Case Study 4

Alex Christmann

2022-07-27

Question 1

Diving into the college.csv dataset, I've identified 17 distinct variables across 1,269 rows. Each row represents a unique college institution, with columns capturing key metrics like admission rates, tuition costs, and regional information. This comprehensive dataset gives us a solid foundation for exploring patterns in higher education across the country.

```
setwd("C:/Users/alexc/OneDrive/Documents/Summer 2022/Descriptive Analytics/R Markdowns")
college <- read.csv("college.csv", header = TRUE, stringsAsFactors = TRUE, na.strings = "")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.1
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
ncol(college)
```

```
## [1] 17
```

```
colnames(college)
```

```
##  [1] "id"                 "name"               "city"
##  [4] "state"              "region"             "highest_degree"
##  [7] "control"            "gender"             "admission_rate"
## [10] "sat_avg"            "undergrads"         "tuition"
## [13] "faculty_salary_avg" "loan_default_rate"  "median_debt"
## [16] "lon"                "lat"
```

```
nrow(college)
```

```
## [1] 1269
```

Question 2

Before conducting any analysis, I've performed data quality checks to identify missing values that could impact our findings. By combining the is.na() function with sum(), I can quickly assess the completeness of our dataset - an essential first step in any robust analysis.

```
sum(is.na(college))
```

```
## [1] 2
```

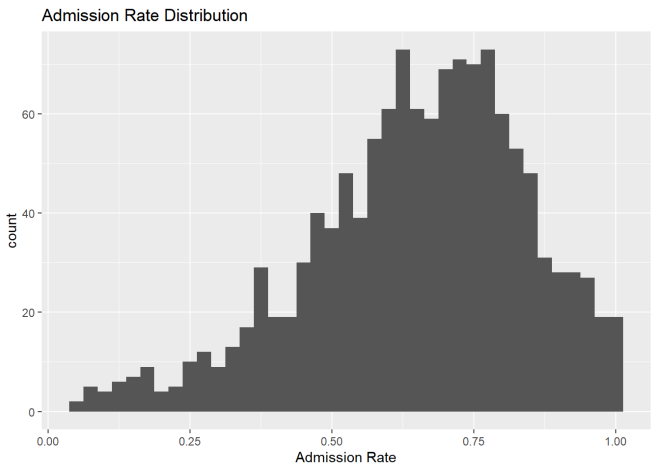
The analysis reveals only 2 missing values in the entire dataset of over 21,500 data points (17 variables × 1,269 institutions), indicating exceptional data completeness at 99.99%. This gives us confidence that our subsequent analyses won't be significantly affected by missing data issues.

Question 3

To understand the landscape of higher education, I'll first explore key variables individually. Below, I've created visualizations for both continuous and categorical variables: admission rates show us how selective institutions are across the spectrum, while the state distribution reveals geographic concentrations of higher education institutions.

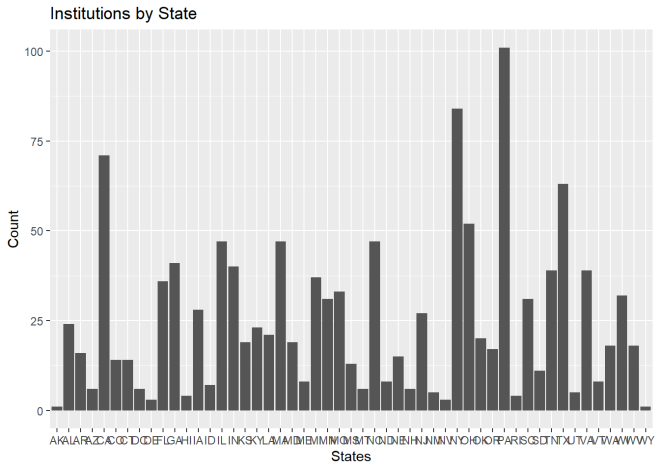
```
admission_rate_plot <- ggplot(college, aes(x=admission_rate))+geom_histogram(binwidth = 0.025)+ggtitle("Admission Rate Distr
ibution")+labs(x="Admission Rate")

admission_rate_plot
```



```
institutions_by_state <- ggplot(college, aes(x=state))+ geom_bar()+ggtitle("Institutions by State")+labs(x="States",y="Coun
t")

institutions_by_state
```



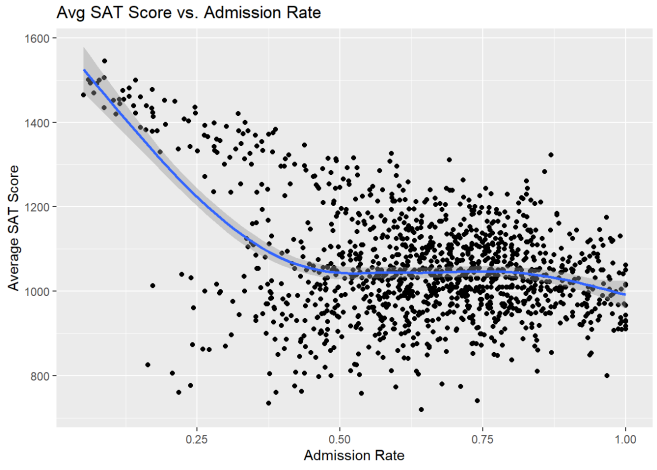
The admission rate histogram reveals an interesting left-skewed distribution, showing that most institutions accept a majority of applicants (60-80% admission rate). This pattern reflects the reality that highly selective institutions (with rates below 30%) represent a relatively small portion of the higher education landscape. The geographic distribution shows clear “education hubs” - Pennsylvania leads with the highest concentration of institutions, followed closely by California and New York. This regional clustering likely reflects population density, historical education development patterns, and state-level higher education funding priorities. Pennsylvania had the most institutions out of any state, followed by California and New York. Alaska had the fewest.

Question 4

Now I’ll explore relationships between different types of variables to uncover meaningful patterns in higher education. I’ve selected three distinct comparison types: admission rates vs. SAT scores (continuous-continuous), tuition by region (continuous-categorical), and gender distribution by institutional control (categorical-categorical).

```
SAT_vs_Admission <- ggplot(data = college, aes(x = admission_rate, y = sat_avg)) +
  geom_point()+
  geom_smooth()+
  labs(title = "Avg SAT Score vs. Admission Rate", x = "Admission Rate", y = "Average SAT Score")
SAT_vs_Admission
```

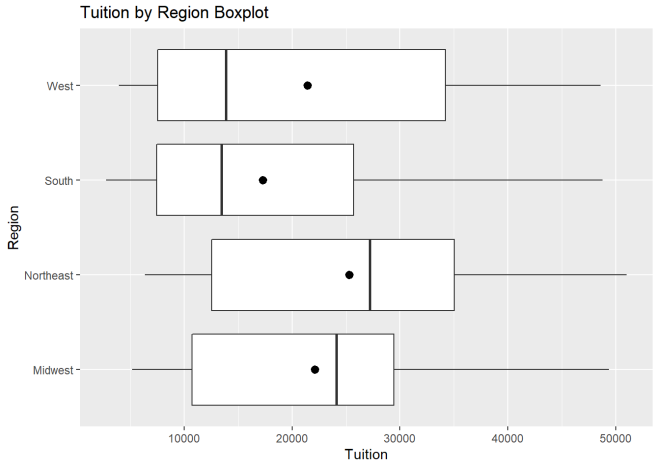
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



This visualization reveals the strong inverse relationship between selectivity and standardized test scores. What’s particularly interesting is that the relationship isn’t linear - there’s a dramatic increase in average SAT scores among the most selective institutions (those with admission rates below 25%). This suggests two distinct “tiers” in higher education: a small group of highly selective institutions where SAT scores are significantly higher, and a larger group where selectivity has a more modest impact on average test scores. This pattern likely reflects both the applicant pool (high-scoring students targeting elite schools) and institutional priorities regarding test scores in admission decisions.

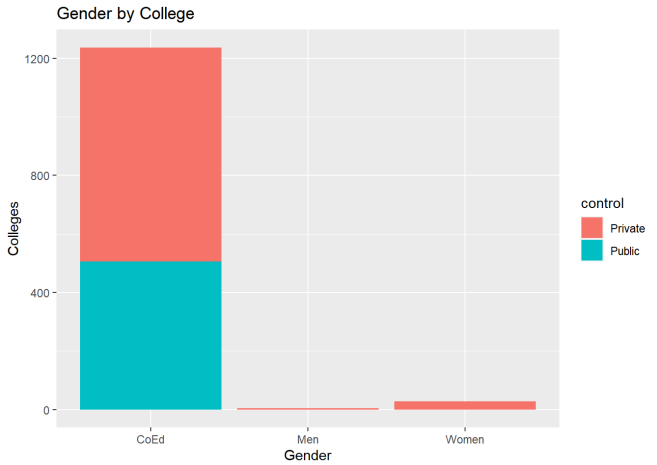
```
tuition_by_region <- ggplot(data = college, aes(x = tuition, y = region))+
  geom_boxplot()+
  stat_summary(fun = "mean")+
  labs(title = "Tuition by Region Boxplot", x = "Tuition", y = "Region")
tuition_by_region
```

```
## Warning: Removed 4 rows containing missing values (geom_segment).
```



The regional tuition comparison reveals clear geographic patterns in higher education costs. Northeast institutions command the highest tuition rates, with the Midwest following as a strong second. The right-skewed distributions in the South and West (where means exceed medians) point to a concentration of elite, high-tuition institutions in these regions that pull the averages upward. These regional price differences likely reflect a combination of factors: different balances of public vs. private institutions, regional cost-of-living variations, and possibly different state funding models for public higher education.

```
gender_by_college <- ggplot(college)+
  geom_bar(aes(x=gender, fill=control))+
  ggtitle("Gender by College")+
  labs(x="Gender",
       y="Colleges")
gender_by_college
```



This visualization highlights how dramatically the higher education landscape has evolved toward co-education. The data shows that single-sex institutions have become rare educational enclaves, with men's colleges being particularly scarce (just 4 remaining institutions nationwide). Notably, all single-sex institutions are privately controlled, suggesting that market forces and public funding models have universally pushed public institutions toward co-education. This represents a significant historical shift, as many of America's oldest institutions began as single-sex schools before transitioning to co-education in the mid-to-late 20th century.

Question 5

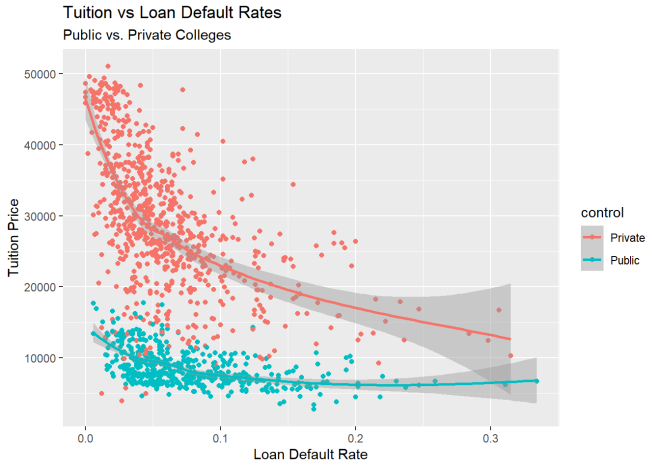
Building on our previous analyses, I'll now introduce a third variable to explore more complex relationships. By examining loan default rates against tuition while distinguishing between public and private institutions, we can uncover important patterns in higher education financing and student outcomes.

```
tuition_vs_defaultrate <- ggplot(data = college, aes(x = loan_default_rate, y = tuition, color = control))+
     geom_point()+
     geom_smooth()+
     labs(title = "Tuition vs Loan Default Rates ", subtitle = "Public vs. Private Colleges", x = "Loan Default Rate", y = "Tuition Price")
tuition_vs_defaultrate
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



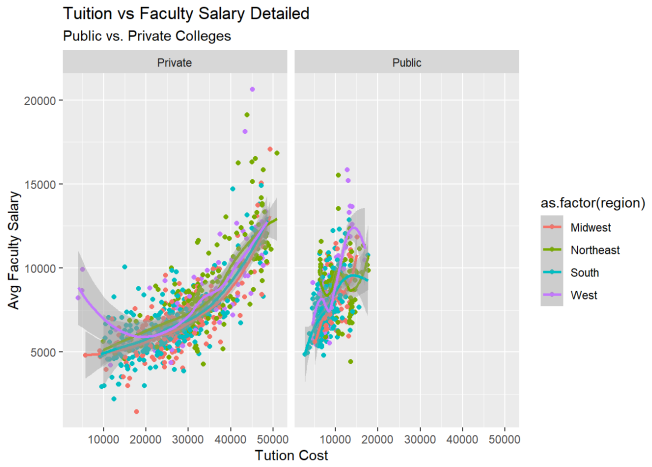
This multi-dimensional analysis reveals a counterintuitive pattern: higher-priced private institutions actually show lower loan default rates. This negative correlation contradicts what we might expect - that higher costs lead to more financial distress. Several factors likely explain this pattern: elite private institutions often attract students from wealthier backgrounds who need fewer loans; these schools typically offer more generous financial aid packages; and their graduates often secure higher-paying jobs that make loan repayment more manageable. Meanwhile, public institutions show a more consistent default rate regardless of tuition level, suggesting that their pricing has less impact on graduates' ability to repay loans - likely because their tuition range is much narrower and generally more affordable.

Question 6

To fully understand the complex relationships in higher education, I've created a four-variable analysis below. This visualization examines how faculty salaries relate to tuition costs across different regions, with separate panels for public and private institutions - providing a comprehensive view of how these key factors interact.

```
ggplot(college, aes(x=tuition, y=faculty_salary_avg, color=as.factor(region))) +
     geom_point() +
     geom_smooth()+
     facet_grid(. ~ control)+
   labs(title = "Tuition vs Faculty Salary Detailed ", subtitle = "Public vs. Private Colleges", x = "Tution Cost", y = "Avg
Faculty Salary")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



This comprehensive visualization reveals fundamentally different economic models between public and private higher education. Private institutions show a clear positive relationship between tuition and faculty compensation across all regions, suggesting that students' tuition dollars directly translate to faculty salary investments. The steepest relationship appears in Southern private schools, where each tuition dollar correlates more strongly with faculty pay. Public institutions tell a different story - their relatively narrow tuition band ($5,000-$15,000) shows much weaker correlation with faculty salaries, particularly in the Northeast. This suggests public institutions' faculty compensation depends more on state funding formulas, endowments, and research grants than on tuition revenue. These distinct patterns highlight how governance structure fundamentally shapes an institution's financial priorities and resource allocation.

Question 7.1

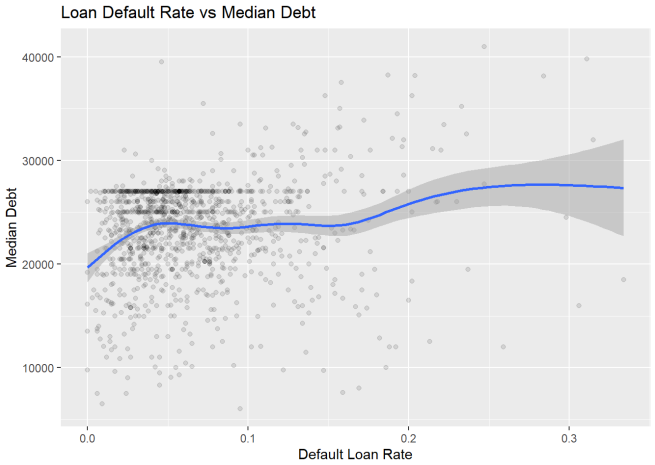
Understanding student debt drivers is crucial for addressing the growing higher education affordability crisis. I'll investigate key factors potentially contributing to student debt burdens, beginning with the relationship between loan default rates and median debt levels.

```
ggplot(data = college, aes(x = loan_default_rate, y = median_debt))+
     geom_point(alpha = 0.1)+
     geom_smooth()+
     labs(title = "Loan Default Rate vs Median Debt", x = "Default Loan Rate", y = "Median Debt")
```

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

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

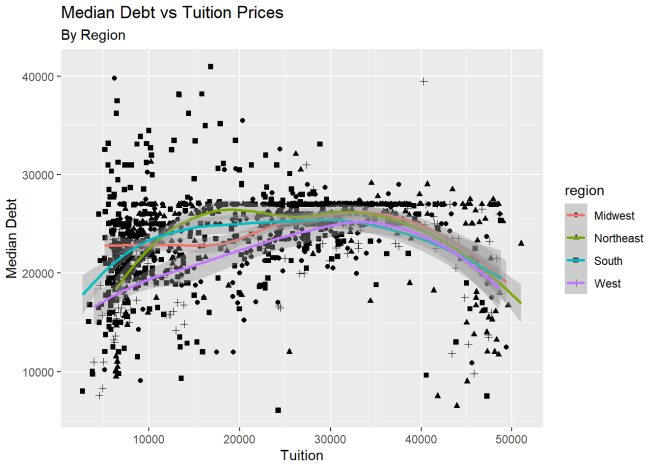


There seems to be very little correlation between loan default rates and median debt, which is surprising as I figured defaulting on loans would directly contribute to higher levels of debt. However, Once the default rate reaches 15% or so, the median debt appears to be climbing higher on average. Overall, loan default rates are not a solid predictor of student debt.

Next, I'll examine whether tuition costs translate directly to student debt levels. The visualization below explores this relationship across different geographic regions to identify any regional variations in how tuition impacts student borrowing.

```
ggplot(data = college, aes(x = tuition, y = median_debt, shape = region))+
  geom_point()+
  geom_smooth(aes(color = region))+
  labs(title = "Median Debt vs Tuition Prices", subtitle = "By Region", x = "Tuition", y = "Median Debt")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



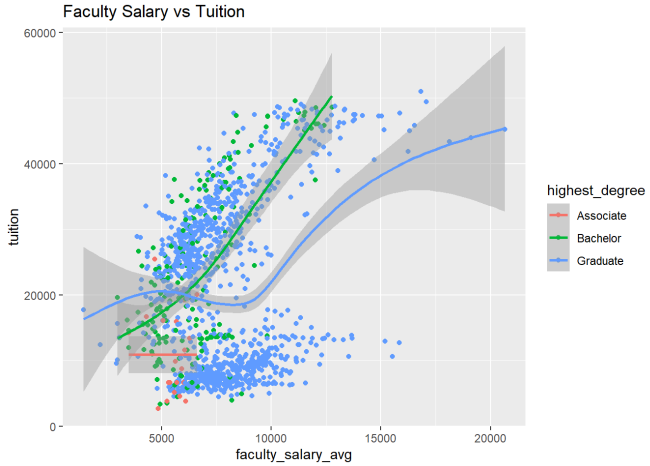
Surprisingly, our analysis reveals that neither loan default rates nor tuition costs strongly predict student debt levels. This counterintuitive finding challenges conventional wisdom that higher tuition directly leads to higher student debt. The weak correlation suggests other factors outside our dataset likely play more significant roles - possibly including family income levels, institutional financial aid policies, degree completion rates, or state-specific scholarship programs. This highlights an important limitation in our current analysis and points to the need for more comprehensive data that includes student socioeconomic backgrounds, institutional aid packages, and post-graduation employment outcomes to fully understand the student debt puzzle.

Question 7.2

Finally, I'll investigate whether institutions with higher tuition costs invest more in faculty compensation. This analysis explores the relationship between tuition revenue and faculty salaries across different types of institutions, providing insight into how schools allocate their financial resources.

```
ggplot(data=college, aes(faculty_salary_avg, tuition ,color = highest_degree)) +
 geom_point() + geom_smooth() + ggtitle("Faculty Salary vs Tuition")
```

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




```
cor(college$tuition,college$faculty_salary_avg)
```

```
## [1] 0.2807547
```

```
degrees <- group_by(college,highest_degree)
avg_salary <- summarise(degrees, meansalaries=mean(faculty_salary_avg, na.rm = TRUE))

avg_salary
```

```
## # A tibble: 3 × 2
##   highest_degree meansalaries
##   <fct>                <dbl>
## 1 Associate            5490.
## 2 Bachelor             6689.
## 3 Graduate             7882.
```

The analysis reveals a modest positive correlation (0.28) between tuition and faculty salaries, indicating that while higher-tuition institutions tend to pay faculty more, this relationship explains only a small portion of salary variation. When broken down by degree level, we see a clear academic hierarchy - institutions offering more advanced degrees consistently compensate faculty at higher levels. This pattern likely reflects the greater research expectations, specialized expertise, and competitive market for faculty at graduate-level institutions. The modest overall correlation suggests that other factors beyond tuition revenue significantly influence faculty compensation, potentially including research funding, endowment size, institutional prestige, and cost-of-living in different locations. This nuanced relationship highlights the complex financial ecosystem within higher education institutions.