Project Part 2

Description of the Data Set

Research Questions

The research questions that I will explore are based on undergraduate university data in the United States, focused on various attributes of universities and state allocated resources to higher education in order to gain a better understanding of how universities in the United States work in relation to the plethora of factors that affect them. The research questions are:
1) How does the average proportion of first generation students who completed all four years of their undergraduate degree at the university they are enrolled in compare across public and private universities?
2) How does the average in-state tuition of public universities compare across states that have allocated more funds towards higher education and states that have not allocated as much?
3) How does the average acceptance rate of public and private universities with higher net tuition revenue per student compare to public and private universities with lower net tuition revenue per student?

Data Context and Collection

The data were collected from two primary sources, the 2019-2020 U.S. Department of Education College Scorecard and the 2020 The State Higher Education Executive Officers Association (SHEEO) State Higher Education Finance (SHEF) Report, and compiled into one comprehensive data set. The College Scorecard is a platform launched by the U.S. Department of Education which contains thousands of measures such as acceptance rate, in-state tuition, and graduation rate for 6,694 universities in the United States so that prospective students and anyone who wants to learn more about the specifics of a certain university can directly access this information (Reference 5). The College Scorecard data are a population, as it contains information on all nationally recognized universities in the United

States. The College Scorecard data were collected through direct federal reporting from institutions, federal financial aid data from Federal Student Aid (which is an office directly under the U.S. Department of Education), and other partnered federal reporting such as the National Students Loan Data System and the Integrated Postsecondary Education Data System (Reference 6). The State Higher Education Finance (SHEF) Report is an annual report that depicts the patterns and consequences of state funding for higher education of the 50 states, with attributes such as education appropriations, net tuition revenue, and state public financial aid (Reference 7). The SHEF Report data are a population, as it contains statistics on every state in the United States. The data were collected from SHEEO's (the parent organization for SHEF) member agencies across multiple sectors that directly report to SHEOO (Reference 8).

Data Manipulation and Contents

Both the College Scorecard and the SHEF Report had many variables that were not needed for this analysis, so the desired variables for each data set were first selected and saved in Excel and then read into R. The College Scorecard data set's observations are individual universities, while the SHEF Report data set's observations are individual states. The SHEF Report data set was first subset to the 2020 fiscal year, and the two data sets were then merged on the basis of state in a data set called "colleges". Thus, the corresponding variables for the SHEF Report for each state were added on to every college that was from the respective state and the observations of this merged data set are individual colleges.There were other data cleaning techniques performed on the data, such as removing rows with missing values and variables that were no longer needed, that cleaned the data set down to 1,433 observations (universities) with 10 variables. Because the data set no longer contains all the nationally recognized universities, the data are now a sample. The variables are:
1) the state that the university is in (state)
2) name of the university (name)
3) whether the university if Public or Private (type)
4) acceptance rate (adm_rate)
5) in-state tuition (tuition_in)
6) net tuition revenue per full-time equivalent student, which is how much profit a university makes per student (net_tut_rev)
7) the percentage of first-generation students who completed within 4 years at the university (firstgen_yr4)
8) the graduation rate of the university (grad_rate)

9) the total monetary amount of state support for higher education (state_support)
10) education appropriations, which is the monetary amount of state and local support to public higher education institutions including state funded financial aid (edu_approp)

Data Issues

A potential issue with the data is that some states have many more universities than other states, so the data is not equally representative of each state. I attempted to combat this issue by averaging and using means for statistics that have to do with states or state specific data, but it could still be affecting my findings. Another potential issue is that some universities are part of university systems, such as the University of California and University of Wisconsin systems. It was not clear whether the statistics for the universities in a system are completely independent of each other, but since it wasn't explicitly noted anywhere, I assumed that the statistics were independent of each other.

Data Appropriateness

The data are appropriate to address the research questions because they look at a sample of universities and include the variables needed to answer the questions about the universities, such as the percentage of first-generation students who completed within 4 years at the university and whether the university is public or private for question 1, and the net tuition revenue per student, the acceptance rate, and whether the university is public or private for question 3. The data also includes the appropriate variables (the in-state tuition and the amount of state and local support to public higher education) to answer question 2, that focuses more on the individual state funding than universities.

Data Summaries

Numerical Summary

```
# Relating to Research Question #1

# Subsetting only public universities
public <- colleges[colleges$type == "Public",]
# Finding the average percentage of first-generation students
# who stayed at public universities all 4 years
```

```
mean(public$firstgen_yr4)
```

```
## [1] 0.4602974
```

```
# Subsetting only private universities
private <- colleges[colleges$type == "Private",]
# Finding the average percentage of first-generation students
# who stayed at private universities all 4 years
mean(private$firstgen_yr4)
```

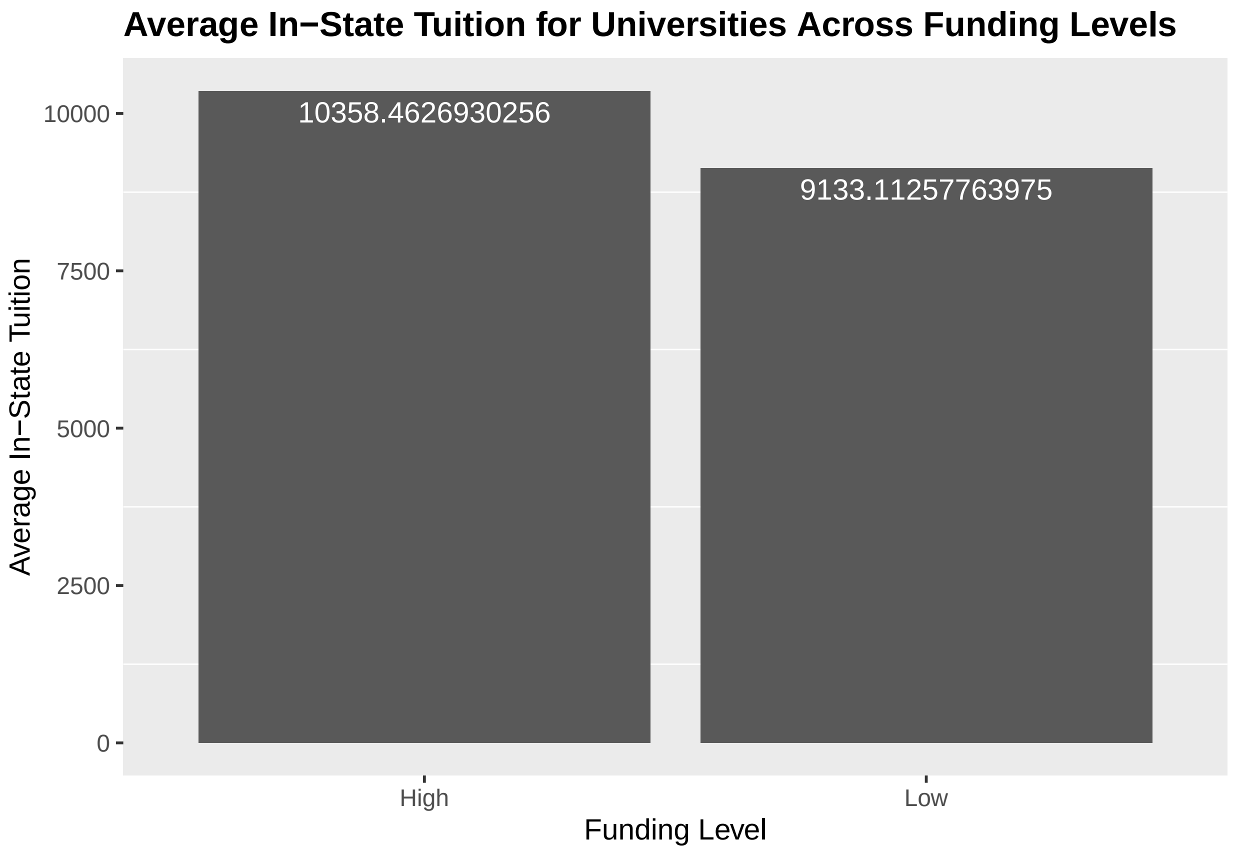
```
## [1] 0.5094866
```

Based on this summary, the average proportion of first generation students who completed all four years of their undergraduate degree at the university they are enrolled in appears to be slightly higher for private universities (50.9%) than public universities (46.0%). This slight difference could be due to private universities possibly having more university specific resources than public universities, which could affect the decision of first-generation students, an often vulnerable group on college students in regards to staying in college, to stay at their university, among many other possibilities.

Graphical Summaries

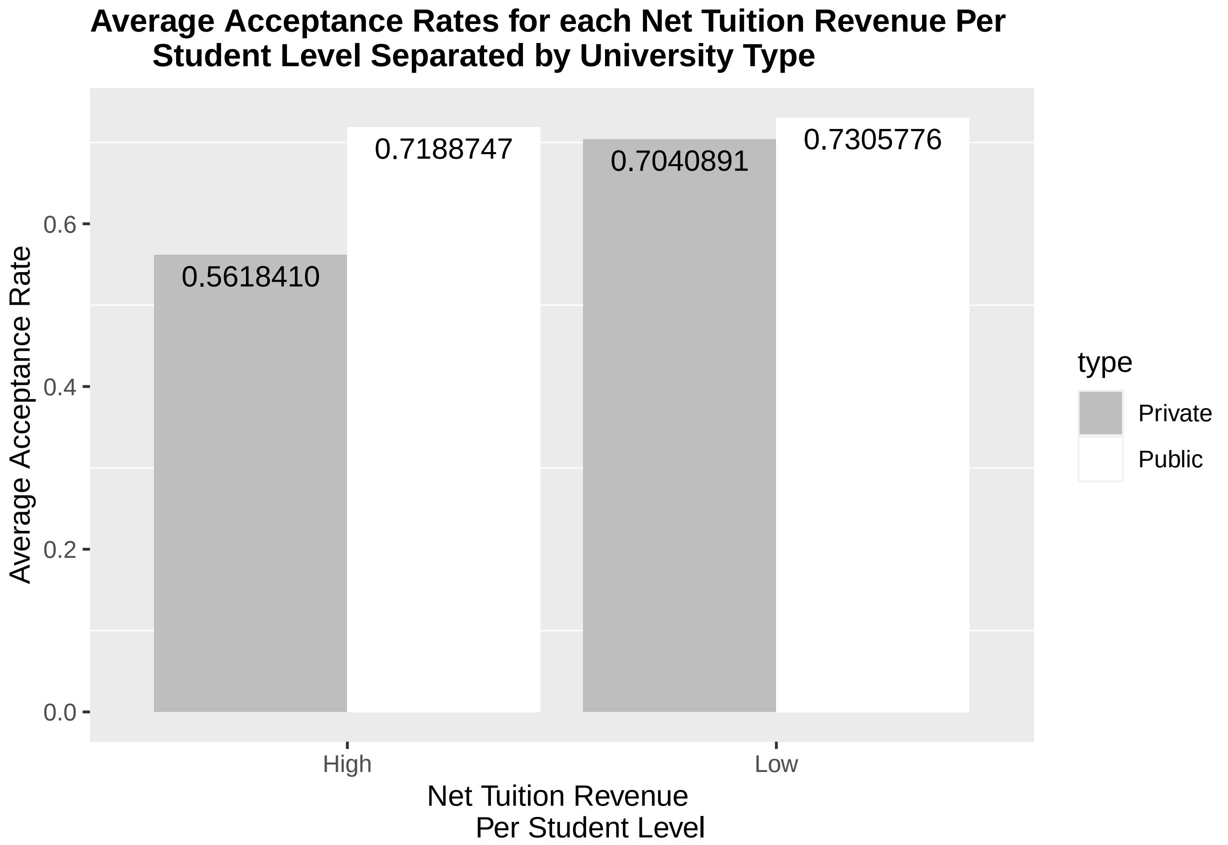
```
# Relating to Research Question #2

# Finding the average in state tuition of public universities for each state
in_state_tuition <- tapply(public$tuition_in, public$state, mean)

# Adding the average values to a dataframe
tuition_instate <- as.data.frame(in_state_tuition)
tuition_instate$state <- row.names(tuition_instate)

# Merging the average in-state tuition values with education
# appropriation values based on state
avg_tuition_state <- unique(merge(tuition_instate,
colleges[,c("state", "edu_approp")], by="state"))
```

```
# Creating a new column that specifies if the state’s
# allocated funds towards higher education is high or low
avg_tuition_state_new <- avg_tuition_state%>%mutate(funding_level=ifelse(
  avg_tuition_state$edu_approp <= 1000000000, "Low", "High"))

# Plotting the average in-state tuition for universities across each funding level
ggplot(avg_tuition_state_new, aes(x=funding_level, y=in_state_tuition)) +
  stat_summary(fun=mean, geom="bar") +
  stat_summary(fun=mean, geom="text", aes(label=..y..),
             vjust=1.5, position=position_dodge(0.9), color="white") +
  labs(title="Average In-State Tuition for Universities Across Funding Levels",
       x="Funding Level", y="Average In-State Tuition") +
  theme(panel.grid.major=element_blank(),
        plot.title=element_text(face="bold", size=13))
```



Based on this summary, the average in-state tuition of public universities for states that have allocated more funds towards higher education (about 10,358 dollars) is slightly higher than that of states that have allocated less funds towards higher education (about 9,133 dollars). Although I originally expected states that have allocated more funds towards higher education to have a lower average in-state tuition because they would have more funds and thus more freedom to decrease students' tuition, there could be many other factors that affect how much tuition a university charges to its students, and the state allocated funds could be applied to aspects of the universities that are independent of tuition.

```
# Relating to Research Question #3

# Creating new data set with new column which assigns the net
# tuition revenue per student as "high" or "low" depending on
# whether the university is public or private and above a certain threshold
colleges_tut_rev <- colleges%>%mutate(tut_rev_level=ifelse(
  colleges$net_tut_rev <= 8000 & colleges$type == "Public", "Low",
  ifelse(colleges$net_tut_rev > 8000 & colleges$type == "Public", "High",
  ifelse(colleges$net_tut_rev <= 20050 & colleges$type == "Private", "Low", "High"))))

# Plotting average admission rate for each net tuition revenue
# level separated by university type (public and private)
ggplot(colleges_tut_rev, aes(x=tut_rev_level, y=adm_rate, fill=type)) +
  stat_summary(fun=mean, geom="bar", position="dodge") +
  stat_summary(fun=mean, geom="text", aes(label=format(..y.., nsmall=0)),
               vjust=1.5, position=position_dodge(0.9)) +
  scale_fill_manual("type", values=c("Private"="gray", "Public"="white")) +
  labs(title="Average Acceptance Rates for each Net Tuition Revenue Per
       Student Level Separated by University Type", x="Net Tuition Revenue
       Per Student Level", y="Average Acceptance Rate") +
  theme(panel.grid.major=element_blank(),plot.title=element_text(face="bold", size=12))
```



Note: The level of net tuition revenue per student (how much profit a college makes) is standardized by whether the university is public or private, as it is usually significantly higher for private universities (Reference 4).

Based on this summary, the average acceptance rate of public universities for both net tuition revenue levels as well as those of private universities for the low net tuition revenue level are roughly the same at around 70%. The average acceptance rate of private universities for the high net tuition revenue level, however, is significantly lower, at 56%. This could be due to private universities generally charging higher tuition, thus having a higher net tuition revenue per student, as well as sometimes being more selective, as seen in the lower acceptance rate.

References

1. https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/
2. https://www.statology.org/remove-dollar-sign-in-r/#:~:text=You%20can%20easily%20remove%20dollar,by%20using%20gsub()%20function.
3. https://statisticsglobe.com/r-remove-data-frame-rows-with-some-or-all-na
4. https://nces.ed.gov/programs/coe/indicator/cud
5. https://collegescorecard.ed.gov/data/
6. https://collegescorecard.ed.gov/assets/FieldOfStudyDataDocumentation.pdf
7. https://shef.sheeo.org/about/
8. https://shef.sheeo.org/faq/