Job Placement Analysis

IST 687 Introduction to Data Science

Allison Deming, Brandon Griffing and Marissa Hucke

Objective

Team A completed the analysis of the organization's student database to help identify different factors that contribute to a student's job placement status after college graduation. These findings will assist the organization with future recommendations on current academic programs and student success projections.

Dataset

The analysis was conducted across a population of 21,285 students in a foreign education institution and included academic records ranging from high school to college using this dataset: https://www.kaggle.com/niki188/campus-recruitment

Below are the data definitions used to complete the analysis.

Attribute	Definition
UniqueID	Unique student identification
Gender	Male or Female
SSC Placement Test Percentage 10th Grade	10th grade test scores
SSC 10th Grade Location	Central: 10th grade student located in metro area Other: student located in rural area
HSC 12th Grade Percentage	12th grade test scores
HSC 12th Grade Location	Central: 12th grade student located in metro area Other: 12th grade student located in rural area
HSC 12th Grade Track	Studied track in 12th grade: Arts, Commerce or Science
HS Placement Percentage	Post-High School Placement for College
HS Track	Studied high school track: Comm&Mgmt, Sci&Tech or Others
Work Experience	Yes or No on whether a student had work experience during school years
College Placement Percentage	College placement percentage

College Track	Studied college track: Mkt&Fin or Mkt&HR
Graduate Level Placement Percentage	GRE/GMAT Like placement for Graduate School
Placement Status	Placed or Not Placed indicating whether a student placed in a job after college graduation
Reported Salary	A student's reported earned salary if they placed

Figure 1: Data Dictionary.

Dataset Questions

To facilitate the analysis the team identified these dataset questions to explore.

- Within each high school or college track, which track had the most placed students?
    - High school Tracks:
        - Arts
        - Commerce
        - Comm&Mgmt
        - Science
        - Sci&Tech
        - Other
    - College Tracks:
        - Mkt&Fin
        - Mkt&HR
- Does a student's 10th or 12th grade high school test scores predict whether they will be placed for a job or not?
- Does a student's location of residency affect whether they will be placed for a job or not?
- What is the average Placed status by Gender?
- What is the average Reported Salary by Gender?
- Does a student's Work Experience prior to receiving a job placement affect how much their Reported Salary is?

Dataset Analysis

Does an individual's location affect their placement

Analysis type: Filter, Rename, ggplot geom bar

Sample R Code:

```
# Import dataset
data <- file.choose("C:\\Users\\brandonjgriffing\\Desktop\\IST647\\Project\\
Placement_Data_Full_Class_FINAL_projectExcel.csv")
data <- read.csv(data)

# Drop columns you don't need
data <- data[ -c(1,2,3,5,7,8,9,10,11,12,13,15)]

# rename columns to better read/understand
data <- data %>% rename(TenthGradeLocation = 1, TwelfthGradeLocation = 2, PlacementStatus = 3)

# drop columns for tenth grade dataset
data <- data[ -c(2)]

# create new columns for 10th grade Others and are Placed or Not, graph, and explain
P10Others <- data[data$TenthGradeLocation == 'Others' & data$PlacementStatus == "Placed",]
NP10Others <- data[data$TenthGradeLocation == 'Others' & data$PlacementStatus == "Not Placed",]

# create df with those "Others" but placed
tp10Others <- P10Others %>% count(PlacementStatus)
# create df with those "Others" not placed
np10Others <- NP10Others %>% count(PlacementStatus)

# bind the columns of "Others" Placed and Not Placed and rename for coherence
total10Others <- rbind(tp10Others, np10Others)
total10Others <- total10Others %>% rename(Location_Others = 2)

# plot the data using ggplot
plotOthers <-ggplot(data=total10Others, aes(x=PlacementStatus, y=Location_Others)) + geom_bar(fill="light
blue", stat="identity")
plotOthers
```

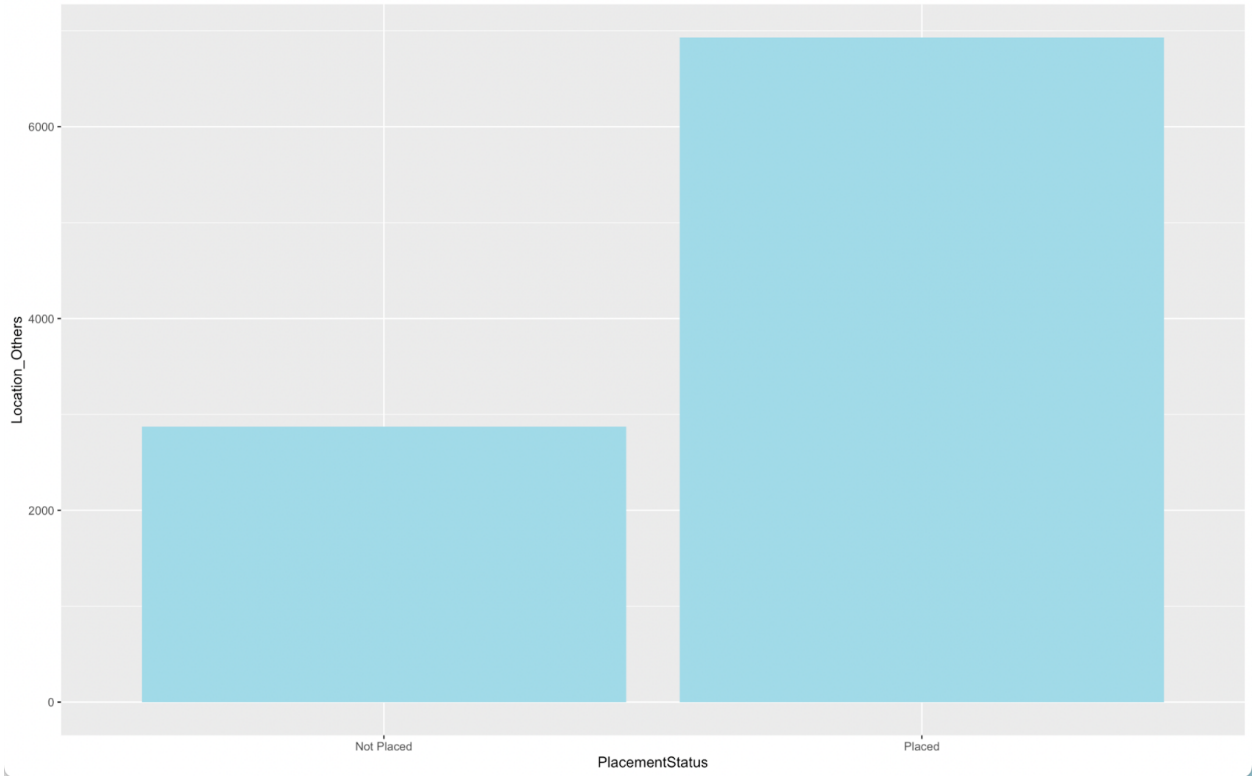

Figure 2: Placed by not placed barchart.

```
# Keep graph in mind..

# create new columns for 10th grade Central and are Placed or Not, graph, and explain
P10Central <- data[data$TenthGradeLocation == 'Central' & data$PlacementStatus == "Placed",]
NP10Central <- data[data$TenthGradeLocation == 'Central' & data$PlacementStatus == "Not Placed",]

# create df with those "Central" but placed
tp10Central <- P10Central %>% count(PlacementStatus)
# create df with those "Central" not placed
np10Central <- NP10Central %>% count(PlacementStatus)

# bind the columns of "Others" Placed and Not Placed and rename for coherence
total10Central <- rbind(tp10Central, np10Central)
total10Central <- total10Central %>% rename(Location_Central = 2)

# plot the data using ggplot
plotCentral <-ggplot(data=total10Central, aes(x=PlacementStatus, y=Location_Central)) + geom_bar(fill="light
blue", stat="identity")
plotCentral
```

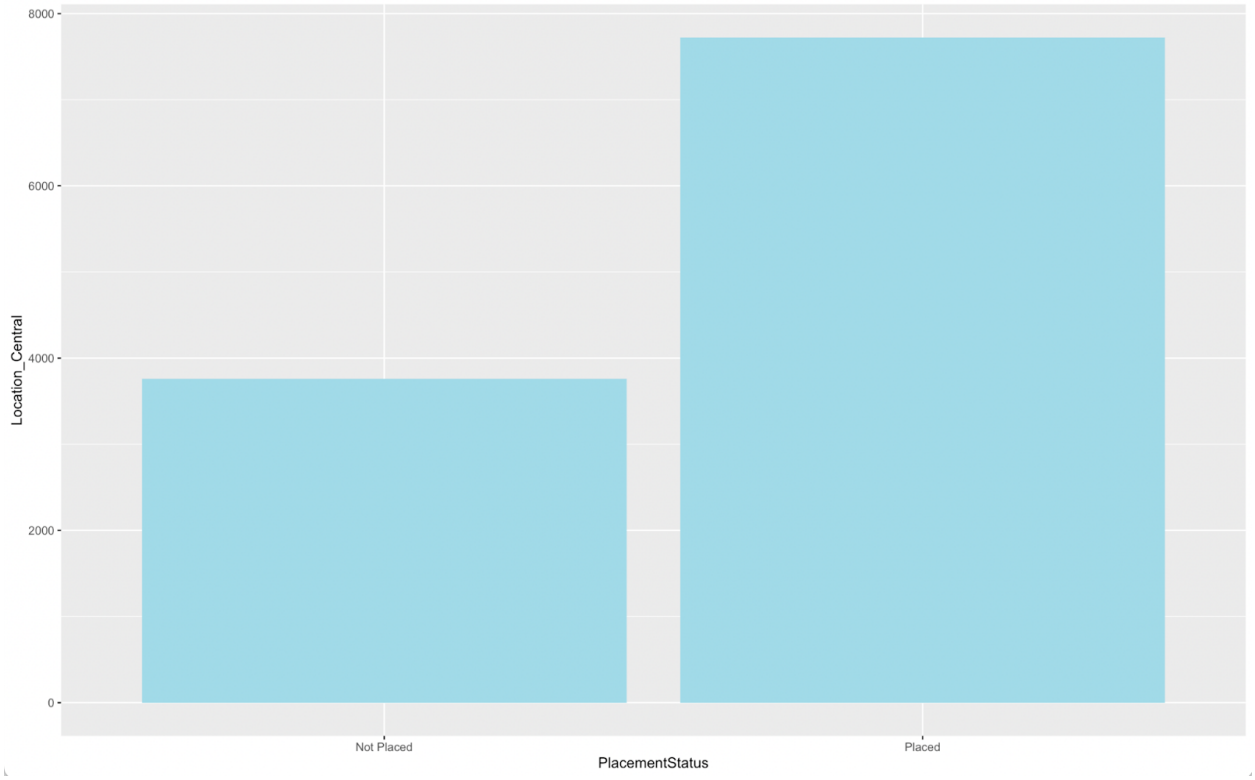

Figure 3: Placement Status by Location

 # Keep graph in mind and make analysis

 # Make analysis
glimpse(total10Others)
glimpse(total10Central)

 # It appears that location plays a slight part in placement for tenth grade given one from Central location has about a 59% higher chance of being placed from those in Other locations where individuals have about a 52% chance.

~ Twelfth Grade~

 # Re-import dataset or reattach columns you need
data <- file.choose("C:\\Users\\brandonjgriffing\\Desktop\\IST647\\Project
Placement_Data_Full_Class_FINAL_projectExcel.csv")
data <- read.csv(data)

 # Drop columns you don't need
data <- data[-c(1,2,3,5,7,8,9,10,11,12,13,15)]

 # rename columns to better read/understand
data <- data %>% rename(TenthGradeLocation = 1, TwelfthGradeLocation = 2, PlacementStatus = 3)

```
# drop columns for twelfth grade dataset
data <- data[ -c(1)]

# create new columns for 12th grade Others Placed or Not, graph, and explain
P10Others2 <- data[data$TwelfthGradeLocation == 'Others' & data$PlacementStatus == "Placed",]
NP10Others2 <- data[data$TwelfthGradeLocation == 'Others' & data$PlacementStatus == "Not Placed",]

# create df with those "Others" but placed
tp10Others2 <- P10Others2 %>% count(PlacementStatus)
# create df with those "Others" not placed
np10Others2 <- NP10Others2 %>% count(PlacementStatus)

# bind the columns of "Others" Placed and Not Placed and rename for coherence
total10Others2 <- rbind(tp10Others2, np10Others2)
total10Others2 <- total10Others2 %>% rename(Location_Others = 2)

# plot the data using ggplot
plotOthers2 <-ggplot(data=total10Others2, aes(x=PlacementStatus, y=Location_Others)) + geom_bar(fill="light
blue", stat="identity")
plotOthers2
```

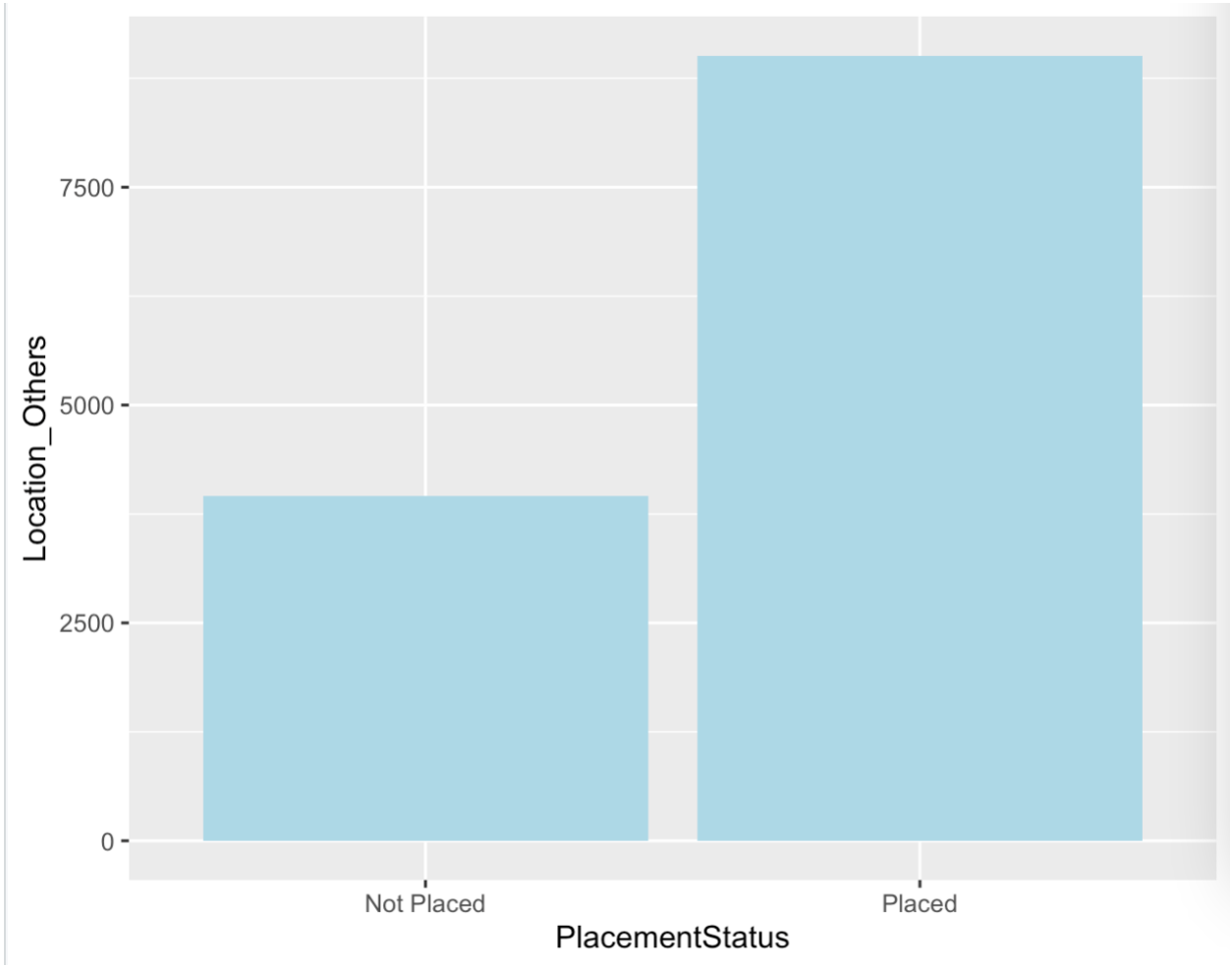

Figure 4: Placement Status by Location

```
# Keep graph in mind and make analysis

# create new columns for 12th grade Central and are Placed or Not, graph, and explain
P10Central2 <- data[data$TwelfthGradeLocation == 'Central' & data$PlacementStatus == "Placed",]
NP10Central2 <- data[data$TwelfthGradeLocation == 'Central' & data$PlacementStatus == "Not Placed",]

# create df with those "Central" but placed
tp10Central2 <- P10Central2 %>% count(PlacementStatus)
# create df with those "Central" not placed
np10Central2 <- NP10Central2 %>% count(PlacementStatus)

# bind the columns of "Others" Placed and Not Placed and rename for coherence
total10Central2 <- rbind(tp10Central2, np10Central2)
total10Central2 <- total10Central2 %>% rename(Location_Central = 2)

# plot the data using ggplot
```

plotCentral2 <-ggplot(data=total10Central2, aes(x=PlacementStatus, y=Location_Central)) + geom_bar(fill="light blue", stat="identity")
plotCentral2

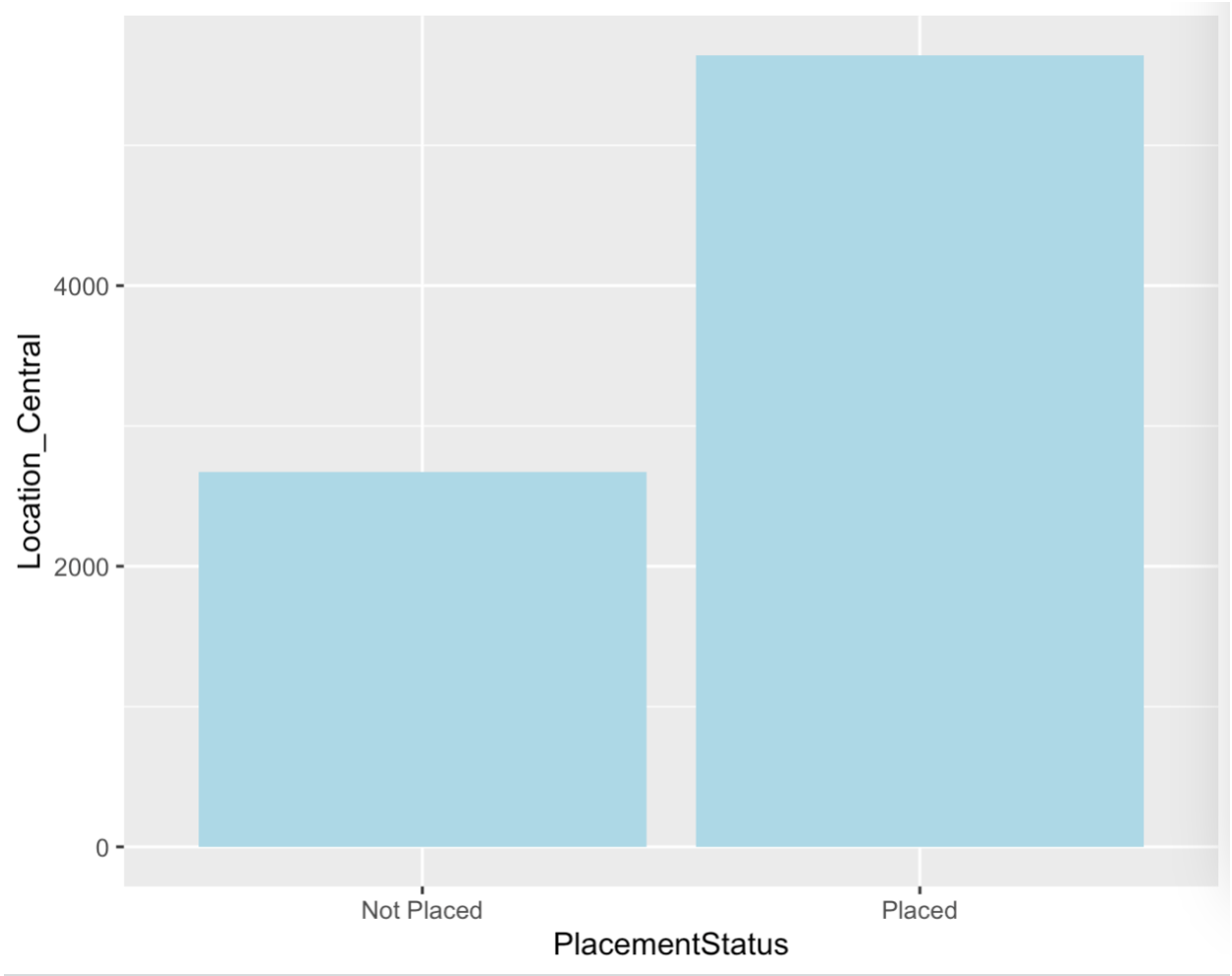

Figure 5: Placement Status by Central Location

 # Make analysis
glimpse(total10Others2)
glimpse(total10Central2)

 # It appears that location plays a slight part in placement for twelfth grade given one from Central location has about a 56% higher chance of being placed from those in Other locations where individuals have about a 53% chance.

Does an individual's work experience affect how much their reported salary is?

Analysis type: Drop, Rename, ggplot geom_boxplot

Sample R Code:

```
# Reattach original dataset and Drop columns you don't need
data2 <- data[ -c(1,2,3,4,5,6,7,8,9,11,12,13,14)]

# rename columns to better read/understand
data2 <- data2 %>% rename(WorkExperience = 1, Salary = 2)

# Check for na's and remove if needed
sum(is.na(data$Salary))
data2 <- data2 %>% drop_na(Salary)

# Visualize the data and make an inference in a comment below with your code
options(scipen = 999) # to disable scientific notation
ggplot(data2, aes(x=WorkExperience, y=Salary)) + geom_boxplot(color="blue", fill="blue",
alpha=0.2, notch=TRUE, notchwidth = 0.8, outlier.colour="light blue", outlier.fill="black",
outlier.size=3) + coord_flip() + ggtitle("Work Experience and Salary")

# It appears having work experience does impact ones salary due to having a higher starting salary
```

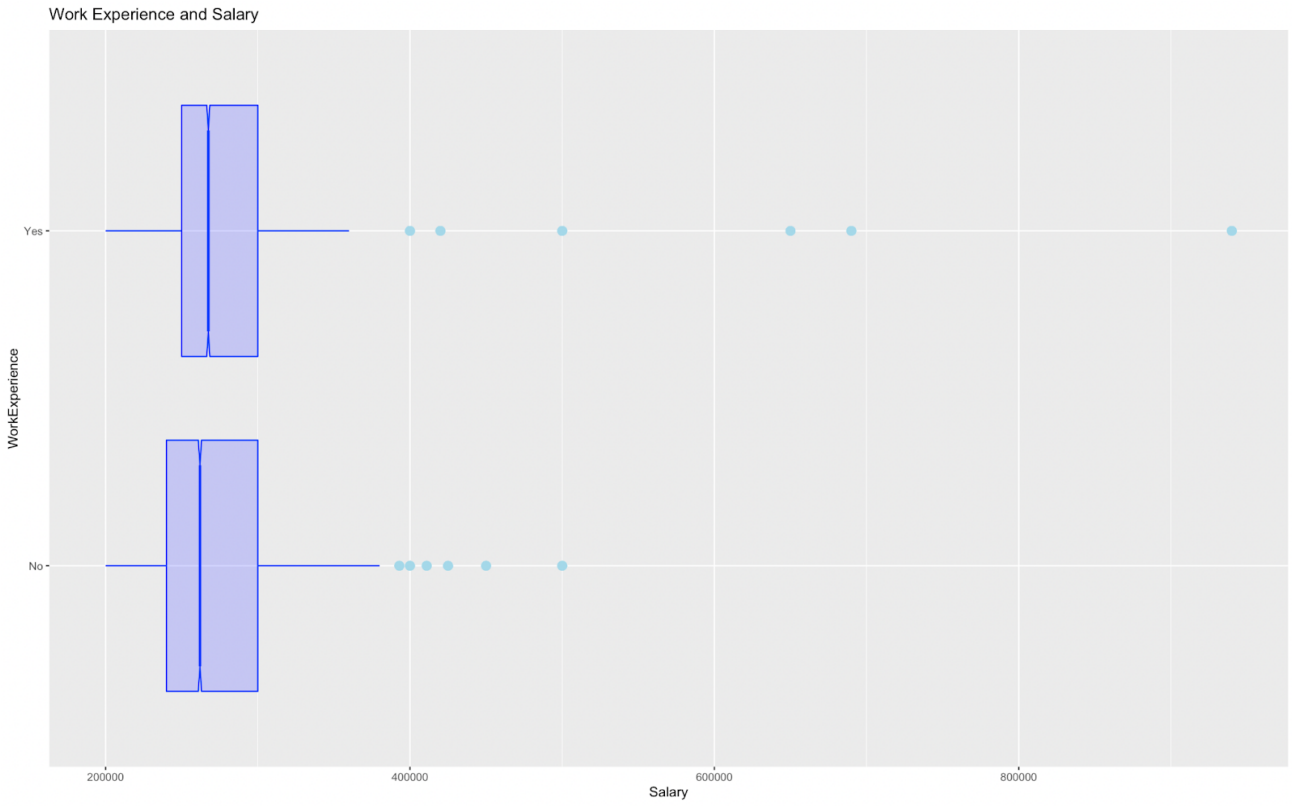

Figure 6: Boxplot of Work experience and Salary

Placement by High School Tracks

Analysis type: Function, Filter, Average and ggplot geom bar

Sample R Code:

```
# Total number of students
TotalNumStudents <- 21285

# Function for average
avg_num <- function(num1) {
  print(num1/TotalNumStudents)
}

# Total number of students who placed per High School track
# Commerce
placedCommerce <- filter(campusRecruitment, campusRecruitment$`HSC 12th Grade Track` == "Commerce" &
campusRecruitment$`Placement Status` == "Placed")
placedCommerceTotal <- nrow(placedCommerce)
placedCommerceTotal

# 37% of students who studied Commerce placed
avgPlacedCommerce <- avg_num(placedCommerceTotal)

# HS Track Bar Chart
avgPlacedPerTrack <- c(avgPlacedCommerce, avgPlacedArts, avgPlacedScience, avgPlacedScienceTech,
avgPlacedCommMgmt, avgPlacedOthers)
Tracks <- c("Commerce", "Arts", "Science", "Sci&Tech", "Com&Mgt", "Others")
dfTracks <- data.frame(Tracks,avgPlacedPerTrack)
dfTracks <- dfTracks[order(-dfTracks$avgPlacedPerTrack),]

ggplot(data=dfTracks, aes(x=reorder(Tracks,-avgPlacedPerTrack), y=avgPlacedPerTrack)) +
 geom_bar(stat="identity", fill="steelblue", width=0.5) +
 theme_minimal() +
 xlab("Tracks") +
 ylab("Frequency")          +
 ggtitle("Average Placement by Studied High School Track")
```

Results: On average, students who studied Commerce Management placed the most out of all of the other tracks with Commerce closely leading behind at 37%.

- 47% of students who studied CommMgmt placed
- 37% of students who studied Commerce placed
- 29% of students who studied Science placed

- 19% of students who studied Sci&Tech placed
- 3% of students who studied Arts placed
- 2% of students who did not specify a track “Others” placed

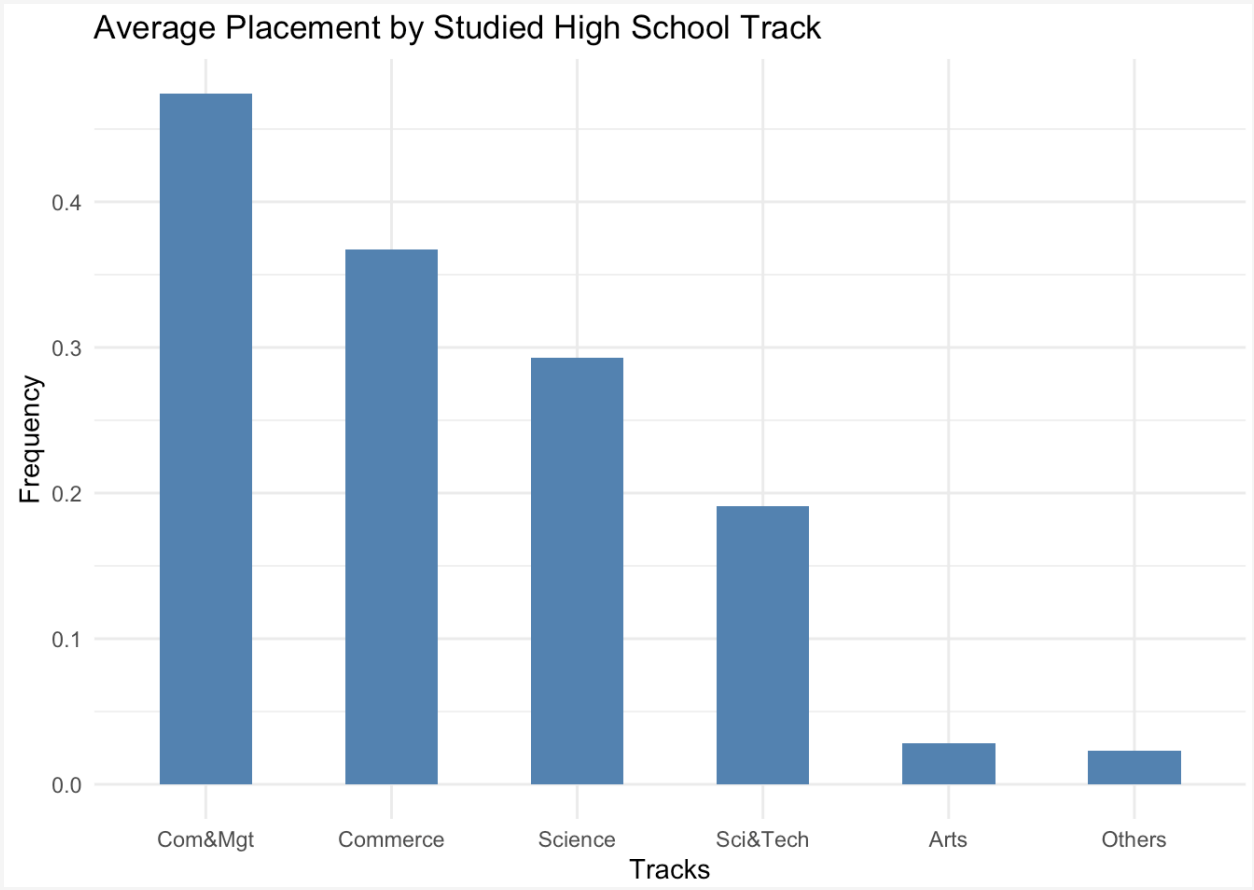

Figure 7: Tracks Barchart

Placement by College Tracks

Analysis type: Function, Filter, Average and ggplot geom bar

Sample R Code:

```
# Mkt&Fin
placedMktFin <- filter(campusRecruitment, campusRecruitment$`College Track` == "Mkt&Fin" &
campusRecruitment$`Placement Status` == "Placed")
placedMktFinTotal <- nrow(placedMktFin)
placedMktFinTotal

avgPlacedMktFin <- avg_num(placedMktFinTotal)

# College Track Bar Chart
avgPlacedPerTrack2 <- c(avgPlacedMktHR, avgPlacedMktFin)
Tracks <- c("Mkt&HR", "Mkt&Fin")
dfTracks2 <- data.frame(Tracks,avgPlacedPerTrack2)
dfTracks2 <- dfTracks2[order(-dfTracks2$avgPlacedPerTrack2),]

ggplot(data=dfTracks2, aes(x=reorder(Tracks,-avgPlacedPerTrack2), y=avgPlacedPerTrack2)) +
 geom_bar(stat="identity", fill="steelblue", width=0.5) +
 theme_minimal() +
 xlab("Tracks") +
 ylab("Frequency") +
 ggtitle("Average Placement by Studied College Track")
```

Results: On average, students who studied Marketing & Finance placed the most compared to those who studied Marketing & HR.

- 44% of students who studied Mkt&Fin placed
- 25% of students who studied Mkt&HR placed

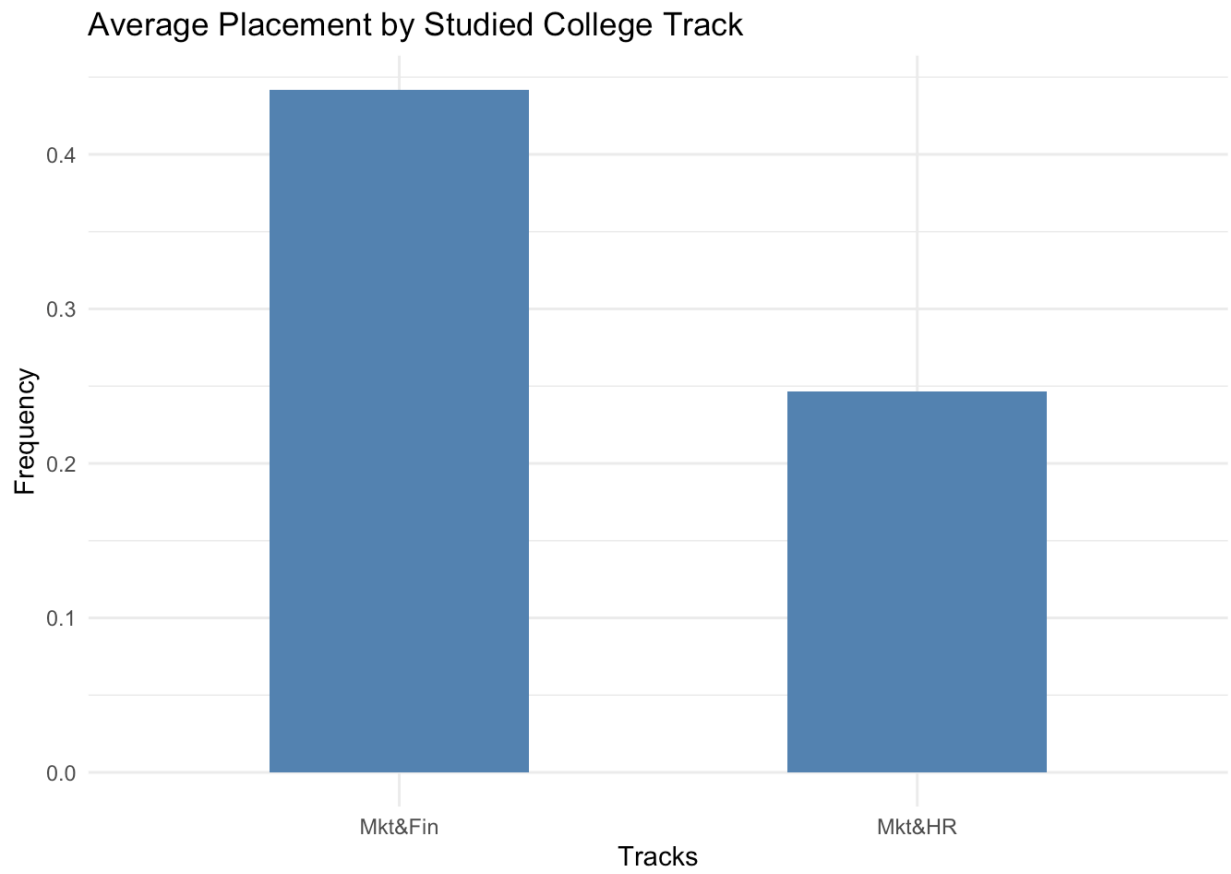

Figure 8: Advanced Placement Status by Studied College Track

Placement by Gender

Analysis type: Filter, Average and ggplot geom bar

Sample R Code:

```
# Total number of students by gender who placed (proportioned)

# Total number of female students 7,524
totalFemaleStudents <- filter(campusRecruitment, campusRecruitment$`Gender` == "F")
totalFemales <- nrow(totalFemaleStudents)
totalFemales

# Total number of male students 13,761
totalMaleStudents <- filter(campusRecruitment, campusRecruitment$`Gender` == "M")
totalMales <- nrow(totalMaleStudents)
totalMales

# Bar Chart (Avg Female vs Male)
avgPlacedPerGender2 <- c(avgFemalesPlacedOfFemales, avgMalesPlacedOfMales)
Gender <- c("Female", "Male")
```

```
dfGender2 <- data.frame(Gender,avgPlacedPerGender2)
dfGender2 <- dfGender2[order(-dfGender2$avgPlacedPerGender2),]

ggplot(data=dfGender2, aes(x=reorder(Gender,-avgPlacedPerGender2), y=avgPlacedPerGender2)) +
 geom_bar(stat="identity", fill="steelblue", width=0.5) +
 theme_minimal() +
 xlab("Gender") +
 ylab("Frequency") +
 ggtitle("Average Placement by Total Within Each Gender")
```

Results: When taking the total number of female students and male students who placed out of the total number of students (21,285), the data looks biased or skewed as it is not proportionate to the total number of females vs males where there are less female students in the dataset. When you take the total number of placed female students out of the total number of female students (7,524), and the same with male students (13,761), you can see that the average placement between gender is more comparative.

- 22% of female students placed compared to 47% of male students who placed (disproportionate)
- 63% of female students placed compared to 72% of male students who placed (proportionate)

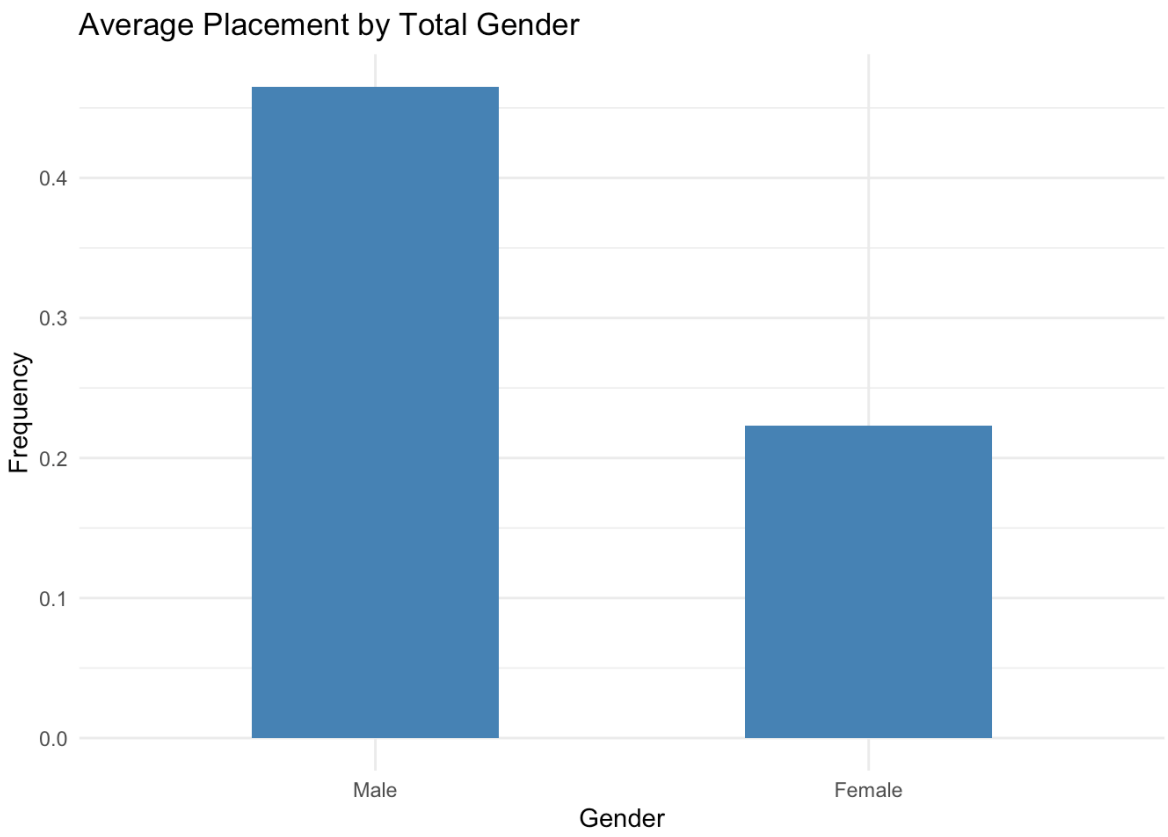

Figure 9: Gender by Placement Status

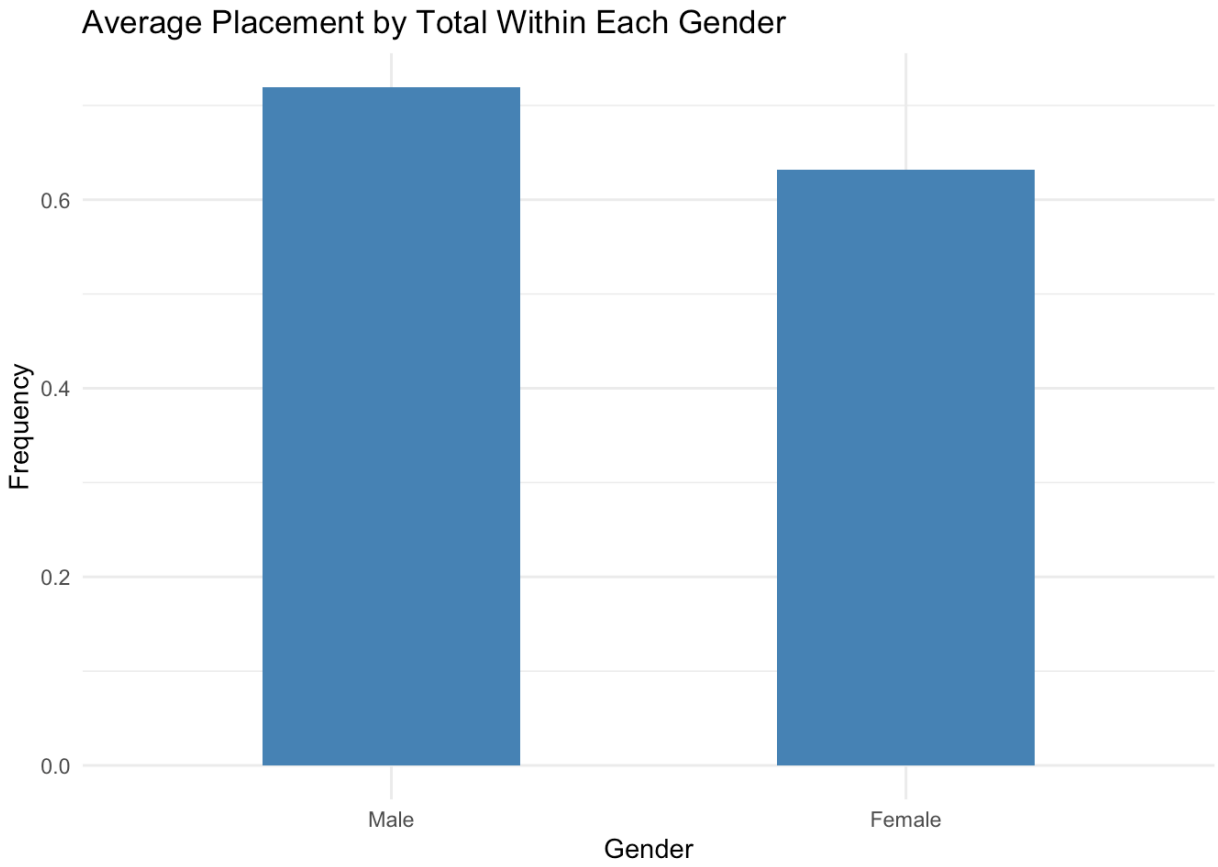

Figure 10: Placement Status by Gender (Proportional)

Linear Regression Models

It was important to take a look at the variables to see how they are correlated and if the relationships predict anything down the line, in particular, predicted salary. As school is built on the basics, and then those combine into more complex materials, it is possible that early success can predict later placement and placement test scores. Therefore, we performed pairwise comparisons of our numeric variables and looked at linear regression models and correlations.These are shown below and the code for the correlations and models are in the appendix. It should be noted that with over 10,000 data points all t-tests, linear models, and correlations are significant at an alpha < .05 level (due to sample size). Others that are more practically significant, and have linear relationships will be presented.

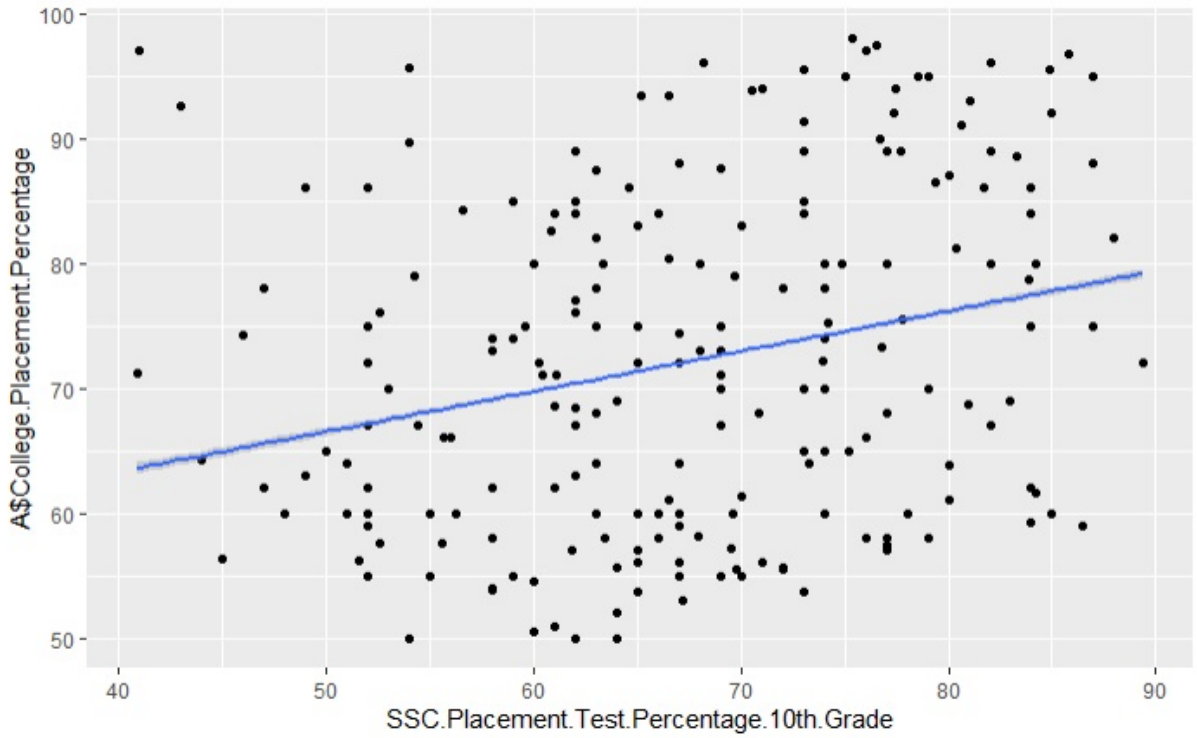

Figure 11: Correlation = .26, Linear Model Adjusted R Squared = .214; α <.05. This shows the correlation and linear relationship between the 10th grade placement test and the college placement score.

Example R Code:

```
cor.test(A$SSC.Placement.Test.Percentage.10th.Grade,A$College.Placement.Percentage)
LM00<-lm(formula=A$SSC.Placement.Test.Percentage.10th.Grade ~
A$College.Placement.Percentage, data=A)
summary(LM00)
plotB <- ggplot(data = A, aes(x = SSC.Placement.Test.Percentage.10th.Grade, y
=College.Placement.Percentage)) + geom_point()
plotB + geom_smooth(formula = y ~ x, method = "lm")
```

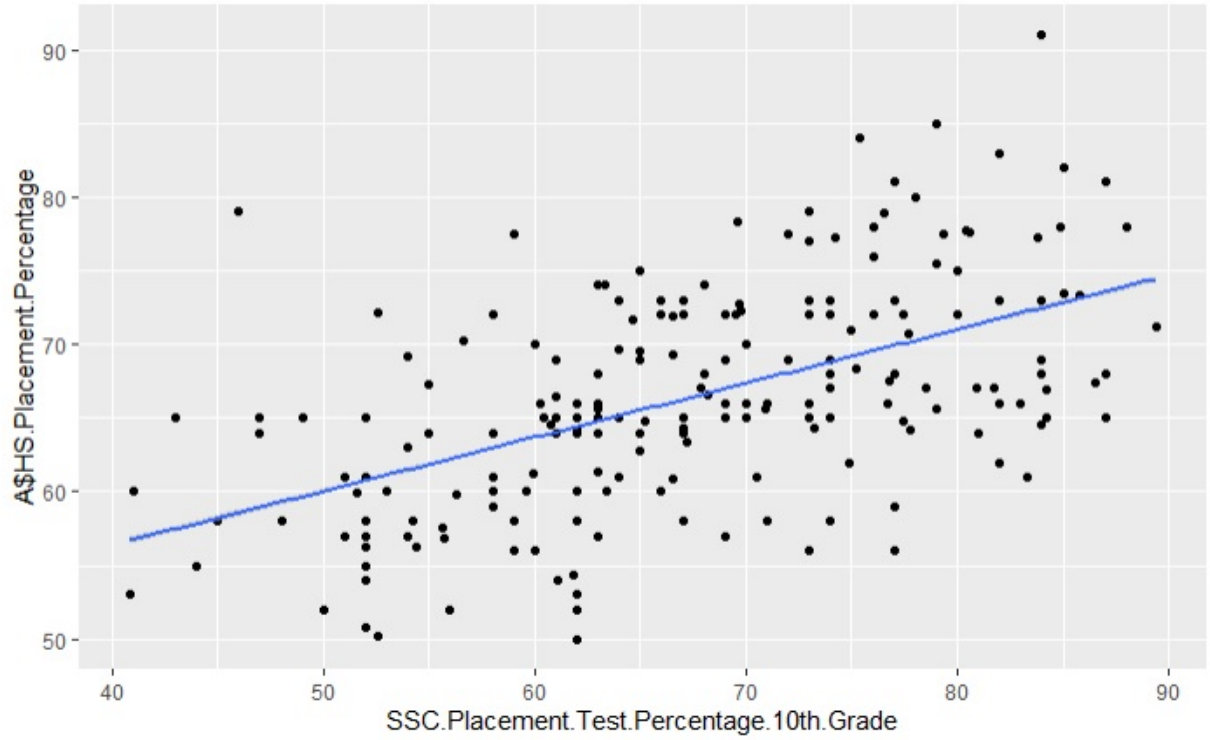

Figure 12: Correlation = .26. Linear Model Adjusted R Squared = 0.0686) α <.05. This shows the correlation and linear relationship between the 10th grade placement test and the end of high school placement score.

Example R Code:

```
cor.test(A$SSC.Placement.Test.Percentage.10th.Grade, A$HS.Placement.Percentage)
LM000<-lm(formula=A$SSC.Placement.Test.Percentage.10th.Grade ~
A$HS.Placement.Percentage, data=A)
summary(LM000)
library(ggplot2)
plotC <- ggplot(data = A, aes(x = SSC.Placement.Test.Percentage.10th.Grade, y
=HS.Placement.Percentage)) + geom_point()
plotC + geom_smooth(formula = y ~ x, method = "lm")
```

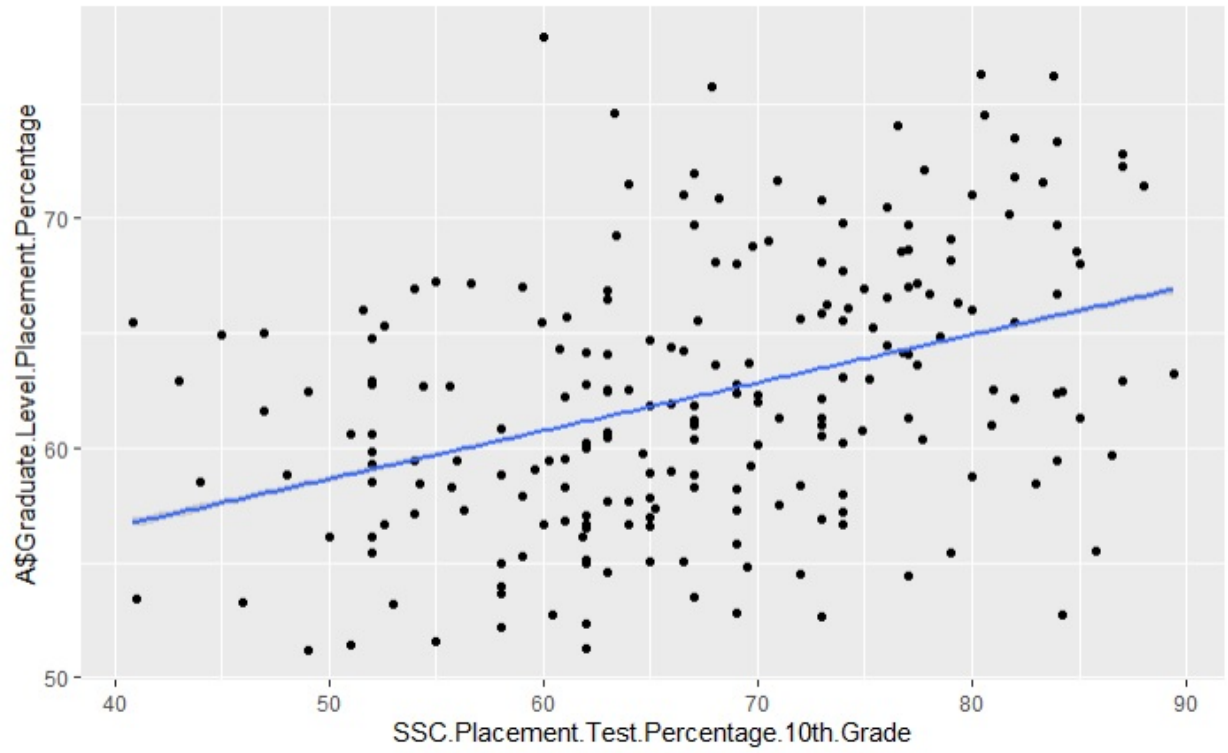

Figure 13: Correlation = .39, Linear Model Adjusted R Squared = 0.1509; α <.05. This shows the correlation and linear relationship between the 10th grade placement test and the graduate level placement score.

Example R Code:

```
cor.test(A$SSC.Placement.Test.Percentage.10th.Grade,A$Graduate.Level.Placement.Percentage)
LM0000<-lm(formula=A$SSC.Placement.Test.Percentage.10th.Grade ~
A$Graduate.Level.Placement.Percentage, data=A)
summary(LM0000)
plotD <- ggplot(data = A, aes(x = SSC.Placement.Test.Percentage.10th.Grade, y
=Graduate.Level.Placement.Percentage)) + geom_point()
plotD + geom_smooth(formula = y ~ x, method = "lm")
```

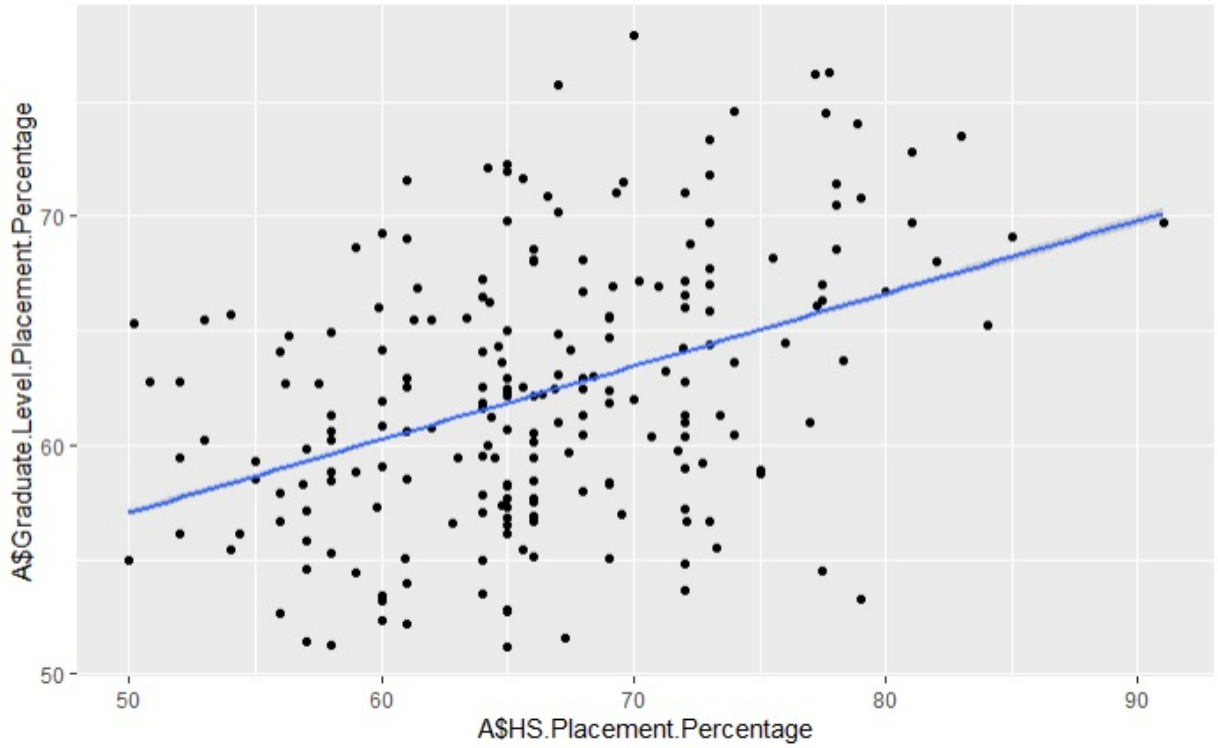

Figure 14: Correlation = .22, Linear Model Adjusted R Squared =0.0475; α <.05. This shows the correlation and linear relationship between the end of high school placement test and the graduate placement score.

Example R Code:

```
cor.test(A$HSC.12th.Grade.Percentage, A$HS.Placement.Percentage)
LM00000<-lm(formula=A$HSC.12th.Grade.Percentage ~ A$HS.Placement.Percentage, data=A)
summary(LM00000)
plotE <- ggplot(data = A, aes(x = HSC.12th.Grade.Percentage, y =HS.Placement.Percentage)) +
geom_point()
plotE + geom_smooth(formula = y ~ x, method = "lm")
```

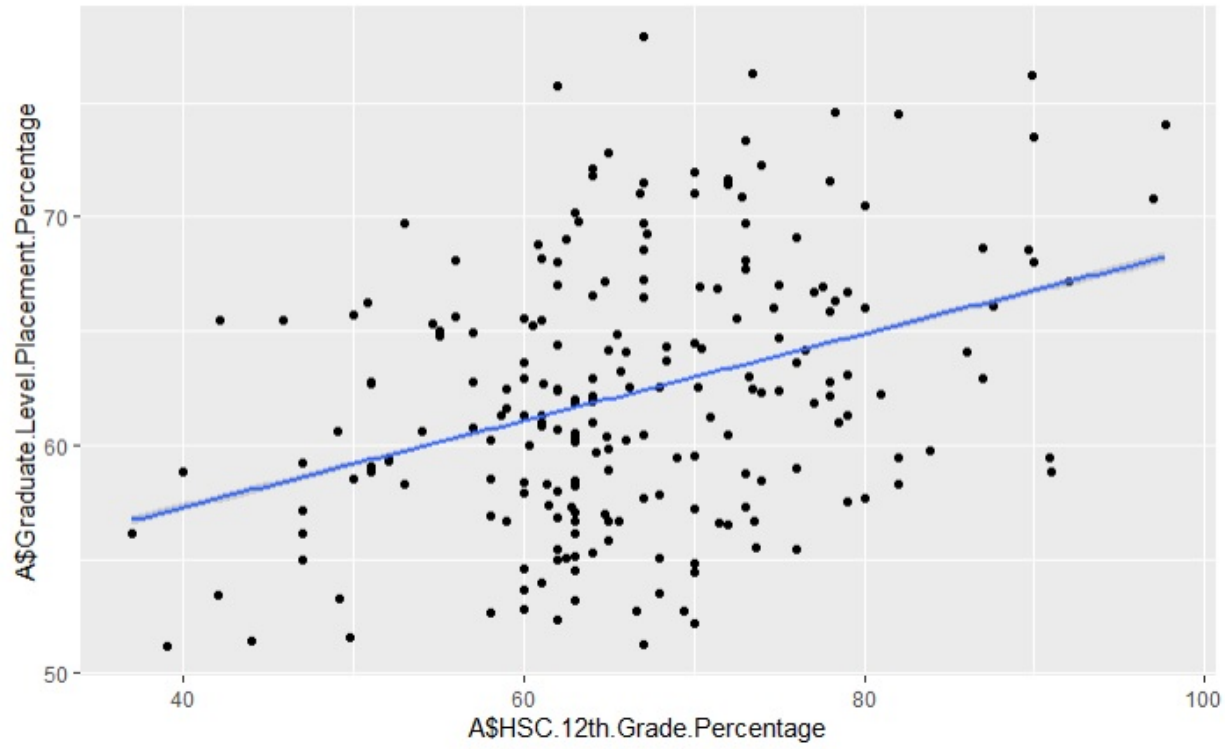

Figure 15: Correlation = .40, Linear Model Adjusted R Squared = 0.1619;α <.05. This shows the correlation and linear relationship between the 12th grade placement test and the graduate placement score.

Example R Code:

```
cor.test(A$HSC.12th.Grade.Percentage, A$Graduate.Level.Placement.Percentage)
LM010<-lm(formula=A$HSC.12th.Grade.Percentage ~ A$Graduate.Level.Placement.Percentage, data=A)
summary(LM010)
plotG <- ggplot(data = A, aes(x = HSC.12th.Grade.Percentage, y =Graduate.Level.Placement.Percentage)) + geom_point()
plotG + geom_smooth(formula = y ~ x, method = "lm")
```

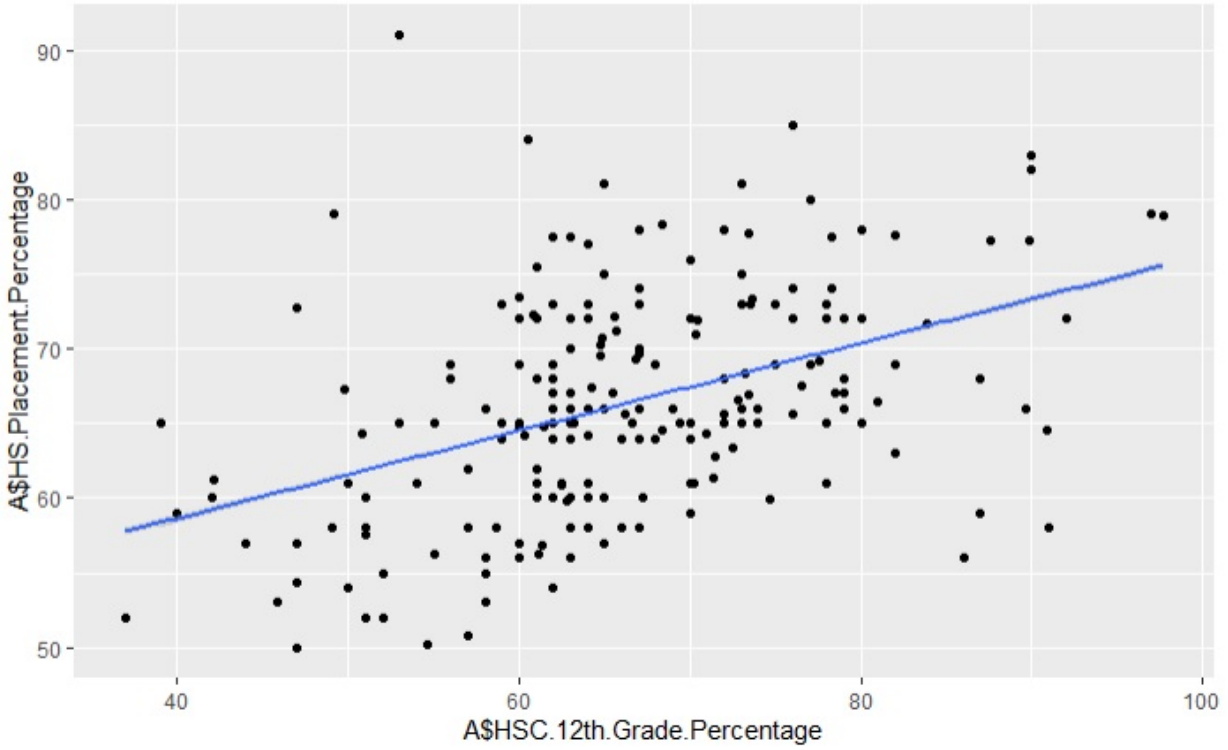

Figure 16: Correlation = .434, Linear Model Adjusted R Squared = 0.1885;α <.05. This shows the correlation and linear relationship between the 12th grade placement test and the high school placement score.

Example R Code:

```
cor.test(A$HSC.12th.Grade.Percentage, A$HS.Placement.Percentage)
LM00000<-lm(formula=A$HSC.12th.Grade.Percentage ~ A$HS.Placement.Percentage, data=A)
summary(LM00000)
plotE <- ggplot(data = A, aes(x = HSC.12th.Grade.Percentage, y =HS.Placement.Percentage)) +
geom_point()
plotE + geom_smooth(formula = y ~ x, method = "lm")
```

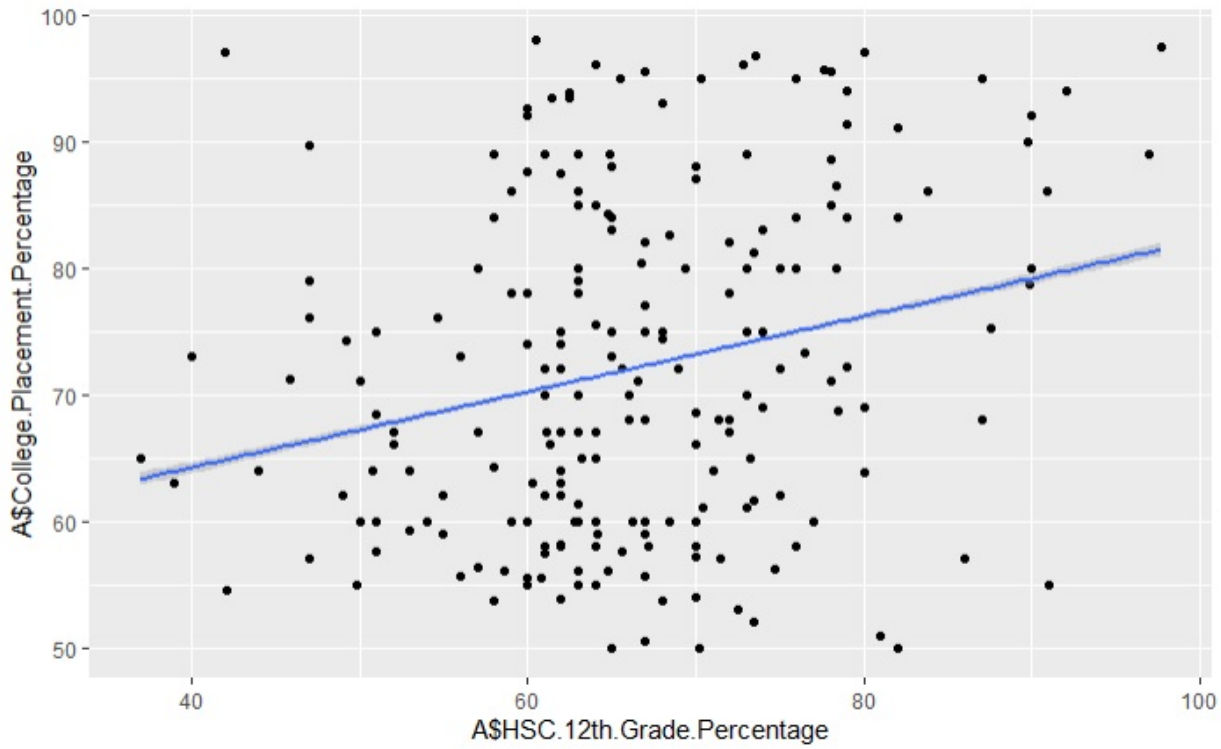

Figure 17: Correlation = .25, Linear Model Adjusted R Squared = 0.06004;α <.05. This shows the correlation and linear relationship between the 12th grade placement test and the college placement score.

Example R Code:

```
cor.test(A$HSC.12th.Grade.Percentage, A$College.Placement.Percentage)
LM01<-lm(formula=A$HSC.12th.Grade.Percentage ~ A$College.Placement.Percentage, data=A)
summary(LM01)
plotF <- ggplot(data = A, aes(x = HSC.12th.Grade.Percentage, y =College.Placement.Percentage))
+ geom_point()
plotF + geom_smooth(formula = y ~ x, method = "lm")
```

```
Call:
lm(formula = Reported.Salary ~ Gender + HSC.12th.Grade.Location +
    HS.Placement.Percentage + SSC.Placement.Test.Percentage.10th.Grade +
    College.Placement.Percentage + Graduate.Level.Placement.Percentage,
    data = A)

Residuals:
    Min      1Q  Median      3Q     Max
-120069  -47131  -16837   21006  611017

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               97186.43    9934.38   9.783  < 2e-16 ***
GenderM                                   35158.71    1674.15  21.001  < 2e-16 ***
HSC.12th.Grade.LocationOthers             -4127.62    1523.68  -2.709 0.006757 **
HS.Placement.Percentage                   -1791.68     133.38 -13.433  < 2e-16 ***
SSC.Placement.Test.Percentage.10th.Grade   -360.70      99.22  -3.635 0.000278 ***
College.Placement.Percentage                821.35      59.04  13.912  < 2e-16 ***
Graduate.Level.Placement.Percentage        4140.79     154.46  26.809  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 88550 on 14645 degrees of freedom
  (6633 observations deleted due to missingness)
Multiple R-squared:  0.09661,   Adjusted R-squared:  0.09624
F-statistic:   261 on 6 and 14645 DF,  p-value: < 2.2e-16
```

Figure 18: Multiple regression model, with adjusted r-squared of about 10%. While statistically significant, I am not sure it is practically significant. The extremely high values of N make it likely that we have variance inflation.