Final Project

Vi Nguyen

2020-05-12

Question of Interest

Any graduating senior in high school would want to know where they should invest their future. To many, success is measured by whether they get a stable full-time job upon graduation or quantitatively, whether their earnings are sufficient to support a living and at the same time, pay off college debts. Is it true that graduating from an Ivy League college gives students an advantage over others and a higher chance of succeeding in the workplace? How does the choice of college regarding its cost of attendance factor into students' earnings after graduation? A hypothesis test is used to help answer this question.

Columns: - Cost: COSTT4_A, COSTT4_P - Earnings: MN_EARN_WNE_P6, MN_EARN_WNE_P7, MN_EARN_WNE_P8, MN_EARN_WNE_P9, MN_EARN_WNE_P10

This is a worthwhile question to explore because it is important for students especially seniors in high school to consider when making their college investment.

Preprocessing

```
## Extract columns
college_red <- college %>%
  select(c(INSTNM, # Name
           COSTT4_A, COSTT4_P, # Costs
           MN_EARN_WNE_P6, MN_EARN_WNE_P7,
           MN_EARN_WNE_P8, MN_EARN_WNE_P9, MN_EARN_WNE_P10)) # Earnings
```

```
## COSTT4_A: cost of attendance for public institutions
## COSTT4_P: private (NAs)

college_red <- college_red %>%
  mutate(
    # Replace NA's in public with values from private
    COSTT4_A = if_else(
      condition = is.na(COSTT4_A),
      true = COSTT4_P,
      false = COSTT4_A
)
)%>%
  # Drop COSTT4_P
  select(-COSTT4_P) %>%

  # Rename columns
  rename(
    COST = COSTT4_A,
```

```
    `6` = MN_EARN_WNE_P6,
    `7` = MN_EARN_WNE_P7,
    `8` = MN_EARN_WNE_P8,
    `9` = MN_EARN_WNE_P9,
    `10` = MN_EARN_WNE_P10) %>%
  # "gather" into 1 column
  gather(`6`:`10`, key = 'YRS_AFTER_GRAD', value = 'EARNING')
# Change column type to numeric
college_red$YRS_AFTER_GRAD <- as.numeric(college_red$YRS_AFTER_GRAD)
```

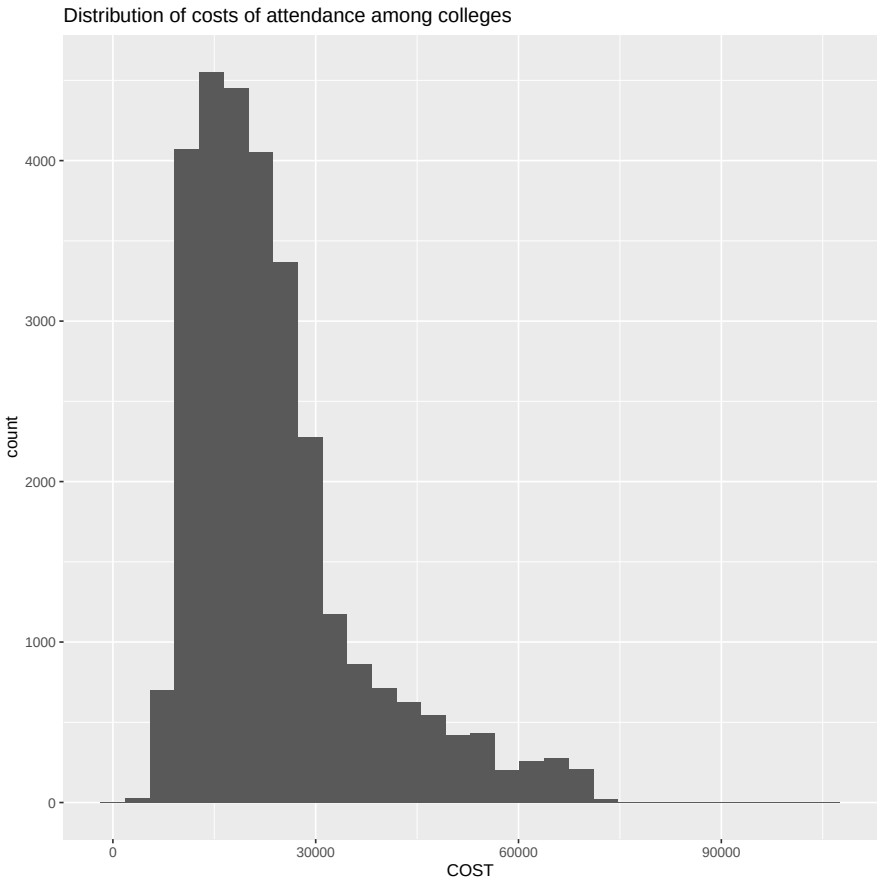
```
## Group contiunous variable into categories
## Below the 25th percentile = low
## Higher than the 75th percentile = high
## Anything in between is medium

## 25th percentile
cost25 <- quantile(college_red$COST, na.rm = TRUE)[2]
## 75th percentile
cost75 <- quantile(college_red$COST, na.rm = TRUE)[4]

college_red <- college_red %>%
  ## Create new categorical variable
  mutate(
    TUITION = case_when(
      COST < cost25 ~ 'low cost',
      between(COST, cost25, cost75) ~ 'medium cost',
      COST > cost75 ~ 'high cost'
  )
  )
```

Visualization

```
## Histogram using continuous COST
college_red %>%
  ggplot() +
  geom_histogram(mapping = aes(COST)) +
  labs(title = 'Distribution of costs of attendance among colleges')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

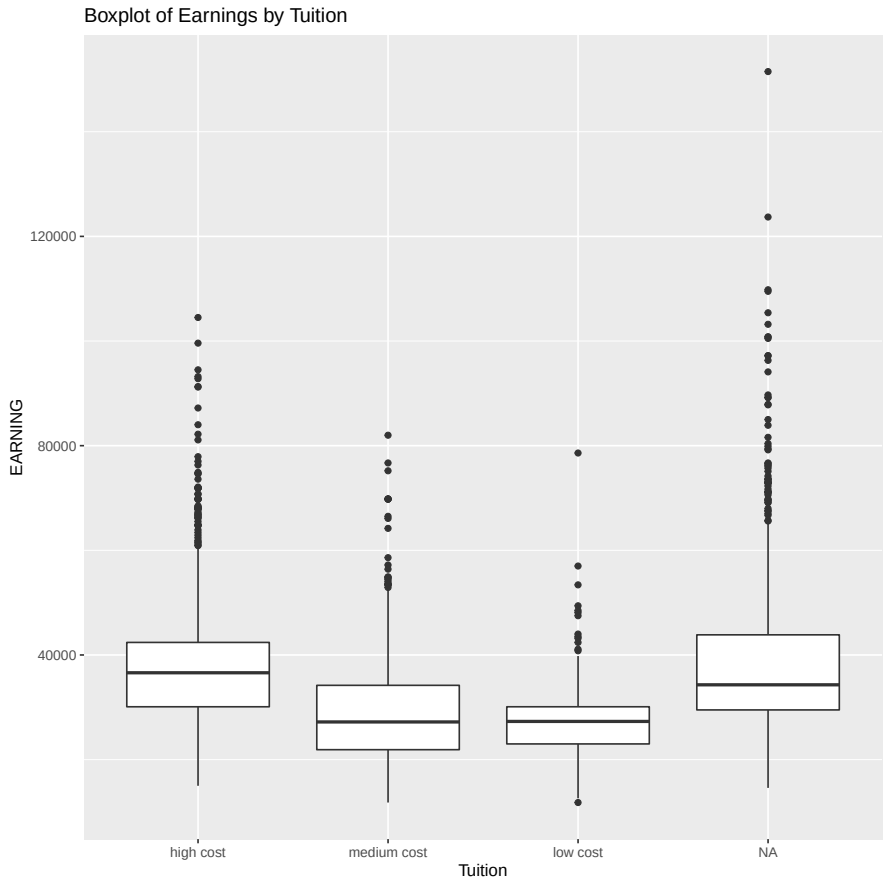
```
## Warning: Removed 6020 rows containing non-finite values (stat_bin).
```



This histogram shows that the majority of the colleges cost from $15000 to $30000. The graph is right-skewed with a few prestigous colleges that are extremely expensive.

```
## Boxplot using categorical TUTION
college_red %>%
  ## Filter out duplicates from the other years
  filter(YRS_AFTER_GRAD == 6) %>%
  ggplot() +
  geom_boxplot(mapping = aes(
    x = fct_relevel(TUITION, 'high cost', 'medium cost', 'low cost'),
    y = EARNING)) +
  labs(title = 'Boxplot of Earnings by Tuition',
       x = 'Tuition')
```

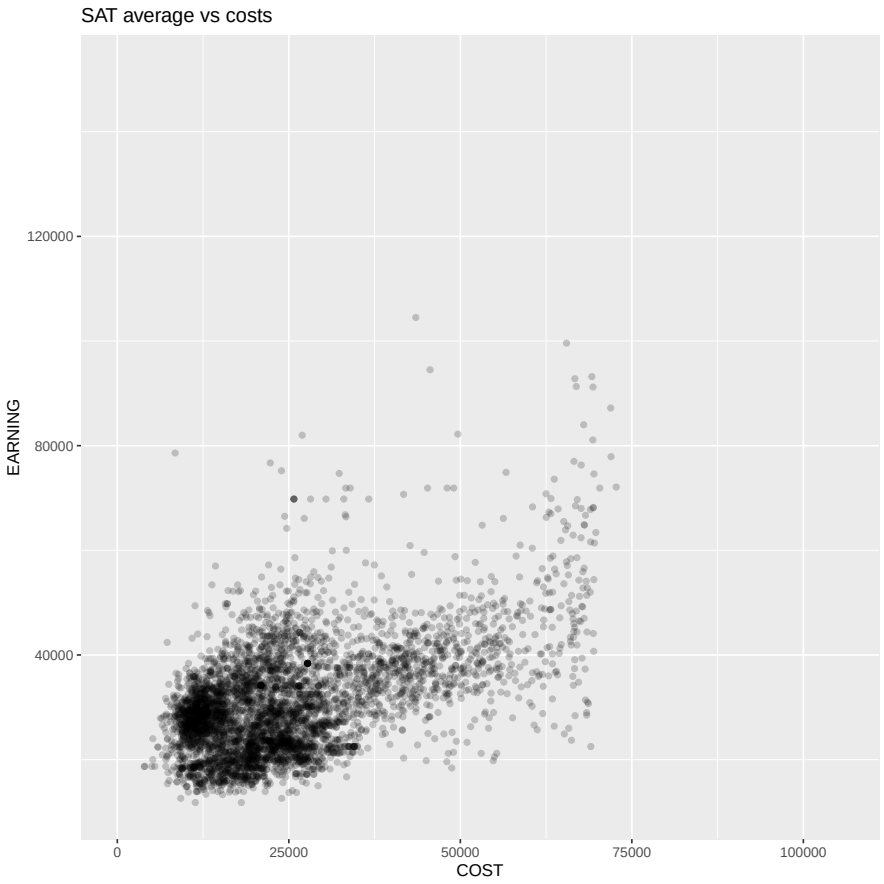
```
## Warning: Removed 1720 rows containing non-finite values (stat_boxplot).
```



From this boxplot, we can make the inference that students graduating from colleges with higher tuitions tend to earn more 6 years after graduation. There seems to be less variation in the low-cost group as its interquaritle range is narrower than those of other groups. Mean earnings of students graduating from colleges that are classified, in this analysis, as low-cost are closer together and don't show a lot of deviation from the mean. There are a number of outliers in the all groups, suggesting that cost may be not the only factor that contributes to the mean earning.

```
## Scatterplot of Earnings vs. Cost
college_red %>%
  # Remove duplicates of INSTNM from other years
  filter(YRS_AFTER_GRAD == 6) %>%
  ggplot(mapping = aes(x = COST, y = EARNING)) +
  geom_point(alpha = 0.2) +
  labs(title = 'SAT average vs costs')
```

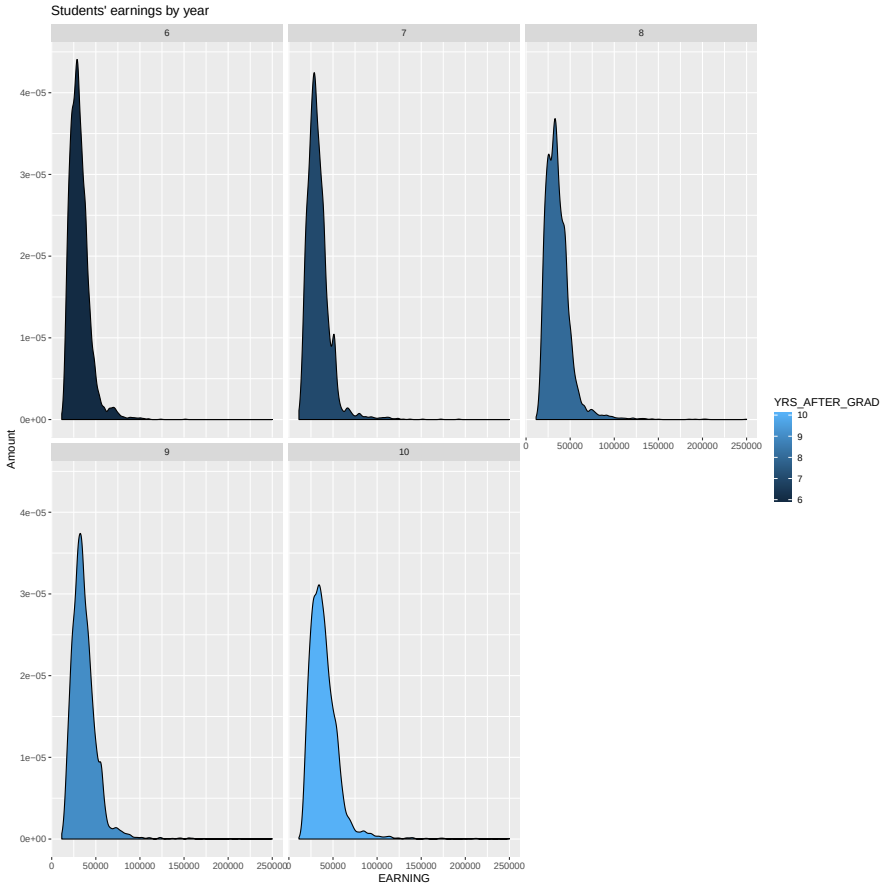
```
## Warning: Removed 2282 rows containing missing values (geom_point).
```



There is a slight correlation between EARNING and COST. There seems to be a cluster on the lower left of the graph as the majority of the colleges cost less than $25000 a year. By visually inspecting this scatterplot, we can see that the mean earnings of students graduating from higher-end and more prestigous colleges are higher. However, there are a few points above the cluster on the graph that may suggest otherwise.

```
## Comparison of earnings by year
college_red %>%
  ggplot(mapping = aes(x = EARNING, fill = YRS_AFTER_GRAD)) +
  geom_density() +
  facet_wrap(~YRS_AFTER_GRAD) +
  labs(title = "Students' earnings by year",
       y = 'Amount')
```

```
## Warning: Removed 10221 rows containing non-finite values (stat_density).
```



Earnings seem to be more spread-out in year 10. In year 6, most graduates earn less than $50000. However, as time goes on some of them start to earn more money, moving to the right of the graph, thus the curve is more spread-out.

```
## Expensive colleges vs students' earnings

# Remove NA's and sort by descending tuition
sorted_df <- college_red[complete.cases(college_red), ] %>%
  arrange(desc(COST))

## high-cost colleges
top3 <- head(unique(sorted_df$INSTNM), 3)

## low-cost colleges
bot3 <- tail(unique(sorted_df$INSTNM), 3)

## medium
med_df <- college_red %>%
  filter(TUITION == 'medium cost')
med3 <- head(med_df$INSTNM, 3)
```

```
## Examine mean earnings of colleges classified as high-cost, medium-cost, and low-cost over 5
college_red %>%
```

```
  filter(INSTNM %in% c(top3, med3, bot3)) %>%
  ggplot(mapping = aes(x = YRS_AFTER_GRAD, y = EARNING, fill = TUITION)) +
  geom_col() +
  facet_wrap(
    fct_relevel(TUITION, 'high cost', 'medium cost', 'low cost') ~ INSTNM) +
  labs(title = "Average earnings years after graduation by cost of attendance",
       x = 'Years after graduation')
```

This graph shows earnings of students graduating from prestigous colleges versus lower-cost colleges. Students from low-cost and medium-cost colleges seem to earn less. There is not a lot of fluctuation in earnings over years as the increase in salary seems subtle and gradual compared to higher cost colleges where the mean earnings tend to go up drastically in year 10.

Summary Statistics

```
## Statistics of earnings in years 6, 7, 8, 9

college_red %>%
  group_by(YRS_AFTER_GRAD) %>%
  summarize(
    mean = mean(EARNING, na.rm = TRUE),
    median = median(EARNING, na.rm = TRUE),
```

```
    std = sd(EARNING, na.rm = TRUE),
    range = max(EARNING, na.rm = TRUE) - min(EARNING, na.rm = TRUE),
    IQR = IQR(EARNING, na.rm = TRUE)
  )
```

YRS_AFTER_GRAD	mean	median	std	range	IQR
6	31623.83	29600	11679.68	139700	13400
7	33329.96	31000	13422.60	181200	13600
8	35989.38	33600	15012.99	236800	16300
9	36893.26	34300	15318.48	238200	15500
10	39707.16	36600	17976.96	235400	18400

As expected, there is more variation in students' earnings 10 years after college as the standard deviation is larger than those of other groups.

```
college_red %>%
  filter(YRS_AFTER_GRAD == 6) %>%
  group_by(TUITION) %>%
  summarize(
    total = n()
  )
```

TUITION	total
high cost	1463
low cost	1463
medium cost	2928
NA	1204

Data Analysis

The null hypothesis is that there is no difference between different colleges of varying costs in regard to the mean earning of students 6 years after graduation. H(0): slope = 0 The alternative hypothesis is that is a correlation between college tuition and mean earnings. H(a): slope $\neq$ 0

This will be a two-way hypothesis test, examining whether the sample slope is different from 0 and whether this difference is significant. The test statistic for this hypothesis test is slope of the relationship between tuition and earnings.

```
## Only test earnings after 6 years
sub_colleges <- college_red %>%
  filter(YRS_AFTER_GRAD == 6)
```

```
## Create a null distribution
null_distn <- sub_colleges %>%
   specify(EARNING ~ COST) %>%
   hypothesize(null = "independence") %>%
   generate(reps = 10000, type = "permute") %>%
```

```
   calculate(stat = "slope")
```

```
## Warning: Removed 2282 rows containing missing values.
```

```
## Calculate the test statistic
obs_stat <- sub_colleges %>%
   specify(EARNING ~ COST) %>%
   calculate(stat = "slope")
```

```
## Warning: Removed 2282 rows containing missing values.
```

```
obs_stat
```

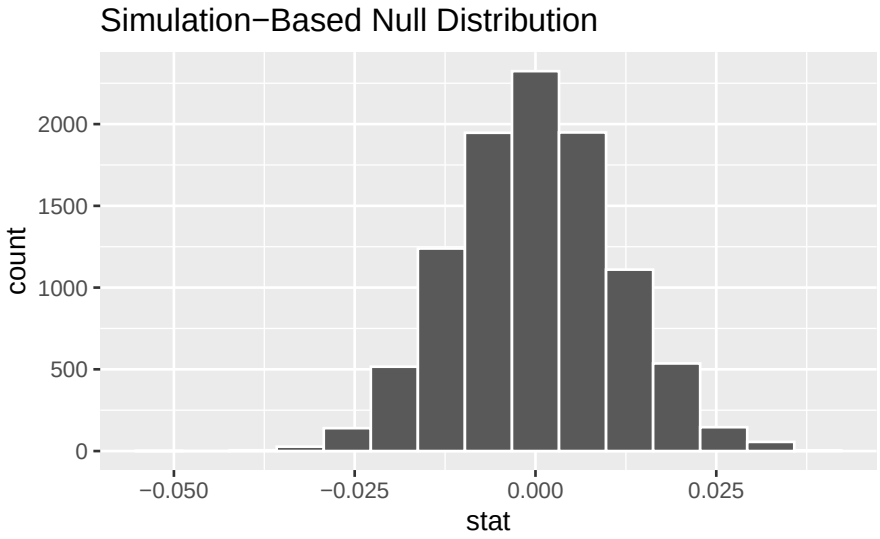
stat
0.4047326

The test statistic is 0.405.

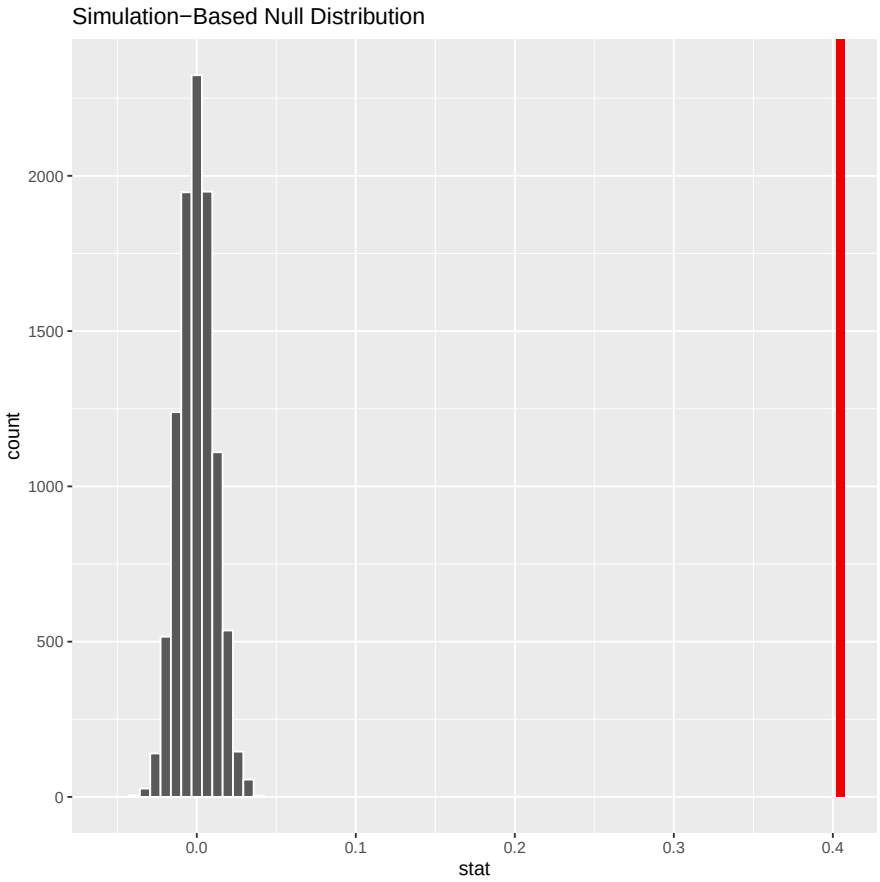
```
## p-value
null_distn %>%
  get_p_value(obs_stat = obs_stat, direction = "both")
```

p_value
0

```
null_distn %>%
  visualize()
```




```
null_distn %>%
  visualize() +
  shade_p_value(obs_stat = obs_stat, direction = "both")
```



We reject the null hypothesis because the p-value is less than the significance value of 0.05. There is statistically significant evidence to suggest that college tuition is correlated to the mean earnings of graduates. The results of this analysis support the claim that colleges with higher tuitions graduate higher-paid individuals to the workforce.

Conclusion

This analysis only examines the effect of college tuition on students' earnings and thus neglects other confounding factors that might have been responsible for the variation in students' earnings. Interactions between tuition and these confounding factors could in turn have an impact on the response variable. Colleges that come with higher tuitions might offer better coursework or they might be more selective in their admissions process and thus graduate students that are better equipped in the workforce. In addition, the college's most popular major could have also significantly skewed the mean earning of all students at that particular college due to how that major is applied in the workforce and its starting salary. For example, Johns Hopkins University's most famous major is neuroscience which, at its core one, is of the most high-paying fields in the workforce. The definition of a prestigous college does not simply involve its cost of attendance but also these other factors that make that college stand out among the rest. Overall, all analyses from different sections support each other as we see big gaps in mean earnings of students graduating from colleges of varying costs. The outcome of this analysis reinforces and factualizes the assumption that the choice of college plays in a role in one's earnings after graduation. This analysis aims to aid students in their college decision-making process and in the overall understanding of how college investment could potentially transform one's success.