Maaz Nadeem Khalil
Professor Dr. Vivek Singh
17:610:561:90
Data Analytics for Information Professionals

Assignment 4 - Final Class Project
Title: Predicting US News & World Report University Rankings in the United States of America

Background and Predictive Analysis Problem

An area I have identified that can utilize applied data science especially in a big data context is that of university rankings. University rankings come from many sources and publishers but one name stands above the rest as the benchmark standard and that is U.S News & World Report. The U.S News & World report annually ranks the top universities in America utilizing a variety of metrics ranging from graduation rate to exclusivity and more. University rankings are important as they have a significant impact on the success of alumni later in life, this is why I will be investigating the top 25 universities in the United States, evaluating various variables and their impact on the final university rankings. Ultimately, this will help university stakeholders understand what their institution needs to do to achieve a greater rank and thus a greater degree of success.

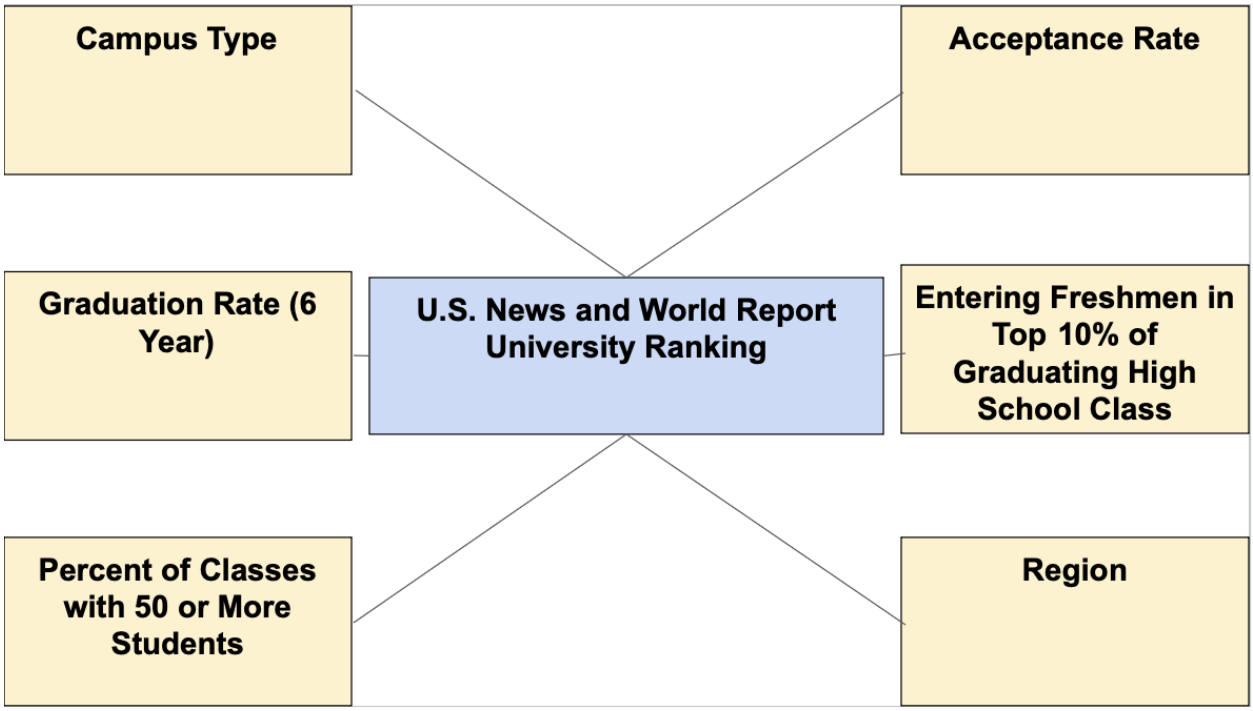
Predictive Model

For an accurate and insightful model to be developed to predict university ranking, a multi-linear regression model must be utilized in assessing the impact important variables have on the final ranking. The Y variable is the final ranking and there are multiple X variables detailed below. The usage of time series analysis is applied over the course of a 10 year period to accurately identify trends and patterns.

The proposed model is represented by a regression equation as follows:

$$\textbf{(Rank)}_i = B_0 + \textbf{(Campus Type)}_i B_1 + \textbf{(NE)}_i B_2 + \textbf{(SE)}_i B_3 + \textbf{(MW)}_i B_4 + \textbf{(W)}_i B_5 + \textbf{(SW)}_i B_6 + \textbf{(Acceptance Rate)}_i B_7 + \textbf{(Graduation Rate (6 Years))}_i B_8 + \textbf{(Entering Freshman in Top 10\% of Class)}_i B_9 + \textbf{(Percent of Classes with 50 or More Students)}_i B_{10}$$

The visualization below illustrates the model being applied for this study:



Unit of Analysis: Sample of Top 25 Universities (2003-2012)

Variables

Dependent Variable (Y):

- U.S News and World Report University Ranking - (#)

Independent Variables (X):

- Campus Type (Urban/Suburban)
- Region
    - NE
    - SE
    - MW
    - W
    - SW
- Acceptance Rate - (%)
- Graduation Rate (6 year) - (%)
- Entering Freshman in Top 10% of Graduating High School Class - (%)
- Percent of Classes with 50 or More Students - (%)

The “Ranking” variable is an ordinal data type that falls under the wider qualitative umbrella. The “Campus Type” variable is a binary (nominal) category of qualitative data, representing whether a campus is urban (1) or suburban (0). The region variables (NE, SE, MW, W, SW) are also nominal, indicating the geographical area of each university. In order to effectively analyze these regions, I have encoded them using binary to indicate whether a university is in that region or not. “Acceptance Rate”, “Graduation Rate”, “Entering Freshman in Top 10% of Class”, and “Percent of Classes with 50 or More Students” are quantitative variables measured in percentages. These data points are continuous, but for this project, percentages will be rounded to the nearest whole number, effectively making these variables and their data points discrete. Both binary encoding of region and transformation of continuous variables to discrete are methodological choices to facilitate analysis.

Data Source

The primary data source for this project is the U.S World News & Report Annual University Rankings. Unfortunately, U.S World News & Report does not have archival data accessible to the public in an easy manner so I had to find a reputable secondary source who has the data year by year. Fortunately, I came across the website of Dr. Andrew G. Reiter, an associate professor of politics and international relations at Mount Holyoke College who had compiled U.S News and World Report National University Rankings throughout many years.

A link to Dr. Reiter’s Website.
A link to Dr. Reiter’s Datasets.

Curation Process

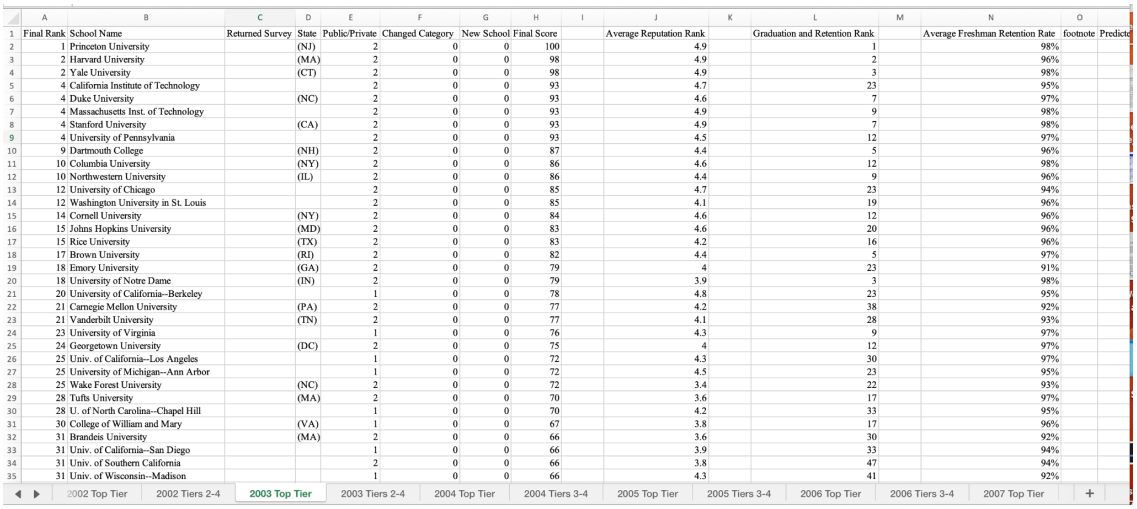
The goal of data curation is to enhance the value of data by ensuring its quality, reliability, readability, accessibility, and shareability. Effective data curation transforms a potentially messy dataset into a central asset for data analysis.

The curation process began with downloading the dataset from Dr. Reiter’s website in Excel format. The multi-sheet file indicated the need for a comprehensive curation approach. While all the necessary data was in one source, several challenges were apparent:

1. **Data Spread Across Multiple Sheets:** Each sheet corresponded to a different year, requiring consolidation.
2. **Wide Format with Unnecessary Variables:** The data included extraneous variables and lacked others, notably the 'region' and 'campus type'.

3. **Inconsistency Across Years/Sheets:** Some variables were not consistently available in all years.

The initial state of the data in Dr. Reiter's sheet was as follows:



Final Rank	School Name	Returned Survey	State	Public/Private	Changed Category	New School	Final Score	Average Reputation Rank	Graduation and Retention Rank	Average Freshman Retention Rate	footnote	Predicte
1	Princeton University		(NJ)	2	0	0	100	4.9	1	98%		
2	Harvard University		(MA)	2	0	0	98	4.9	2	96%		
2	Yale University		(CT)	2	0	0	98	4.9	3	98%		
4	California Institute of Technology			2	0	0	93	4.7	23	95%		
4	Duke University		(NC)	2	0	0	93	4.6	7	97%		
4	Massachusetts Inst. of Technology			2	0	0	93	4.9	9	98%		
4	Stanford University		(CA)	2	0	0	93	4.9	7	98%		
4	University of Pennsylvania			2	0	0	93	4.5	12	97%		
9	Dartmouth College		(NH)	2	0	0	87	4.4	5	96%		
10	Columbia University		(NY)	2	0	0	86	4.6	12	98%		
10	Northwestern University		(IL)	2	0	0	86	4.4	9	96%		
12	University of Chicago			2	0	0	85	4.7	23	94%		
12	Washington University in St. Louis			2	0	0	85	4.1	19	96%		
14	Cornell University		(NY)	2	0	0	84	4.6	12	96%		
15	Johns Hopkins University		(MD)	2	0	0	83	4.6	20	96%		
15	Rice University		(TX)	2	0	0	83	4.2	16	96%		
17	Brown University		(RI)	2	0	0	82	4.4	5	97%		
18	Emory University		(GA)	2	0	0	79	4	23	91%		
18	University of Notre Dame		(IN)	2	0	0	79	3.9	3	98%		
20	University of California--Berkeley			1	0	0	78	4.8	23	95%		
21	Carnegie Mellon University		(PA)	2	0	0	77	4.2	38	92%		
21	Vanderbilt University		(TN)	2	0	0	77	4.1	28	93%		
23	University of Virginia			1	0	0	76	4.3	9	97%		
24	Georgetown University		(DC)	2	0	0	75	4	12	97%		
25	Univ. of California--Los Angeles			1	0	0	72	4.3	30	97%		
25	University of Michigan--Ann Arbor			1	0	0	72	4.5	23	95%		
25	Wake Forest University		(NC)	2	0	0	72	3.4	22	93%		
28	Tufts University		(MA)	2	0	0	70	3.6	17	97%		
28	U. of North Carolina--Chapel Hill			1	0	0	70	4.2	33	95%		
30	College of William and Mary		(VA)	1	0	0	67	3.8	17	96%		
31	Brandeis University		(MA)	2	0	0	66	3.6	30	92%		
31	Univ. of California--San Diego			1	0	0	66	3.9	33	94%		
31	Univ. of Southern California			2	0	0	66	3.8	47	94%		
31	Univ. of Wisconsin--Madison			1	0	0	66	4.3	41	92%		

2002 Top Tier | 2002 Tiers 2-4 | 2003 Top Tier | 2003 Tiers 2-4 | 2004 Top Tier | 2004 Tiers 3-4 | 2005 Top Tier | 2005 Tiers 3-4 | 2006 Top Tier | 2006 Tiers 3-4 | 2007 Top Tier

The cleaning process involved filtering, sorting, reformatting, and defining the scope of the study. The scope was narrowed down to the top 25 universities over a ten-year period (2003 - 2012), chosen for consistency in available metrics.

The following steps were taken in the curation process:

1. **Identification and Removal of Irrelevant Variables:** After pinpointing the necessary variables for analysis, I discarded any that were not pertinent.
2. **Consolidation into a Single, Long-Format Sheet:** The diverse sheets were amalgamated into one comprehensive long-format sheet through an arduous process of copying and pasting. This painstaking transfer, encompassing 311 records, was pivotal to maintaining the integrity of the data and laying the groundwork for a more efficient analysis.
3. **Creation of a Regional Breakdown:** The dataset originally lacked regional data, this prompted the segmentation of the United States into five standard regions: Northeast (NE), Southeast (SE), Midwest (MW), Southwest (SW), and West (W). These regions were then represented as binary dummy variables to facilitate analysis.
4. **Addition of Campus Type:** Though not available on Dr. Reiter's site, this information was available on the U.S. News and World Report website.
5. **Transformation of Percentage Data to Decimal Format:** For variables originally expressed as percentages, including the acceptance rate, graduation rate, percentage of entering freshmen in the top 10% of their graduating high school class, and percentage of

classes with 50 or more students, I implemented a two-step process. Firstly, I removed the percentage sign from the original columns. Then, I introduced adjacent columns to apply the formula "=(cell)/100" for converting these values into decimal format, simplifying the subsequent analytical procedures.

The final curated dataset included the following variables: university name, year, campus type, NE, SE, MW, SW, W (regional indicators), rank, acceptance rate (in decimal), graduation rate (6-year, in decimal), entering freshmen in the top 10% of their class (in decimal) and percent of classes with 50 or more students (in decimal).

This meticulous curation process was in two parts. The first part being focused on making the dataset consistent and comprehensive (changes 1,2, and 4), the second part being priming the dataset for analysis through R (changes 3 and 5). This curated data is exactly what is needed for a robust analysis of the factors influencing the U.S. News and World Report university rankings.

The state of the updated dataset is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	University	Year	Campus_Type	NE	SE	MW	W	SW	Rank	Acceptance_Rate	Acceptance_Rate	Graduation_Rate_(6 years)	Graduation_Rate	Entering_Freshmen_in_Top_10%_of_Class	Freshmen_Top_10_Percent	Percent_of_Classes_with_50_or_More_Students	Percent_of_Classes_with_50_or_More_Students
2	Princeton University	2003	0	1	0	0	0	0	1	12	0.12	97	0.97	93	0.93	10	0.1
3	Princeton University	2004	0	1	0	0	0	0	1	11	0.11	97	0.97	95	0.95	11	0.11
4	Princeton University	2005	0	1	0	0	0	0	1	10	0.10	97	0.97	94	0.94	11	0.11
5	Princeton University	2006	0	1	0	0	0	0	1	13	0.13	97	0.97	94	0.94	11	0.11
6	Princeton University	2007	0	1	0	0	0	0	1	11	0.11	97	0.97	94	0.94	10	0.1
7	Princeton University	2008	0	1	0	0	0	0	1	10	0.10	96	0.96	94	0.94	10	0.1
8	Princeton University	2009	0	1	0	0	0	0	2	10	0.10	95	0.95	96	0.96	10	0.1
9	Princeton University	2010	0	1	0	0	0	0	1	10	0.10	96	0.96	97	0.97	9	0.09
10	Princeton University	2011	0	1	0	0	0	0	2	10	0.10	96	0.96	95	0.95	11	0.11
11	Princeton University	2012	0	1	0	0	0	0	1	9	0.09	96	0.96	99	0.99	11	0.11
12	Harvard University	2003	1	1	0	0	0	0	2	11	0.11	96	0.96	90	0.90	13	0.13
13	Harvard University	2004	1	1	0	0	0	0	1	11	0.11	98	0.98	90	0.90	13	0.13
14	Harvard University	2005	1	1	0	0	0	0	1	10	0.10	98	0.98	90	0.90	13	0.13
15	Harvard University	2006	1	1	0	0	0	0	1	11	0.11	98	0.98	96	0.96	13	0.13
16	Harvard University	2007	1	1	0	0	0	0	2	9	0.09	98	0.98	96	0.96	13	0.13
17	Harvard University	2008	1	1	0	0	0	0	2	9	0.09	98	0.98	95	0.95	13	0.13
18	Harvard University	2009	1	1	0	0	0	0	1	9	0.09	97	0.97	95	0.95	9	0.09
19	Harvard University	2010	1	1	0	0	0	0	1	8	0.08	98	0.98	95	0.95	8	0.08
20	Harvard University	2011	1	1	0	0	0	0	1	7	0.07	98	0.98	95	0.95	8	0.08
21	Harvard University	2012	1	1	0	0	0	0	1	7	0.07	97	0.97	95	0.95	8	0.08
22	Yale University	2003	1	1	0	0	0	0	2	14	0.14	94	0.94	95	0.95	9	0.09
23	Yale University	2004	1	1	0	0	0	0	3	13	0.13	95	0.95	95	0.95	8	0.08
24	Yale University	2005	1	1	0	0	0	0	3	11	0.11	96	0.96	99	0.99	8	0.08
25	Yale University	2006	1	1	0	0	0	0	3	10	0.10	96	0.96	95	0.95	8	0.08
26	Yale University	2007	1	1	0	0	0	0	3	10	0.10	96	0.96	95	0.95	8	0.08
27	Yale University	2008	1	1	0	0	0	0	3	9	0.09	96	0.96	95	0.95	8	0.08
28	Yale University	2009	1	1	0	0	0	0	3	10	0.10	96	0.96	97	0.97	8	0.08
29	Yale University	2010	1	1	0	0	0	0	3	9	0.09	97	0.97	97	0.97	7	0.07
30	Yale University	2011	1	1	0	0	0	0	3	8	0.08	98	0.98	96	0.96	7	0.07
31	Yale University	2012	1	1	0	0	0	0	3	8	0.08	96	0.96	97	0.97	7	0.07
32	California Institute of Technology	2003	0	0	0	0	1	0	4	15	0.15	81	0.81	98	0.98	6	0.06
33	California Institute of Technology	2004	0	0	0	0	1	0	5	21	0.21	85	0.85	99	0.99	7	0.07
34	California Institute of Technology	2005	0	0	0	0	1	0	8	17	0.17	89	0.89	94	0.94	9	0.09
35	California Institute of Technology	2006	0	0	0	0	1	0	7	21	0.21	88	0.88	93	0.93	9	0.09
36	California Institute of Technology	2007	0	0	0	0	1	0	4	20	0.20	90	0.90	94	0.94	8	0.08
37	California Institute of Technology	2008	0	0	0	0	1	0	5	17	0.17	89	0.89	88	0.88	8	0.08
38	California Institute of Technology	2009	0	0	0	0	1	0	6	17	0.17	89	0.89	99	0.99	8	0.08
39	California Institute of Technology	2010	0	0	0	0	1	0	4	17	0.17	88	0.88	97	0.97	6	0.06
40	California Institute of Technology	2011	0	0	0	0	1	0	7	15	0.15	89	0.89	98	0.98	9	0.09
41	California Institute of Technology	2012	0	0	0	0	1	0	5	13	0.13	90	0.90	96	0.96	10	0.1
42	Duke University	2003	0	0	1	0	0	0	4	26	0.26	94	0.94	86	0.86	7	0.07
43	Duke University	2004	0	0	1	0	0	0	5	25	0.25	93	0.93	89	0.89	5	0.05
44	Duke University	2005	0	0	1	0	0	0	5	16	0.16	92	0.92	97	0.97	15	0.15
45	Duke University	2006	0	0	1	0	0	0	5	24	0.24	94	0.94	87	0.87	5	0.05
46	Duke University	2007	0	0	1	0	0	0	8	24	0.24	93	0.93	88	0.88	5	0.05
47	Duke University	2008	0	0	1	0	0	0	8	23	0.23	94	0.94	89	0.89	6	0.06
48	Duke University	2009	0	0	1	0	0	0	8	23	0.23	94	0.94	90	0.90	5	0.05
49	Duke University	2010	0	0	1	0	0	0	10	22	0.22	95	0.95	90	0.90	5	0.05
50	Duke University	2011	0	0	1	0	0	0	9	19	0.19	95	0.95	90	0.90	6	0.06
51	Duke University	2012	0	0	1	0	0	0	10	16	0.16	94	0.94	95	0.95	6	0.06
52	Massachusetts Institute of Technology	2003	1	1	0	0	0	0	4	17	0.17	92	0.92	98	0.98	15	0.15
53	Massachusetts Institute of Technology	2004	1	1	0	0	0	0	4	16	0.16	91	0.91	99	0.99	11	0.11

Correlation Analysis

To understand the factors that contribute to the U.S. News and World Report rankings, a Pearson's correlation analysis was conducted through the R coding language. It is essential to note that in these rankings, a lower numerical value signifies a higher rank; therefore, a negative correlation actually implies a higher ranking and a positive correlation implies a lower ranking. The following screenshots include the input code and the a summary output of the results.

Here is the code:

```r
# Load Data
data <- read.csv("DA_Assign_4_Data.csv")

# Define Input Variables
input_variables <- c("Campus_Type", "NE", "SE", "MW", "W", "SW", "Acceptance_Rate", "Graduation_Rate", "Freshmen_Top_10_Percent", "Percent_of_Classes_with_50_or_More_Students")

# Loop through Input Variables for Correlation Analysis
for (var in input_variables) {
  cat("\nCorrelation Analysis for:", var, "\n")

  # Perform Correlation Test
  test_result <- cor.test(data[[var]], data[["Rank"]], method = "pearson")

  # Display Results
  cat("Correlation Coefficient (r):", test_result$estimate, "\n")
  cat("P-value:", test_result$p.value, "\n")
  cat("Direction:", ifelse(test_result$estimate > 0, "Positive", "Negative"), "\n")
}
```

Here are the results from the code:

```
Correlation Analysis for: Campus_Type
Correlation Coefficient (r): -0.03491257
P-value: 0.5402714
Direction: Negative

Correlation Analysis for: NE
Correlation Coefficient (r): -0.4401449
P-value: 4.038986e-16
Direction: Negative

Correlation Analysis for: SE
Correlation Coefficient (r): 0.4040207
P-value: 1.333889e-13
Direction: Positive

Correlation Analysis for: MW
Correlation Coefficient (r): 0.01030323
P-value: 0.8566194
Direction: Positive

Correlation Analysis for: W
Correlation Coefficient (r): 0.05288994
P-value: 0.3533498
Direction: Positive

Correlation Analysis for: SW
Correlation Coefficient (r): 0.0469627
P-value: 0.4256069
Direction: Positive

Correlation Analysis for: Acceptance_Rate
Correlation Coefficient (r): 0.7007924
P-value: 4.269443e-47
Direction: Positive

Correlation Analysis for: Graduation_Rate
Correlation Coefficient (r): -0.602976
P-value: 4.478489e-32
Direction: Negative

Correlation Analysis for: Freshmen_Top_10_Percent
Correlation Coefficient (r): -0.5452203
P-value: 2.105613e-25
Direction: Negative

Correlation Analysis for: Percent_of_Classes_with_50_or_More_Students
Correlation Coefficient (r): 0.09349045
P-value: 0.1003788
Direction: Positive
```

a. Correlation Scores ('r' score):
    - Campus Type: r = -0.0349
    - NE: r = -0.4401
    - SE: r = 0.4040
    - MW: r = 0.0103
    - W: r = 0.0529
    - SW: r = 0.0470
    - Acceptance Rate: r = 0.7008
    - Graduation Rate: r = -0.6030
    - Freshmen Top 10 Percent: r = -0.5452
    - Percent of Classes with 50 or More Students: r = 0.0935
b. Significance (p-value):
    - Campus Type: p = .5403 (Not Significant)
    - NE: $p < 0.001$ (Significant)
    - SE: $p < 0.001$ (Significant)
    - MW: p = 0.8566 (Not Significant)
    - W: p = 0.3533 (Not Significant)
    - SW: p = 0.4256
    - Acceptance Rate: $p < 0.001$ (Significant)
    - Graduation Rate: $p < 0.001$ (Significant)
    - Freshmen Top 10 Percent: $p < 0.001$ (Significant)
    - Percent of Classes with 50 or More Students: p = 0.1004 (Not Significant)
c. Direction
    - Campus Type: Negative
    - NE: Negative
    - SE: Positive
    - MW: Positive
    - SW: Positive
    - Acceptance Rate: Positive
    - Graduation Rate: Negative
    - Freshmen Top 10 Percent: Negative
    - Percent of Classes with 50 or More Students: Positive

d. Interpretation of Findings
    - Campus Type: The negative correlation, though not significant, shows that urban campuses are slightly higher in rankings than suburban campuses. However this is not statistically significant.

- NE Region: The negative correlation shows us that universities in the Northeast tend to be ranked higher, this aligns with the expectation as this region has a world renowned academic reputation (e.g. Ivy League).
- SE Region: The positive correlation suggests that universities in the Southeast are associated with lower rankings. This is interesting as these universities are known throughout the nation for having some of the most acclaimed athletic reputations.
- Acceptance Rate: The positive correlation here shows us that universities with higher acceptance rates are associated with lower rankings. This makes sense as these universities are less selective.
- Graduation Rate: The negative correlation shows us that higher graduation rates are associated with higher rankings. This is probably a very important factor when ranking universities.
- Freshmen Top 10 Percent: The negative correlation shows us that a greater percentage of top-performing incoming freshmen correlates with higher rankings. The best students are ultimately going to the best schools.
- Percent of Classes with 50 or More Students: This is an insignificant positive correlation. Meaning this variable has minimal impact on rankings, but could potentially mean that larger class sizes do not correlate with lower university rankings.

These findings show us a few things. Firstly, they underscore the multifaceted nature of the ranking system. Traditional indicators of academic excellence such as selectivity (shown by lower acceptance rates) and student performance (higher percentage of top performing incoming students) are closely aligned to higher rankings. Additionally, we can see that certain regions have higher rankings compared to others.

Multiple Linear Regression Model Analysis

A multiple linear regression model was used to analyze the influence of various factors on the rankings by U.S. News and World Report. As a reminder, The model's code and output from R Studio is shown in the screenshot below:

```
> # Building the Multiple Linear Regression Model
> model <- lm(Rank ~ Campus_Type + NE + SE + MW + W + SW + Acceptance_Rate + Graduation_Rate + Freshmen_Top_10_Percent + Percent_of_Classes_with_50_or_More_Students, data = data)
>
> # Summary of the Model
> model_summary <- summary(model)
> model_summary

Call:
lm(formula = Rank ~ Campus_Type + NE + SE + MW + W + SW + Acceptance_Rate +
    Graduation_Rate + Freshmen_Top_10_Percent + Percent_of_Classes_with_50_or_More_Students,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6548  -4.0244  -0.6221   3.8153  16.5248

Coefficients: (1 not defined because of singularities)
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                44.9726    14.4668   3.109  0.00207 **
Campus_Type                                 0.9900     0.7623   1.299  0.19512
NE                                         -2.8133     1.8675  -1.506  0.13308
SE                                         -1.4777     1.9324  -0.765  0.44510
MW                                         -4.2171     2.0753  -2.032  0.04309 *
W                                           1.5285     2.1946   0.696  0.48670
SW                                              NA         NA      NA       NA
Acceptance_Rate                            39.3591     6.4405   6.111 3.31e-09 ***
Graduation_Rate                           -20.6178    14.4490  -1.427  0.15471
Freshmen_Top_10_Percent                   -25.3579     6.1254  -4.140 4.60e-05 ***
Percent_of_Classes_with_50_or_More_Students  23.7229     8.7083   2.724  0.00685 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.525 on 280 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.5913,    Adjusted R-squared:  0.5781
F-statistic:    45 on 9 and 280 DF,  p-value: < 2.2e-16
```

a. **Variance Explained:**

The model's R-squared value is 0.5913, this means that approximately 59.13% of the variance in university rankings is explained by the model. Overall, that shows this model is a moderate-to-strong fit, because more than half of the variability in rankings can be accounted for by the predictors used.

b. **Significance of Variables:**

- *Acceptance Rate:* A coefficient of 39.3591 and a p-value well below 0.001, the acceptance rate is a highly significant predictor, it suggests that more selective universities (those with lower acceptance rates) are ranked higher. This shows us that prestige is a large indicator of ranking.
- *Graduation Rate:* The negative coefficient of -20.6178 shows that higher graduation rates might be associated with higher rankings, however this is not statistically significant as the p-value is 0.15471. This does make sense however, since graduation rates are great indicators of whether a university is fulfilling its role as students aren't transferring/dropping out.
- *Freshmen in Top Ten Percent of Graduating High School Class:* This has a significant negative coefficient of -25.3579 and p-value that is less than 0.001.

This means that a higher proportion of freshmen in the top 10% of their high school class is associated with higher rankings, reinforcing expectations that higher ranked institutions have rigorous acceptance requirements.

- *Percent of Classes with 50 or More Students:* A positive coefficient of 23.7229 with a significant p-value of 0.00685 shows that universities with more large classes have lower rankings. This makes sense as smaller class sizes are often associated with more personalized education and better student-faculty relationships.

Ultimately, this shows us that the most important and significant variables in this model show us that higher ranking universities are generally smaller, have a very capable student body fueled by accepting only the best of the best, have high graduation rates, and low acceptance rates.

Interpretation of Findings

The first step of interpreting these statistical findings is always to examine the significance of each variable. After gauging whether or not a variable is statistically significant, then we can confidently examine its impact on the final ranking. Anything that is not statistically significant must be disregarded. For this reason, acceptance rate, graduation rate, percent of classes with 50 or more students and incoming freshmen in the top ten percent of graduating high school class will be examined as the key determinants.

The implications of these findings are multifaceted. For university administrators and policymakers, understanding the fact that the quality of the student body and the selectivity of an institution are closely correlated with higher rankings could help inform their decision making for improving educational offerings and marketing to prospective students. These policymakers should have a focus on increasing the academic rigor along with pushing their students to achieve the best they can, to feed into a positive feedback loop. Additionally, the non-significant impact of class-size on rankings challenges the notion that small is better, and that there are other factors at play that matter more in ranking (i.e., research output, faculty reputation).

Many variables seem to impact each other and measure the same things, namely academic rigor. Acceptance rate and the measure of entering freshman who were in the top 10% of their graduating high school class essentially measure the same thing, selectivity. The selectivity impacts the graduation rates as the student body already consists of high achievers.

Limitations, Ethical Implications, and Future Directions

Limitations

The primary limitation of this study is the partial understanding of the U.S. News & World Report's ranking methodology. Without comprehensive insight into each variable used and its respective weight in the ranking algorithm, the model used in this study can only approximate the impact of the factors we have considered. An additional limitation is the lack of data, the dataset is constrained to the top 25 universities over a ten-year span, providing a limited view that measures the elite universities instead of the broader landscape of higher education institutions. The limited number of variables measured is another constraint, as additional factors could offer deeper insights into the ranking mechanics. Finally, the data's outdated timeframe (2003-2012), does not reflect recent changes that have occurred in universities, possibly affecting current rankings.

Ethical Implications

While the study uses publicly available data, there are still some ethical questions to take under consideration. It is essential that the interpretation of this data does not misguide university policymakers or students when making educational decisions. The study should not be used to influence university policies in a manner that prioritizes rankings over educational quality and student well-being. The number one concern is safeguarding the integrity of the study and its conclusions.

Future Directions

Given more time and resources (namely data), the scope of this study would have expanded. A larger, more complete dataset including more universities and extending to the present would be used to provide a more representative and up-to-date analysis. Incorporating additional variables would be another factor that would help create a more complex model that captures the multifaceted nature of the rankings.

If given the opportunity through more time and resources, this study would be expanded beyond the U.S. News and World Report rankings, including different ranking systems for a more nuanced understanding of how various factors are valued across different evaluations, thus painting a clearer picture of the higher education landscape. This comprehensive approach would lead to a more impactful and richer analysis, supporting stakeholders in making informed decisions that reflect their goals and values.

In conclusion, while this study has shed light on some of the factors that contribute to university rankings, it also opens the door to future exploration, serving as a stepping stone for more extensive research in the field of higher education analytics.