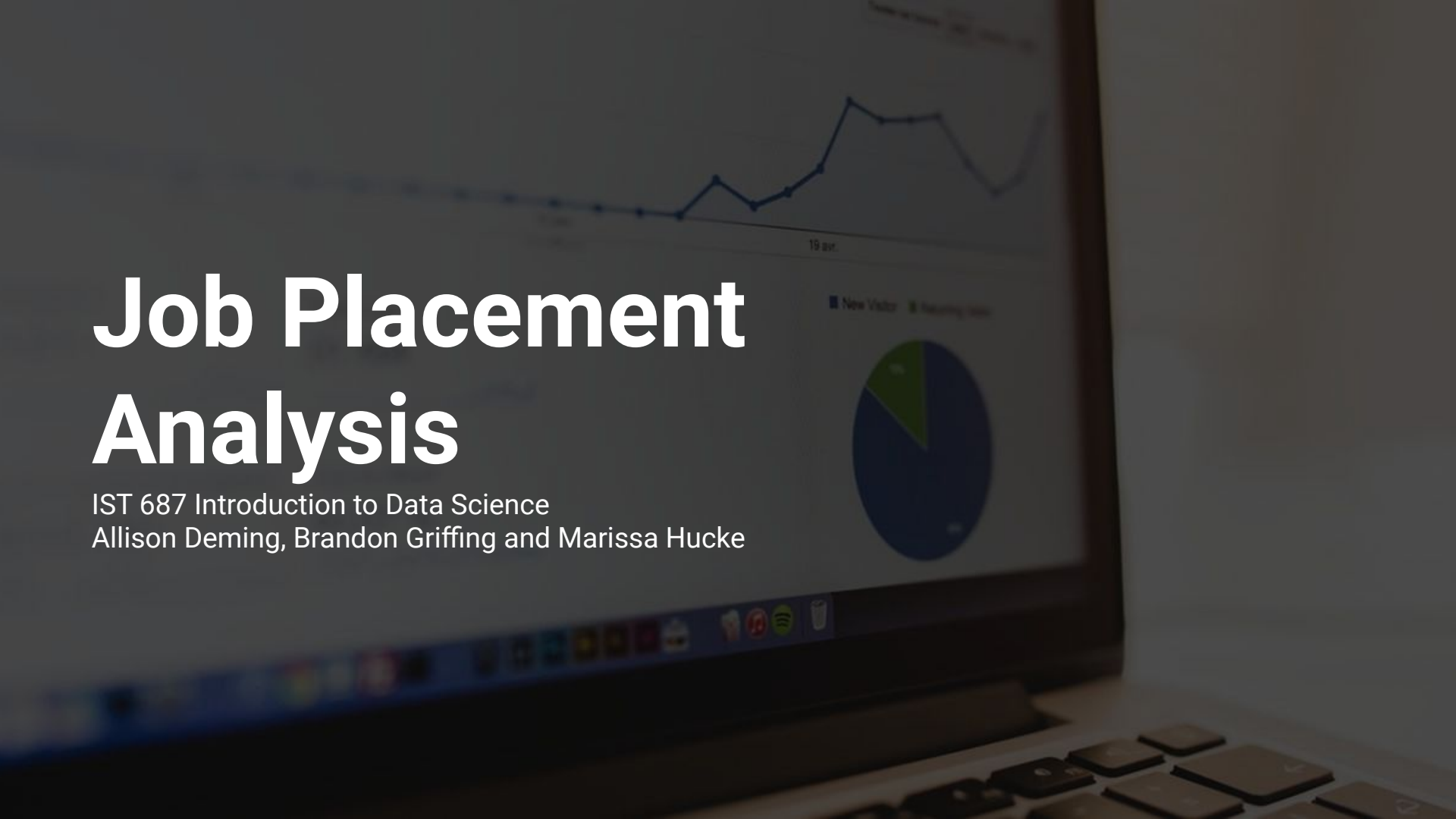


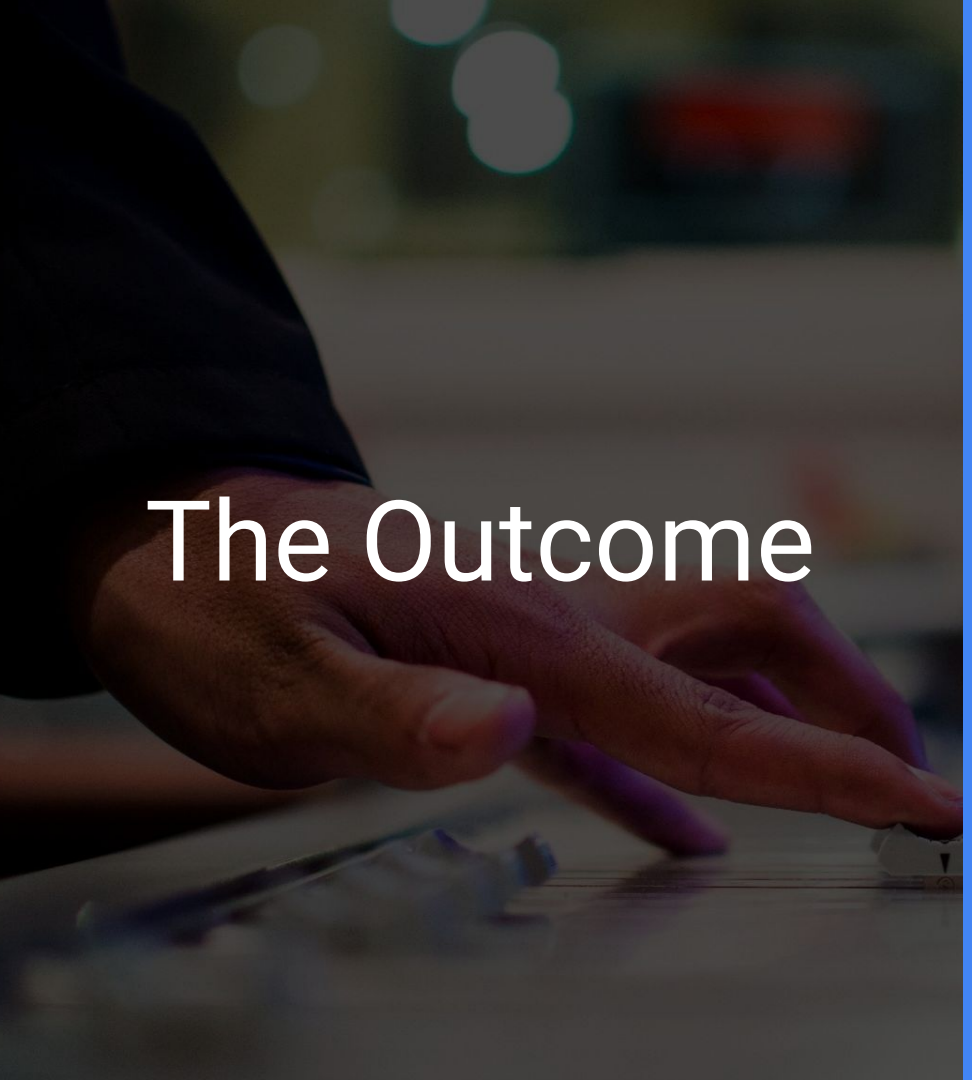
Job Placement Analysis

IST 687 Introduction to Data Science

Allison Deming, Brandon Griffing and Marissa Hucke

Objective

Analyze the organization’s student database to help identify different factors that contribute to a student's job placement status after college graduation

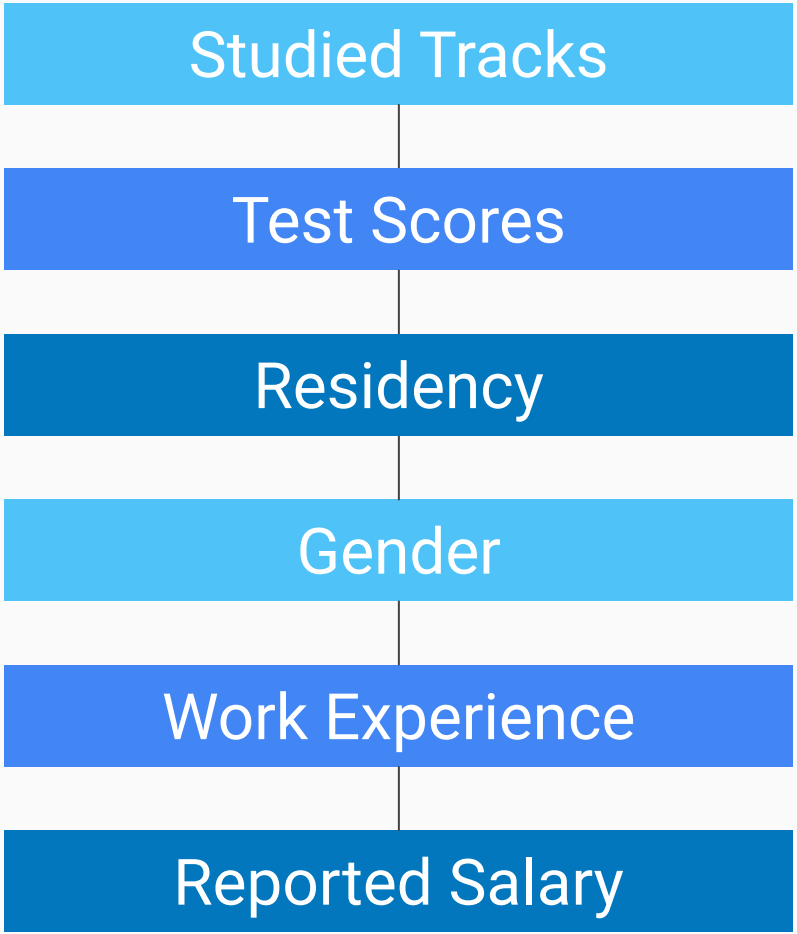


The Outcome

Provide the organization with future **insight and recommendations** on current academic programs and **student success projections**


The Dataset

The dataset included **12 attributes** focusing in these areas using 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?select=Placement_Data_Full_Class.csv

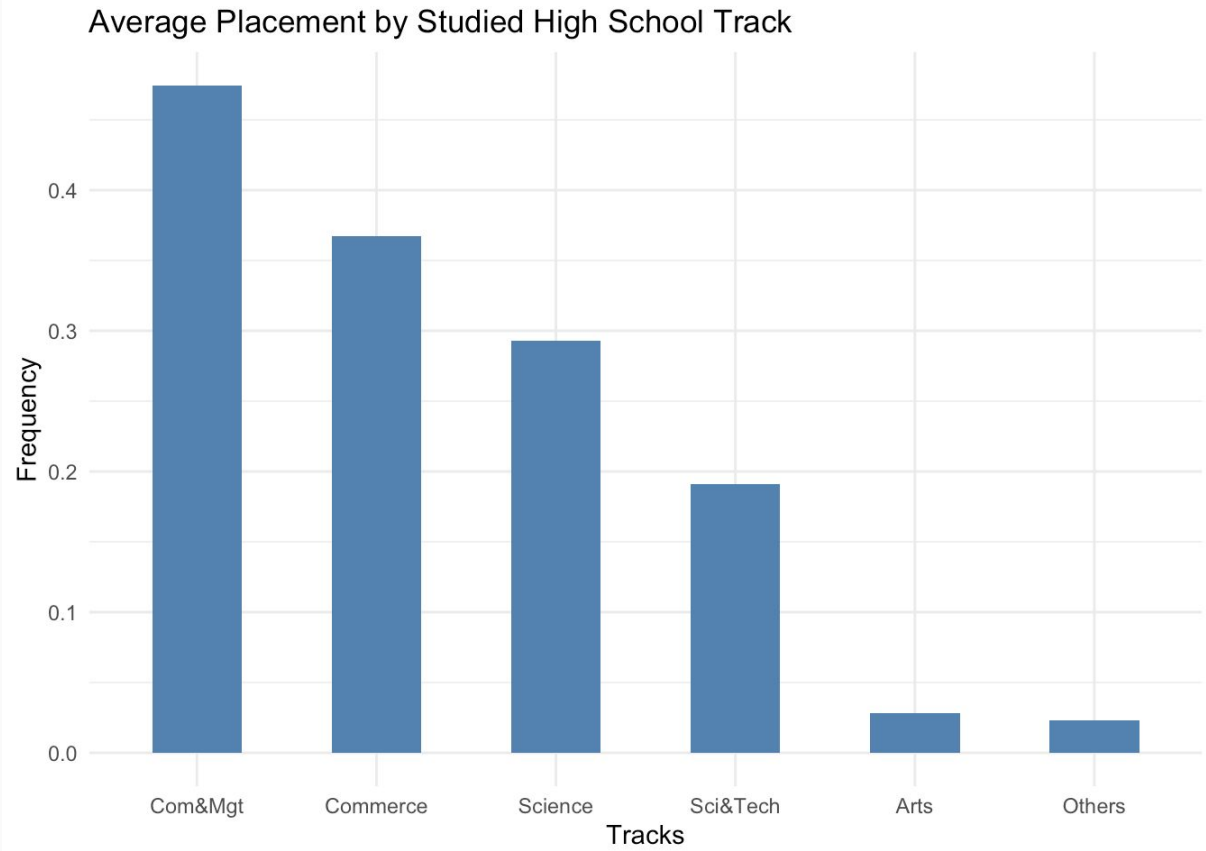


Placement by High School Tracks

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

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

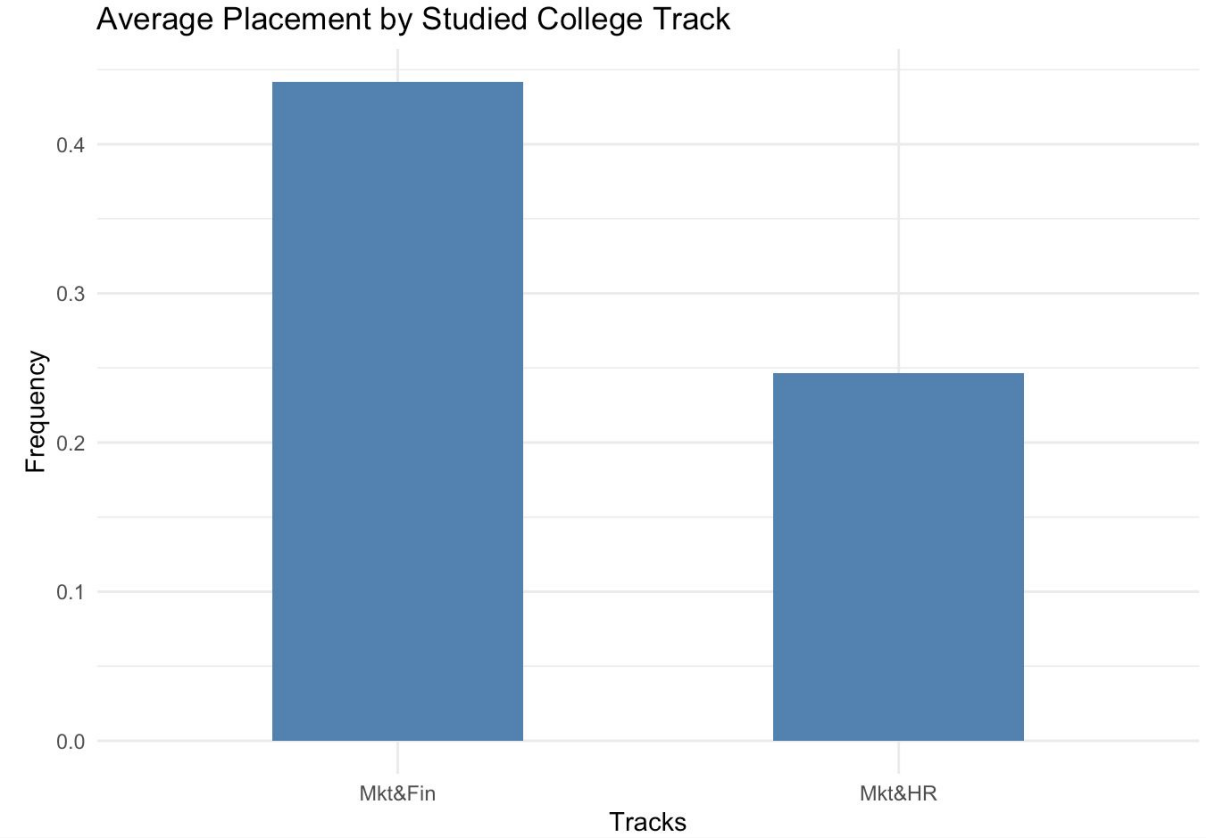


Placement by College Tracks

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

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

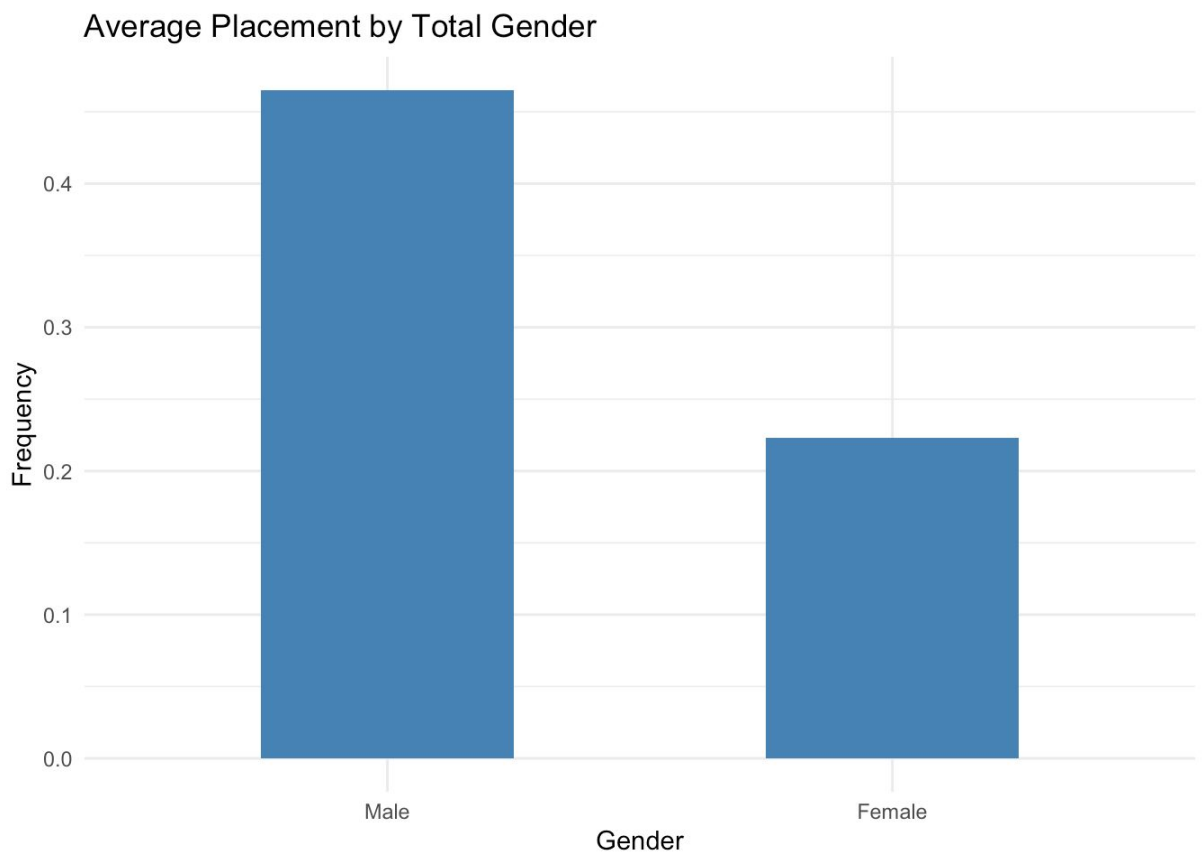


Placement by Gender (Disproportionate)

Analysis type: Filter, Average and ggplot geom bar

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.

- **22%** of female students placed compared to **47%** of male students who placed (disproportionate)

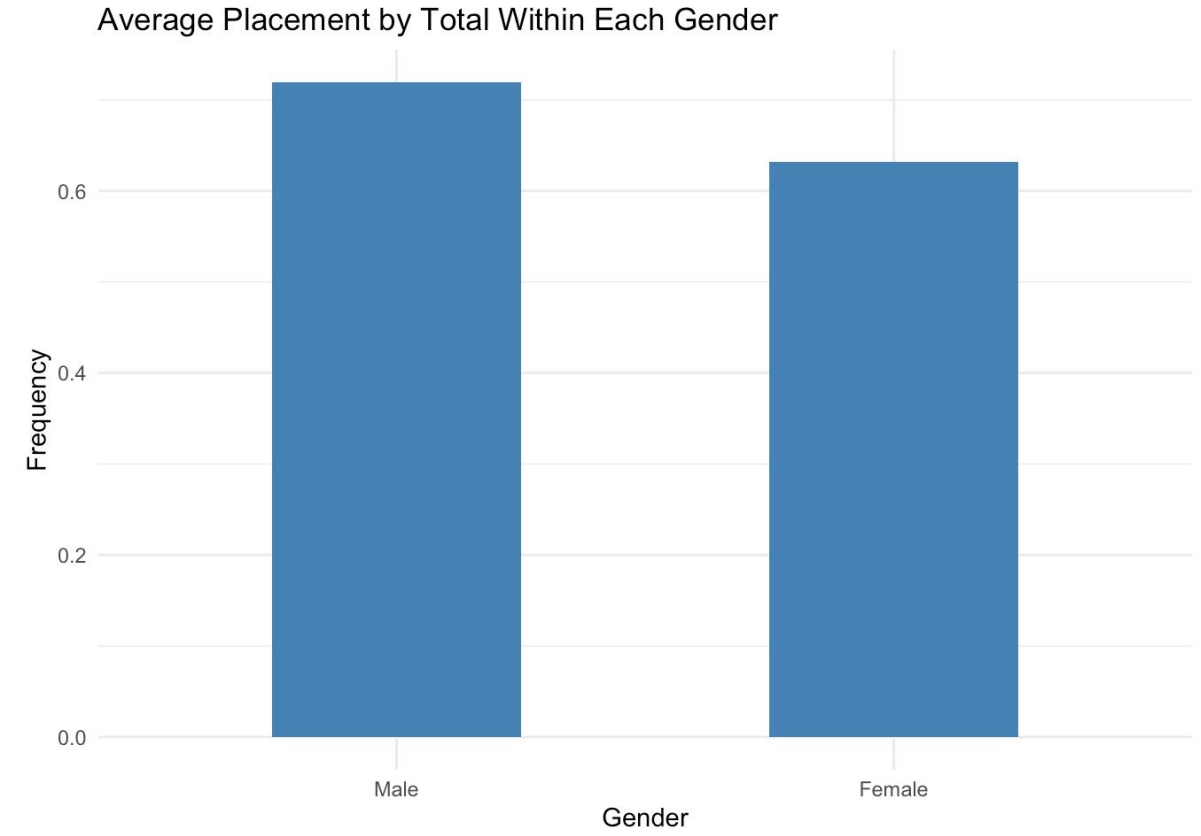


Placement by Gender (Proportionate)

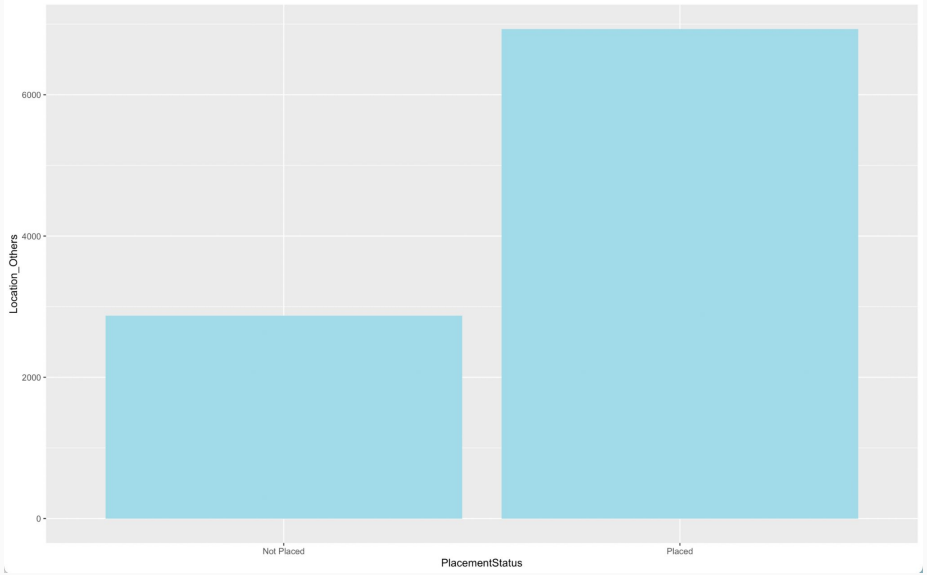
Analysis type: Filter, Average and ggplot geom bar

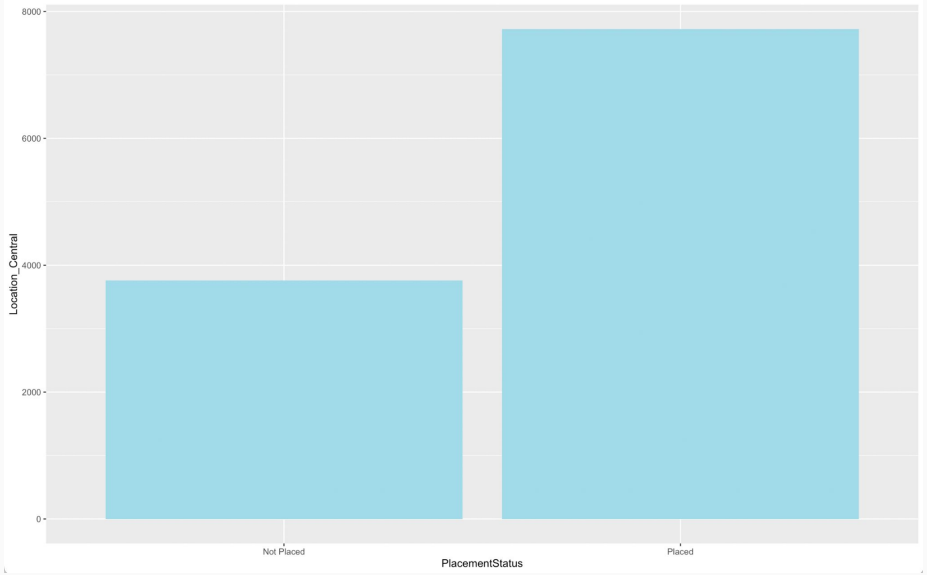
Results: When you take the total number of placed female students out of the total number of **female students (7,524)**, and the same with **male students (13,761)**, you can see that the average placement between gender is more comparative.

- **63%** of female students placed compared to 72% of male students who placed (proportionate)



Placement by Location: Tenth Grade

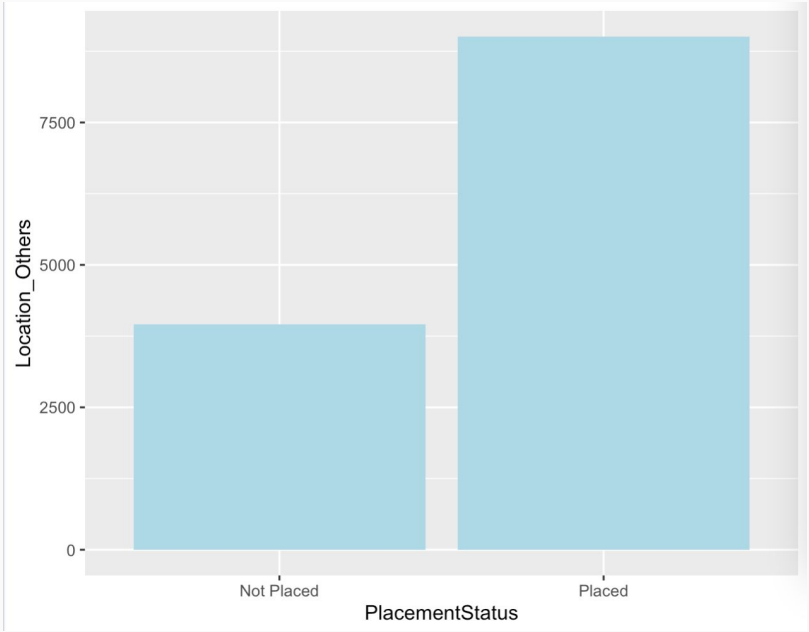


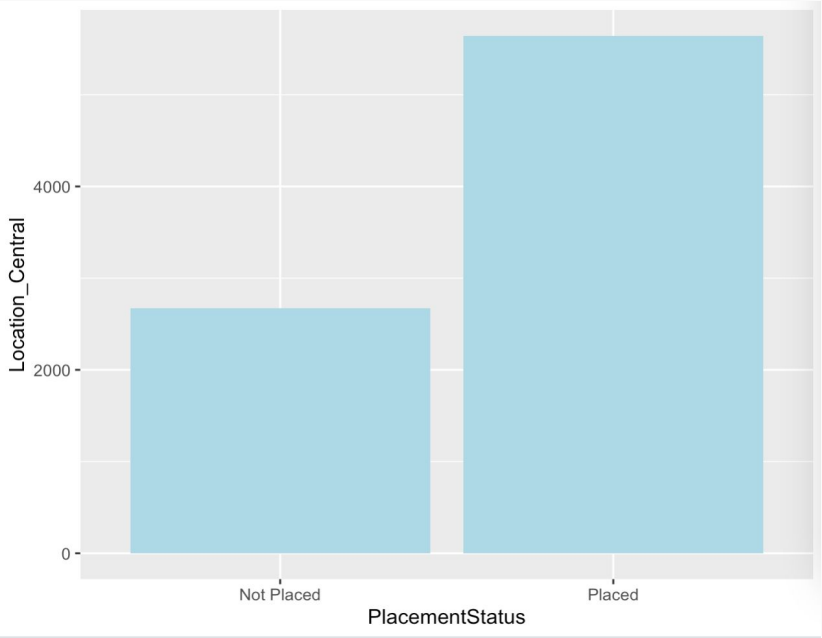


Analysis type: Filter, Rename, ggplot geom bar

Results: 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.

Placement by Location: Twelfth Grade





Analysis type: Filter, Rename, ggplot geom bar

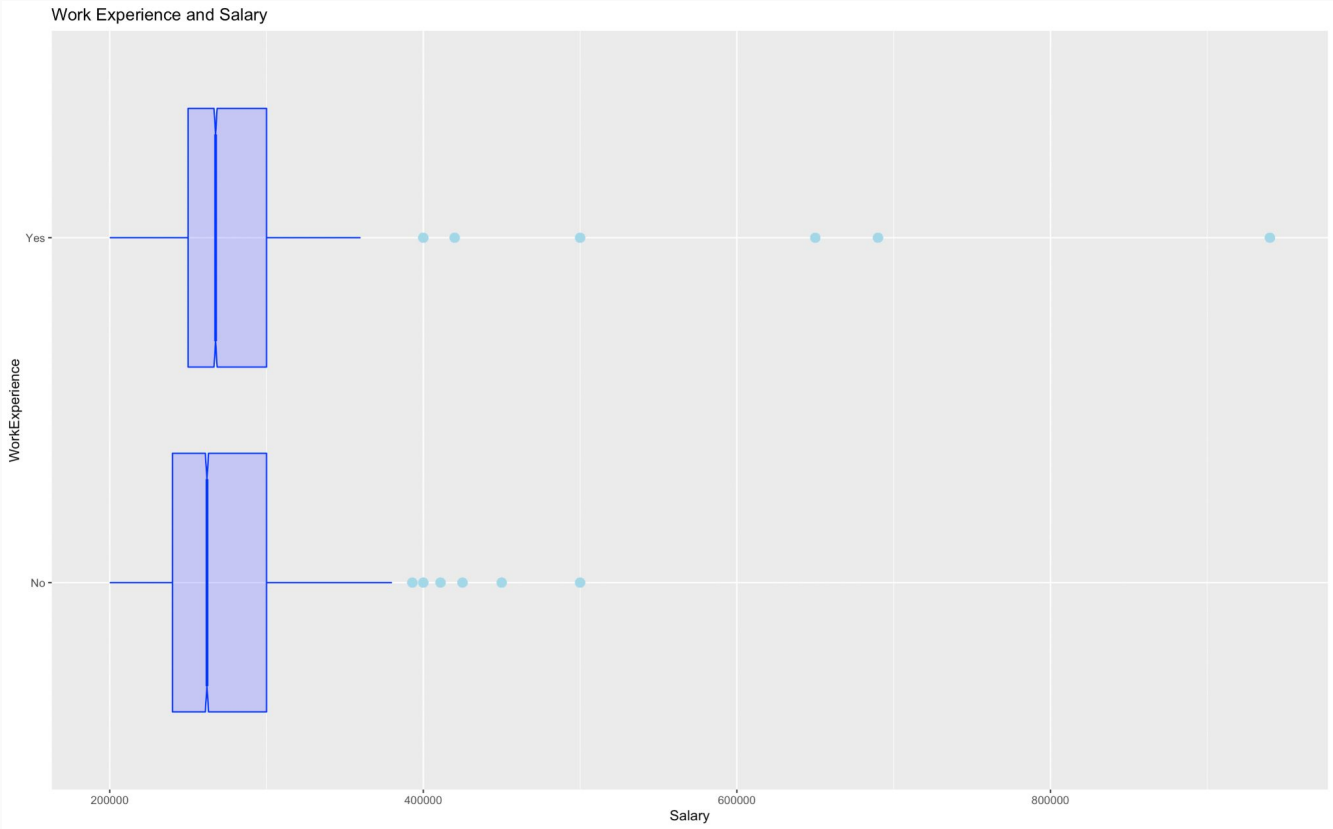
Results: Location plays a slight part in placement for twelfth grade given one from Central location has about a 56% higher chance of being placed from those in Other locations where individuals have about a 53% chance.

Reported Salary and Work Experience

Analysis type: Drop, Rename, ggplot geom_boxplot

Results:

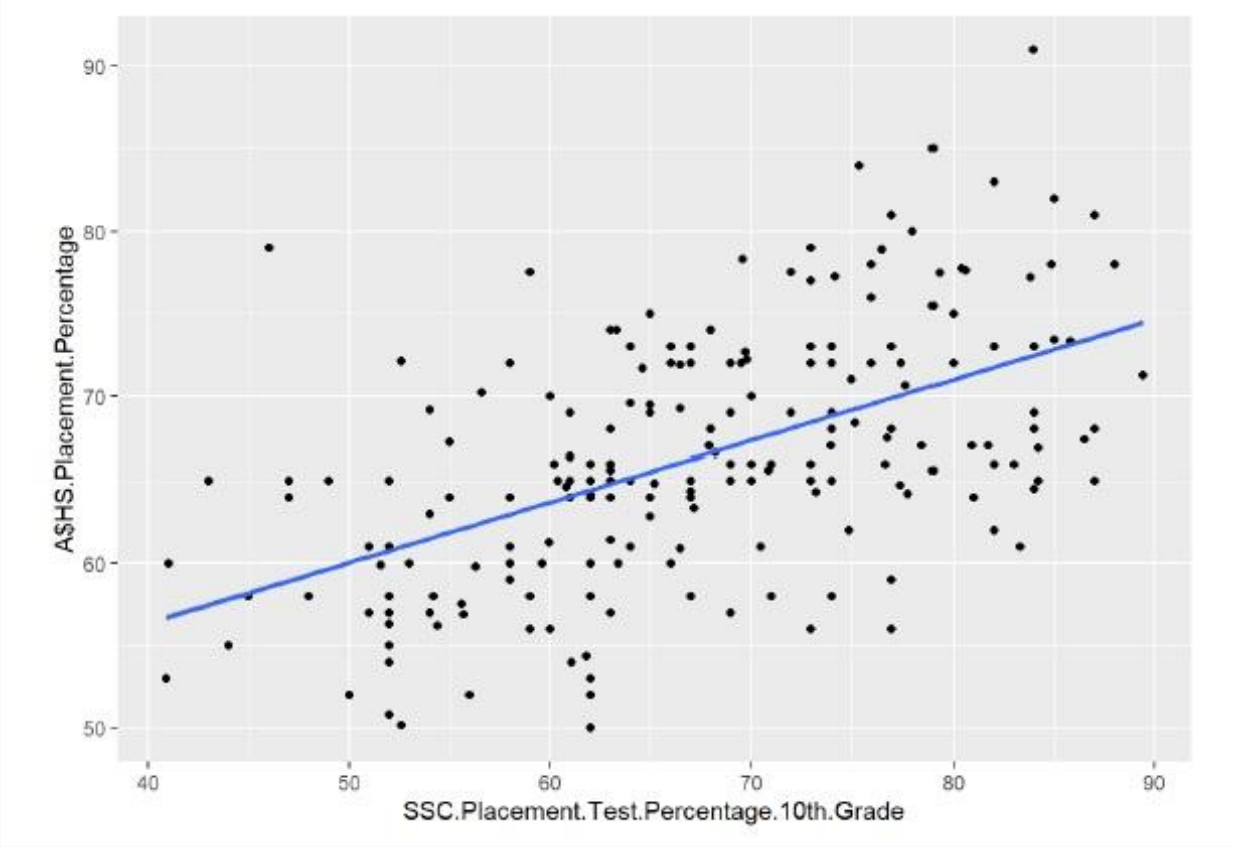
- Having work experience does impact one's salary due to having a higher starting salary with Work Experience (Q1)
- Maximum salary for one with "No" experience is $5,003.51 (380000 INR) as oppose to someone with work experience which is $4,740.17 (360000 INR)



Correlation, Linear Models, and Scatterplots - Test Scores / Grade Level

Analysis type: Linear Model, Correlation, Plots

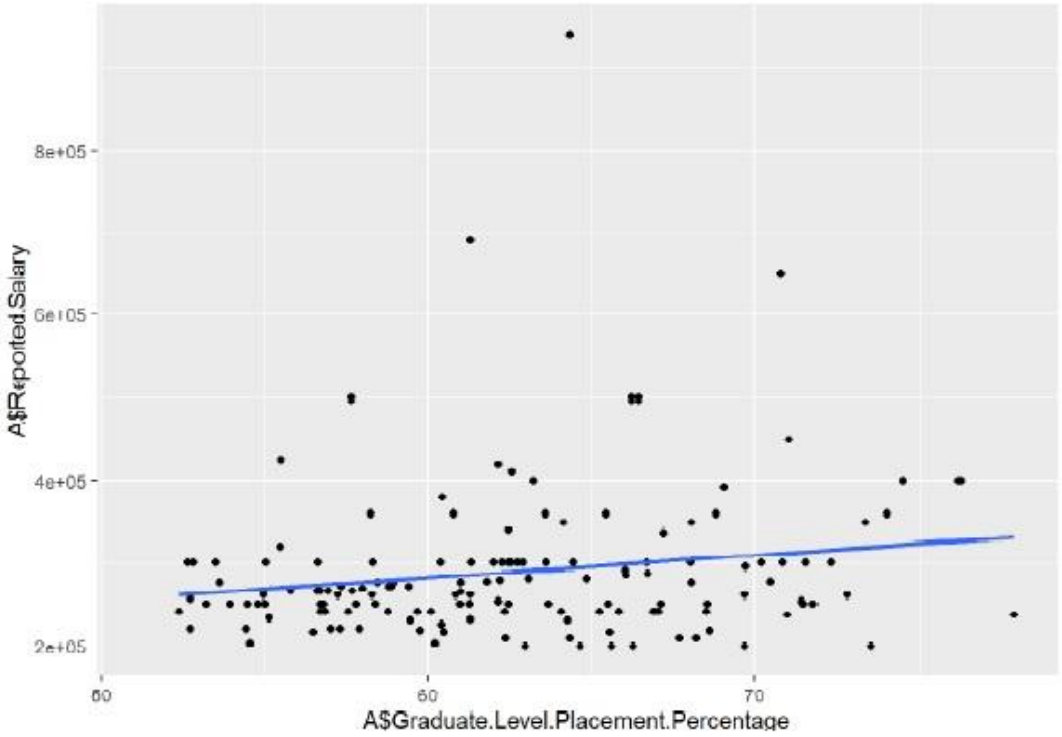
Results: Here the correlation was positive, and equal to .54, with the adjusted r-squared equal to .29 (so about 30% of the variance).



Correlation, Linear Models, and Scatterplots - Test Scores / Reported Salary

Analysis type: Linear Model, Correlation, Plots

Results: The correlation is equal to .18, but the regression model had an adjusted r-squared of less than 4%.




```
#Here there is a positive correlation between the two variables, with r=.18, and a very small am
ount of explained variance with no practical significance.
```

Correlation, Linear Models, and Scatterplots - All Significant Predictors for Salary

Analysis type: Multivariate Model

Results: The multivariate model had an adjusted r-squared of less than 10%

This data excludes students who were not placed which accounts for more than 6,000 of the 10,000 observations. Without this data it makes it more difficult to predict outcomes such as Reported Salary. Based on this analysis we recommend that the dataset includes additional information pertaining to the population of students not placed.

```
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
```

Placement Test Percentage By Being Placed (Graduate)

Population of Missing Data:

```
aggregate(A$Graduate.Level.Placement.Percentage, by=list(A$Placement.Status), FUN=mean)
```

```
##      Group.1        x
## 1 Not Placed 61.61284
## 2     Placed 62.57939
```

```
#This is really close, really, really close.
```

Thank You

Questions?

Appendix

Placement by High School or College Tracks Sample R Code

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")
```

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")
```

Placement by Location: Tenth Grade

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# 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
```

Placement by Location: Twelfth Grade

Sample R Code:

```
# 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
# 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
```

Reported Salary and Work Experience

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
```

Example R Code: Correlation, Linear Modeling, Scatterplots

```
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")
```