Final Project

Nathan Nguyen

2024-05-02

Introduction

For the final project, I have chosen to investigate whether there is a difference in median earnings after graduation between colleges with different admission standards. The columns I will be using are:

- MN_EARN_WNE_P10(Median Earning After 10 Years). This would be the response variable.
- ADM_RATE(Admission Rate). This would be the explanatory Variable.

This is an interesting question to see if having a higher admission rate means that students who graduate after 10 years will earn more or less than another student. Many of us would assume that by going to a school with a lower acceptance rate, you would get a higher-paying job. I want to be able to prove whether this is true or not based on the given data. I will use a hypothesis test to answer my question during this project.

Preprocessing

```
college_sum <- college %>%
  select(ADM_RATE, MN_EARN_WNE_P10, INSTNM, CONTROL)

college_sum <- college_sum %>%
  mutate(
    CONTROL = recode(CONTROL,
                     "1" = "Public",
                     "2" = "Private_NP",
                     "3" = "Private_FP"))

college_sum <- college_sum %>%
  rename(
    Admission_Rate = ADM_RATE,
    Median_Earnings = MN_EARN_WNE_P10,
    College_Name = INSTNM,
    School_Type = CONTROL
  )
college_sum <- college_sum %>%
  mutate(Admission_Rate = Admission_Rate * 100)
```

i. In this section, I have selected certain columns to analyze and created my data set with those columns. Following this, School types were organized to understand what type of the different types of college. Subsequently, the columns were renamed to make it easier to use the variables without confusion. Additionally, the values from the Admission Rate column were multiplied by 100 so that the admission rate values were already in percentage and easier to understand.

```
college_sum <- college_sum %>%
  drop_na(Median_Earnings, Admission_Rate) %>%
  mutate(admission_category = cut(Admission_Rate,
                                  breaks = c(-Inf, 45, Inf),
                                  labels = c("Highly competitive",
                                             "Less competitive"),
                                  right = FALSE))
```

i. In this section, I have created another column to be able to compare schools with higher admission rates to schools with lower admission rates. I also took out any schools with NA values in them to not interfere with the data set. This will help me find out whether different admission rates have anything to do with the median salaries.

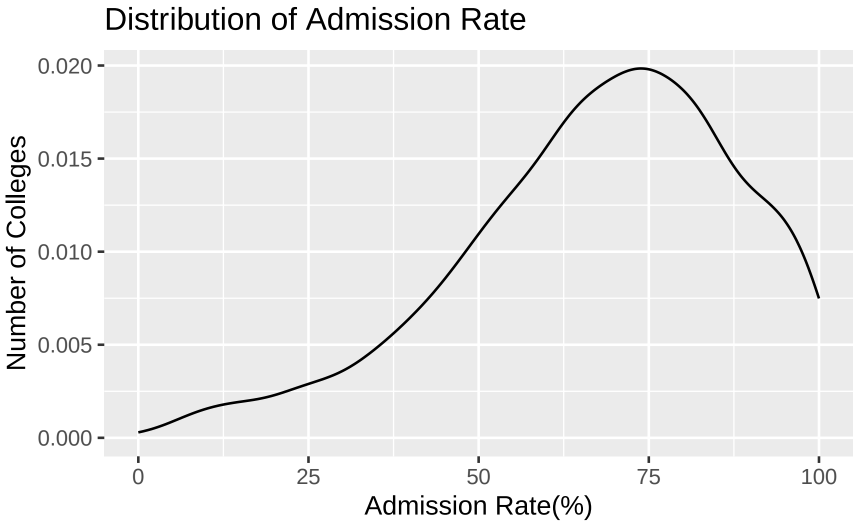
```r
head(college_sum)
```

Admission_Rate	Median_Earnings	College_Name	School_Type	admission_category
90.27	35500	Alabama A & M University	Public	Less competitive
91.81	48400	University of Alabama at Birmingham	Public	Less competitive
81.23	52000	University of Alabama in Huntsville	Public	Less competitive
97.87	30600	Alabama State University	Public	Less competitive
53.30	51600	The University of Alabama	Public	Less competitive
82.54	38000	Auburn University at Montgomery	Public	Less competitive

i. This is the data set created to be used for this project.

Visualization

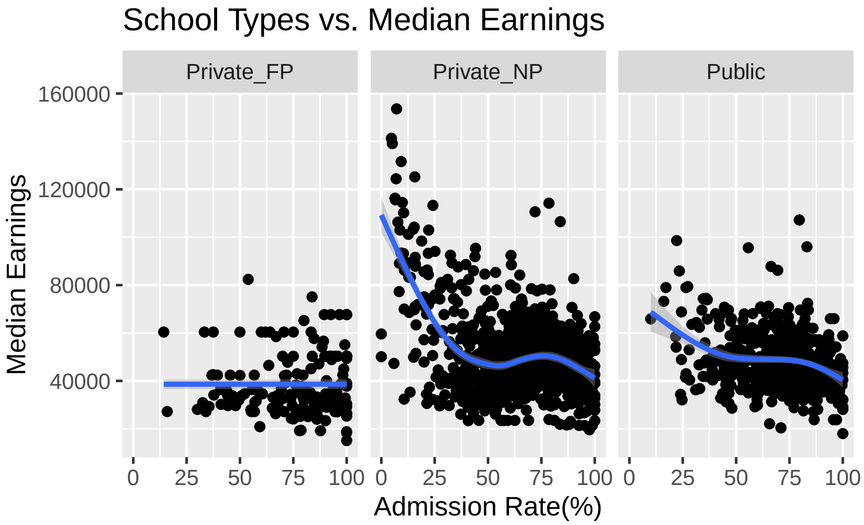
```
college_sum %>%
  ggplot() +
  geom_density(mapping = aes(x = Admission_Rate)) +
  labs(title = "Distribution of Admission Rate",
       x = "Admission Rate(%)",
       y = "Number of Colleges")
```



i. I chose this line graph to help see where the majority of the university's admission rates are. The graph demonstrates what the average acceptance rate is and be able to compare it to the average earnings. This graph is seen to be skewed to the left. This represents that there are more colleges with higher acceptance rates, than lower acceptance rates. The peak of this line graph looks to be around 75% acceptance rate.

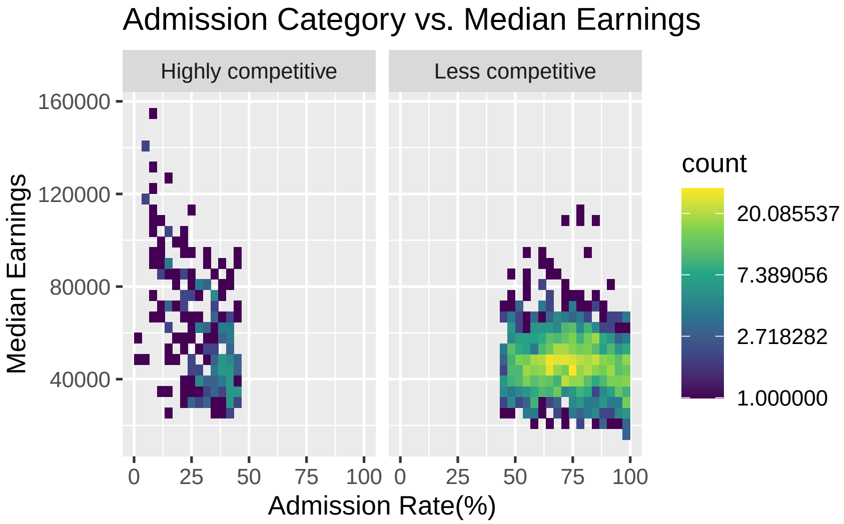
```
college_sum %>%
  ggplot() +
  geom_point(mapping = aes(x = Admission_Rate, y = Median_Earnings)) +
  facet_wrap(~School_Type) +
  geom_smooth(mapping = aes(x = Admission_Rate, y = Median_Earnings)) +
  labs(title = "School Types vs. Median Earnings",
       x = "Admission Rate(%)",
       y = "Median Earnings")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



i. I chose to use the scatter plot for this data set because it would help see the pattern of each school type. The graph demonstrates the relationship between the admission rate and median earnings for public, and private schools. Notably, individuals who went to a private non-profit school with a lower acceptance rate had higher-paying jobs. The private(FP) school didn't show the same pattern as the other two categories did; Demonstrating that the admission rate did not affect the median earnings. The other two categories have a similar curve and showed that the higher the admission rate was, the less the median earning was. This dataset provides insightful information about whether differences in admission rates between college types affect median earnings ten years after graduation.

```r
college_sum %>%
ggplot() +
  geom_bin2d(mapping = aes(x = Admission_Rate, y = Median_Earnings)) +
  scale_fill_viridis_c(trans = "log") +
  facet_wrap(~admission_category) +
  labs(title = "Admission Category vs. Median Earnings",
       x = "Admission Rate(%)",
       y = "Median Earnings")
```



i. For this data set, I have also decided to use a facet wrap with a histogram to help see where the plot is mostly dense. The graph demonstrates the relationship between admission rate and median earnings. When comparing the two variables, it is evident that people who went to a school with an admission rate of 45% to 90% have very similar median earnings. There are few instances where people who graduated from a higher acceptance-rated school made more money. There is also a drastic change in those who attended a lower admission-rated school, showing that these individuals have a higher median earning. The graph split into two helps show how competitive the school's admission rate is to have a difference in the median salaries.

Summary Statistics

```
college_sum %>%
  group_by(admission_category) %>%
  summarise(
    Mean = mean(Median_Earnings, na.rm = TRUE),
    Median = median(Median_Earnings, na.rm = TRUE),
    Range = max(Median_Earnings, na.rm = TRUE) - min(Median_Earnings, na.rm = TRUE),
    SD= sd(Median_Earnings, na.rm = TRUE),
    IQR = IQR(Median_Earnings, na.rm = TRUE)
  )
```

admission_category	Mean	Median	Range	SD	IQR
Highly competitive	59329.72	53000	130200	24686.77	33900
Less competitive	46939.45	46600	99100	11516.44	13500

i. The summarise() function was use to find the mean, median, range, standard deviation, and interquartile range.

- Mean: The mean represents the average value of median earnings for each of the admission categories. Based on the table, colleges that have a higher competitive rate have a higher mean value than the less competitive colleges.
- Median: The median demonstrates the middle value of our data set. Based on the table, the median value of the highly competitive category is higher than the median value of the less competitive category.
- Range: The range is the difference between the maximum and minimum values of median earnings. For the highly competitive category, the range of median earning value at $130,000, which suggests that there is a wide variation in what the graduates make.
- Standard Deviation: The standard deviation measures how dispersed the data points are around the mean. The higher the standard deviation, the greater the variability of the data. The median earning for the highly competitive category is higher than the less competitive category with a value of $22660.73. This suggests that there is somewhat of a variability in earning outcomes with the graduates.
- Interquartile Range: The IQR represents the mid-spread of the data set. The higher the IQR value is, the more spread out of the middle set of the data. The IQR value is higher in the highly competitive category than the less competitive value, which indicates that there is a notable difference in the distribution between the two categories.

Data Analysis

Null Hypothesis: There is no difference in median earnings after graduation between colleges with different admission standards. Alternative Hypothesis: There is a difference in median earnings after graduation between colleges with different admission standards.

I will be conducting a two-sided test demonstrating whether highly competitive colleges have higher median earnings or less competitive colleges have higher median earnings.

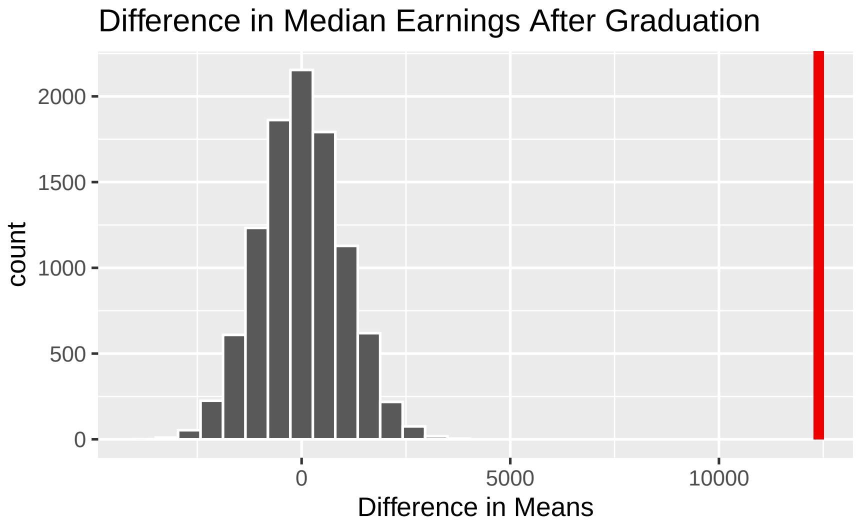
```r
college_null <- college_sum %>%
  specify(Median_Earnings ~ admission_category) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("Highly competitive", "Less competitive"))


college_obs <- college_sum %>%
  specify(response = Median_Earnings, explanatory = admission_category) %>%
  calculate(stat = "diff in means", order = c("Highly competitive", "Less competitive"))


p_value <- college_null %>%
  get_p_value(obs_stat = college_obs, direction = "both")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
## based on the number of ‘reps‘ chosen in the ‘generate()‘ step.
## i See ‘get_p_value()‘ (‘?infer::get_p_value()‘) for more information.
```

```r
head(p_value)
```

p_value
0

```r
visualize(college_null) +
  shade_p_value(obs_stat = college_obs, direction = "both") +
  labs(
    title = "Difference in Median Earnings After Graduation",
    x = "Difference in Means"
  )
```



i. The p-value has a really small value, somewhere that is less than 0.001. This indicates that the observation made between the difference in median earnings and colleges with high or low competitive admission rates has no difference compared to the actual data. The small p values reject the null hypothesis and indicate a significant difference in median earnings after graduating between colleges with high or low competitive admission rates. This is something that was inferred earlier based on the previous graphs was looked at.

Conclusion

Numerous inferences about the connection between college and university median earnings after graduation and admission competitiveness can be made based on the research performed for this project.

The exploratory data analysis first revealed median wage differences between admission competitiveness levels. Interestingly, median incomes at highly competitive colleges were typically greater than those at less competitive colleges. Additionally, a comparison of different school types had an impact on median salaries. In particular, graduates from Private Non-Profit universities showed greater wages than other groups, particularly those categorized as extremely competitive. Data summaries that showed notable differences in median incomes between colleges classified as highly competitive and less competitive supported this conclusion.

The hypothesis test conducted using the infer package confirmed the presence of a statistically significant difference in median earnings based on admission competitiveness. Strong evidence was found against the null hypothesis that there is no difference in median wages across different levels of admission competitiveness, as indicated by the permutation test's p-value, which was nearly zero.

The results imply a correlation between median wages following graduation and admission competitiveness. Graduates from more competitively admitted colleges typically have higher median incomes. These results highlight the significance of considering admission competitiveness as a factor impacting post-graduation outcomes and could help students make college selection decisions. It also emphasizes the need for more research into the variables causing the observed differences in median salaries amongst schools and institutions.