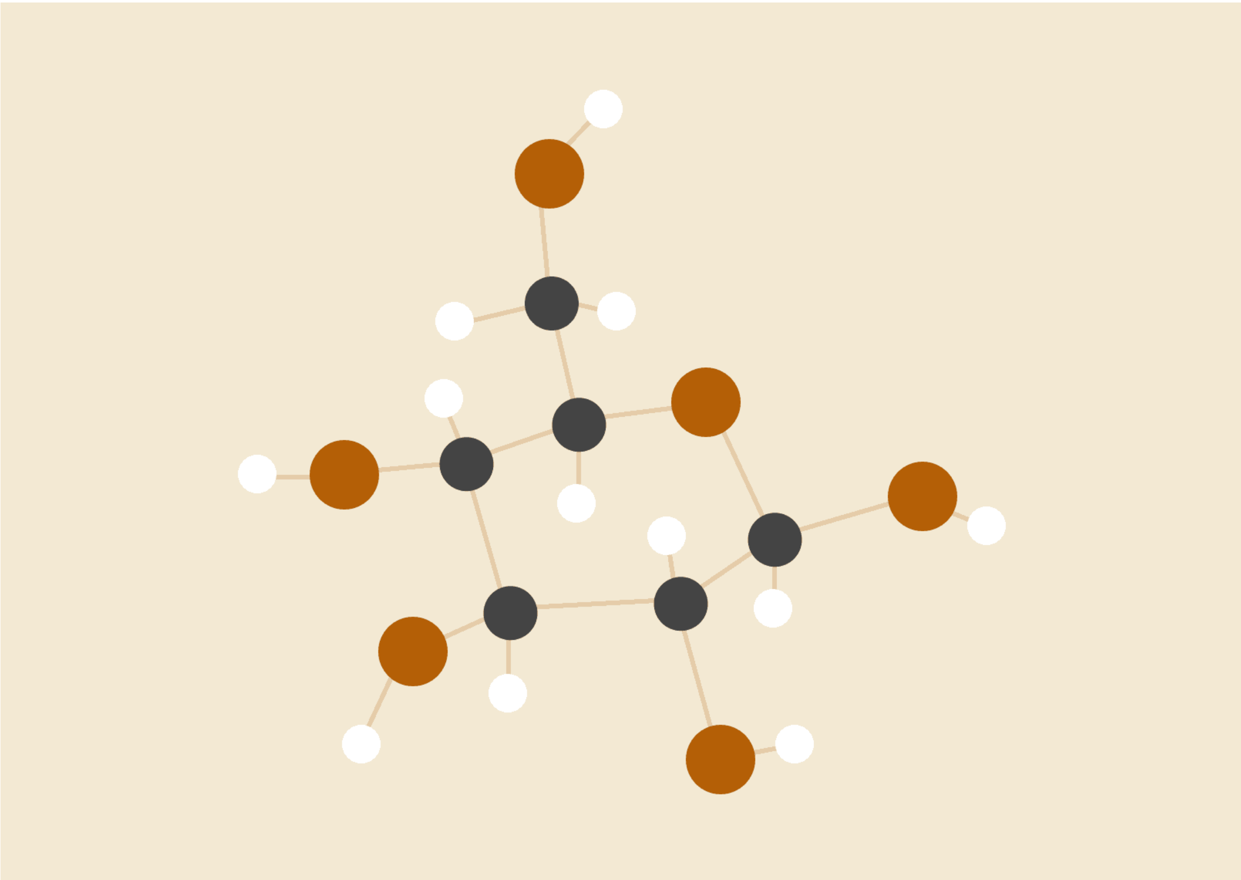
DEGREES THAT PAY YOU BACK



Zhaochuan Lu, Michelle Ortiz, and Emily Nguyen

05.14.2021

STAT 495: Introduction to R Programming

Section 01

Class Number 10072

INTRODUCTION

For our project, we chose a dataset from Kaggle that describes the relationship between undergraduate major and salary post-commencement. It also contains information regarding starting and mid-career median salary, percent change from starting to mid-career salary, as well as the 10th, 25th, 50th, and 75th percentile mid-career salaries. In our analysis, we will carry out certain methods to explore the relationship between starting and mid-career salaries, which degrees make the most money, and what the average starting median salary is for any degree.

We used the following variables:

UMajor: The majors of degrees earned.

Start_Med_Sal: The median of starting salaries for each major.

Mid_Med_Sal: The median of mid-career salaries for each major.

Perc_Change: Changes between starting salaries and mid-career salaries in percentage.

Mid_10_Sal: D1 of mid-career salaries for each major.

Mid_25_Sal: Q1 of mid-career salaries for each major.

Mid_75_Sal: Q3 of mid-career salaries for each major.

Mid_90_Sal: D9 of mid-career salaries for each major.

Degree: The field of the major (STEM, Business, Humanity).

QUESTIONS OF INTEREST

Throughout our project, we aim to answer the following questions:

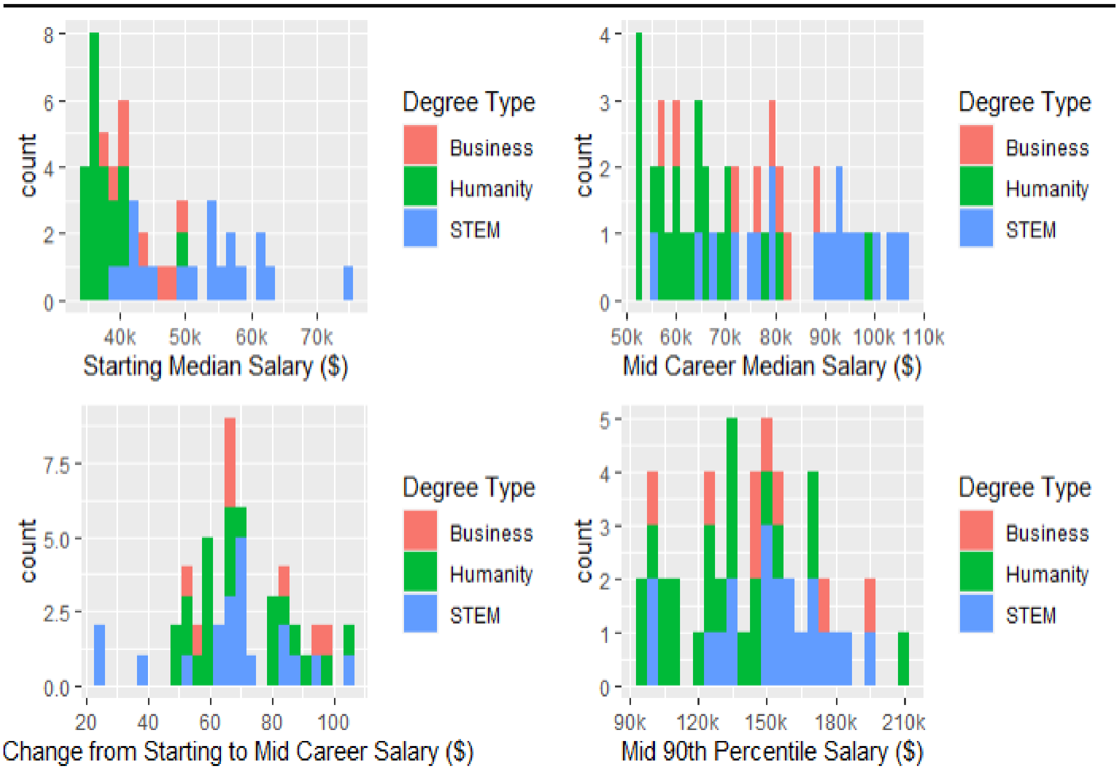
1. What is the average starting median salary for a degree?

2. What is the relationship between starting median salary and mid-career median salary?
3. What are the degrees that make the most money?

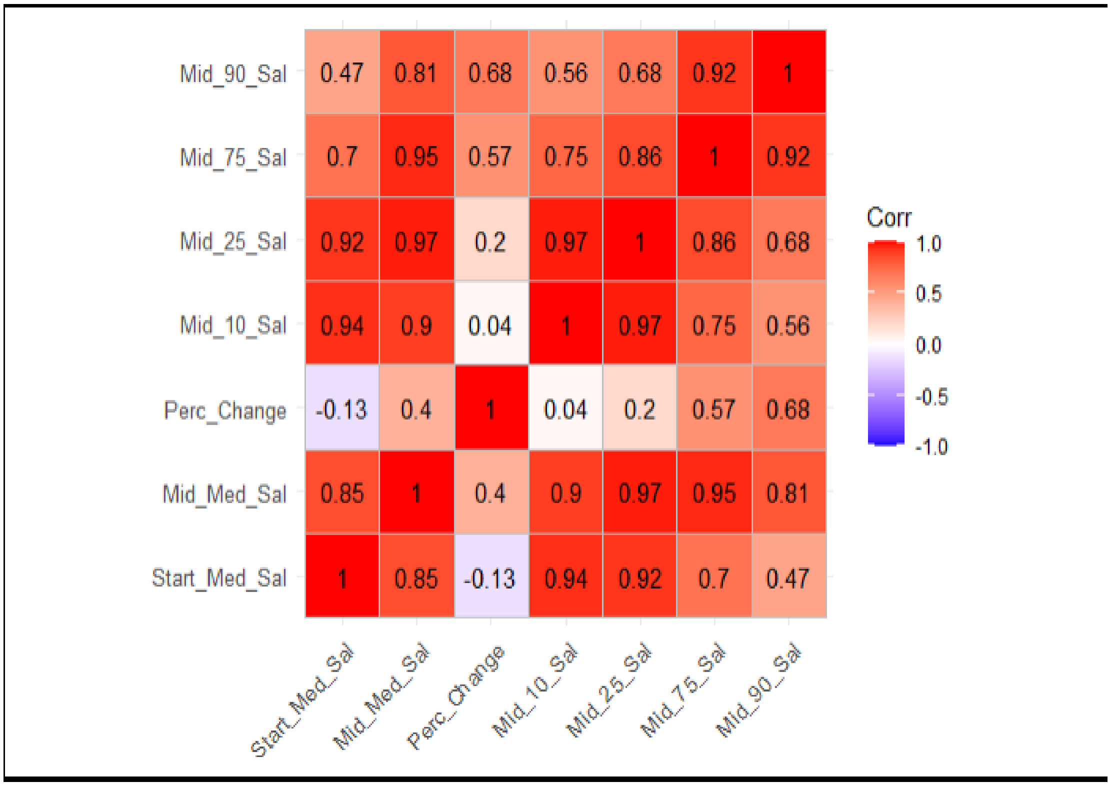
ANALYSIS

I. Exploratory Analysis

After conducting a set of exploratory analyses, a few conclusions about this dataset can be made. First, STEM majors seem to fare generally well with their starting salary. In mid-career salaries, STEM majors are still on the higher end of the spectrum, while business and humanities major salaries are more spread out. All three groups generally have the same percent change. However, humanities degrees scored the highest in the mid 90th percentile salary, suggesting that majors in this group have opportunities for immense salary growth. Even so, these interpretations are not conclusive and we must complete further analysis to confirm they are correct.



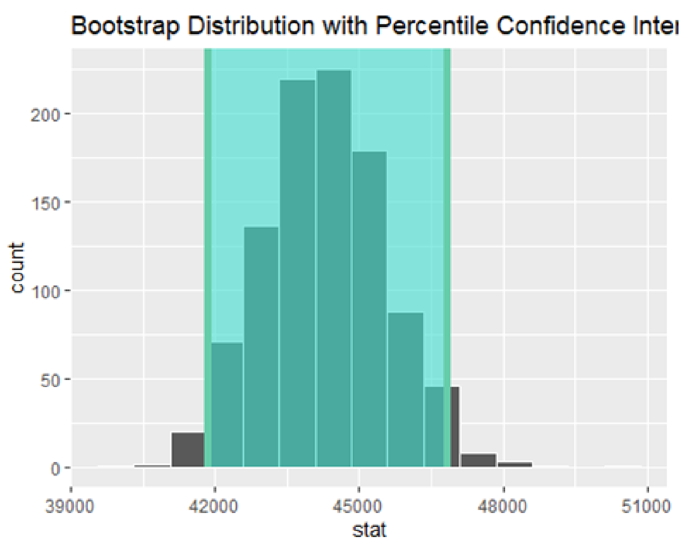
As for mid-career salaries, more than half of undergraduate majors receive a median salary of less than $80,000. Twelve majors are able to get a 70% increase from starting median salary to mid-career salary, while two majors are able to achieve more than a 100% increase from starting to mid-career salaries. There are no outliers in mid-career starting salary; however, there is one major that has a 23.4% increase from starting to mid-career salary.

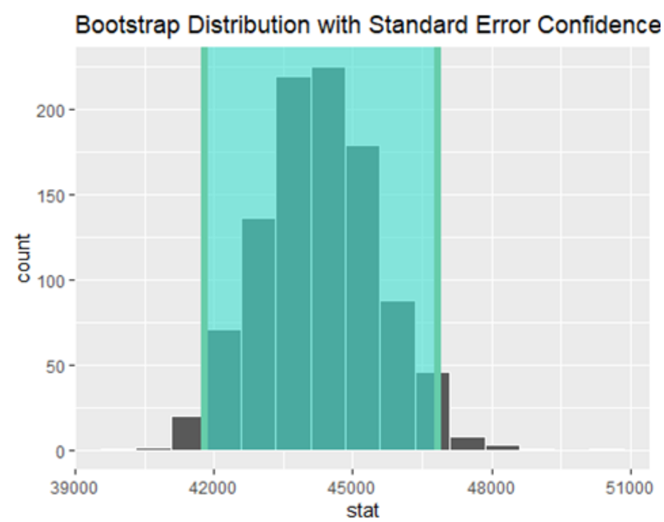


Judging from our correlation matrix, starting median salary seems to be highly correlated with mid 10th and 25th percentile salaries, and the mid-career median salary is highly correlated with mid 25th and mid 75th percentile salaries. Since they have a high correlation with other variables, Mid 10th, 25th, and 75th percentiles will be dropped.

II. Calculating Average Starting Median Salary for a Degree

We performed a bootstrap distribution in order to sample our estimates and get a better feel of our dataset. The distribution shows that most of the majors have a starting median salary of fewer than $60,000. First, we resampled our data 1,000 times. Our results show that the lowest starting median salary for college graduates is around $42,000, while the highest starting median salary is $47,000 per year.





III. Relationship Between Starting Median Salary and Mid-Career Median Salary

From the previous heatmap and scatterplot, there is a positive correlation between starting median salary and mid-career median salary. To find out the quantitative relationship between these two variables, the dataset is fitted into a linear regression model.

```
#fit regression model
salary_model <- lm(Mid_Med_Sal ~ Start_Med_Sal, data = degrees)
#get regression table
get_regression_table(salary_model)

## # A tibble: 2 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     10172.    5944.         1.71   0.093 -1778.   22122.
## 2 Start_Med_Sal     1.46     0.131     11.1    0          1.19     1.72

#observed/fitted values and residuals
regression_points <- get_regression_points(salary_model)
regression_points

## # A tibble: 50 x 5
##       ID Mid_Med_Sal Start_Med_Sal Mid_Med_Sal_hat residual
##    <int>       <dbl>         <dbl>           <dbl>    <dbl>
##  1     1       77100         46000          77250.    -150.
##  2     2      101000         57700          94312.    6688.
##  3     3       71900         42600          72292.    -392.
##  4     4       61500         36800          63835.   -2335.
##  5     5       76800         41600          70834.    5966.
##  6     6       64900         35800          62376.    2524.
##  7     7       64800         38800          66751.   -1951.
##  8     8       72100         43000          72876.    -776.
##  9     9      107000         63200         102332.    4668.
```

The summary of the regression model shows that the coefficient of the variable is 1.46 and the intercept of the regression is 10,172. So, the regression model can be written as follows:

$$\text{Mid\_Med\_Sal} = 1.46 \times \text{Start\_Med\_Sal} + 10172$$

The adjusted R-square of the regression model is 0.71, which indicates relatively high goodness of fit. Intuitively, the regression results show that if one gets one unit ($) more on the starting salary, they would expect a 1.46 dollar increase in the mid-career stage.

IV. Majors that Make the Most

When sorting the dataset according to different columns, different answers can be made when it comes to different career stages. For starting median salary, Physician Assistants led the charts by $74,300.

```
## # A tibble: 6 x 6
##   UMajor          Start_Med_Sal Mid_Med_Sal Perc_Change Mid_90_Sal `Degree
Type`
##   <chr>                   <dbl>       <dbl>       <dbl>      <dbl> <chr>
## 1 Physician Assi~         74300       91700        23.4     124000 STEM
## 2 Chemical Engin~         63200      107000        69.3     194000 STEM
## 3 Computer Engin~         61400      105000        71       162000 STEM
## 4 Electrical Eng~         60900      103000        69.1     168000 STEM
## 5 Mechanical Eng~         57900       93600        61.7     163000 STEM
## 6 Aerospace Engi~         57700      101000        75       161000 STEM
```

For mid-career median salary, Chemical Engineering majors had the highest median salary of $63,200.

```
## 1 Chemical Engin~         63200      107000        69.3     194000 STEM
## 2 Computer Engin~         61400      105000        71       162000 STEM
## 3 Electrical Eng~         60900      103000        69.1     168000 STEM
## 4 Aerospace Engi~         57700      101000        75       161000 STEM
## 5 Economics               50100       98600        96.8     210000
Humanity
## 6 Physics                 50300       97300        93.4     178000 STEM
```

However, Math and Philosophy majors had the highest potential for growth of 104%. The starting median salary for these majors are $45,400 and $39,900 respectively, and the mid-career median salary for these majors are $92,400 and $81,200 respectively. The top ten majors for salary growth from starting to mid-career is as follows:

```
## 1 Math                    45400       92400       104.      183000 STEM
## 2 Philosophy              39900       81200       104.      168000
Humanity
## 3 International ~         40900       80900        97.8     157000
Business
## 4 Economics               50100       98600        96.8     210000
Humanity
## 5 Marketing               40800       79600        95.1     175000
Business
## 6 Physics                 50300       97300        93.4     178000 STEM
```

CONCLUSION

After conducting different statistical methods such as linear regression and bootstrapping, it is clear that different undergraduate majors can lead to different ranges of salaries. However, the overall range is relatively concentrated regardless of the major, with the exception of a few outliers. The salaries of the mid-career stage are heavily influenced by the salaries of the early stage, namely, the starting median salary. If you want to choose majors that make the most money in entry-level jobs, Physician Assistants would be the best choice. If you want an overall stable career, a Chemical Engineering degree will pay the most mid-career. However, if you wish to pursue the potential for large salary growth, consider becoming a Math or Philosophy major, as they contain the highest percent change from starting to mid-career salary.

APPENDIX

```
# **1. Exploratory Data Analysis**
#load library
library(readr)
library(tibble)
library(tidyverse)
#read in data
degrees <- as.tibble(read_csv('degrees-that-pay-back.csv'))
#remove $ and , from columns
degrees <- data.frame(lapply(degrees, function(x) {
            gsub("[$,]", "", x)}))
#convert character columns to numeric
degrees[, 2:8] <- sapply(degrees[, c(2:8)], as.numeric)
#shorter variable names
degrees <- degrees %>%
  rename(
    UMajor = Undergraduate.Major,
    Start_Med_Sal = Starting.Median.Salary,
```

```
    Mid_Med_Sal = Mid.Career.Median.Salary,
    Perc_Change = Percent.change.from.Starting.to.Mid.Career.Salary,
    Mid_10_Sal = Mid.Career.10th.Percentile.Salary,
    Mid_25_Sal = Mid.Career.25th.Percentile.Salary,
    Mid_75_Sal = Mid.Career.75th.Percentile.Salary,
    Mid_90_Sal = Mid.Career.90th.Percentile.Salary
  )
#Add variable
`Degree Type` <- c("Business", "STEM", "STEM", "Humanity", "STEM",
"Humanity", "STEM", "Business", "STEM", "STEM", "STEM", "Humanity",
"STEM", "STEM", "STEM", "Humanity", "Humanity", "Humanity",
"Humanity", "STEM", "Humanity", "Humanity", "Business", "Humanity",
"Humanity", "STEM", "Humanity", "Business", "Humanity", "Business",
"STEM", "STEM", "Humanity", "Business", "Humanity", "Business",
"Business", "STEM", "STEM", "Humanity", "STEM", "STEM", "Humanity",
"STEM", "STEM", "Humanity", "Humanity", "Humanity", "Humanity",
"Humanity")
degrees <-  degrees %>%
  add_column(`Degree Type`)
#convert to tibble
degrees <- as.tibble(degrees)
#print dataframe
str(degrees)
```

1.1 Summary Statistics Table

```
#load libraries
library(skimr)
#summary table
skim_without_charts(degrees)
```

1.2 Check Correlation Between Continuous Feature Variables

```
#load libraries
library(ggplot2)
library(ggcorrplot)
#subset continuous variables
noncontinuous <- names(degrees) %in% c("UMajor", "Degree Type")
degrees_continuous <- degrees[!noncontinuous]
#calculate correlations
```

```
degrees_correlation = cor(degrees_continuous)

#plot correlations

ggcorrplot(degrees_correlation, tl.cex = 10, lab = TRUE)

#dropping highly correlated variables

drop <- names(degrees) %in% c("Mid_10_Sal", "Mid_25_Sal",
"Mid_75_Sal")

degrees <- degrees[!drop]
```

1.3 Histograms, Scatterplot Matrix, Boxplots

```
#load libraries

library(gridExtra)

#histograms

d1 <- ggplot(degrees, aes(x = Start_Med_Sal)) +

  geom_histogram(binwidth = 1500) +

  aes(fill = `Degree Type`) +

  xlab("Starting Median Salary ($)") +

  scale_x_continuous(breaks = c(40000, 50000, 60000, 70000),

                     labels = c("40k", "50k", "60k", "70k"))

d2 <- ggplot(degrees, aes(x = Mid_Med_Sal)) +

  geom_histogram(binwidth = 1500) +

  aes(fill = `Degree Type`) +

  xlab("Mid Career Median Salary ($)") +

  scale_x_continuous(breaks = c(50000, 60000, 70000, 80000, 90000,
100000, 110000),

                     labels = c("50k", "60k", "70k", "80k", "90k",
"100k", "110k"))

d3 <- ggplot(degrees, aes(x = Perc_Change)) +

  geom_histogram(binwidth = 3.5) +

  aes(fill = `Degree Type`) +

  xlab("% Change from Starting to Mid Career Salary ($)")

d4 <- ggplot(degrees, aes(x = Mid_90_Sal)) +

  geom_histogram(binwidth = 5000) +

  aes(fill = `Degree Type`) +

  xlab("Mid 90th Percentile Salary ($)") +
```

```
  scale_x_continuous(breaks = c(90000, 120000, 150000, 180000, 
210000),
                     labels = c("90k", "120k", "150k", "180k", 
"210k"))

grid.arrange(d1, d2, d3, d4, ncol = 2, nrow = 2)

#scatterplots

drop <- names(degrees_continuous) %in% c("Mid_10_Sal", "Mid_25_Sal", 
"Mid_75_Sal")

degrees_continuous <- degrees_continuous[!drop]

pairs(degrees_continuous, lower.panel = NULL, cex.labels = .8, cex = 
.2)

#boxplots

d5 <- ggplot(degrees, aes(x=Start_Med_Sal)) +

  geom_boxplot() +

  xlab("Starting Median Salary ($)") +

  xlab("Starting Median Salary ($)") +

  scale_x_continuous(breaks = c(40000, 50000, 60000, 70000),

                     labels = c("40k", "50k", "60k", "70k"))

d6 <- ggplot(degrees, aes(x=Mid_Med_Sal)) +

  geom_boxplot() +

  xlab("Mid Career Median Salary ($)") +

  scale_x_continuous(breaks = c(50000, 60000, 70000, 80000, 90000, 
100000, 110000),

                     labels = c("50k", "60k", "70k", "80k", "90k", 
"100k", "110k"))

d7 <- ggplot(degrees, aes(x=Perc_Change)) +

  geom_boxplot() +

  xlab("% Change from Starting to Mid Career ($)")

d8 <- ggplot(degrees, aes(x=Mid_90_Sal)) +

  geom_boxplot() +

  xlab("Mid 90th Percentile Salary ($)") +

  scale_x_continuous(breaks = c(90000, 120000, 150000, 180000, 
210000),

                     labels = c("90k", "120k", "150k", "180k", 
"210k"))

grid.arrange(d5, d6, d7, d8, ncol = 2, nrow = 2)
```

```
# **2. Answering Questions**

**2.1 What is the average starting median salary for a degree?**

#load libraries

library(infer)

#average median salary for a degree

x_bar <- degrees %>%

  summarise(mean_start_med_sal = mean(Start_Med_Sal))

#specify variables, generate reps and calculate summary stats

bootstrap_dist <- degrees %>%

  specify(response = Start_Med_Sal) %>%

  generate(reps = 1000) %>%

  calculate(stat = "mean")

#visualize results

visualize(bootstrap_dist) +

  ggtitle("Bootstrap Distribution of Average Starting Median Salary
($)")

#calculate percentile confidence interval

percentile_ci <- bootstrap_dist %>%

  get_confidence_interval(level = 0.95, type = "percentile")

percentile_ci

#visualize percentile interval

visualize(bootstrap_dist) +

  shade_confidence_interval(endpoints = percentile_ci) +

  ggtitle("Bootstrap Distribution with Percentile Confidence
Interval")

#calculate standard error confidence interval

standard_error_ci <- bootstrap_dist %>%

  get_confidence_interval(level = 0.95, type = "se", point_estimate =
x_bar)

standard_error_ci

#visualize standard error interval

visualize(bootstrap_dist) +

  shade_confidence_interval(endpoints = standard_error_ci) +
```

```
  ggtitle("Bootstrap Distribution with Standard Error Confidence 
Interval")
```

****2.2 What is the relationship between Starting Median Salary and Mid Career Median Salary?****

```
#load libraries

library(scales)

library(moderndive)

#scatterplot

ggplot(degrees, aes(x = Start_Med_Sal, y = Mid_Med_Sal)) +

  geom_point(color = "navy") +

  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +

  labs(x = "Starting Median Salary ($)", y = "Mid Career Median 
Salary ($)") +

  ggtitle("Scatterplot of Starting Median Salary Vs Mid Career Median 
Salary") +

  scale_y_continuous(labels = comma) +

  scale_x_continuous(labels = comma)

#fit regression model

salary_model <- lm(Mid_Med_Sal ~ Start_Med_Sal, data = degrees)

#get regression table

get_regression_table(salary_model)

#observed/fitted values and residuals

regression_points <- get_regression_points(salary_model)

regression_points
```

****2.3 What are the degrees that make the most?****

```
#sorted by starting median salary

degrees_sorted1 <- degrees %>%

  arrange(desc(Start_Med_Sal))

head(degrees_sorted1)

#sorted by mid career median salary

degrees_sorted2 <- degrees %>%

  arrange(desc(Mid_Med_Sal))

head(degrees_sorted2)

#sorted by percent change
```

```
degrees_sorted3 <- degrees %>%
  arrange(desc(Perc_Change))
head(degrees_sorted3)
```