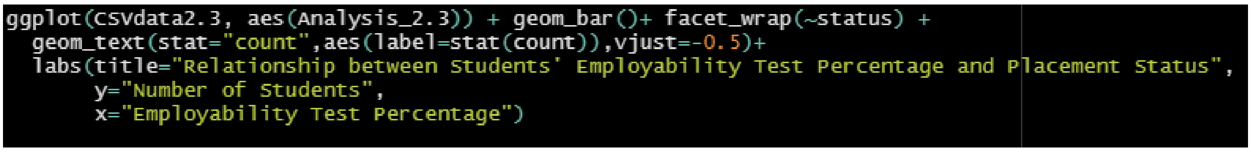
Figure 2.3.1: Step 1: Identify lowest and highest employability test percentage, duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of employability test percentage student should score in order to secure higher chance to receive placement. According to the source code above, it is necessary to identify lowest and highest employability test percentage by using data exploration concept in order to decide the range to be filter. Then, data manipulation and transformation concept are applied to duplicate a new data frame (CSVdata2.3) to avoid alter the original data frame and create new column (Analysis_2.3) by using mutate function and case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the employability percentage.



```
> min(CSVdata$etest_p)
[1] 50
> max(CSVdata$etest_p)
[1] 98
```

Figure 2.3.2: The lowest and highest employability test percentage



```
ggplot(CSVdata2.3, aes(Analysis_2.3)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Employability Test Percentage and Placement Status",
       y="Number of Students",
       x="Employability Test Percentage")
```

Figure 2.3.3: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_2.3) and placement status. The x-axis displays the range of Students' Employability Percentage while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

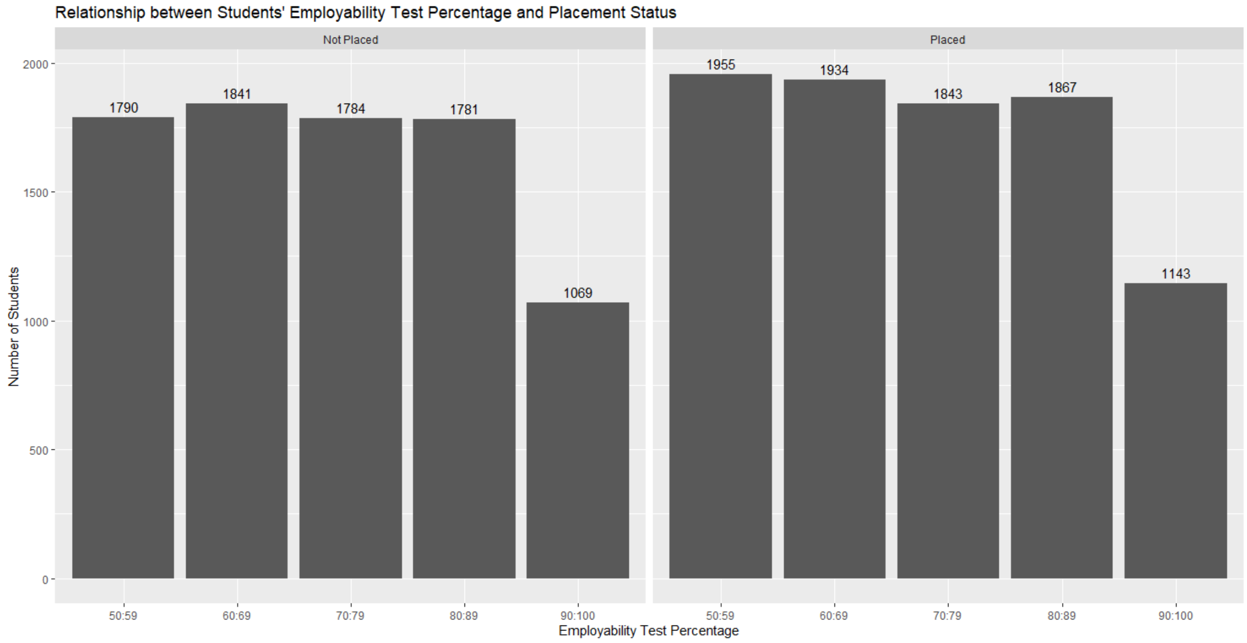


Figure 2.3.4: Relationship between Students' Employability Test Percentage and Placement Status

Based on the bar chart, the total number of students according to the employability test percentage which are: ~

50:59 (1955 + 1790 = **3745**), 60:69 (1934 + 1841 = **3775**) ,70:79 (1843 + 1784 = **3627**)

80:89 (1867 + 1781 = **3648**), 90:100 (1143 + 1069 = **2212**)

The percentage of students who receive placement according to the employability test percentage which are: ~

50:59 (1955/3745 * 100 = **52.20%**), 60:69 (1934/3775 * 100 = **51.23%**)

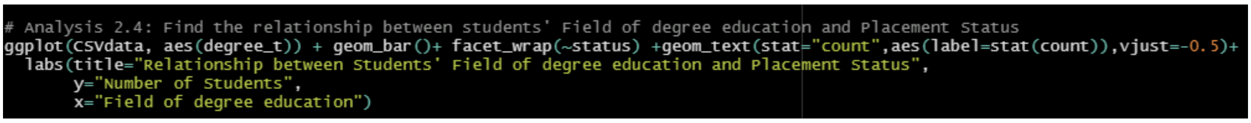
70:79 (1843/3627 * 100 = **50.81%**), 80:89 (1867/3648 * 100 = **51.18%**)

90:100 (1143/2212 * 100 = **51.67%**)

In conclusion, the percentage of students who receive placement according to the range of employability test percentage are quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless any range of employability test percentage that achieved.

Analysis 2.4: Find the relationship between students' Field of degree education and Placement Status

Source Code



```
# Analysis 2.4: Find the relationship between students' Field of degree education and Placement Status
ggplot(CSVdata, aes(degree_t)) + geom_bar()+ facet_wrap(~status) +geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Field of degree education and Placement Status",
       y="Number of Students",
       x="Field of degree education")
```

Figure 2.4.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between students' field of degree education and placement status. The x-axis displays the field of degree education while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

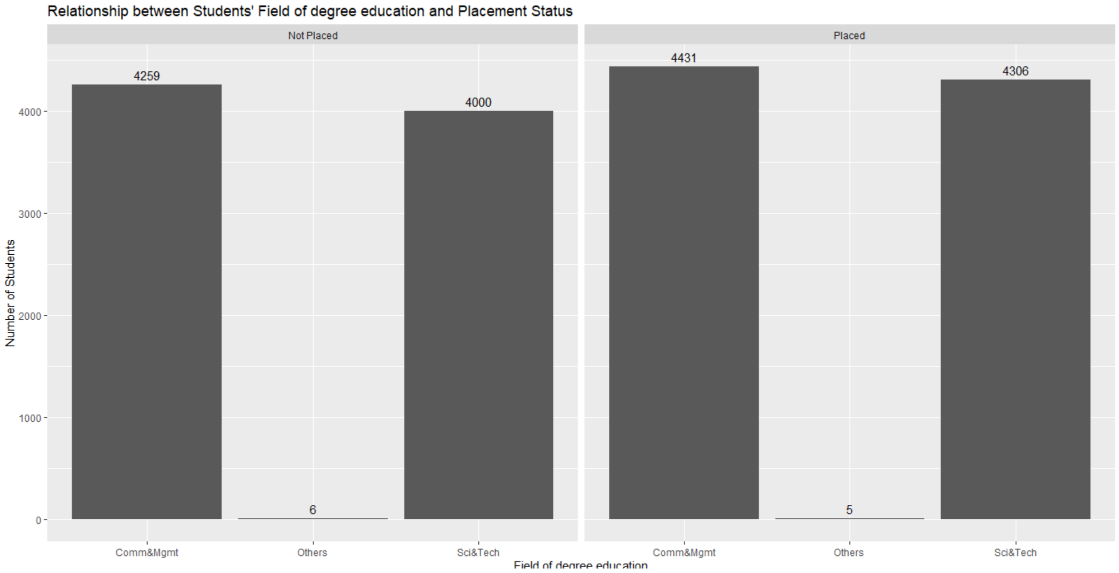


Figure 2.4.2: Relationship between Students' Field of degree education and Placement Status

Based on the bar chart, the total number of students according to the field of degree education which are: ~ Comm&Mgmt (4431 + 4259 = **8690**), Others (5 + 6 = **11**),

Sci&Tech (4306 + 4000 = **8306**)

The percentage of students who receive placement according to the field of degree education which are: ~ Comm&Mgmt (4431/8690 * 100 = **50.99%**), Others (5/11 * 100 = **45.45%**)

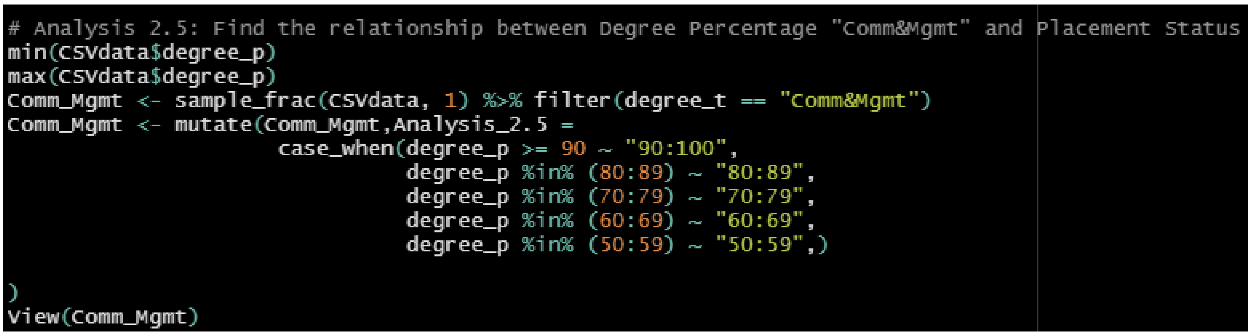
Sci&Tech (4306/4000 * 100 = **51.84%**)

In conclusion, the percentage of students who receive placement according to the field of degree education Comm& Mgmt and Sci&Tech are quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless any field of degree education chosen.

However, student who chose Others as degree education has very high risk to not receiving placement perhaps due to the course is being less demand or causes by other factors. Thus, more analysis needs to be conducted investigate why Others has lesser chance receiving placement.

Analysis 2.5: Find the relationship between Degree Percentage "Comm&Mgmt" and Placement Status

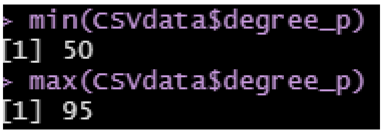
Source Code



```
# Analysis 2.5: Find the relationship between Degree Percentage "Comm&Mgmt" and Placement Status
min(CSVdata$degree_p)
max(CSVdata$degree_p)
Comm_Mgmt <- sample_frac(CSVdata, 1) %>% filter(degree_t == "Comm&Mgmt")
Comm_Mgmt <- mutate(Comm_Mgmt,Analysis_2.5 =
                     case_when(degree_p >= 90 ~ "90:100",
                               degree_p %in% (80:89) ~ "80:89",
                               degree_p %in% (70:79) ~ "70:79",
                               degree_p %in% (60:69) ~ "60:69",
                               degree_p %in% (50:59) ~ "50:59",)

)
View(Comm_Mgmt)
```

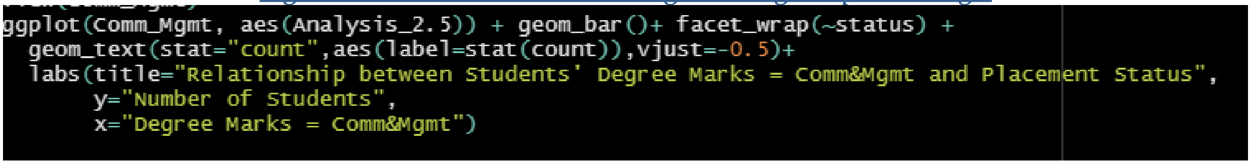
Figure 2.5.1: Step 1: Identify lowest and highest degree percentage, filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of degree percentage student should score in “Comm_Mgmt” in order to secure higher chance to receive placement. According to the source code above, it is necessary to identify lowest and highest degree percentage by using data exploration concept in order to decide the range to be filter. Then, data manipulation and transformation concept are applied to duplicate a new data frame (Comm_Mgmt) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_2.5), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the degree percentage.



```
> min(CSVdata$degree_p)
[1] 50
> max(CSVdata$degree_p)
[1] 95
```

Figure 2.5.2: The lowest and highest degree percentage



```
ggplot(Comm_Mgmt, aes(Analysis_2.5)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Degree Marks = Comm&Mgmt and Placement Status",
       y="Number of Students",
       x="Degree Marks = Comm&Mgmt")
```

Figure 2.5.3: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_2.5) and placement status. The x-axis displays the range of Degree Percentage = “Comm&Mgmt” while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

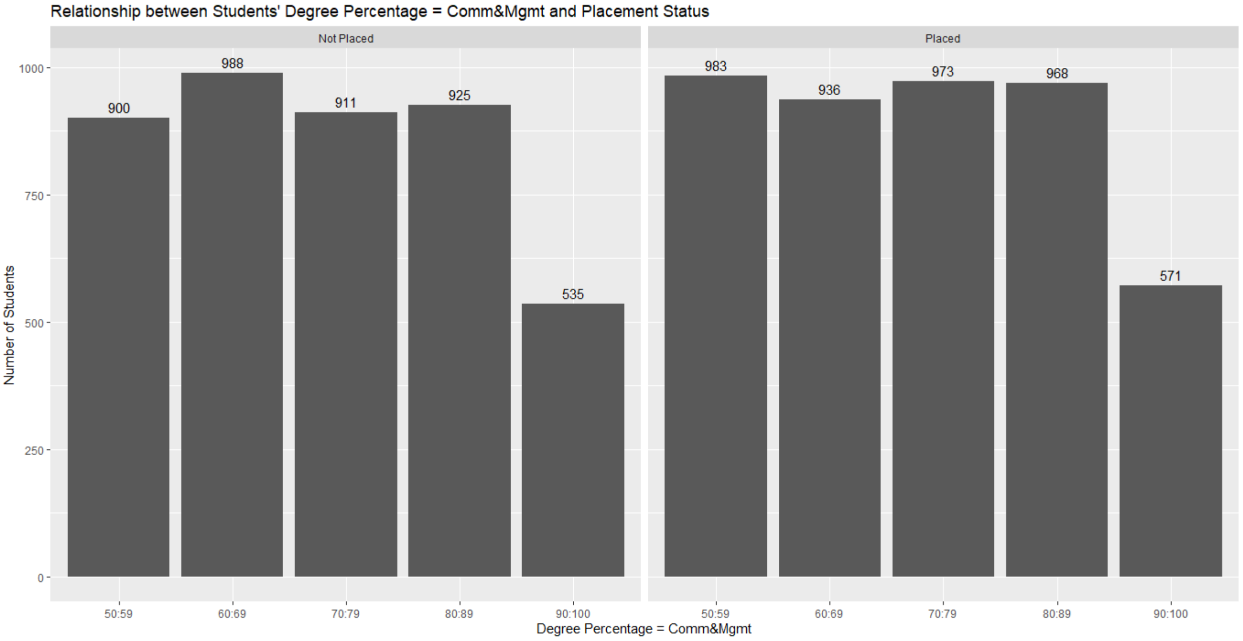


Figure 2.5.4: Relationship between Students' Degree Percentage = Comm&Mgmnt and Placement Status

Based on the bar chart, the total number of students according to the Students' Degree Percentage = Comm&Mgmnt which are: ~

50:59 (983 + 900 = **1883**), 60:69 (936 + 988 = **1924**) ,70:79 (973 + 911 = **1884**)

80:89 (968 + 925 = **1893**), 90:100 (571 + 535 = **1106**)

The percentage of students who receive placement according to the Students' Degree Percentage = Comm&Mgmnt which are: ~

50:59 (983/1883 * 100 = **52.20%**), 60:69 (936/1924 * 100 = **48.65%**)

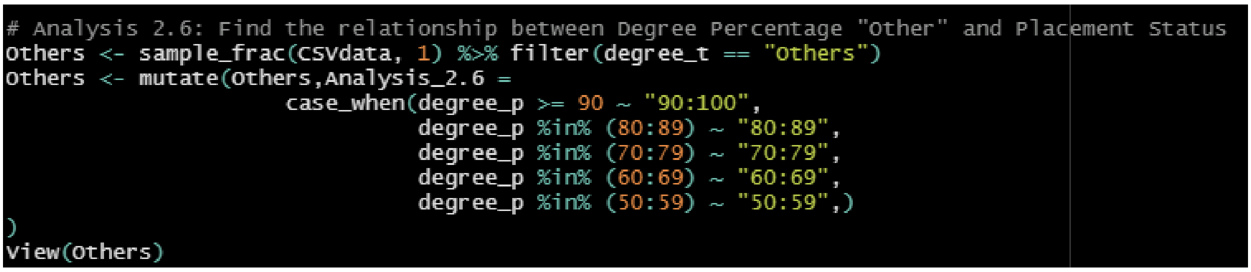
70:79 (973/1884 * 100 = **51.65%**), 80:89 (968/1893 * 100 = **51.14%**)

90:100 (571/1106 * 100 = **51.63%**)

In conclusion, the percentage of students who receive placement according to the range of degree percentage = Comm&Mgmt are quite close to each other except student who score range between 60:69 has slightly lower chance receive placement due to various reasons such as caused by other factors. Overall, the bar chart shows each student has a rather fair chance to receive placement regardless any range of degree percentage that achieved in Comm&Mgmt. Therefore, more analysis needs to be conducted.

Analysis 2.6: Find the relationship between Degree Percentage "Other" and Placement Status

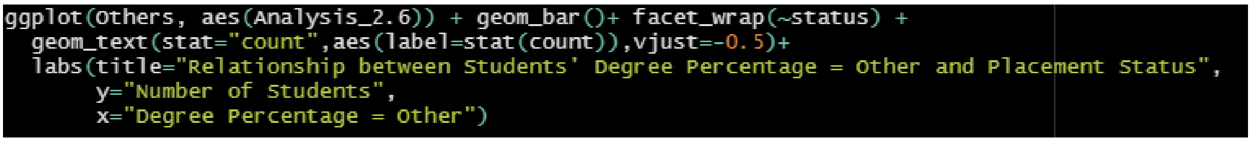
Source Code



```
# Analysis 2.6: Find the relationship between Degree Percentage "Other" and Placement Status
Others <- sample_frac(CSVdata, 1) %>% filter(degree_t == "Others")
Others <- mutate(Others,Analysis_2.6 =
                     case_when(degree_p >= 90 ~ "90:100",
                               degree_p %in% (80:89) ~ "80:89",
                               degree_p %in% (70:79) ~ "70:79",
                               degree_p %in% (60:69) ~ "60:69",
                               degree_p %in% (50:59) ~ "50:59",)
)
View(Others)
```

Figure 2.6.1: Step 1: Filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of degree percentage student should score in “Others” in order to secure higher chance to receive placement. According to the source code above data manipulation and transformation concept are applied to duplicate a new data frame (Others) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_2.6), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the degree percentage.



```
ggplot(Others, aes(Analysis_2.6)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Degree Percentage = Other and Placement Status",
       y="Number of Students",
       x="Degree Percentage = Other")
```

Figure 2.6.2: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_2.6) and placement status. The x-axis displays the range of Degree Percentage = “Others” while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

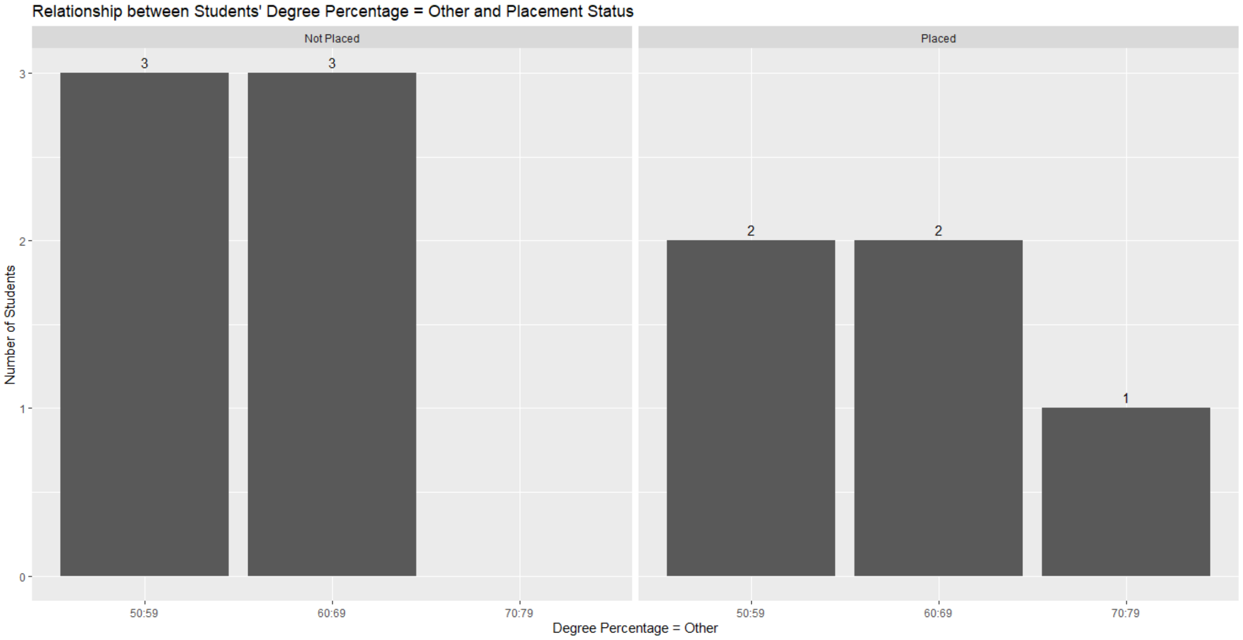
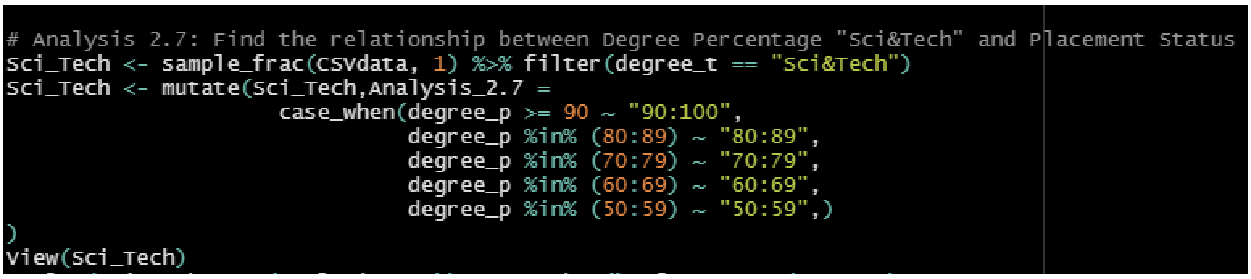


Figure 2.6.3: Relationship between Students' Degree Percentage = Others and Placement Status

Based on the bar chart, student who score between 70:79 has full chance receive placement due to only one student who score in that particular range percentage, meanwhile student who score between 50:59 and 60:69 has less than 50% chance receive placement. This is because the only 11 student who studies in “Others”. Thus, only 7 students can receive placement while another 6 students cannot receive placement is decided in order to fulfill the condition of only half of the student can receive placement while the other half unable to receive placement regardless any percentage scored during degree.

Analysis 2.7: Find the relationship between Degree Percentage "Sci&Tech" and Placement Status

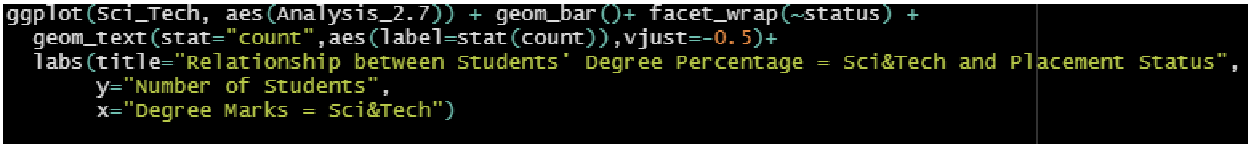
Source Code



```
# Analysis 2.7: Find the relationship between Degree Percentage "Sci&Tech" and Placement Status
Sci_Tech <- sample_frac(CSVdata, 1) %>% filter(degree_t == "Sci&Tech")
Sci_Tech <- mutate(Sci_Tech,Analysis_2.7 =
                     case_when(degree_p >= 90 ~ "90:100",
                               degree_p %in% (80:89) ~ "80:89",
                               degree_p %in% (70:79) ~ "70:79",
                               degree_p %in% (60:69) ~ "60:69",
                               degree_p %in% (50:59) ~ "50:59",)
)
View(Sci_Tech)
```

Figure 2.7.1: Step 1: Filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of degree percentage student should score in “Sci&Tech” in order to secure higher chance to receive placement. According to the source code above data manipulation and transformation concept are applied to duplicate a new data frame (Sci_Tech) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_2.7), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the degree percentage.



```
ggplot(Sci_Tech, aes(Analysis_2.7)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Degree Percentage = Sci&Tech and Placement Status",
       y="Number of Students",
       x="Degree Marks = Sci&Tech")
```

Figure 2.7.2: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_2.7) and placement status. The x-axis displays the range of Degree Percentage = “Sci&Tech” while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

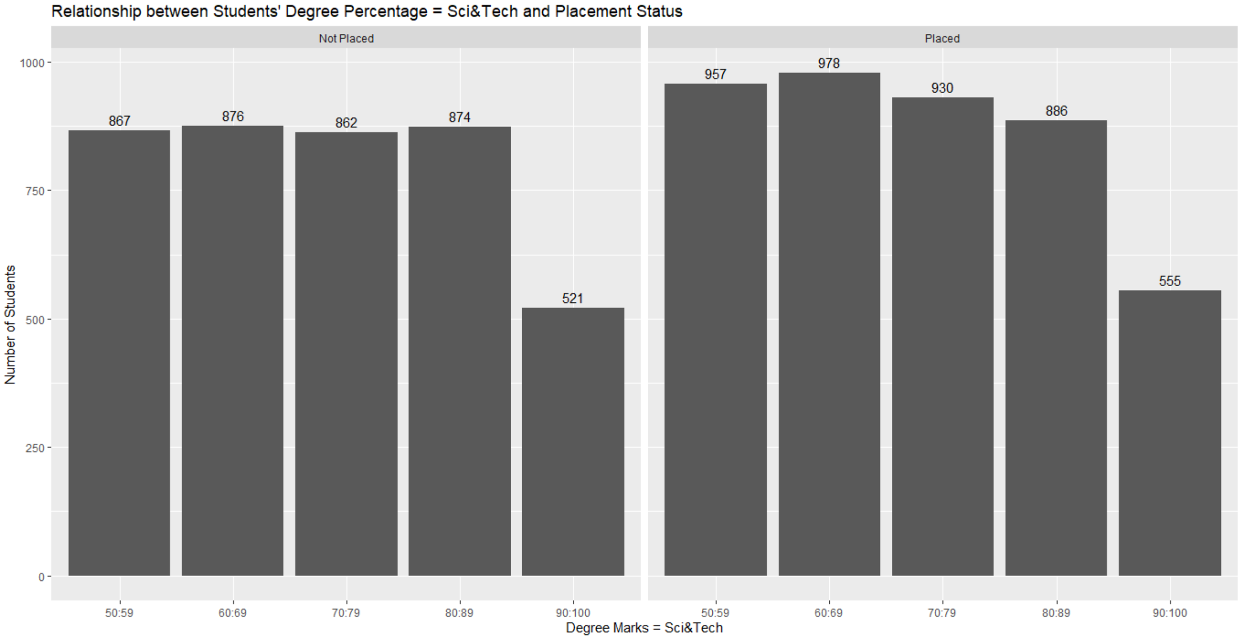


Figure 2.7.3: Relationship between Students' Degree Percentage = Sci&Tech and Placement Status

Based on the bar chart, the total number of students according to the Students' Degree Percentage = Sci&Tech which are: ~

50:59 (957 + 867 = **1824**), 60:69 (978 + 876 = **1854**) ,70:79 (930 + 862 = **1854**)

80:89 (886 + 974 = **1860**), 90:100 (555 + 521 = **1076**)

The percentage of students who receive placement according to the Students' Degree Percentage = Sci&Tech which are: ~

50:59 (957/1824 * 100 = **52.47%**), 60:69 (978/1854 * 100 = **52.75%**)

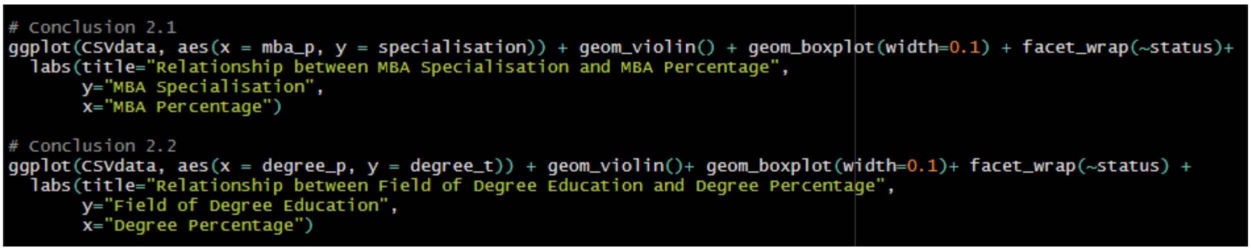
70:79 (930/1792 * 100 = **51.90%**), 80:89 (886/1860 * 100 = **47.63%**)

90:100 (555/1076 * 100 = **51.58%**)

In conclusion, the percentage of students who receive placement according to the range of degree percentage = Sci&Tech are quite close to each other except student who score range between 80:89 has slightly lower chance receive and the students being reject perhaps due to other factors causes. Thus, the bar chart shows each student has a rather fair chance to receive placement regardless any range of degree percentage that achieved in Sci&Tech. Therefore, more analysis needs to be conducted.

Conclusion for Question 2

Based on the various analysis conducted in Question 2, there are two important clues has been discovered from the data visualization: The first clue is that student has fair chance to receive placement regardless any range of MBA percentage scored, regardless any range of employability test percentage scored and regardless any range of degree percentage scored. The second clue is that student studied any MBA specialization and any degree specialization also has fair chance to receive placement regardless any percentage student achieved.



```
# Conclusion 2.1
ggplot(csvdata, aes(x = mba_p, y = specialisation)) + geom_violin() + geom_boxplot(width=0.1) + facet_wrap(~status)+
  labs(title="Relationship between MBA Specialisation and MBA Percentage",
       y="MBA Specialisation",
       x="MBA Percentage")

# Conclusion 2.2
ggplot(csvdata, aes(x = degree_p, y = degree_t)) + geom_violin()+ geom_boxplot(width=0.1)+ facet_wrap(~status) +
  labs(title="Relationship between Field of Degree Education and Degree Percentage",
       y="Field of Degree Education",
       x="Degree Percentage")
```

Figure 2.8.1 – Source Code to create violin plot and boxplot

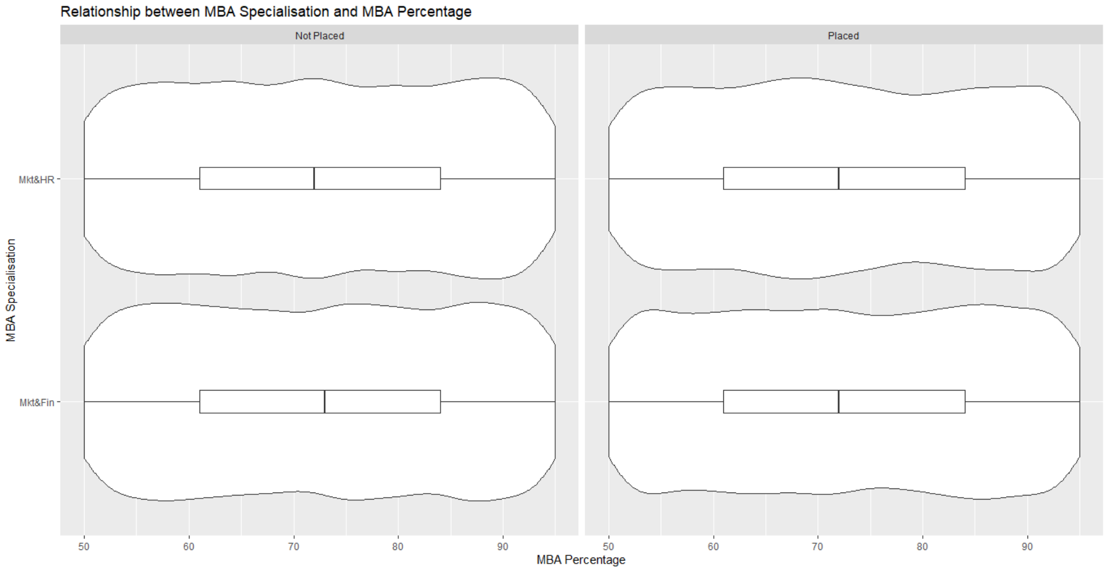


Figure 2.8.2 – Violin plot and Box plot of Relationship between MBA Specialisation and MBA Percentage

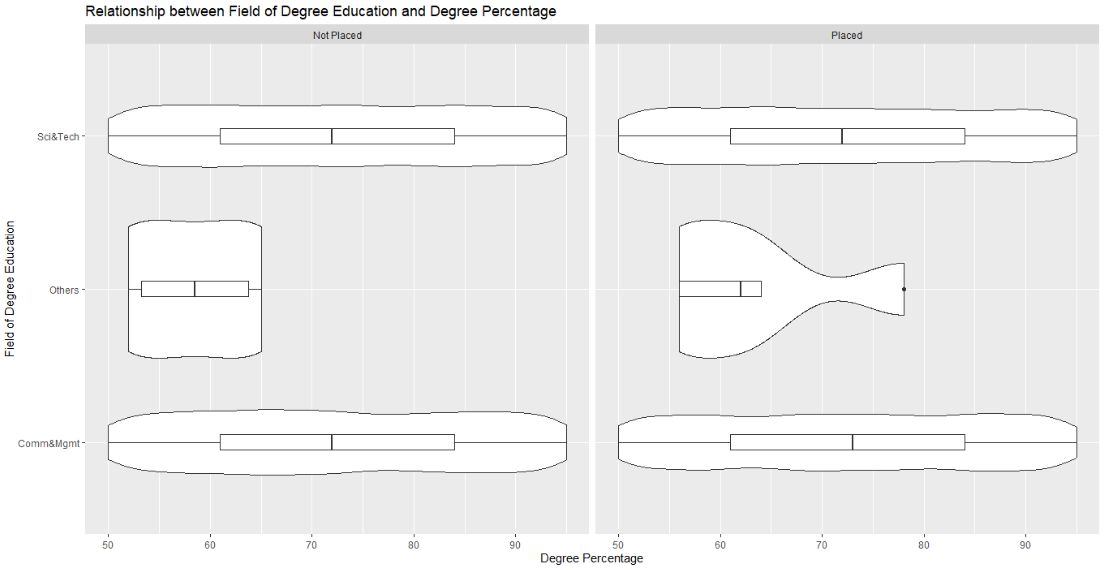


Figure 2.8.3 – Violin plot and Box plot of Relationship between Field of Degree Education and Degree Percentage

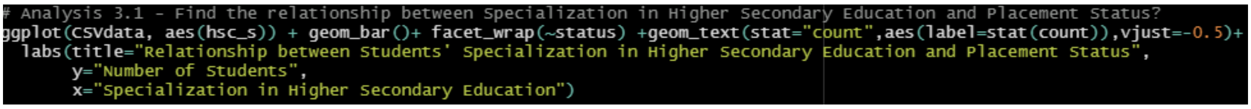
Based on the source code in Figure 2.8.1, the geom_violin and box_plot function had been utilized to construct violin plot and box plot in order to investigate the distribution of the data and the graph statistics. According to Figure 2.8.2, the size of the violin plot and the median in the boxplot between two MBA specialisation (Mkt&Fin and Mkt&HR) are almost equal in both “Placed” and “Not Placed” graph. Thus, this explains that student who study different MBA specialisation still has the fair chances in receiving placement regardless any range of percentage is being scored by the students.

Similar to Figure 2.8.3, the size of the violin plot and the median in the boxplot in both degree specialization (Comm&Mgmt and Sci Tech) are almost equal in both “Placed” and “Not Placed” graph except for “Others” specialization due to lesser number of students study in that course (refer to Analysis 2.6). Thus, degree specialization is also using the same concept like MBA specialization: Student who study different degree specialisation still has the fair chances in receiving placement regardless any range of percentage is being scored by the students. However, more question and analysis should be conducted for further detail investigation in order to discover the condition in receiving placement.

Question 3: Does Student's Past Education Record Affect Student's Placement Status?

Analysis 3.1 - Find the relationship between Specialization in Higher Secondary Education and Placement Status

Source Code



```
# Analysis 3.1 - Find the relationship between Specialization in Higher Secondary Education and Placement Status?
ggplot(CSVdata, aes(hsc_s)) + geom_bar()+ facet_wrap(~status) +geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Specialization in Higher Secondary Education and Placement Status",
       y="Number of Students",
       x="Specialization in Higher Secondary Education")
```

Figure 3.1.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between students' Specialization in Higher Secondary Education and placement status. The x-axis displays the Specialization in Higher Secondary Education while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart. By using bar chart, it is easier to identify which specialization has more students received placement.

Data Visualization

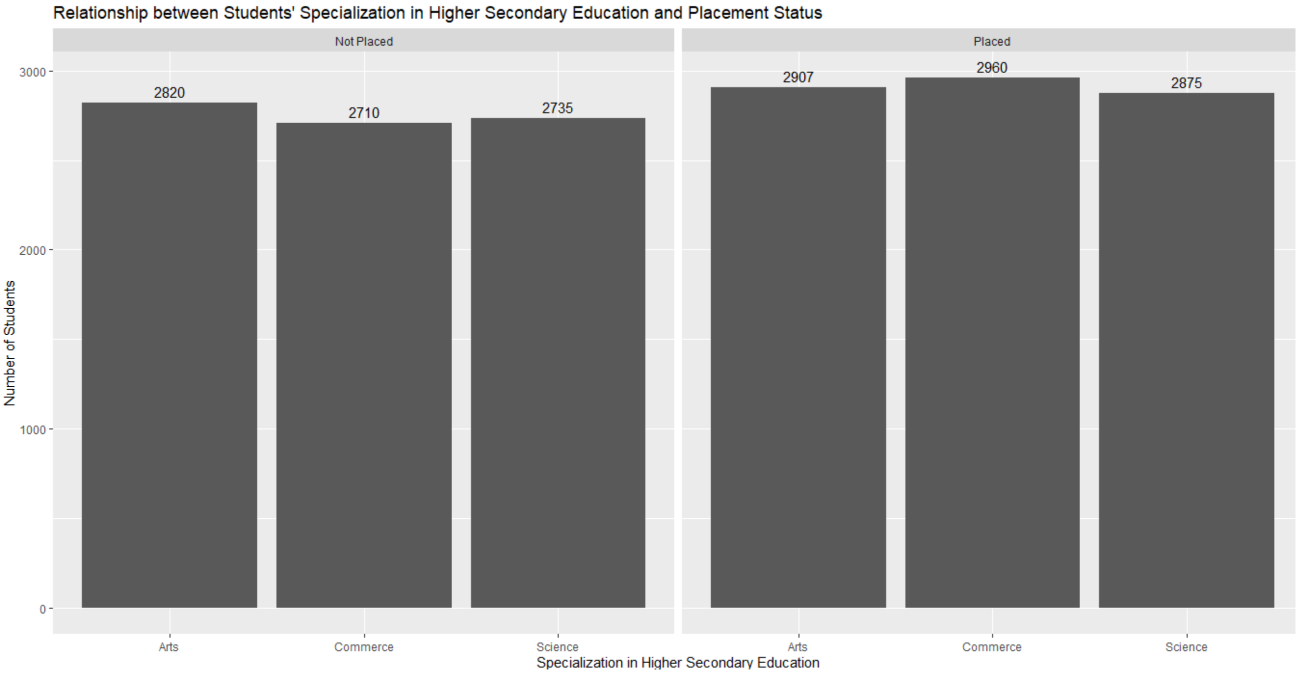


Figure 3.1.2: Relationship between Students' Specialization in Higher Secondary Education and Placement Status

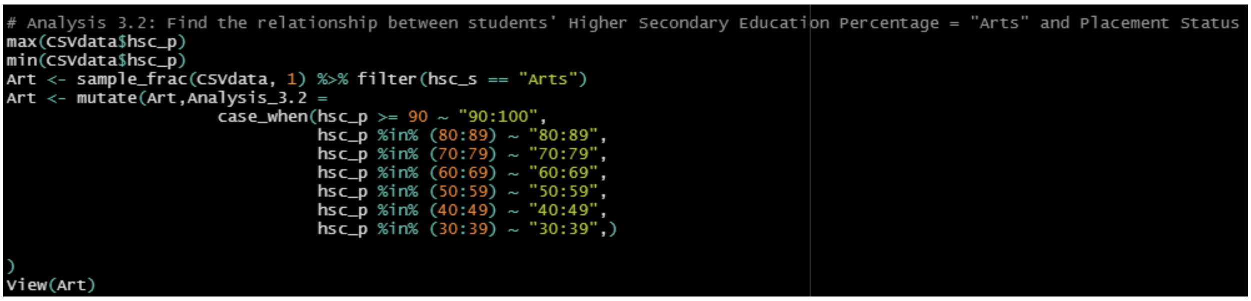
	Placed	Not Placed	Total Students	Placed Percentage
Arts	2907	2820	5727	50.76 %
Commerce	2960	2710	5670	52.20%
Science	2875	2375	8306	51.25%

In conclusion, the percentage of students who receive placement according to the Specialization in Higher Secondary Education are quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless any field of degree education chosen.

However, student who chose Arts has very high risk to not receiving placement perhaps due to other factor causes. Thus, more analysis needs to be conducted investigate why Arts has lesser chance receiving placement.

Analysis 3.2 - Find the relationship between students' Higher Secondary Education Percentage = "Arts" and Placement Status

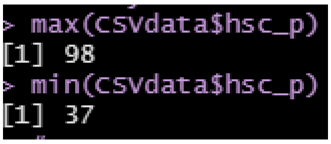
Source Code



```
# Analysis 3.2: Find the relationship between students' Higher Secondary Education Percentage = "Arts" and Placement Status
max(CSVdata$hsc_p)
min(CSVdata$hsc_p)
Art <- sample_frac(CSVdata, 1) %>% filter(hsc_s == "Arts")
Art <- mutate(Art,Analysis_3.2 =
                    case_when(hsc_p >= 90 ~ "90:100",
                              hsc_p %in% (80:89) ~ "80:89",
                              hsc_p %in% (70:79) ~ "70:79",
                              hsc_p %in% (60:69) ~ "60:69",
                              hsc_p %in% (50:59) ~ "50:59",
                              hsc_p %in% (40:49) ~ "40:49",
                              hsc_p %in% (30:39) ~ "30:39",)

)
View(Art)
```

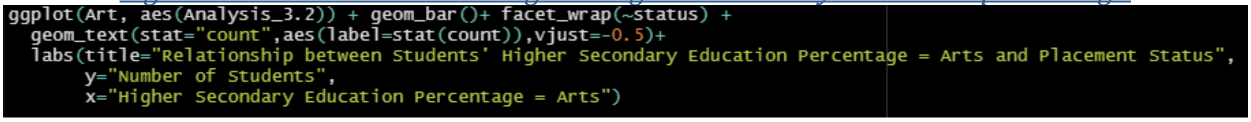
Figure 3.2.1: Step 1: Identify lowest and highest higher secondary education percentage, filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of higher secondary education percentage student should score in “Arts” in order to secure higher chance to receive placement. According to the source code above, it is necessary to identify lowest and highest higher secondary education percentage by using data exploration concept in order to decide the range to be filter. Then, data manipulation and transformation concept are applied to duplicate a new data frame (Art) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_3.2), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the higher secondary education percentage.



```
> max(CSVdata$hsc_p)
[1] 98
> min(CSVdata$hsc_p)
[1] 37
```

Figure 3.2.2: The lowest and highest higher secondary education percentage



```
ggplot(Art, aes(Analysis_3.2)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Higher Secondary Education Percentage = Arts and Placement Status",
       y="Number of Students",
       x="Higher Secondary Education Percentage = Arts")
```

Figure 3.2.3: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_3.2) and placement status. The x-axis displays the range of Higher Secondary Education Percentage = “Arts" while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

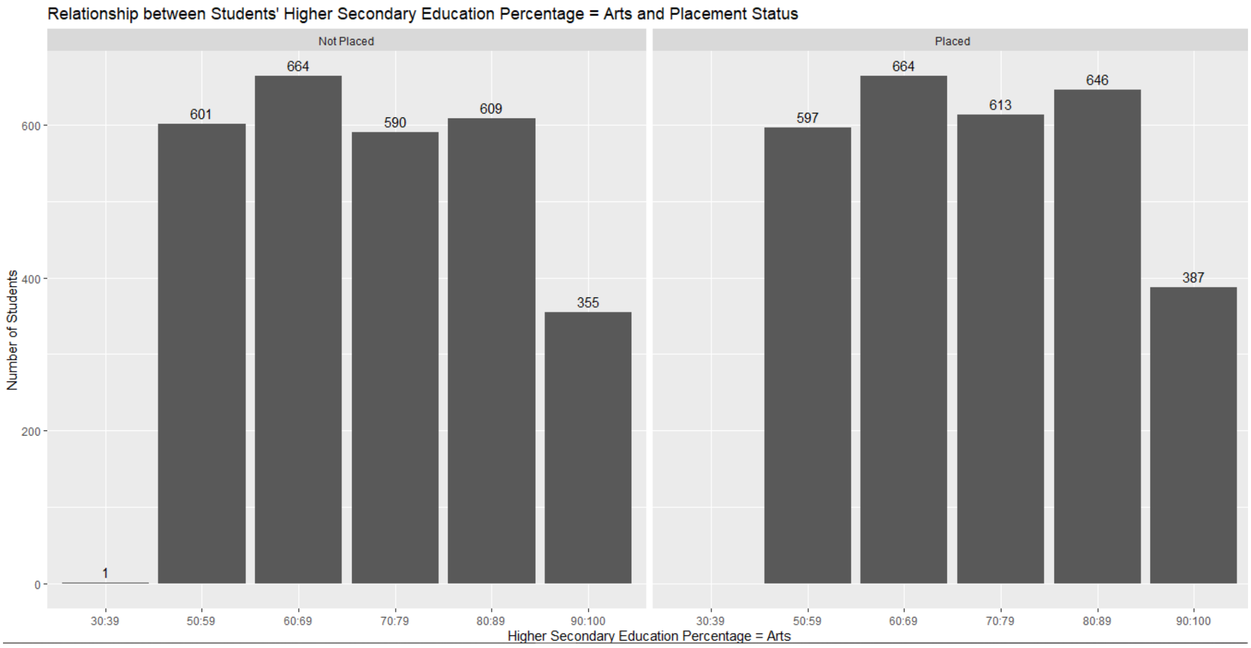


Figure 3.2.4: Relationship between Students' Higher Secondary Education Percentage = Arts and Placement Status

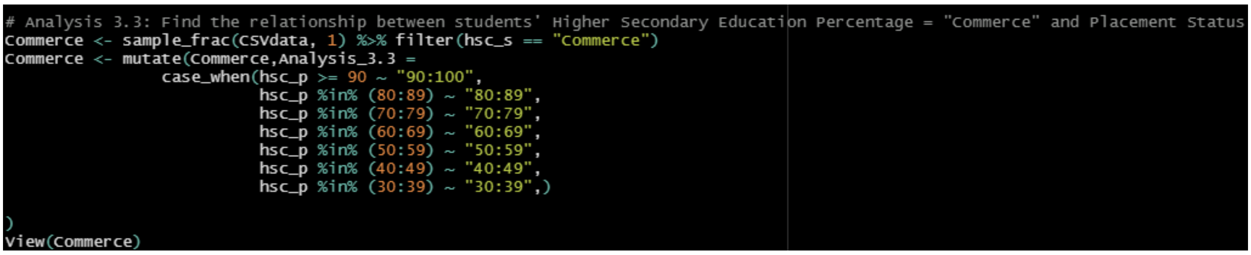
	Placed	Not Placed	Total Students	Placed Percentage
30:39	0	1	1	0.00 %
50:59	597	601	1198	49.83%
60:69	664	664	1328	50.00%
70:79	613	590	1203	50.95%
80:89	646	609	1255	51.47%
90:100	387	355	742	52.16%

In conclusion, the percentage of students who receive placement according to the placed percentage in Higher Secondary Education Percentage = Arts are quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless of any range of percentage score by the student.

However, the bar chart also suggested that student should not score lower than 50 percent as the placed percentage is 0%. Meanwhile. Student who scores between 50:59 has slightly lower chance receive placement, it is perhaps due to other factors cause, thus more analysis on the other factor is required to executed.

Analysis 3.3 - Find the relationship between students' Higher Secondary Education Percentage = "Commerce" and Placement Status

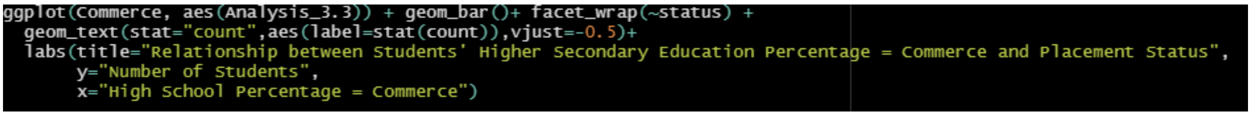
Source Code



```
# Analysis 3.3: Find the relationship between students' Higher Secondary Education Percentage = "Commerce" and Placement Status
Commerce <- sample_frac(CSVdata, 1) %>% filter(hsc_s == "Commerce")
Commerce <- mutate(Commerce,Analysis_3.3 =
                case_when(hsc_p >= 90 ~ "90:100",
                          hsc_p %in% (80:89) ~ "80:89",
                          hsc_p %in% (70:79) ~ "70:79",
                          hsc_p %in% (60:69) ~ "60:69",
                          hsc_p %in% (50:59) ~ "50:59",
                          hsc_p %in% (40:49) ~ "40:49",
                          hsc_p %in% (30:39) ~ "30:39",)

)
View(Commerce)
```

Figure 3.3.1: Step 1: Filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of higher secondary education percentage student should score in “Commerce” in order to secure higher chance to receive placement. According to the source code above, data manipulation and transformation concept are applied to duplicate a new data frame (Commerce) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_3.3), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the higher secondary education percentage.



```
ggplot(Commerce, aes(Analysis_3.3)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Higher Secondary Education Percentage = Commerce and Placement Status",
       y="Number of Students",
       x="High School Percentage = Commerce")
```

Figure 3.3.3: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_3.3) and placement status. The x-axis displays the range of Higher Secondary Education Percentage = “Commerce" while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

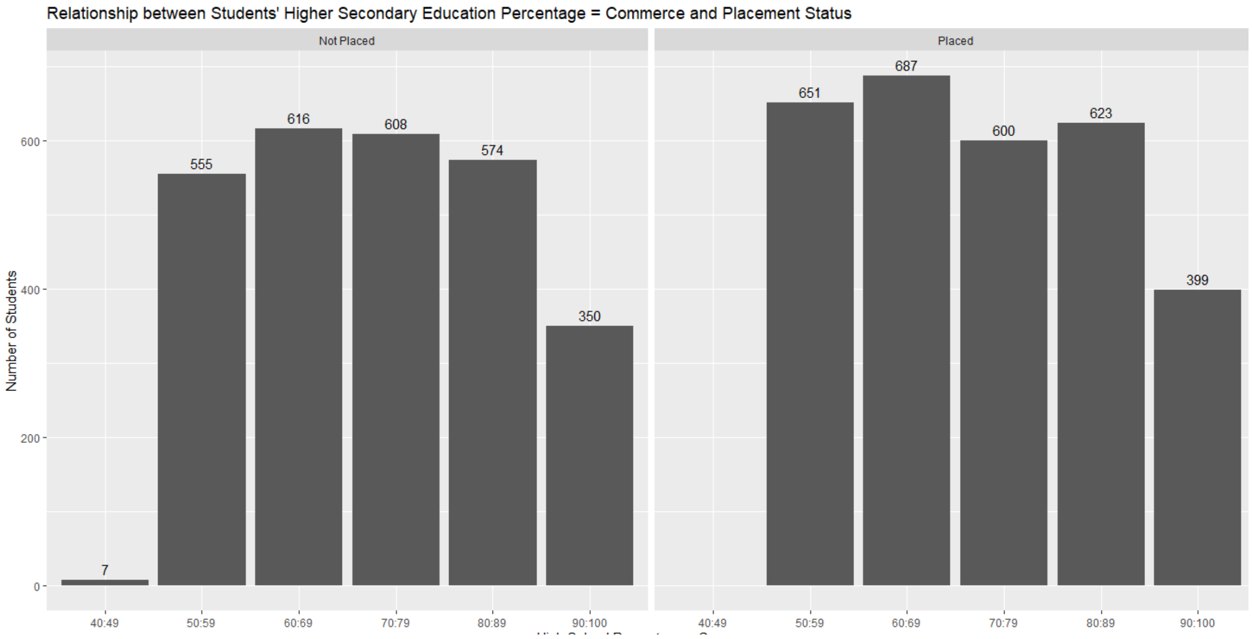


Figure 3.3.3: Relationship between Students' Higher Secondary Education Percentage = Commerce and Placement Status

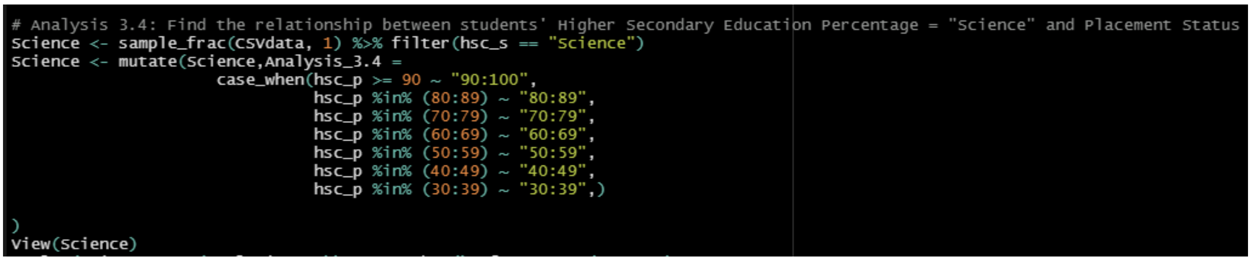
	Placed	Not Placed	Total Students	Placed Percentage
40:49	0	7	7	0.00%
50:59	651	555	1206	53.98%
60:69	687	616	1303	52.72%
70:79	600	608	1208	49.67%
80:89	623	574	1197	52.05%
90:100	399	350	749	53.27%

In conclusion, the percentage of students who receive placement according to the placed percentage Higher Secondary Education = Commerce is quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless of any range of percentage achieved by the student.

However, the bar chart also suggested that student should not score lower than 50 percent as the placed percentage is 0%. Meanwhile, student who scores between 70:79 has slightly lower chance receive placement, it is perhaps due to other factors cause, thus more analysis on the other factor is required to implement.

Analysis 3.4 - Find the relationship between students' Higher Secondary Education Percentage = "Science" and Placement Status

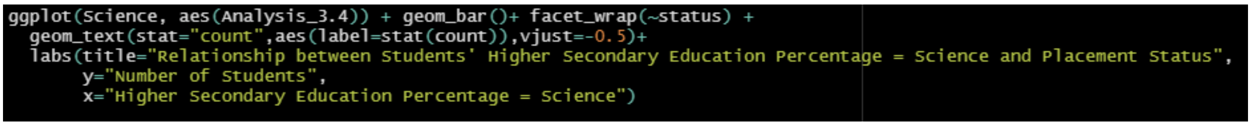
Source Code



```r
# Analysis 3.4: Find the relationship between students' Higher Secondary Education Percentage = "Science" and Placement Status
Science <- sample_frac(CSVdata, 1) %>% filter(hsc_s == "Science")
Science <- mutate(Science,Analysis_3.4 =
                    case_when(hsc_p >= 90 ~ "90:100",
                              hsc_p %in% (80:89) ~ "80:89",
                              hsc_p %in% (70:79) ~ "70:79",
                              hsc_p %in% (60:69) ~ "60:69",
                              hsc_p %in% (50:59) ~ "50:59",
                              hsc_p %in% (40:49) ~ "40:49",
                              hsc_p %in% (30:39) ~ "30:39",)

)
View(Science)
```

Figure 3.4.1: Step 1: Filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of higher secondary education percentage student should score in "Science" in order to secure higher chance to receive placement. According to the source code above, data manipulation and transformation concept are applied to duplicate a new data frame (Science) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_3.4), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the higher secondary education percentage.



```r
ggplot(Science, aes(Analysis_3.4)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Higher Secondary Education Percentage = Science and Placement Status",
       y="Number of Students",
       x="Higher Secondary Education Percentage = Science")
```

Figure 3.4.2: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_3.4) and placement status. The x-axis displays the range of Higher Secondary Education Percentage = "Science" while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

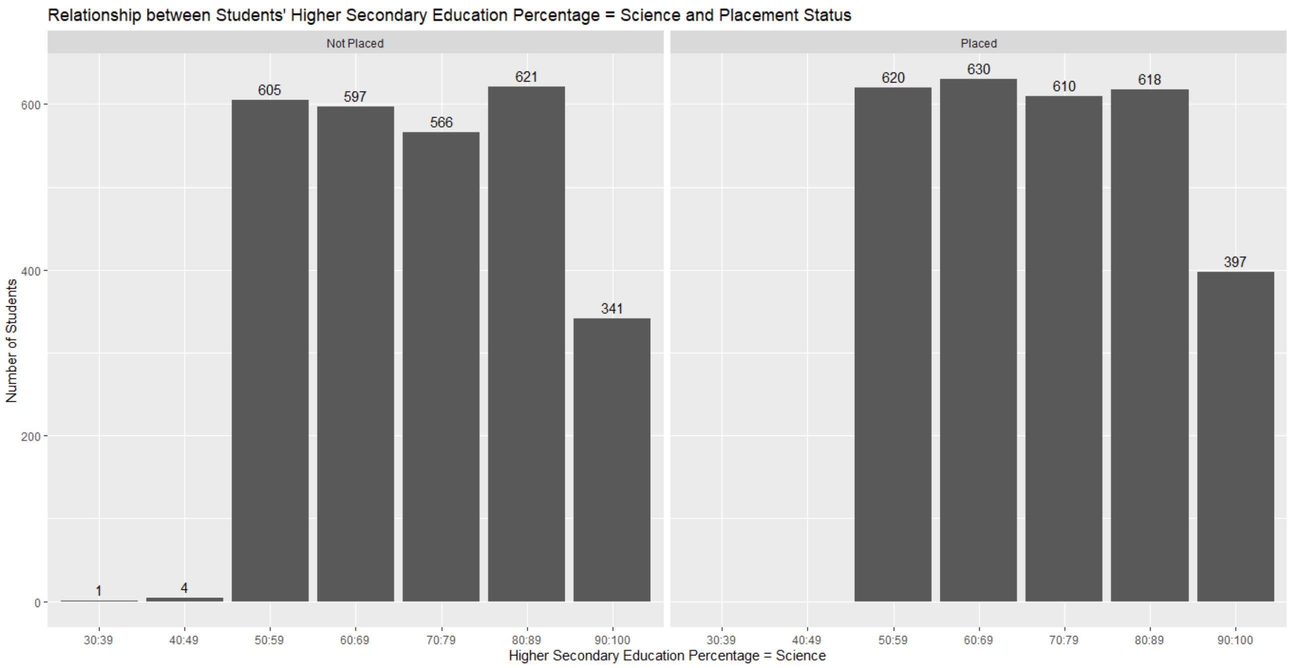


Figure 3.4.3: Relationship between Students' Higher Secondary Education Percentage = Science and Placement Status

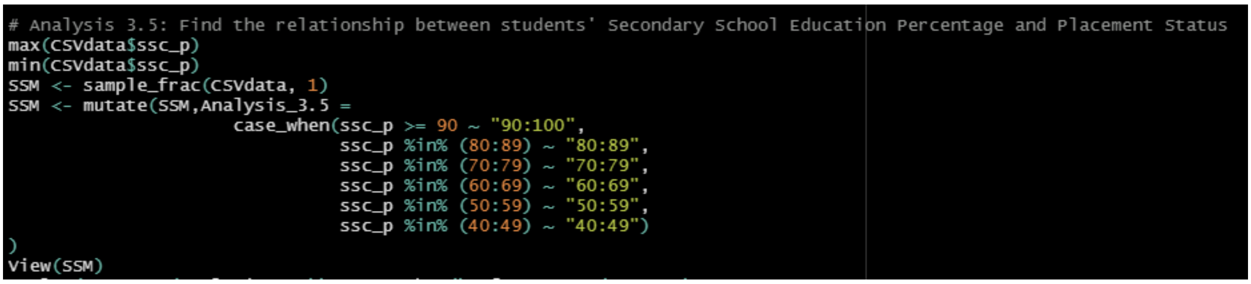
	Placed	Not Placed	Total Students	Placed Percentage
30:39	0	1	1	0.00 %
40:49	0	4	4	0.00%
50:59	620	605	1225	50.61%
60:69	630	597	1227	51.34%
70:79	610	566	1176	51.87%
80:89	618	621	1239	49.88%
90:100	397	341	738	53.79%

In conclusion, the percentage of students who receive placement according to the placed percentage in Higher Secondary Education Percentage = Science is quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless of any range of percentage achieved by the student.

However, the bar chart also suggested that student should not score lower than 50 percent as the placed percentage is 0%. Meanwhile, student who scores between 80:89 has slightly lower chance receive placement, it is perhaps due to other factors cause, thus more analysis on the other factor is required to conducted.

Analysis 3.5: Find the relationship between students' Secondary School Education Percentage and Placement Status

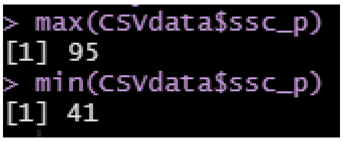
Source Code



```
# Analysis 3.5: Find the relationship between students' Secondary School Education Percentage and Placement Status
max(CSVdata$ssc_p)
min(CSVdata$ssc_p)
SSM <- sample_frac(CSVdata, 1)
SSM <- mutate(SSM,Analysis_3.5 =
                    case_when(ssc_p >= 90 ~ "90:100",
                              ssc_p %in% (80:89) ~ "80:89",
                              ssc_p %in% (70:79) ~ "70:79",
                              ssc_p %in% (60:69) ~ "60:69",
                              ssc_p %in% (50:59) ~ "50:59",
                              ssc_p %in% (40:49) ~ "40:49")
)
View(SSM)
```

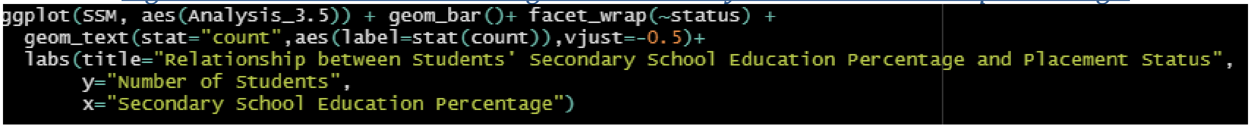
Figure 3.5.1: Step 1: Identify lowest and highest secondary school education percentage, filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which minimum range of secondary school education percentage student should score in order to secure higher chance to receive placement. According to the source code above, it is necessary to identify lowest and highest secondary school education percentage by using data exploration concept in order to decide the range to be filter. Then, data manipulation and transformation concept are applied to duplicate a new data frame (SSM) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_3.2), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else in order to set the percentage range according to the secondary school education percentage.



```
> max(CSVdata$ssc_p)
[1] 95
> min(CSVdata$ssc_p)
[1] 41
```

Figure 3.5.2: The lowest and highest secondary school education percentage



```
ggplot(SSM, aes(Analysis_3.5)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Secondary School Education Percentage and Placement Status",
       y="Number of Students",
       x="Secondary School Education Percentage")
```

Figure 3.5.4: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_3.5) and placement status. The x-axis displays the range of secondary school education percentage while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

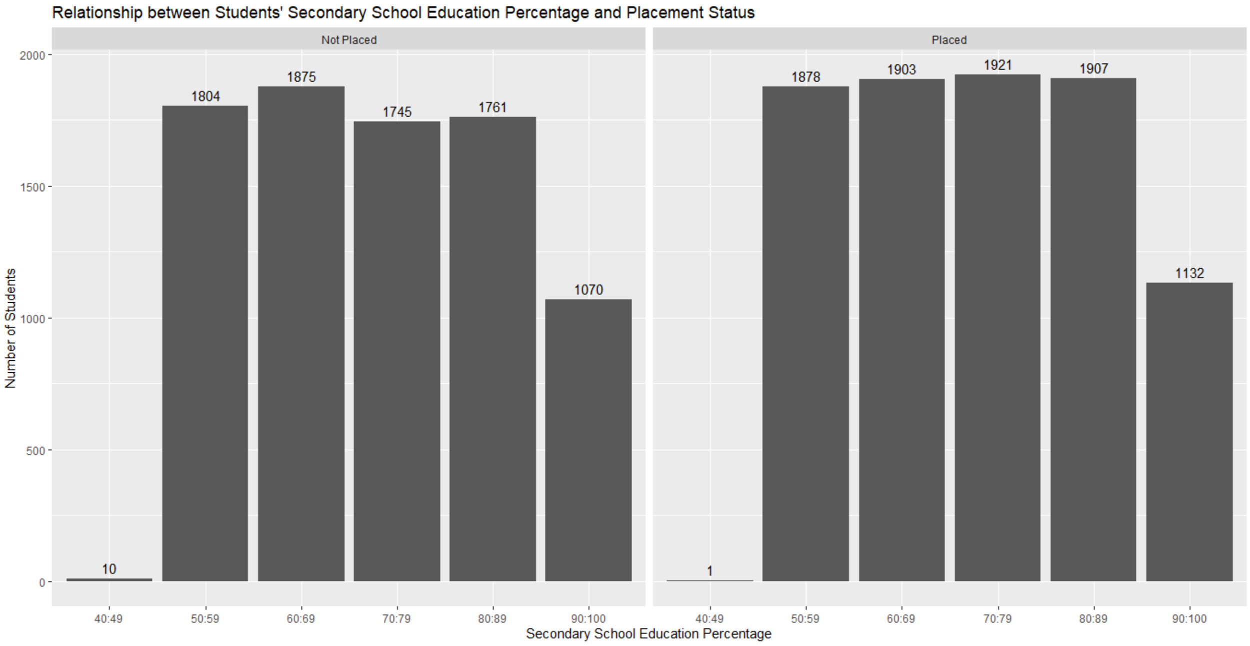


Figure 3.5.4: Relationship between Students' Secondary School Education Percentage = Science and Placement Status

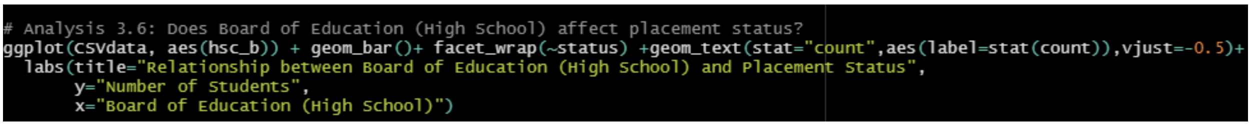
	Placed	Not Placed	Total Students	Placed Percentage
40:49	1	10	11	9.09 %
50:59	1878	1804	3682	51.00 %
60:69	1903	1875	3778	50.37%
70:79	1903	1875	3778	50.37%
80:89	1921	1745	3666	52.40%
90:100	1132	1070	2202	51.41%

In conclusion, the percentage of students who receive placement according to the placed percentage in Secondary School Education Percentage is quite close to each other. Thus, the bar chart shows each student has a fair chance to receive placement regardless of any range of percentage achieved by the student.

However, the bar chart also suggested that student should not score lower than 50 percent as the placed percentage is 9.09%.

Analysis 3.6: Does Board of Education (High School) affect placement status?

Source Code



```
# Analysis 3.6: Does Board of Education (High School) affect placement status?
ggplot(CSVdata, aes(hsc_b)) + geom_bar()+ facet_wrap(~status) +geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Board of Education (High School) and Placement Status",
       y="Number of Students",
       x="Board of Education (High School)")
```

Figure 3.6.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between students' Board of Education (High School) and placement status. The x-axis displays the Board of Education (High School) while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart. By using bar chart, it is easier to identify which board of education has more students received placement.

Data Visualization

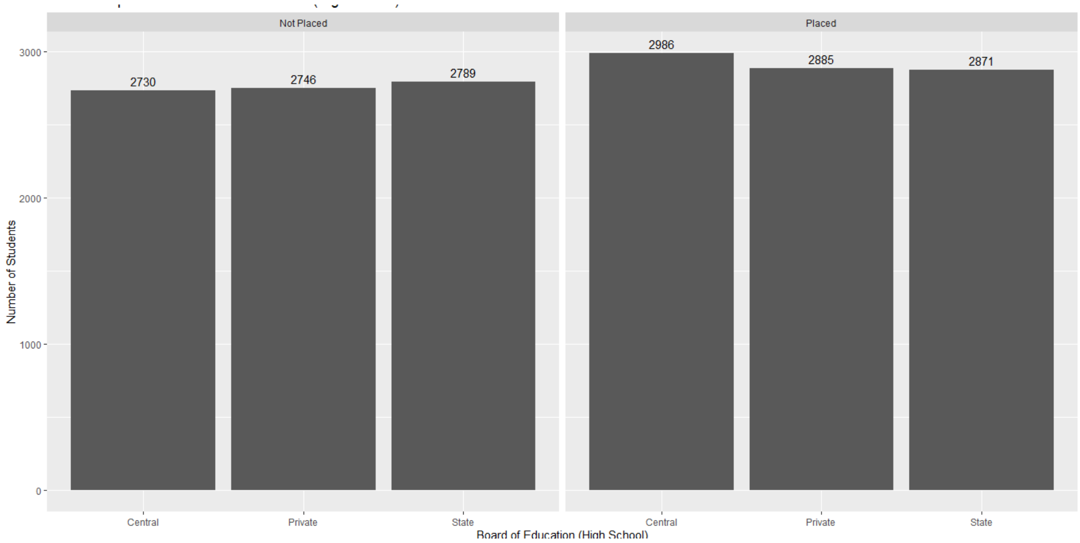


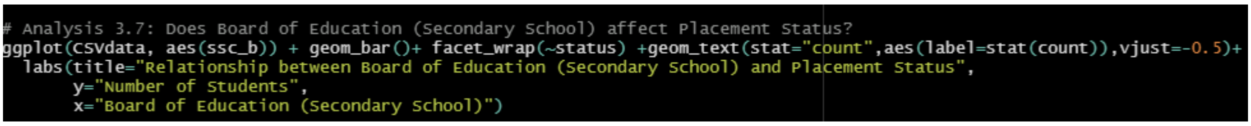
Figure 3.6.2: Relationship between Board of Education (High School) and Placement Status

	Placed	Not Placed	Total Students	Placed Percentage
Central	2986	2730	5716	52.24 %
Private	2885	2746	5631	51.23 %
State	2871	2789	5660	50.72 %

In conclusion, student receive fair chance in placement regardless any board of education chosen by the student for higher secondary school.

Analysis 3.7: Does Board of Education (Secondary School) affect placement status?

Source Code



```
# Analysis 3.7: Does Board of Education (Secondary School) affect Placement Status?
ggplot(CSVdata, aes(ssc_b)) + geom_bar()+ facet_wrap(~status) +geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Board of Education (Secondary School) and Placement Status",
       y="Number of Students",
       x="Board of Education (Secondary School)")
```

Figure 3.7.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between students' Board of Education (Secondary School) and placement status. The x-axis displays the Board of Education (Secondary School) while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart. By using bar chart, it is easier to identify which board of education has more students received placement.

Data Visualization

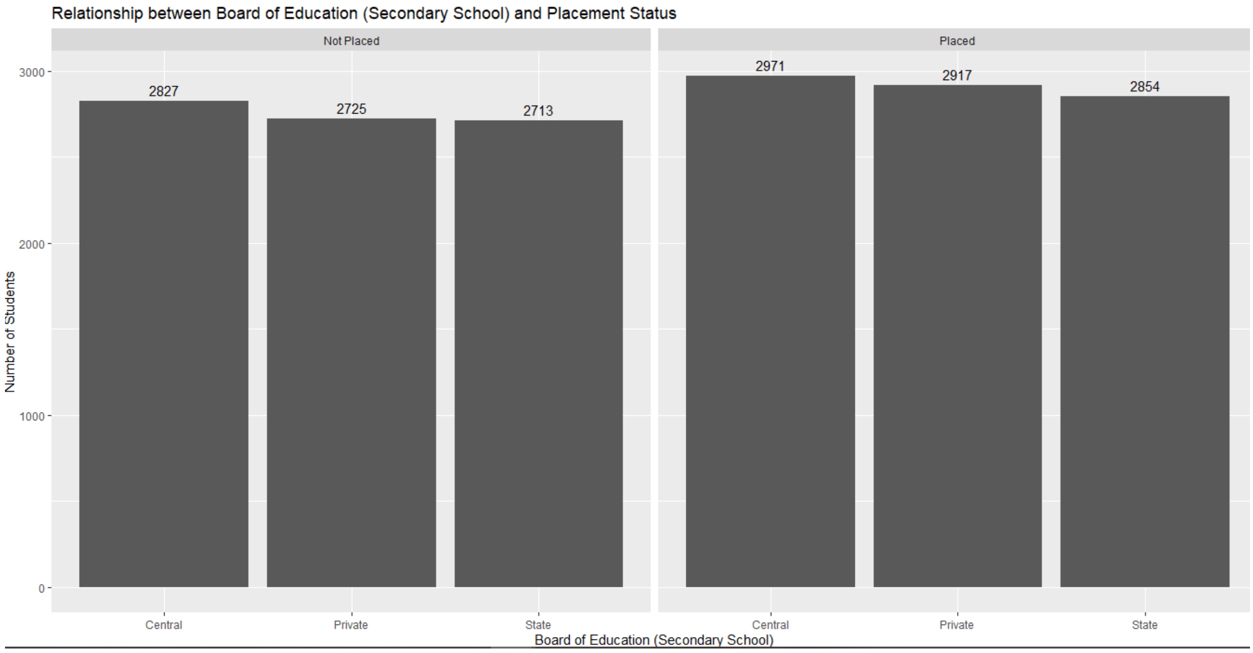


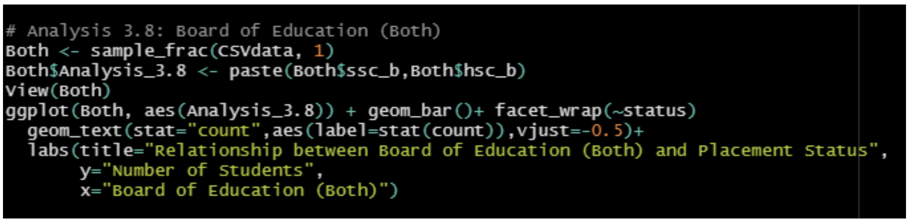
Figure 3.7.2: Relationship between Board of Education (Secondary School) and Placement Status

	Placed	Not Placed	Total Students	Placed Percentage
Central	2971	2827	5798	51.24%
Private	2917	2725	5642	51.70%
State	2854	2713	5567	51.27%

In conclusion, student receive fair chance in placement regardless any board of education chosen by the student for secondary school.

Analysis 3.8: Does Board of Education (Both) affect placement status?

Source Code



```
# Analysis 3.8: Board of Education (Both)
Both <- sample_frac(CSVdata, 1)
Both$Analysis_3.8 <- paste(Both$ssc_b,Both$hsc_b)
View(Both)
ggplot(Both, aes(Analysis_3.8)) + geom_bar()+ facet_wrap(~status)
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Board of Education (Both) and Placement Status",
       y="Number of Students",
       x="Board of Education (Both)")
```

Figure 3.8.1: Using paste function to create new column and ggplot to create bar chart graph

The source code above is applied with paste function to combine both Board of Education in Secondary School Education and Board of Education in Higher Secondary Education to form a new column (Analysis_3.8). Then, ggplot function is used to plot a bar chart graph especially the geom_bar function to analyses the relationship between combination of Board of Education and placement status. The x-axis displays the Board of Education (Both) while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

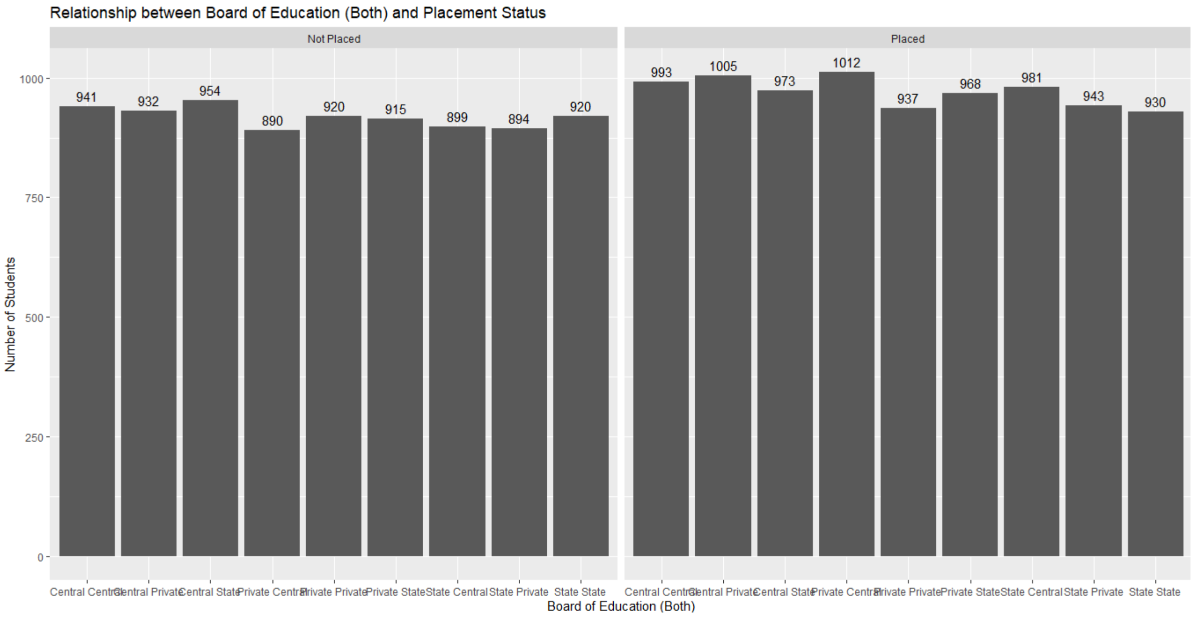
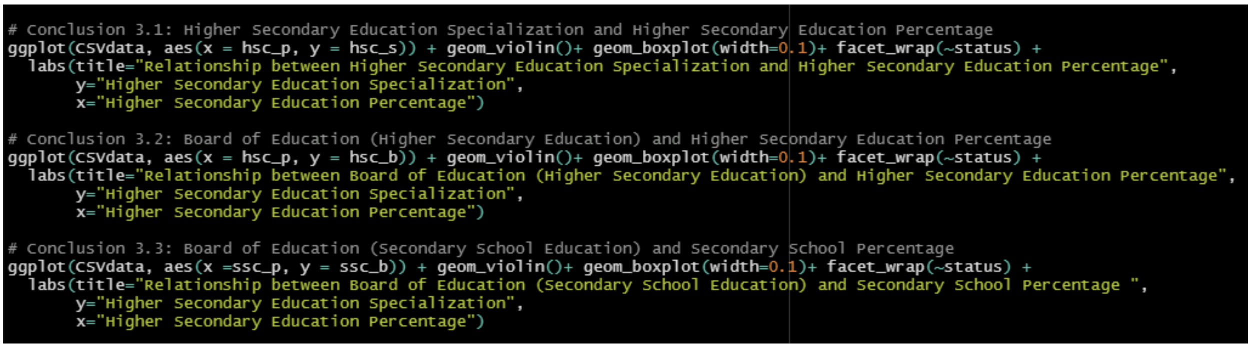


Figure 3.8.2: Relationship between Board of Education (Both) and Placement Status

Based on the bar chart shown in the Figure 3.8.2, the output of the relationship between Board of Education (Both) and Placement Status are in similar range as student will still receive percentage between 50 to 90 marks in their studies and still get a fair chance to receive placement regardless any percentage that they score during the studies. This analysis also has relationship connected with analysis 3.2 until analysis 3.5.

Conclusion for Question 3

Based on the various analysis conducted in Question 3, there are another two important clues has been discovered which nearly identical to the discover in Question 2: The first clue is that student has nearly fair chance to receive placement as the range is nearly equal regardless any range of higher secondary education and secondary education percentage was scored by the student (>= 50), regardless of any board of education is chosen and the combination of board of education. The second clue is that student studied any higher secondary education specialization also has fair chance to receive placement regardless any percentage student achieved.



```
# Conclusion 3.1: Higher Secondary Education Specialization and Higher Secondary Education Percentage
ggplot(csvdata, aes(x = hsc_p, y = hsc_s)) + geom_violin()+ geom_boxplot(width=0.1)+ facet_wrap(~status) +
  labs(title="Relationship between Higher Secondary Education Specialization and Higher Secondary Education Percentage",
       y="Higher Secondary Education Specialization",
       x="Higher Secondary Education Percentage")

# Conclusion 3.2: Board of Education (Higher Secondary Education) and Higher Secondary Education Percentage
ggplot(csvdata, aes(x = hsc_p, y = hsc_b)) + geom_violin()+ geom_boxplot(width=0.1)+ facet_wrap(~status) +
  labs(title="Relationship between Board of Education (Higher Secondary Education) and Higher Secondary Education Percentage",
       y="Higher Secondary Education Specialization",
       x="Higher Secondary Education Percentage")

# Conclusion 3.3: Board of Education (Secondary School Education) and Secondary School Percentage
ggplot(csvdata, aes(x =ssc_p, y = ssc_b)) + geom_violin()+ geom_boxplot(width=0.1)+ facet_wrap(~status) +
  labs(title="Relationship between Board of Education (Secondary School Education) and Secondary School Percentage ",
       y="Higher Secondary Education Specialization",
       x="Higher Secondary Education Percentage")
```

Figure 3.9.1 – Source Code to create violin plot and boxplot

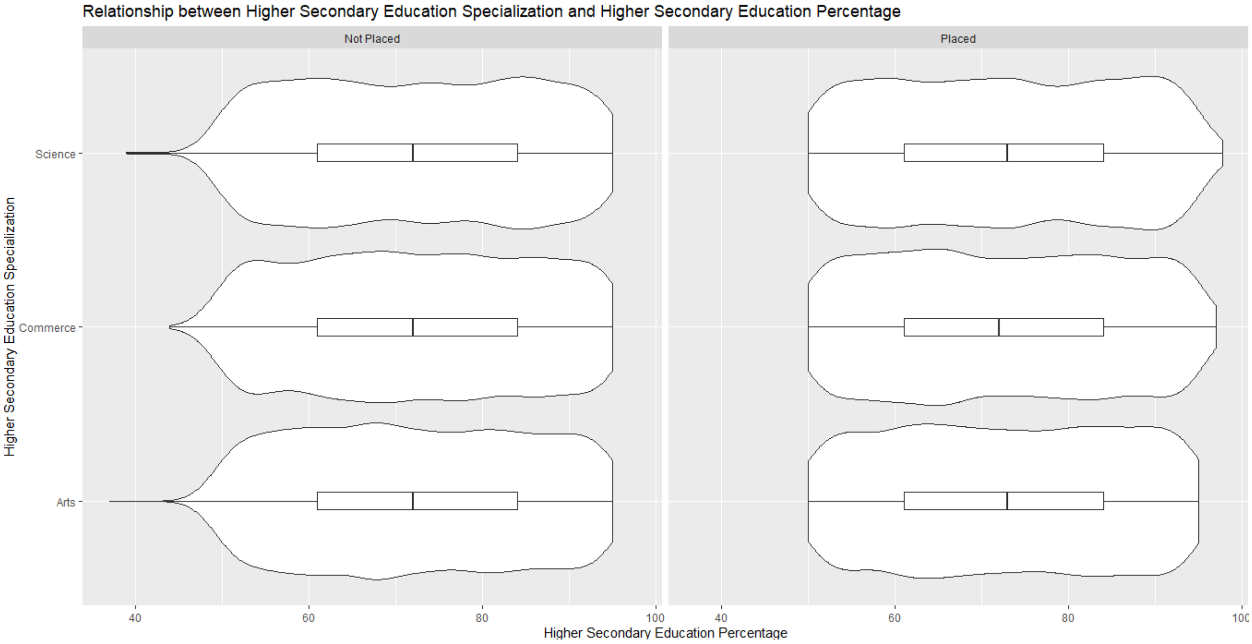


Figure 3.9.2 – Violin plot and Box plot of Relationship between Higher Secondary Education Specialization and Higher Secondary Education Percentage

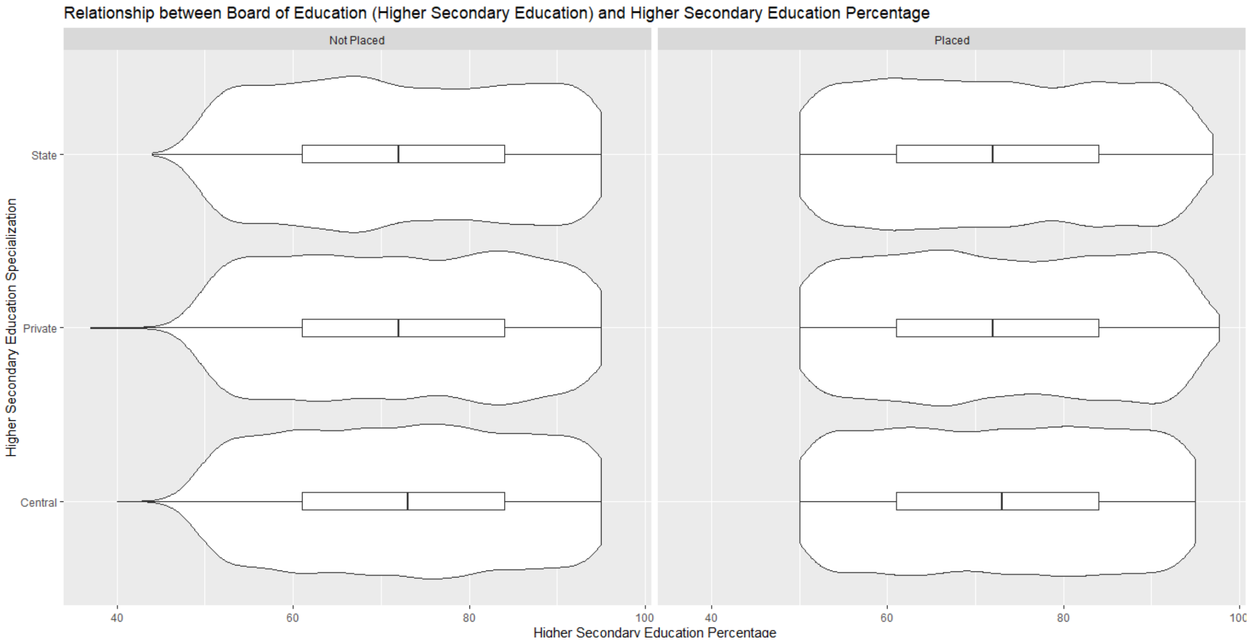


Figure 3.9.3 – Violin plot and Box plot of Relationship between Board of Education (Higher Secondary Education) and Higher Secondary Education Percentage

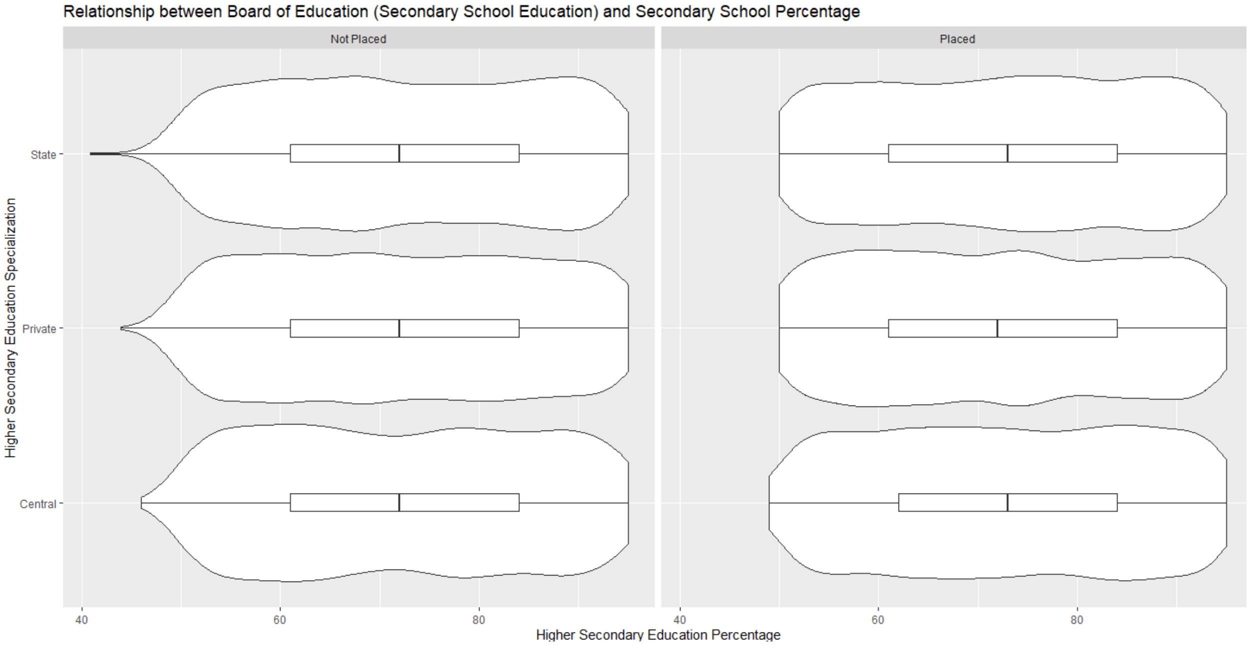


Figure 3.9.4 – Violin plot and Box plot of Relationship between Board of Education (Secondary School Education) and Secondary School Percentage

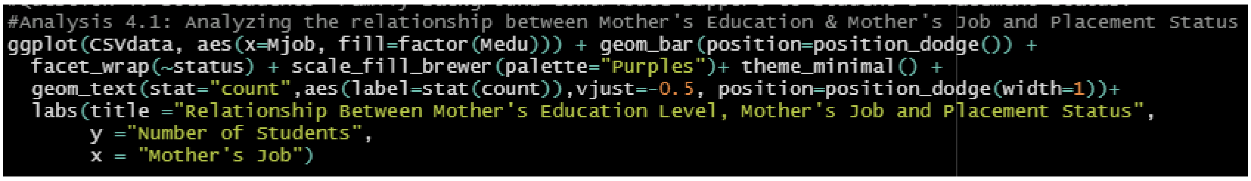
Based on the source code in Figure 3.9.1, the geom_violin and box_plot function had been utilized to construct violin plot and box plot in order to investigate the distribution of the data and the graph statistics. According to Figure 3.9.2 and Figure 3.9.3, the size and distribution of the violin plot and the median in each boxplot are nearly identical to each other (however any student who score below 50 percent will confirm rejected as shown in “Not Placed” graph). Thus, this explains that student who study different specialization in Higher Secondary Education and Board of Education in Higher Secondary Education remain having nearly fair chances in receiving placement regardless any range of percentage is being scored by the students as long as above 50 marks.

According to Figure 3.9.4, the size and distribution of the violin plot and the median in the boxplot in both “Placed” and “Not Placed” graph are nearly same ((however any student who score below 50 percent will confirm rejected as shown in “Not Placed” graph). Thus, this explains that student who study different Board of Education in Secondary School Education remain having nearly fair chances in receiving placement regardless any range of percentage is being scored by the students as long as above 50 marks.
However, more question and analysis should be conducted for further detail investigation in order to discover the other factor’s condition in receiving placement.

Question 4: Does Students' Family Background contribute support to Student's Placement Status?

Analysis 4.1 - Find the relationship between Mother Education & Mother’s Job and Placement Status

Source Code



```
#Analysis 4.1: Analyzing the relationship between Mother's Education & Mother's Job and Placement Status
ggplot(CSVdata, aes(x=Mjob, fill=factor(Medu))) + geom_bar(position=position_dodge()) +
  facet_wrap(~status) + scale_fill_brewer(palette="Purples")+ theme_minimal() +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5, position=position_dodge(width=1))+
  labs(title ="Relationship Between Mother's Education Level, Mother's Job and Placement Status",
       y ="Number of Students",
       x = "Mother's Job")
```

Figure 4.1.1: Construct bar chart by using ggplot function

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between Mother’s Education & Mother’s Job and placement status. The x-axis displays the Mother’s Job while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart while the color is assign based on the factors in Mother’s Education level through fill to enhance the presentation of data.

Data Visualization

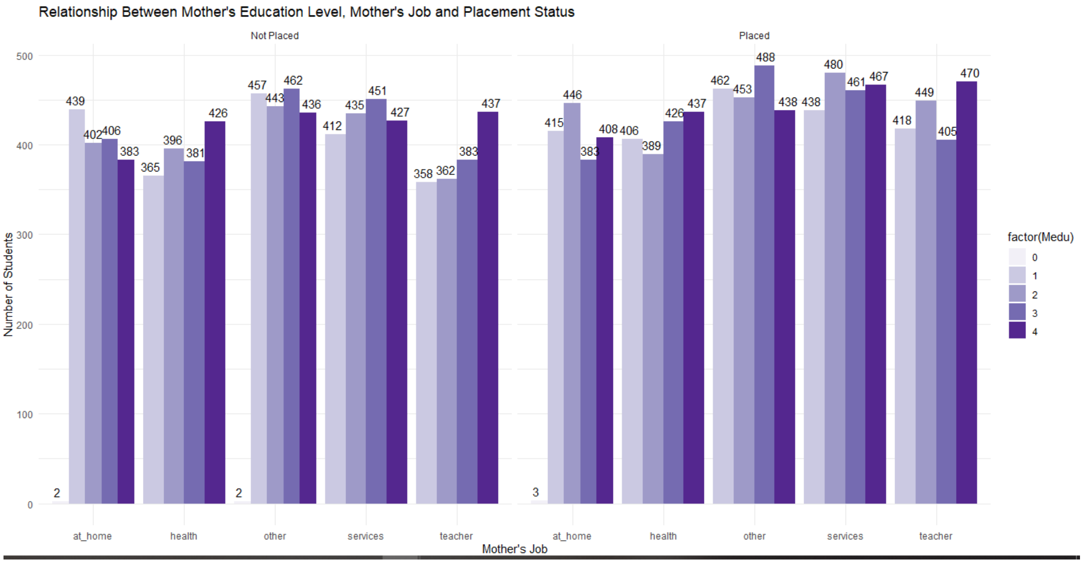


Figure 4.1.2: Relationship Between Mother’s Education Level, Mother’s Job and Placement Status

Despite that Figure 4.1.2 looks complicated, it actually shows that same mother’s education level in different mother’s job has relatively close chance for student to receive placement if detail calculation is performed (For Example: Mother’s Education Level = 1): ~

at home = **48.59%**, health = **52.66 %.** Other = **50.27%,** Services = **51.53%,** teacher = **53.87%**

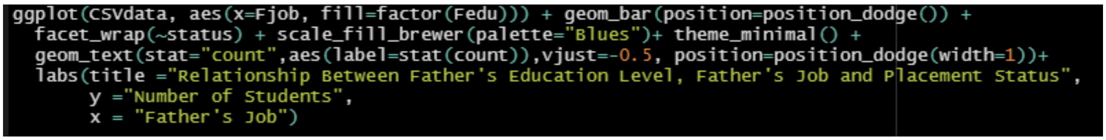
At the same, it also indicated that different mother's education level in the same mother's job also share quite close percentage for student to receive placement. (For Example: Mother's Job = Teacher): ~

Medu 1 = **53.87 %,** Medu 2 = **55.36%,** Medu 3 = **51.40%,** Medu 4 = **51.82%**

Thus, each student has relatively fair chance to receive placement regardless of mother's education level and mother's job.

Analysis 4.2 - Find the relationship between Father Education & Father's Job and Placement Status

Source Code



```
ggplot(csvdata, aes(x=Fjob, fill=factor(Fedu))) + geom_bar(position=position_dodge()) +
  facet_wrap(~status) + scale_fill_brewer(palette="Blues")+ theme_minimal() +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5, position=position_dodge(width=1))+
  labs(title ="Relationship Between Father's Education Level, Father's Job and Placement Status",
       y ="Number of Students",
       x = "Father's Job")
```

Figure 4.2.2: Construct bar chart by using ggplot function

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between Father's Education & Father's Job and placement status. The x-axis displays the Father's Job while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart while the color is assign based on the factors in Father's Education level through fill to enhance the presentation of data.

Data Visualization

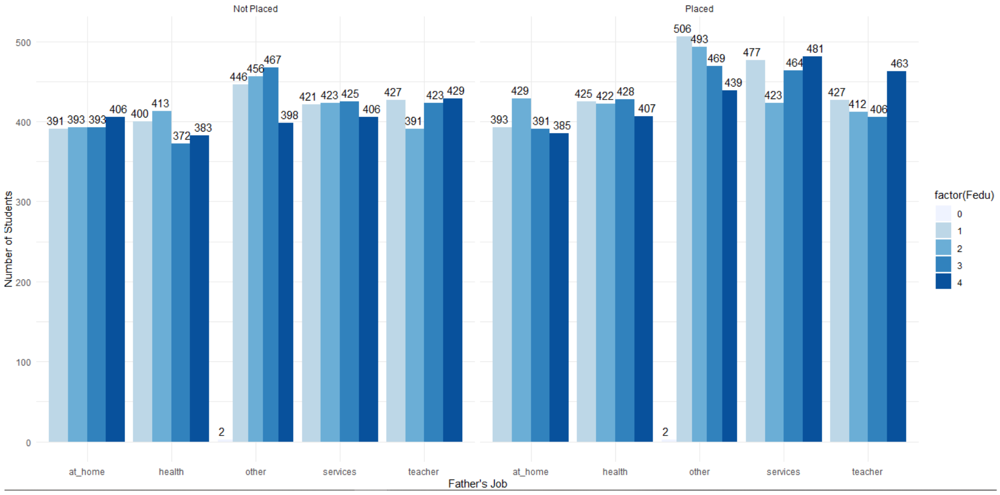
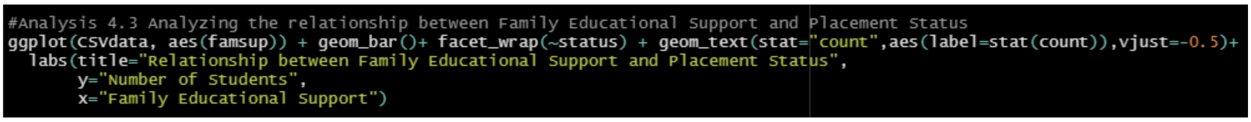


Figure 4.2.2: Relationship Between Father's Education Level & Father's Job and Placement Status

Based on Figure 4.2.2, the statistics is using the same concept similar to Analysis 4.1 as it shows that same father's education level in different father's job has relatively close chance for student to receive placement if detail calculation. At the same, it also indicated that different father's education level in the same father's job also share quite close percentage for student to receive placement. Thus, each student has relatively fair chance to receive placement regardless of father's education level and father's job.

Analysis 4.3: Analyzing the relationship between Family Educational Support and Placement Status

Source Code



```
#Analysis 4.3 Analyzing the relationship between Family Educational Support and Placement Status
ggplot(CSVdata, aes(famsup)) + geom_bar()+ facet_wrap(~status) + geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Family Educational Support and Placement Status",
       y="Number of Students",
       x="Family Educational Support")
```

Figure 4.3.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between family educational support and placement status. The x-axis displays condition of family educational support while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

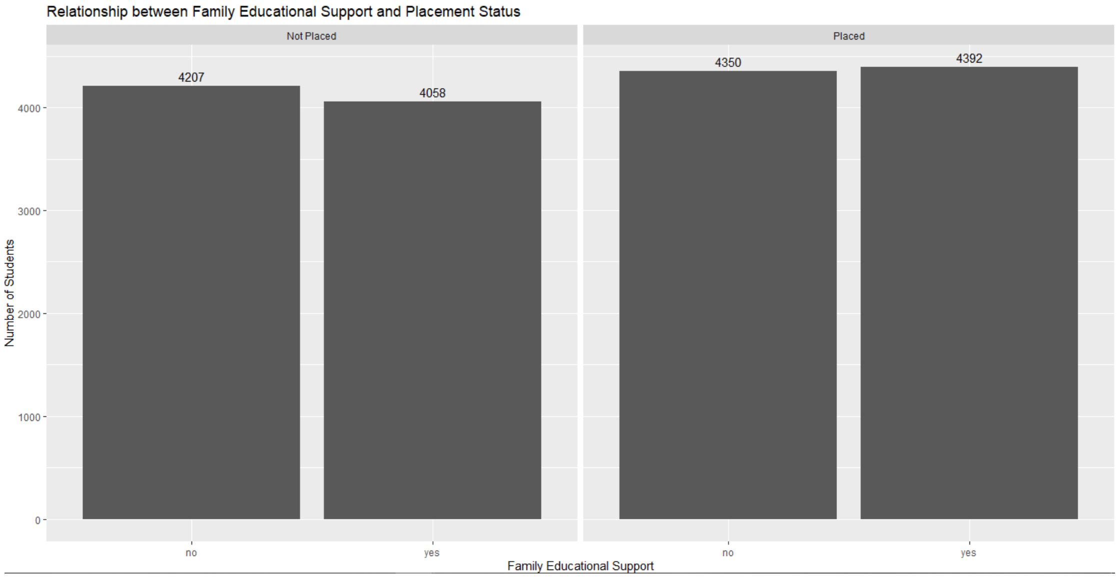
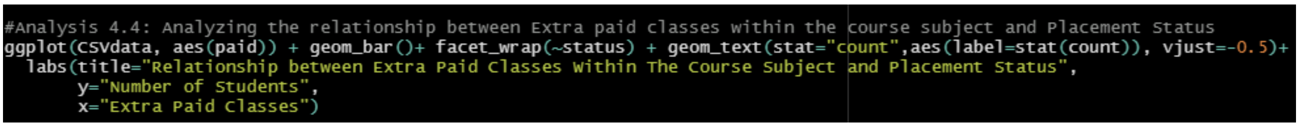


Figure 4.3.2: Relationship Family Educational Support and Placement Status

Based on Figure 4.3.2, students who has family educational support has slightly higher chance received placement then the student who does not has family educational support. Besides that, student who has family educational support has lower chance being rejected. It is possible that various factors cause student who does not receive family educational support has higher chance not placed. Thus, more detail analysis is suggested to be conducted.

Analysis 4.4: Analyzing the relationship between Extra paid classes within the course subject and Placement Status

Source Code



```
#Analysis 4.4: Analyzing the relationship between Extra paid classes within the course subject and Placement Status
ggplot(CSVdata, aes(paid)) + geom_bar()+ facet_wrap(~status) + geom_text(stat="count",aes(label=stat(count)), vjust=-0.5)+
  labs(title="Relationship between Extra Paid Classes Within The Course Subject and Placement Status",
       y="Number of Students",
       x="Extra Paid Classes")
```

Figure 4.4.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between extra paid classes within the course subject and placement status. The x-axis displays condition of extra paid classes within the course subject while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

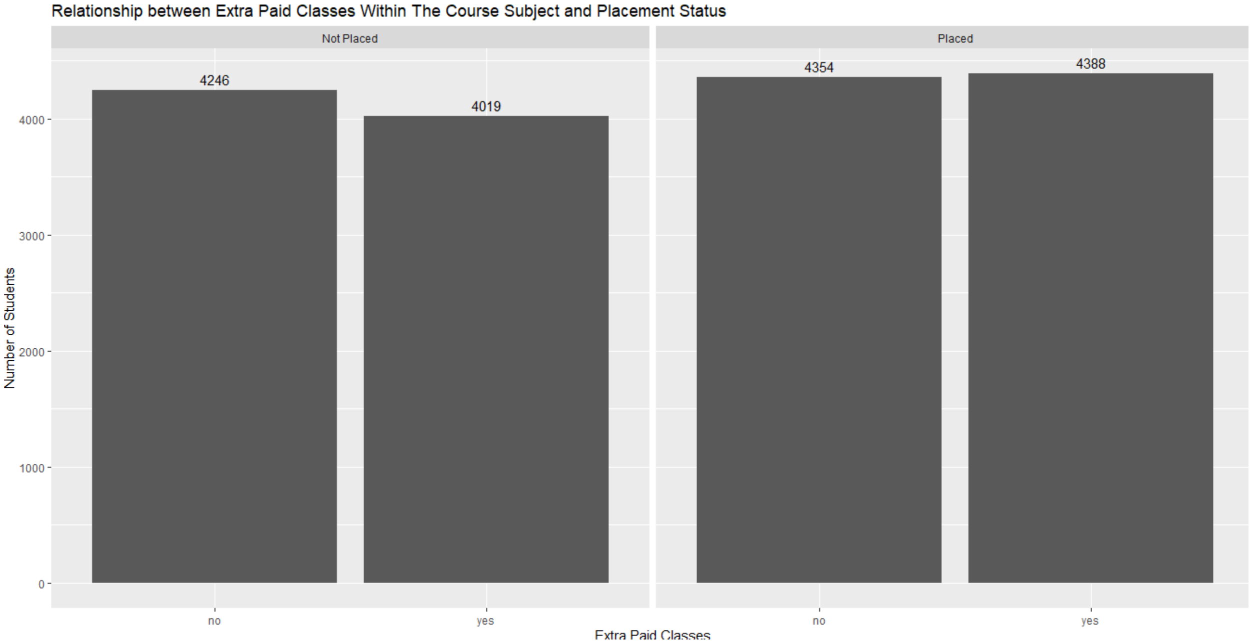
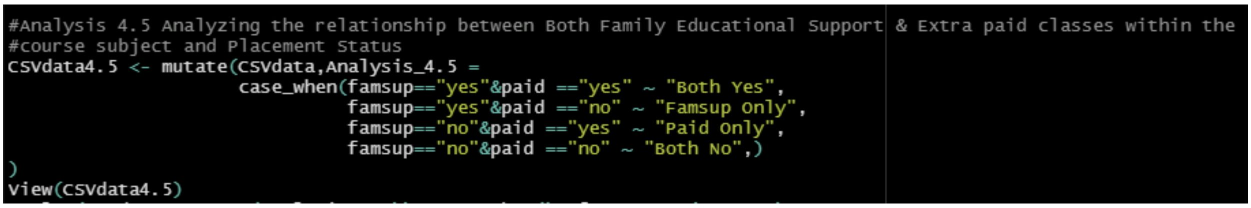


Figure 4.4.2: Relationship Extra Paid Classes within the Course Subject and Placement Status

Based on Figure 4.4.2, students who has taken extra paid classes within the course subject has slightly higher chance received placement then the student who does taken extra paid classes within the course subject and also lower chance being rejected. It is possible that other various factors cause student who does not taken extra paid classes within the course subject has higher chance not placed. Thus, more detail analysis is suggested to be conducted.

Analysis 4.5: Analyzing the relationship between Both Family Educational Support & Extra paid classes within the course subject and Placement Status

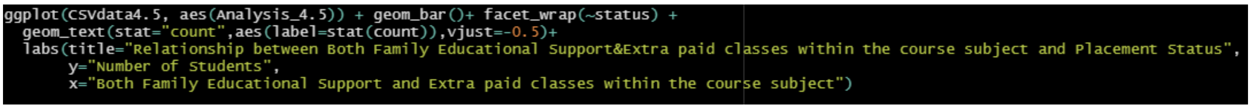
Source Code



```
#Analysis 4.5 Analyzing the relationship between Both Family Educational Support & Extra paid classes within the
#course subject and Placement Status
CSVdata4.5 <- mutate(CSVdata,Analysis_4.5 =
                     case_when(famsup=="yes"&paid =="yes" ~ "Both Yes",
                               famsup=="yes"&paid =="no" ~ "Famsup Only",
                               famsup=="no"&paid =="yes" ~ "Paid Only",
                               famsup=="no"&paid =="no" ~ "Both No",)
)
View(CSVdata4.5)
```

Figure 4.5.1: Step 1: Filter and duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which combination of both family educational support & extra paid classes within the course subject is better in order to secure higher chance to receive placement. According to the source code above, data manipulation and transformation concept are applied to duplicate a new data frame (CSVdata4.5) to avoid alter the original data frame by using sample_frac function alongside with filter function. To create new column (Analysis_4.5), mutate function is applied with case_when function to implement conditional logic like if/else and if/else if/else.



```
ggplot(CSVdata4.5, aes(Analysis_4.5)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Both Family Educational Support&Extra paid classes within the course subject and Placement Status",
       y="Number of Students",
       x="Both Family Educational Support and Extra paid classes within the course subject")
```

Figure 4.5.2: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_4.5) and placement status. The x-axis displays the condition of range of both family educational support & extra paid classes within the course subject while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

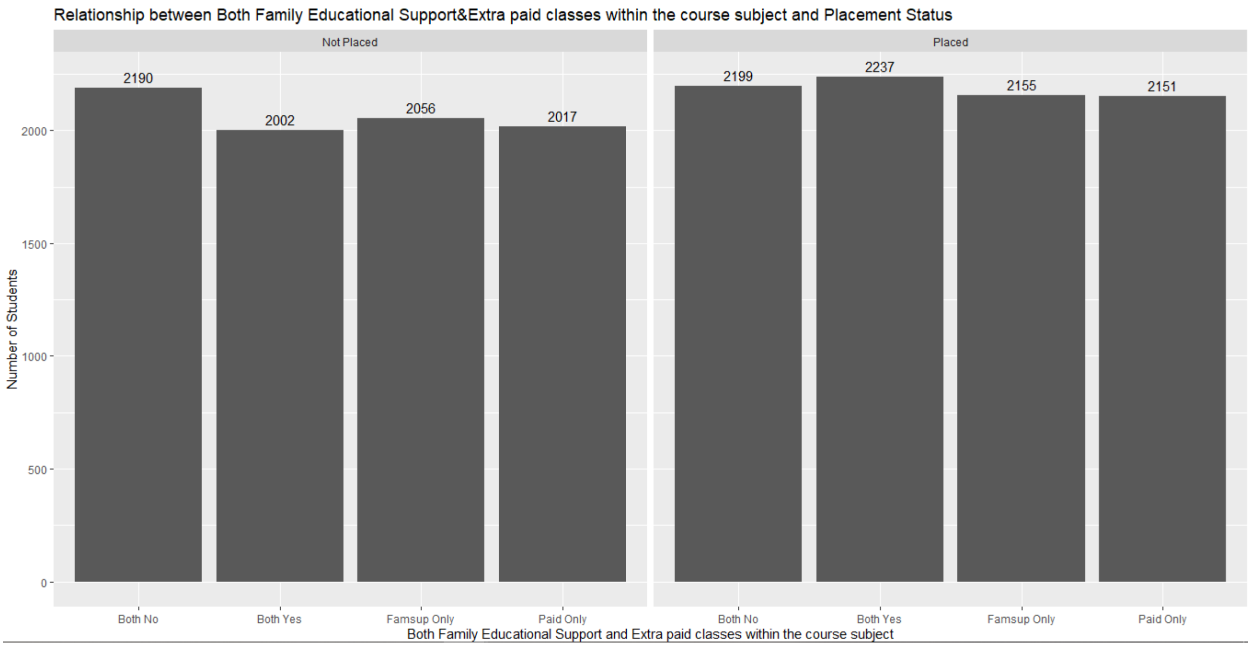


Figure 4.5.3: Relationship between Both Family Educational Support&Extra paid classes within the course subject and Placement Status

Based on Figure 4.5.3, students who has both family educational support and extra paid classes within the course subject has slightly higher chance received placement then other students and also lower chance being rejected. It is possible that other various factors cause student who does not has both family educational support and extra paid classes within the course subject has higher chance not placed. Thus, more detail analysis is suggested to be conducted.

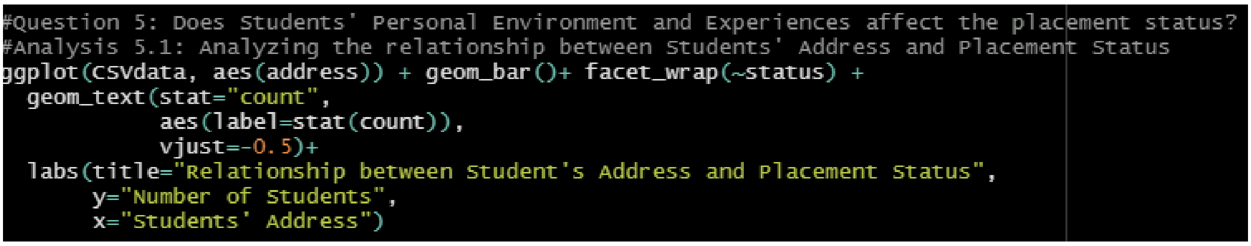
Conclusion for Question 4

After conducting various analysis related to family factors in Question 4, all the result produced in the data visualization shows that the percentage range of receiving placement is largely similar (mostly approximately between 48% to 53% occasion which similar to the conclusion in various analysis from Question 1 until Question 3). As family factors in Question 4 can have connection with student education records in Question 2 to 3, it is discovered that family factors are also one of the factors that can influence student placement status in various ways and produced different input and outcomes that decide the student will receive placement or not in overall.

Question 5: Does Students' Personal Environment and Experiences affect the placement status?

Analysis 5.1: Analyzing the relationship between Students' Address and Placement Status

Source Code



```
#Question 5: Does Students' Personal Environment and Experiences affect the placement status?
#Analysis 5.1: Analyzing the relationship between Students' Address and Placement Status
ggplot(CSVdata, aes(address)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",
            aes(label=stat(count)),
            vjust=-0.5)+
  labs(title="Relationship between Student's Address and Placement Status",
       y="Number of Students",
       x="Students' Address")
```

Figure 5.1.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between student address and placement status. The x-axis displays student address while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

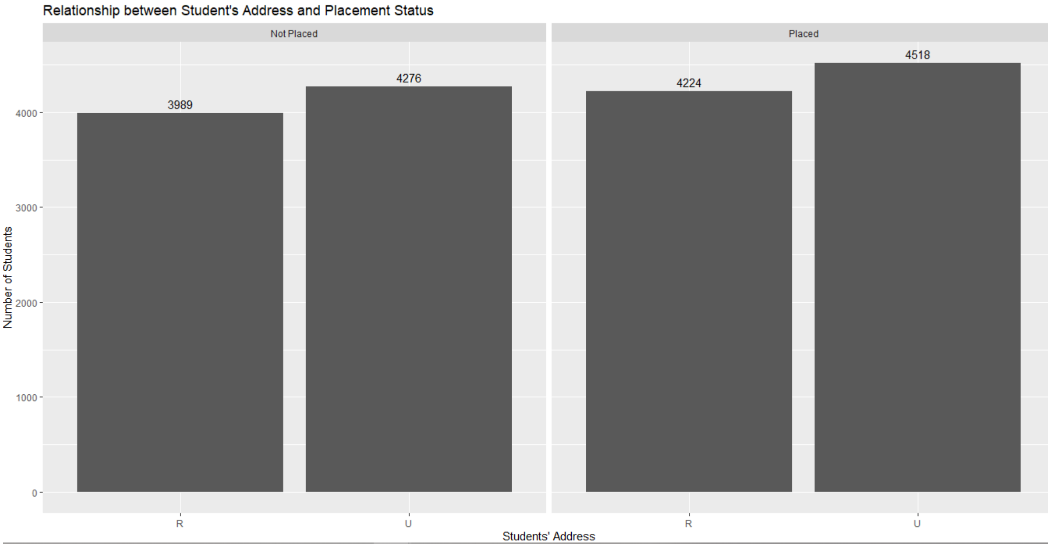


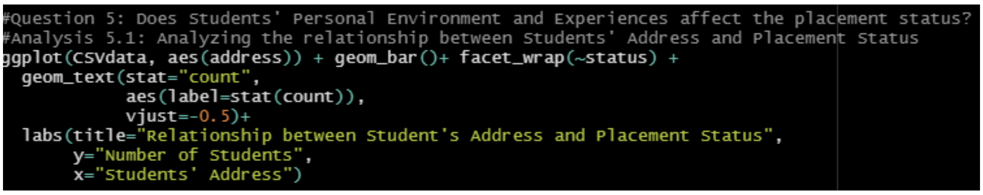
Figure 5.1.2: Relationship between Student Address and Placement Status

	Placed	Not Placed	Total Students	Placed Percentage
Urban (U)	4518	4276	8794	51.38%
Rural (R)	4224	3989	8213	51.43%

In conclusion, student receive fair chance in placement regardless any from any address as the range of placed percentage is very close. Thus, more analysis needs to be conducted

Analysis 5.2: Analyzing the relationship between Internet access and Placement Status

Source Code



```
#Question 5: Does Students' Personal Environment and Experiences affect the placement status?
#Analysis 5.1: Analyzing the relationship between Students' Address and Placement Status
ggplot(CSVdata, aes(address)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",
            aes(label=stat(count)),
            vjust=-0.5)+
  labs(title="Relationship between Student's Address and Placement Status",
       y="Number of Students",
       x="Students' Address")
```

Figure 5.2.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between Internet access and placement status. The x-axis displays condition of internet access while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

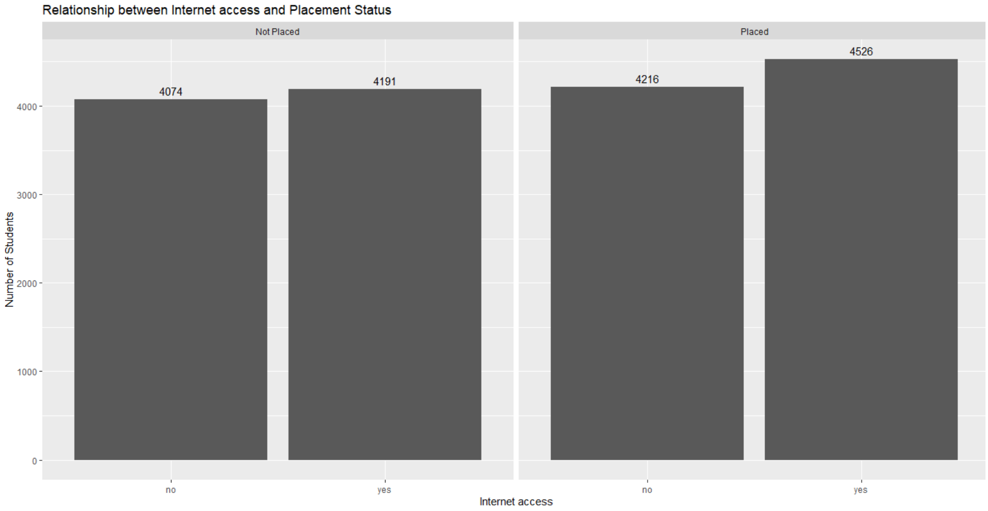


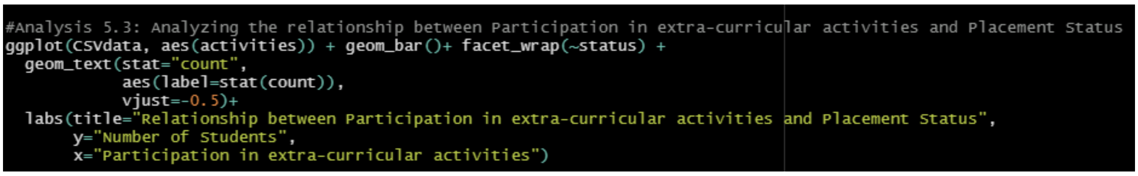
Figure 5.2.2: Relationship between Internet Access and Placement Status

	Placed	Not Placed	Total Students	Placed Percentage
Yes	4526	4191	8717	51.92%
No	4216	4074	8290	50.86%

In conclusion, student receive rather fair chance in placement regardless student has internet access or not as the range of placed percentage is very close. However, student with that has internet access has slightly higher chance received placement and is possible due to other factor's support.

Analysis 5.3: Analyzing the relationship between Participation in extra-curricular activities and Placement Status

Source Code



```
#Analysis 5.3: Analyzing the relationship between Participation in extra-curricular activities and Placement Status
ggplot(CSVdata, aes(activities)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",
            aes(label=stat(count)),
            vjust=-0.5)+
  labs(title="Relationship between Participation in extra-curricular activities and Placement Status",
       y="Number of Students",
       x="Participation in extra-curricular activities")
```

Figure 5.3.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between Participation in extra-curricular activities and placement status. The x-axis displays condition of participation in extra-curricular activities while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

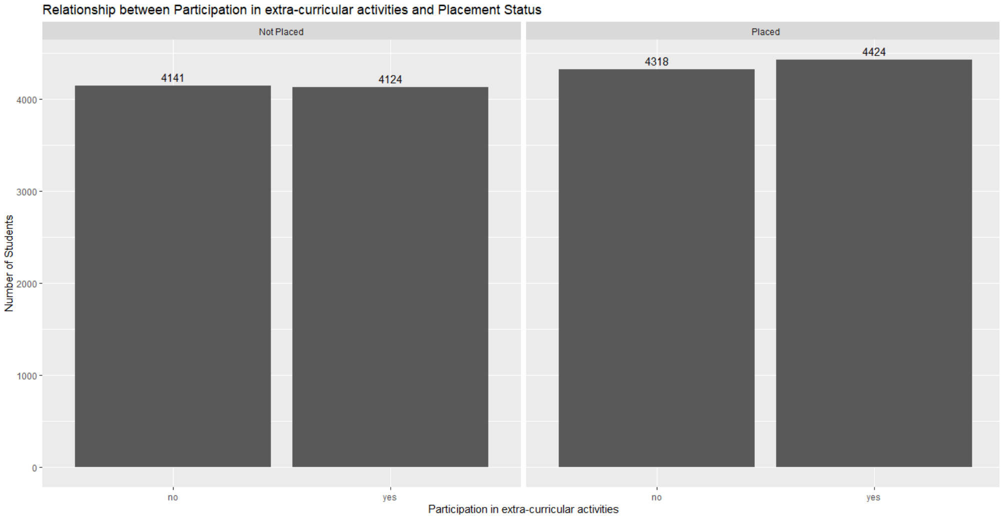


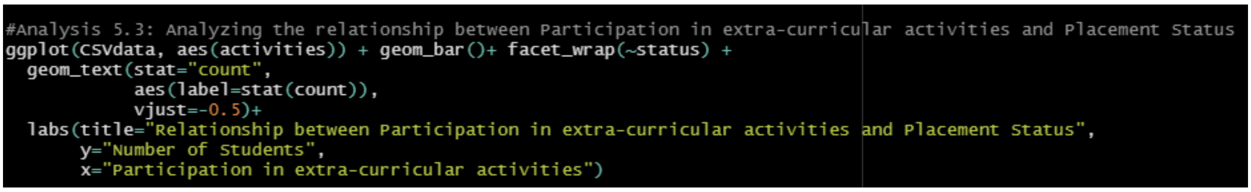
Figure 5.3.2: Relationship between Participation in extra-curricular activities and Placement Status

	Placed	Not Placed	Total Students	Placed Percentage
Yes	4424	4124	8548	51.75%
No	4318	4141	8459	51.05%

In conclusion, student receive rather fair chance in placement regardless student has participated in extra-curricular activities or not due to the range of placed percentage is very close. However, student with that active has slightly higher chance received placement and is possible due to other factor’s support.

Analysis 5.4: Analyzing the relationship between Student's Working Experience and Placement Status

Source Code



```
#Analysis 5.3: Analyzing the relationship between Participation in extra-curricular activities and Placement Status
ggplot(CSVdata, aes(activities)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",
            aes(label=stat(count)),
            vjust=-0.5)+
  labs(title="Relationship between Participation in extra-curricular activities and Placement Status",
       y="Number of Students",
       x="Participation in extra-curricular activities")
```

Figure 5.4.1: Using ggplot to create bar chart graph

The source code above is applied with ggplot function to plot a bar chart graph especially the geom_bar function to analyses the relationship between student's working experience and placement status. The x-axis displays condition of student's working experience while the y-axis displays the number of students who receive and not receive placement. The facet_wrap function is applied in order to separate into multiple graphs based on the factors inside the status column and the geom_text function is applied to display the number of students in the bar chart.

Data Visualization

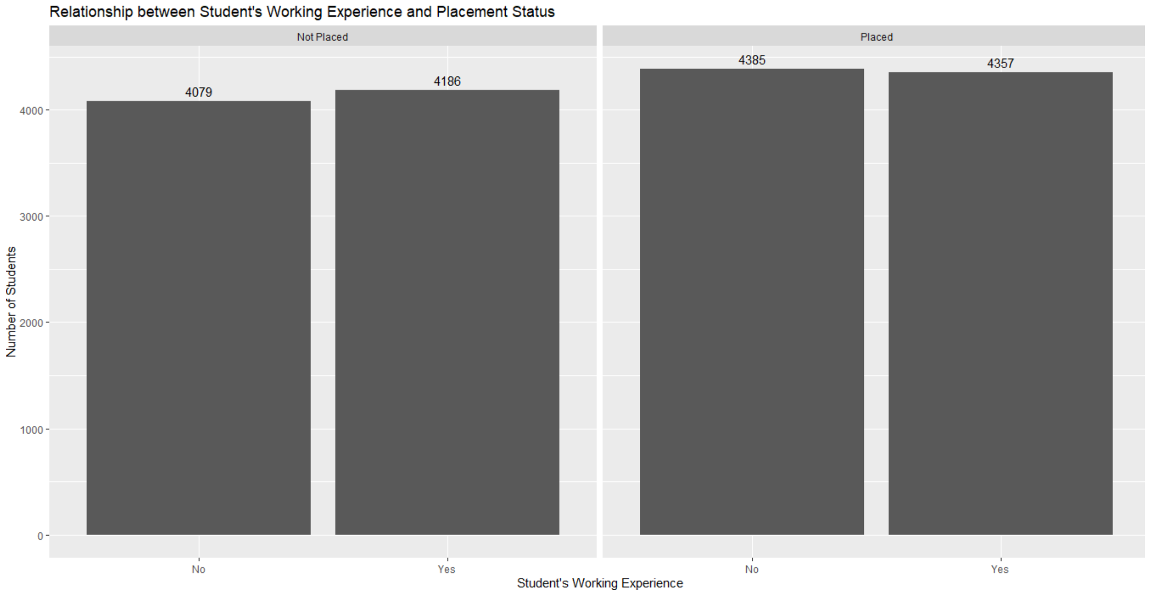


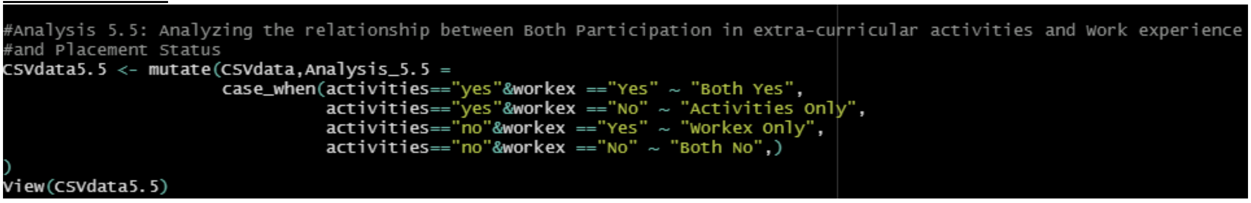
Figure 5.4.2: Relationship between Student’s Working Experience and Placement Status

	Placed	Not Placed	Total Students	Placed Percentage
Yes	4357	4186	8543	50.82%
No	4385	4079	8464	51.81%

In conclusion, student receive rather fair chance in placement regardless student has participated in extra-curricular activities or not due to the range of placed percentage is very close. However, student with no experience has slightly higher chance received placement is possible due to other factor’s support.

Analysis 5.5: Analyzing the relationship between Both Participation in extra-curricular activities and Work experience and Placement Status

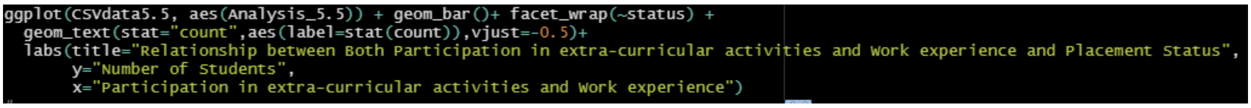
Source Code



```
#Analysis 5.5: Analyzing the relationship between Both Participation in extra-curricular activities and Work experience
#and Placement Status
CSVdata5.5 <- mutate(CSVdata,Analysis_5.5 =
                     case_when(activities=="yes"&workex =="Yes" ~ "Both Yes",
                               activities=="yes"&workex =="No" ~ "Activities Only",
                               activities=="no"&workex =="Yes" ~ "Workex Only",
                               activities=="no"&workex =="No" ~ "Both No",)
)
View(CSVdata5.5)
```

Figure 5.5.1: Step 1: Duplicate data frame and create new column by using mutate and case_when function

This analysis is conducted to investigate which combination of both participation in extra-curricular activities and work experience in order to secure higher chance to receive placement. According to the source code above, the first step is using data manipulation and transformation concept are applied to duplicate a new data frame (CSVdata5.5) to avoid alter the original data frame and create new column (Analysis_5.5) by using mutate function and case_when function to implement conditional logic like if/else and if/else if/ based on the condition input.



```
ggplot(CSVdata5.5, aes(Analysis_5.5)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Both Participation in extra-curricular activities and Work experience and Placement Status",
       y="Number of Students",
       x="Participation in extra-curricular activities and Work experience")
```

Figure 5.5.2: Step 2 – Create bar chart by using ggplot function

The ggplot function is then applied to plot a bar chart graph especially the geom_bar function to analyses the relationship between the new column (Analysis_2.5) and placement status. The x-axis displays the combination of both participation in extra-curricular activities and work experience while the y-axis displays the number of students who receive and not receive placement.

Data Visualization

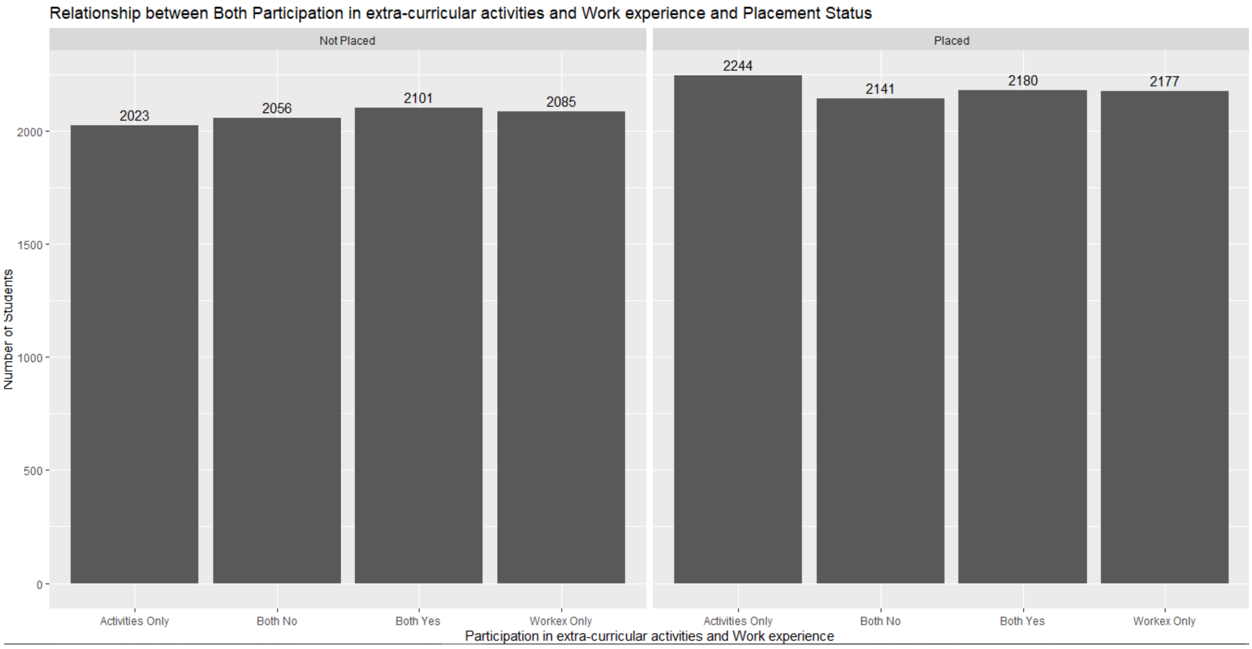


Figure 5.5.3: Relationship between Both Participation in extra-curricular activities and Work experience and Placement Status

	Placed	Not Placed	Total Students	Placed Percentage
Activities Only	2244	2023	4267	52.59%
Both No	2141	2056	4197	51.01%
Both Yes	2180	2101	4281	50.92%
Workex Only	2177	2085	4262	51.08%

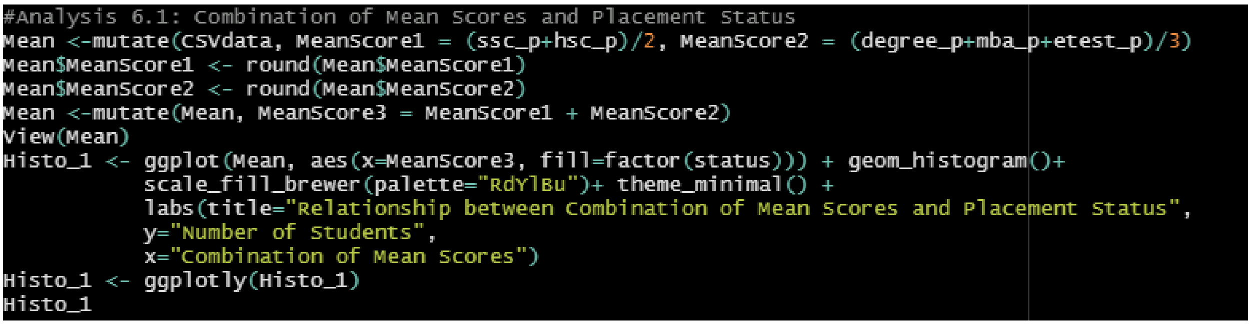
In conclusion, student has fair chance to receive placement regardless they have participated extra-curricular activities or having working experience or both or not as the range of placed percentage is very close.

Conclusion for Question 5

After conducting various analysis related to personal experience factors in Question 5, all the result produced in the data visualization shows that the percentage range of receiving placement is largely similar. Thus, factors in Question 5 shares the nearly same range of placement percentage like Question 1 until Question 4. Thus, it is recommended to conducted one last analysis in order to conclude all Question and Analysis.

Question 6: Does students' overall records affect Student's Placement Status?

Analysis 6.1: Combination of Mean Scores and Placement Status



```
#Analysis 6.1: Combination of Mean Scores and Placement Status
Mean <-mutate(CSVdata, MeanScore1 = (ssc_p+hsc_p)/2, MeanScore2 = (degree_p+mba_p+etest_p)/3)
Mean$MeanScore1 <- round(Mean$MeanScore1)
Mean$MeanScore2 <- round(Mean$MeanScore2)
Mean <-mutate(Mean, MeanScore3 = MeanScore1 + MeanScore2)
View(Mean)
Histo_1 <- ggplot(Mean, aes(x=MeanScore3, fill=factor(status))) + geom_histogram()+
           scale_fill_brewer(palette="RdYlBu")+ theme_minimal() +
           labs(title="Relationship between Combination of Mean Scores and Placement Status",
           y="Number of Students",
           x="Combination of Mean Scores")
Histo_1 <- ggplotly(Histo_1)
Histo_1
```

Figure 6.1.1: Create Histogram by using Data Manipulation, Data Transformation and Data Visualization (ggplot)

According to the source code above, this analysis is conducted to investigate combination of all the scores in order to determine which mean scores has the higher chance receiving placement by using mean concept and assign to duplicate data frame (Mean) through data manipulation and data transformation. Then, ggplot() function is utilize to construct histogram to analyze the overall mean score and the number of student who receive and not receive placement. Last but not least, ggplotly() function is utilize to figure the histogram into interactive and embedding into dash applications to review the histogram in details.

Data Visualization

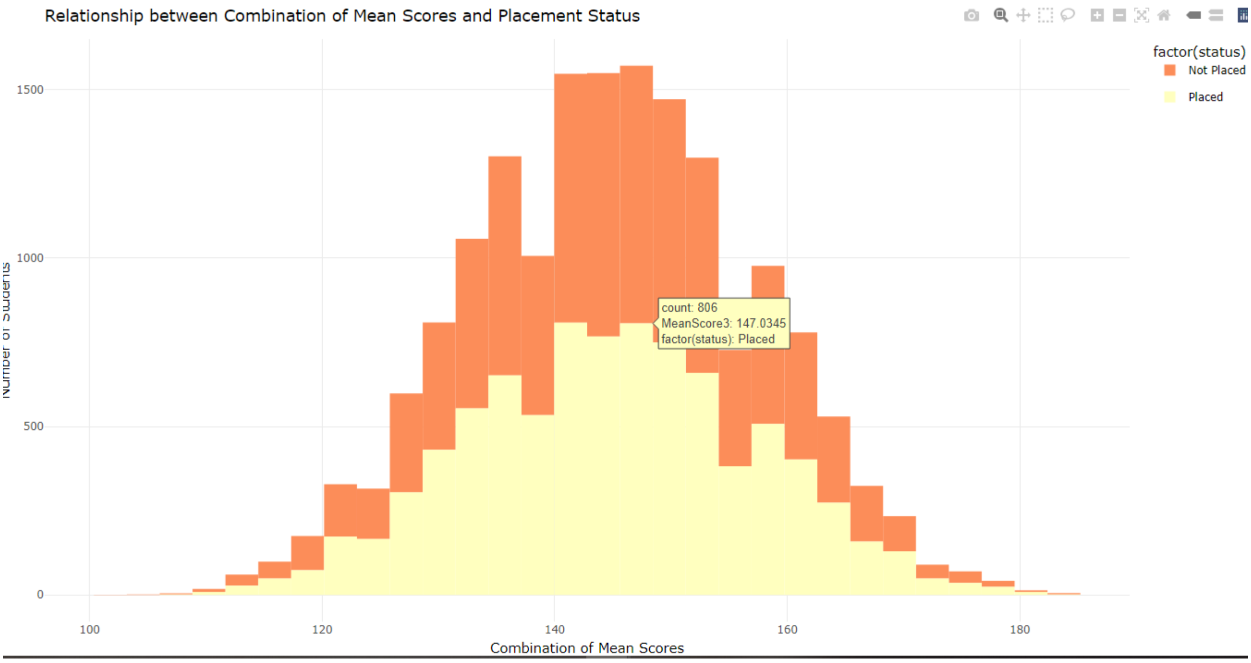
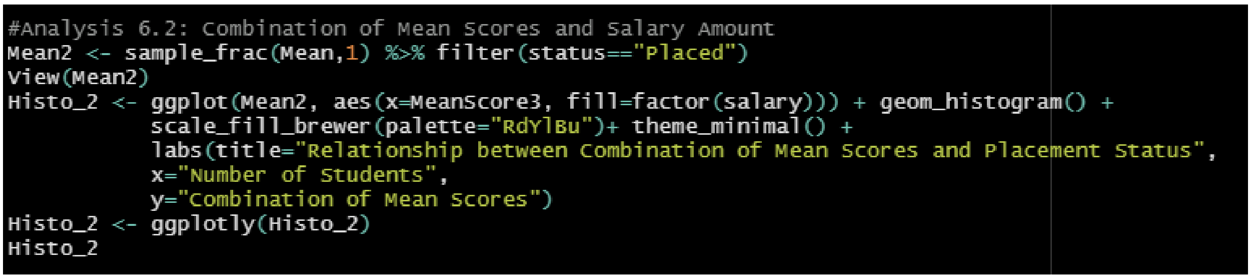


Figure 6.1.2: Histogram about Relationship between Combination of Mean Scores and Placement Status

Based on Figure 6.1.2, the histogram shows that no matter what mean scores that student scores, the chances of receiving placement is nearly the same (approximately 49% to 52%) by using ggplotly to investigate.

Analysis 6.2: Combination of Mean Scores and Salary Amount

Source Code



```
#Analysis 6.2: Combination of Mean Scores and Salary Amount
Mean2 <- sample_frac(Mean,1) %>% filter(status=="Placed")
View(Mean2)
Histo_2 <- ggplot(Mean2, aes(x=MeanScore3, fill=factor(salary))) + geom_histogram() +
           scale_fill_brewer(palette="RdYlBu")+ theme_minimal() +
           labs(title="Relationship between Combination of Mean Scores and Placement Status",
           x="Number of Students",
           y="Combination of Mean Scores")
Histo_2 <- ggplotly(Histo_2)
Histo_2
```

Figure 6.2.1: Create Histogram by using Data Manipulation, Data Transformation and Data Visualization (ggplot)

According to the source code above, this analysis is conducted to investigate combination of all the scores in order to determine which mean scores receive higher amount of salary by using mean concept and assign to duplicate data frame (Mean2) through data manipulation and data transformation. Then, ggplot() function is utilize to construct histogram to analyze the overall mean score and the number of student who receive certain amount of salary. Last but not least, ggplotly() function is utilize to figure the histogram into interactive and embedding into dash applications to review the histogram in details.

Data Visualization

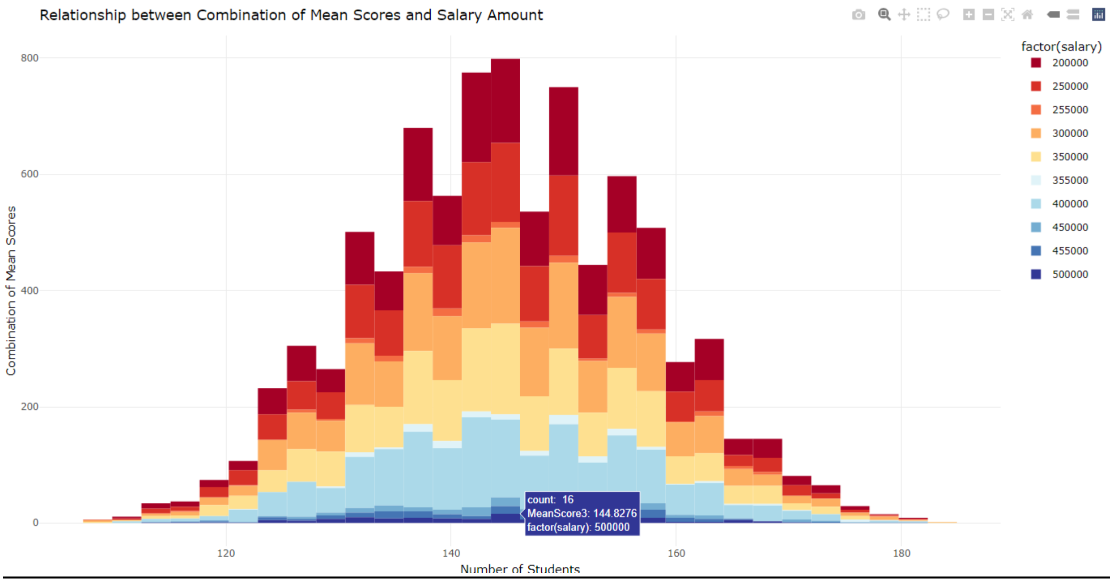
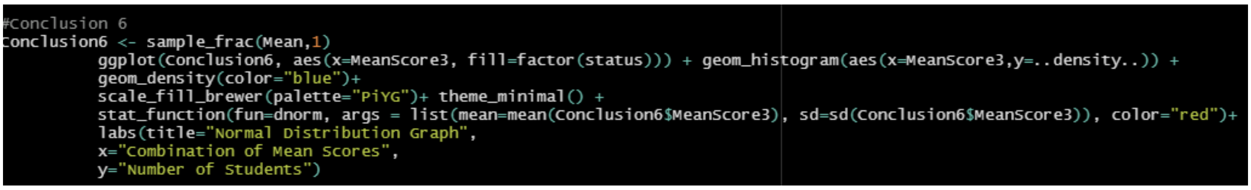


Figure 6.2.2: Histogram about Relationship between Combination of Mean Scores and Salary Amount

Based on Figure 6.2.2, the histogram shows that no matter what mean scores that student scores, the salary amount given has been decided in ratio amount for the student who receive the placement. If the number of students who scored the same mean score is high, the number of students who can receive salary higher salary is increase but in fixed ratio. However, if the number of students who score that particular mean score is low, lesser student can receive high salary amount or even no student can get.

Conclusion for Question 6



```
#Conclusion 6
Conclusion6 <- sample_frac(Mean,1)
        ggplot(Conclusion6, aes(x=MeanScore3, fill=factor(status))) + geom_histogram(aes(x=MeanScore3,y=..density..)) +
        geom_density(color="blue")+
        scale_fill_brewer(palette="PiYG")+ theme_minimal() +
        stat_function(fun=dnorm, args = list(mean=mean(Conclusion6$MeanScore3), sd=sd(Conclusion6$MeanScore3)), color="red")+
        labs(title="Normal Distribution Graph",
        x="Combination of Mean Scores",
        y="Number of Students")
```

Figure 6.3.1: Create Histogram that identify Z-Score and construct Density plot

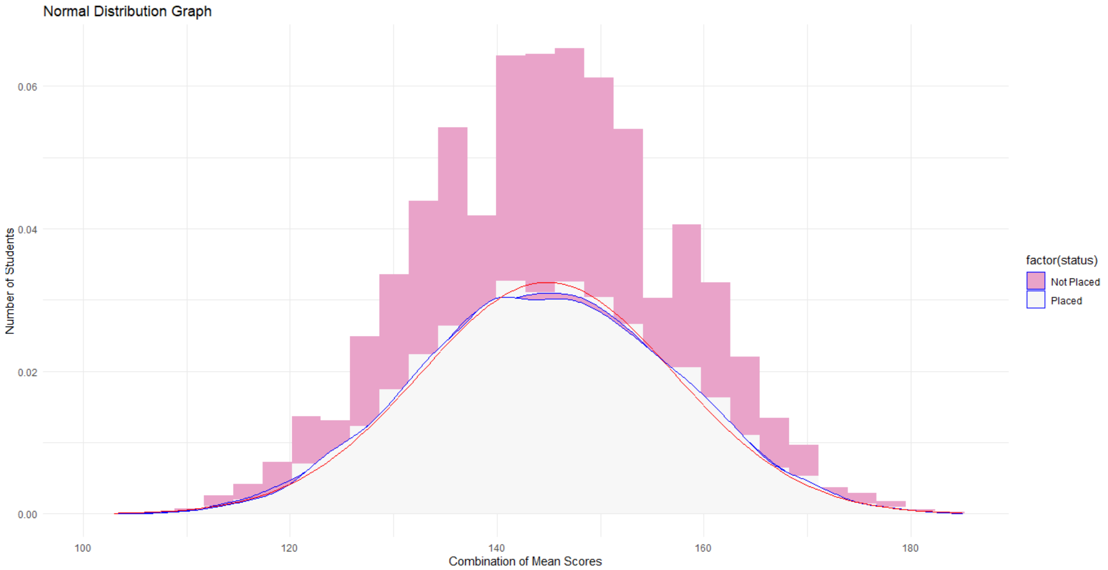


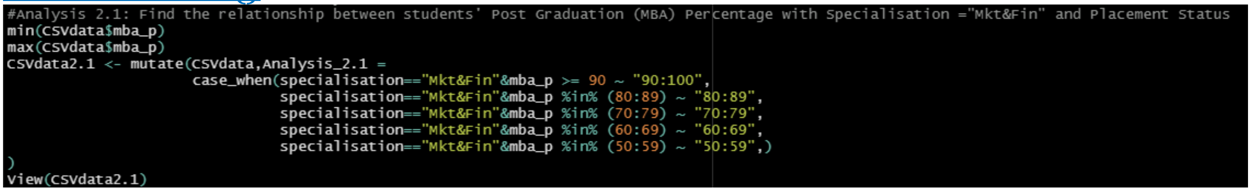
Figure 6.3.2: Normal Distribution Graph

After conducting the last 2 analysis in Question 6, it is discovered that every student is having fair amount of chance to be chosen for placement and salary amount given based on the number and ratio of the student in the mean scores. Figure 6.3.2, a histogram is construct to identify the Z-Score (Red Line) and the distribution of the numeric value (Blue Line). Based on the Figure, the histogram graph is actually normal distribution graph and the Z-Score is located in the center of the graph and its center curve is slightly positive.

Thus, this normal distribution graph proves that the reason why only 51.40% (refer to analysis 1.1) of the students receive placement and all of the factor’s range are nearly similar. This is because employee wants to enroll different type of students, different range of scores, different family background, different personal identity and different personal experience.

4.0 Extra Features

4.1 case_when()

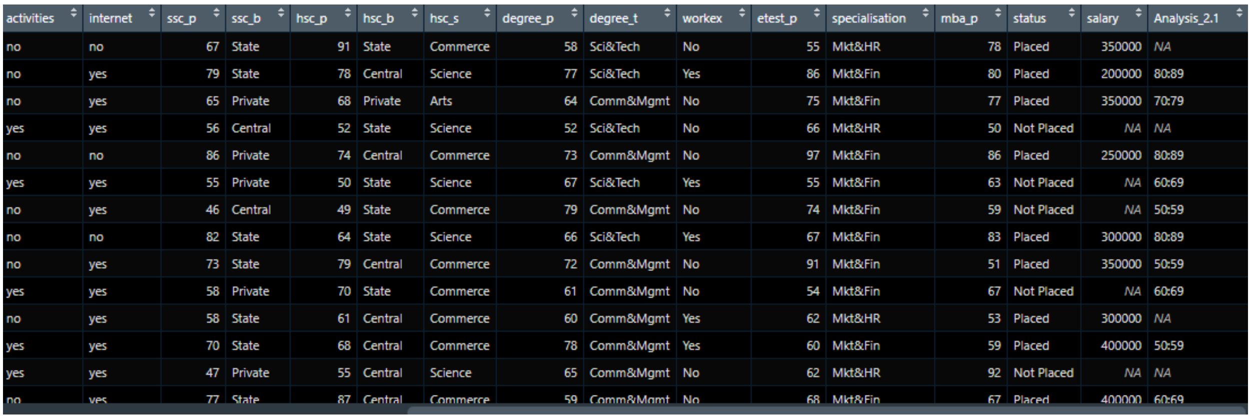


```
#Analysis 2.1: Find the relationship between students' Post Graduation (MBA) Percentage with Specialisation ="Mkt&Fin" and Placement Status
min(CSVdata$mba_p)
max(CSVdata$mba_p)
CSVdata2.1 <- mutate(CSVdata,Analysis_2.1 =
                     case_when(specialisation=="Mkt&Fin"&mba_p >= 90 ~ "90:100",
                               specialisation=="Mkt&Fin"&mba_p %in% (80:89) ~ "80:89",
                               specialisation=="Mkt&Fin"&mba_p %in% (70:79) ~ "70:79",
                               specialisation=="Mkt&Fin"&mba_p %in% (60:69) ~ "60:69",
                               specialisation=="Mkt&Fin"&mba_p %in% (50:59) ~ "50:59",)
)
View(CSVdata2.1)
```

Figure 7: Example of case_ when() function is utilized in Analysis 2.1

The case_when() method is identical to the if/elseif/else statement in other programming language. As a result, it employs a conditional statement to allow the analyst to enter a series of two-sided calculations. Which values correspond to this instance are determined on the left-hand side (LHS). Other the other side, the replacement value is on the right-hand side (RHS). The LHS must be a logical vector to be evaluated. Although the RHS must be reasonable, it must all evaluate to the same type of vector (DataScienceMadeSimple, n.d.).

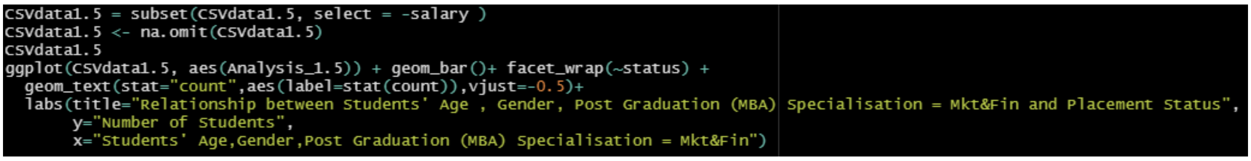
As an example, Figure 7 above shows that case_when() function is implement in Analysis 2.1 in order to employs a conditional statement and assign a new value to the new column with the usage of mutate() function.



activities	internet	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary	Analysis_2.1
no	no	67	State	91	State	Commerce	58	Sci&Tech	No	55	Mkt&HR	78	Placed	350000	NA
no	yes	79	State	78	Central	Science	77	Sci&Tech	Yes	86	Mkt&Fin	80	Placed	200000	80:89
no	yes	65	Private	68	Private	Arts	64	Comm&Mgmt	No	75	Mkt&Fin	77	Placed	350000	70:79
yes	yes	56	Central	52	State	Science	52	Sci&Tech	No	66	Mkt&HR	50	Not Placed	NA	NA
no	no	86	Private	74	Central	Commerce	73	Comm&Mgmt	No	97	Mkt&Fin	86	Placed	250000	80:89
yes	yes	55	Private	50	State	Science	67	Sci&Tech	Yes	55	Mkt&Fin	63	Not Placed	NA	60:69
no	yes	46	Central	49	State	Commerce	79	Comm&Mgmt	No	74	Mkt&Fin	59	Not Placed	NA	50:59
no	no	82	State	64	State	Science	66	Sci&Tech	Yes	67	Mkt&Fin	83	Placed	300000	80:89
no	yes	73	State	79	Central	Commerce	72	Comm&Mgmt	No	91	Mkt&Fin	51	Placed	350000	50:59
yes	yes	58	Private	70	State	Commerce	61	Comm&Mgmt	No	54	Mkt&Fin	67	Not Placed	NA	60:69
no	yes	58	State	61	Central	Commerce	60	Comm&Mgmt	Yes	62	Mkt&HR	53	Placed	300000	NA
yes	yes	70	State	68	Central	Commerce	78	Comm&Mgmt	Yes	60	Mkt&Fin	59	Placed	400000	50:59
yes	yes	47	Private	55	Central	Science	65	Comm&Mgmt	No	62	Mkt&HR	92	Not Placed	NA	NA
no	yes	77	State	87	Central	Commerce	59	Comm&Mgmt	No	68	Mkt&Fin	67	Placed	400000	60:69

Figure 8: The first column from the right shows that the function executed successfully

4.2 na.omit()

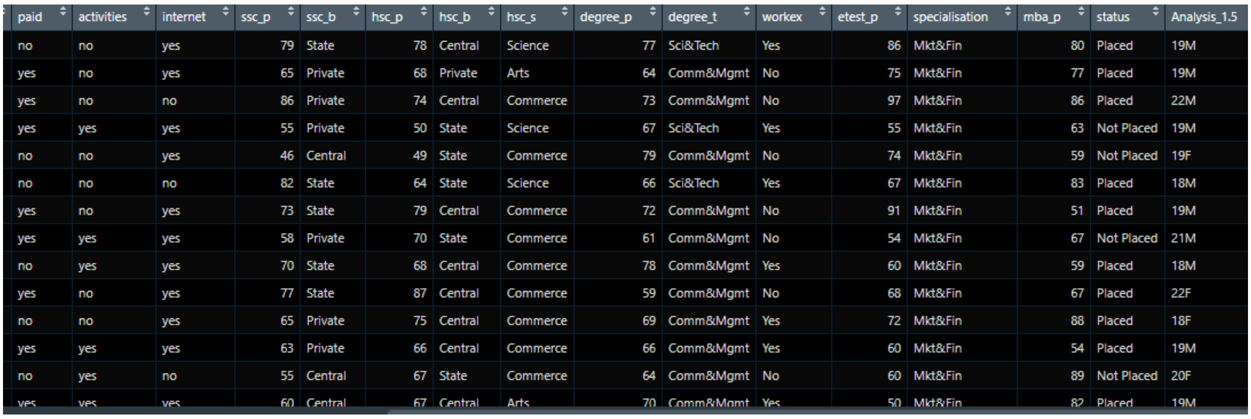


```
CSVdata1.5 = subset(CSVdata1.5, select = -salary )
CSVdata1.5 <- na.omit(CSVdata1.5)
CSVdata1.5
ggplot(CSVdata1.5, aes(Analysis_1.5)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Students' Age , Gender, Post Graduation (MBA) Specialisation = Mkt&Fin and Placement Status",
       y="Number of Students",
       x="Students' Age,Gender,Post Graduation (MBA) Specialisation = Mkt&Fin")
```

Figure 9: Example of na.omit() function is utilized in Analysis 1.5

The na.omit() method is the simplest approach to remove any rows with N/A values from a list, and it produces a list without any rows with na values. The quickest approach to eliminate na rows in R is to route the data frame or matrix through the na.omit() function. This is a straightforward technique to remove incomplete records from the study and it is a quick approach to get rid of na values in r.

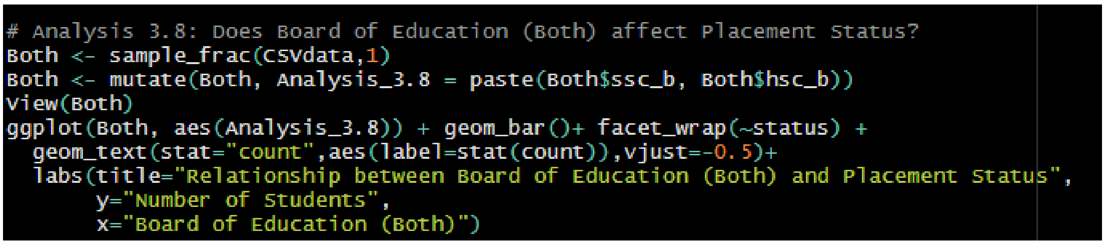
As an example, Figure 9 above shows that na.omit() function is implement in Analysis 1.5 in order to purge any rows that contain na values (after salary column is removed).



paid	activities	internet	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	Analysis_1.5
no	no	yes	79	State	78	Central	Science	77	Sci&Tech	Yes	86	Mkt&Fin	80	Placed	19M
yes	no	yes	65	Private	68	Private	Arts	64	Comm&Mgmt	No	75	Mkt&Fin	77	Placed	19M
yes	no	no	86	Private	74	Central	Commerce	73	Comm&Mgmt	No	97	Mkt&Fin	86	Placed	22M
yes	yes	yes	55	Private	50	State	Science	67	Sci&Tech	Yes	55	Mkt&Fin	63	Not Placed	19M
no	no	yes	46	Central	49	State	Commerce	79	Comm&Mgmt	No	74	Mkt&Fin	59	Not Placed	19F
no	no	no	82	State	64	State	Science	66	Sci&Tech	Yes	67	Mkt&Fin	83	Placed	18M
yes	no	yes	73	State	79	Central	Commerce	72	Comm&Mgmt	No	91	Mkt&Fin	51	Placed	19M
yes	yes	yes	58	Private	70	State	Commerce	61	Comm&Mgmt	No	54	Mkt&Fin	67	Not Placed	21M
no	yes	yes	70	State	68	Central	Commerce	78	Comm&Mgmt	Yes	60	Mkt&Fin	59	Placed	18M
yes	no	yes	77	State	87	Central	Commerce	59	Comm&Mgmt	No	68	Mkt&Fin	67	Placed	22F
no	no	yes	65	Private	75	Central	Commerce	69	Comm&Mgmt	Yes	72	Mkt&Fin	88	Placed	18F
yes	yes	yes	63	Private	66	Central	Commerce	66	Comm&Mgmt	Yes	60	Mkt&Fin	54	Placed	19M
no	yes	no	55	Central	67	State	Commerce	64	Comm&Mgmt	No	60	Mkt&Fin	89	Not Placed	20F
yes	yes	yes	60	Central	67	Central	Arts	70	Comm&Mgmt	Yes	50	Mkt&Fin	82	Placed	19M

Figure 10: Any rows with na values in the Analysis 1.5 removed

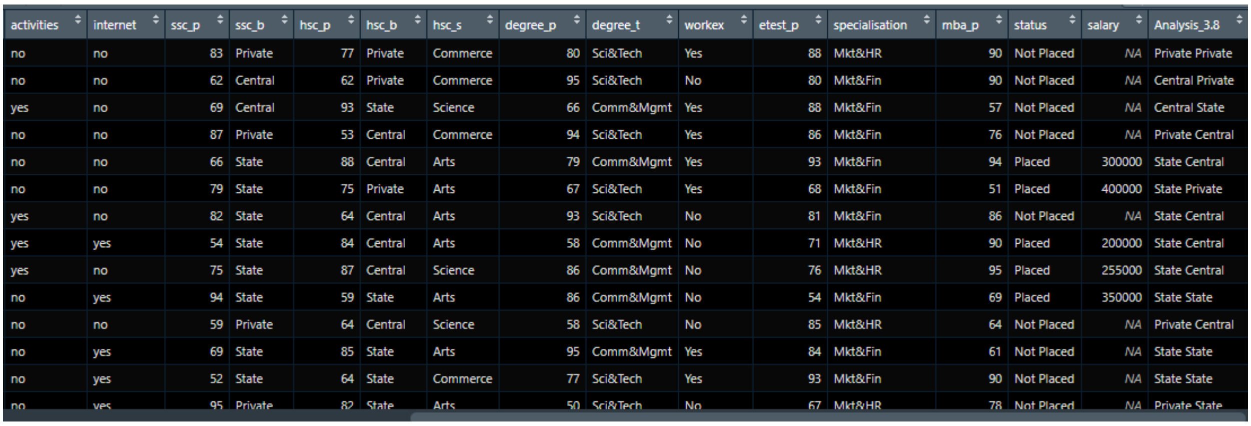
4.3 paste()



```
# Analysis 3.8: Does Board of Education (Both) affect Placement Status?
Both <- sample_frac(CSVdata,1)
Both <- mutate(Both, Analysis_3.8 = paste(Both$ssc_b, Both$hsc_b))
View(Both)
ggplot(Both, aes(Analysis_3.8)) + geom_bar()+ facet_wrap(~status) +
  geom_text(stat="count",aes(label=stat(count)),vjust=-0.5)+
  labs(title="Relationship between Board of Education (Both) and Placement Status",
       y="Number of Students",
       x="Board of Education (Both)")
```

Figure 11: Example of paste() function is utilized in Analysis 3.8

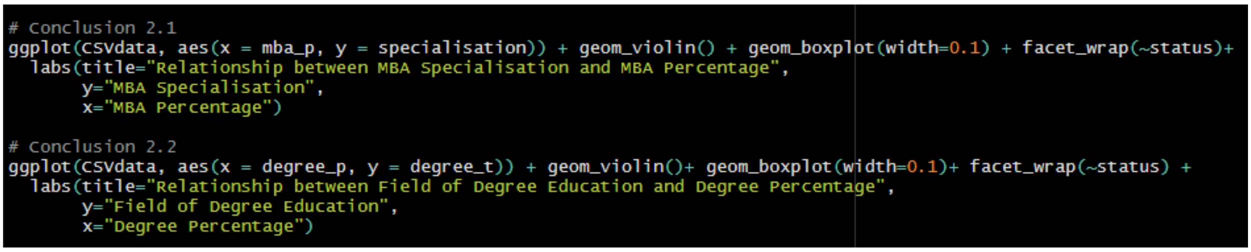
In R, the paste function is used to combine Vectors by transforming them to a new value. In R, the paste0 function simply concatenates a vector without a separator unless it is assigned manually (DataScienceMadeSimple, n.d.).



activities	internet	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary	Analysis_3.8
no	no	83	Private	77	Private	Commerce	80	Sci&Tech	Yes	88	Mkt&HR	90	Not Placed	NA	Private Private
no	no	62	Central	62	Private	Commerce	95	Sci&Tech	No	80	Mkt&Fin	90	Not Placed	NA	Central Private
yes	no	69	Central	93	State	Science	66	Comm&Mgmt	Yes	88	Mkt&Fin	57	Not Placed	NA	Central State
no	no	87	Private	53	Central	Commerce	94	Sci&Tech	Yes	86	Mkt&Fin	76	Not Placed	NA	Private Central
no	no	66	State	88	Central	Arts	79	Comm&Mgmt	Yes	93	Mkt&Fin	94	Placed	300000	State Central
no	no	79	State	75	Private	Arts	67	Sci&Tech	Yes	68	Mkt&Fin	51	Placed	400000	State Private
yes	no	82	State	64	Central	Arts	93	Sci&Tech	No	81	Mkt&Fin	86	Not Placed	NA	State Central
yes	yes	54	State	84	Central	Arts	58	Comm&Mgmt	No	71	Mkt&HR	90	Placed	200000	State Central
yes	no	75	State	87	Central	Science	86	Comm&Mgmt	No	76	Mkt&HR	95	Placed	255000	State Central
no	yes	94	State	59	State	Arts	86	Comm&Mgmt	No	54	Mkt&Fin	69	Placed	350000	State State
no	no	59	Private	64	Central	Science	58	Sci&Tech	No	85	Mkt&HR	64	Not Placed	NA	Private Central
no	yes	69	State	85	State	Arts	95	Comm&Mgmt	Yes	84	Mkt&Fin	61	Not Placed	NA	State State
no	yes	52	State	64	State	Commerce	77	Sci&Tech	Yes	93	Mkt&Fin	90	Not Placed	NA	State State
no	yes	95	Private	82	State	Arts	50	Sci&Tech	No	67	Mkt&HR	78	Not Placed	NA	Private State

Figure 12: The values in ssc_b and hsc_b column has been combine and paste in Analysis_3.8 column

4.4 geom_violin()

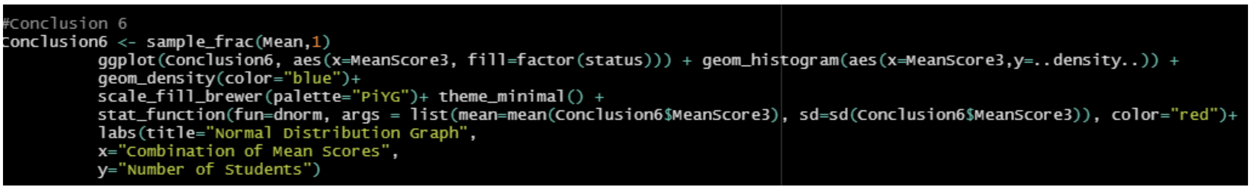


```
# Conclusion 2.1
ggplot(CSVdata, aes(x = mba_p, y = specialisation)) + geom_violin() + geom_boxplot(width=0.1) + facet_wrap(~status)+
  labs(title="Relationship between MBA Specialisation and MBA Percentage",
       y="MBA Specialisation",
       x="MBA Percentage")

# Conclusion 2.2
ggplot(CSVdata, aes(x = degree_p, y = degree_t)) + geom_violin()+ geom_boxplot(width=0.1)+ facet_wrap(~status) +
  labs(title="Relationship between Field of Degree Education and Degree Percentage",
       y="Field of Degree Education",
       x="Degree Percentage")
```

Figure 13: Example of geom_violin() function has been utilized for Question 2 conclusion

A violin plot is a visual representation of a continuous distribution in a small space. A violin plot is a mirrored density plot shown in the same way as a boxplot. It is a combination of geom boxplot() and geom density() (STHDA, n.d.). To construct violin plot, geom_violin() function needs to be utilized as shown in the Figure 13.

4.5 geom_density()

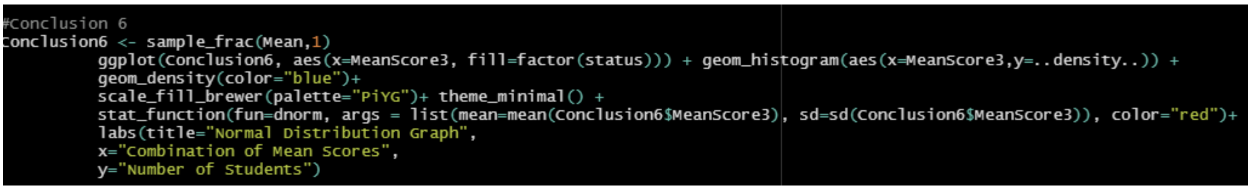


```r
#Conclusion 6
Conclusion6 <- sample_frac(Mean,1)
        ggplot(Conclusion6, aes(x=MeanScore3, fill=factor(status))) + geom_histogram(aes(x=MeanScore3,y=..density..)) +
        geom_density(color="blue")+
        scale_fill_brewer(palette="PiYG")+ theme_minimal() +
        stat_function(fun=dnorm, args = list(mean=mean(Conclusion6$MeanScore3), sd=sd(Conclusion6$MeanScore3)), color="red")+
        labs(title="Normal Distribution Graph",
        x="Combination of Mean Scores",
        y="Number of Students")
```

Figure 14: Example of geom_density() function is utilized in Conclusion for Question 6

A density plot is a visual depiction of a numeric variable's distribution by showing the probability density function of the variable using a kernel density estimate. It can be used alongside the projected line of what the normal distribution would appear like given the mean and standard deviation (STHDA, n.d.).

4.6 scale_fill_brewer()

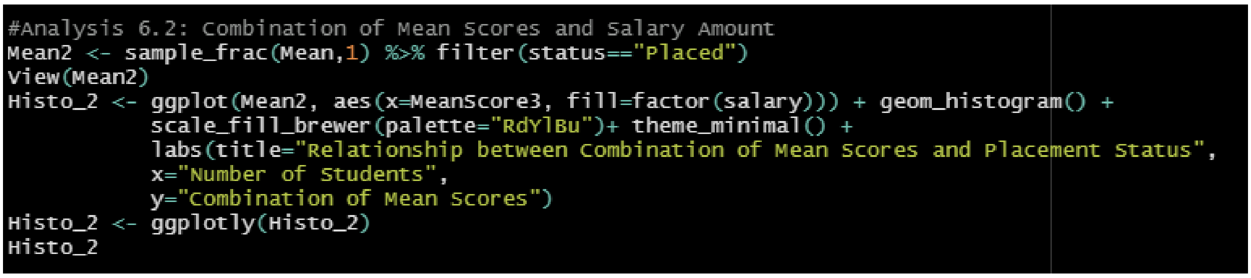


```r
#Conclusion 6
Conclusion6 <- sample_frac(Mean,1)
        ggplot(Conclusion6, aes(x=MeanScore3, fill=factor(status))) + geom_histogram(aes(x=MeanScore3,y=..density..)) +
        geom_density(color="blue")+
        scale_fill_brewer(palette="PiYG")+ theme_minimal() +
        stat_function(fun=dnorm, args = list(mean=mean(Conclusion6$MeanScore3), sd=sd(Conclusion6$MeanScore3)), color="red")+
        labs(title="Normal Distribution Graph",
        x="Combination of Mean Scores",
        y="Number of Students")
```

Figure 15: Example of scale_fill_brewer () function is utilized in Conclusion for Question 6

The scale_fill_brewer() function is utilized as the ColorBrewer's brewer scales allow sequential, divergent, and qualitative colour schemes especially when combining multiple factors under a single column to the chart/graph. The colors are very useful for displaying discrete data on a map by differentiate the factors (ggplot2, n.d.).

4.7 ggplotly ()



```r
#Analysis 6.2: Combination of Mean Scores and Salary Amount
Mean2 <- sample_frac(Mean,1) %>% filter(status=="Placed")
View(Mean2)
Histo_2 <- ggplot(Mean2, aes(x=MeanScore3, fill=factor(salary))) + geom_histogram() +
           scale_fill_brewer(palette="RdYlBu")+ theme_minimal() +
           labs(title="Relationship between Combination of Mean Scores and Placement Status",
           x="Number of Students",
           y="Combination of Mean Scores")
Histo_2 <- ggplotly(Histo_2)
Histo_2
```

Figure 16: ggplotly () function is utilized in Analysis 6.2

By using ggplotly() function, analyst may manipulate the plotly object returned by the ggplotly function specifying the mode of click and drag events, is a basic and practical application of to review the data information in the plot (plotly, n.d.).

5.0 Conclusion

Throughout the study and completing R programming assignment, I have learned how to use Data Exploration, Data Visualization, Data Manipulation and Data Transformation to analyze data and execute R programming code. In some occasion, things become very challenging when coding error occur, unable to decide or figure out which coding should be applied especially during the assignment. Worse, unable to figure out what question and analysis needs to conducted and unable to understand the meaning behind the dataset given.

Thus, it is important to understand the basic concept of the dataset structure given and doing extra research to discover any analysis technique can be used to analyze on the particular analysis. Once a result has been discovered, it is important to conduct further analysis before making hasty conclusion for any existing result. Therefore, this is the biggest lesson that I learned from this assignment and this module: Be analytical and observant but not judgmental in every single situation besides from studies. This value is extremely important for someone who study Cyber Security as they need to be analytical and observant in most of the situations.

Therefore, I wanted to Ms. Minnu Helen Joseph for guiding us through the process of learning about R Programming. I believe I can utilize not only R programming in the future, but also the hidden value that I learned from the lesson during assignment completion.

6.0 References

Cottman, B. H., 2021. *Six Datatype Transformer Functions for Data Pre-Processing for Machine Learning.* [Online]
Available at: https://towardsdatascience.com/six-datatype-transformer-functions-for-data-pre-processing-for-machine-learning-eb9abcce68cd
[Accessed 3 January 2022].

DataScienceMadeSimple, n.d. *CASE WHEN IN R USING CASE_WHEN() DPLYR – CASE_WHEN IN R.* [Online]
Available at: https://www.datasciencemadesimple.com/case-statement-r-using-case_when-dplyr/
[Accessed 15 January 2022].

DataScienceMadeSimple, n.d. *PASTE FUNCTION IN R – PASTE0().* [Online]
Available at: https://www.datasciencemadesimple.com/paste-function-in-r/#:~:text=Paste%20function%20in%20R%20is,()%20function%20for%20the%20dataframe.
[Accessed 22 January 2022].

ggplot2, n.d. *Sequential, diverging and qualitative colour scales from ColorBrewer.* [Online]
Available at: https://ggplot2.tidyverse.org/reference/scale_brewer.html
[Accessed 30 January 2022].

plotly, n.d. *Getting Started with Plotly in ggplot2.* [Online]
Available at: https://plotly.com/ggplot2/getting-started/
[Accessed 31 January 2022].

STHDA, n.d. *ggplot2 density plot : Quick start guide - R software and data visualization.* [Online]
Available at: http://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualization
[Accessed 30 January 2022].

STHDA, n.d. *ggplot2 violin plot : Quick start guide - R software and data visualization.* [Online]
Available at: http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization
[Accessed 23 January 2022].