2016 US News and World Reports Data Exploration

Thomas M. Seeber

thomasmseeeber@gmail.com

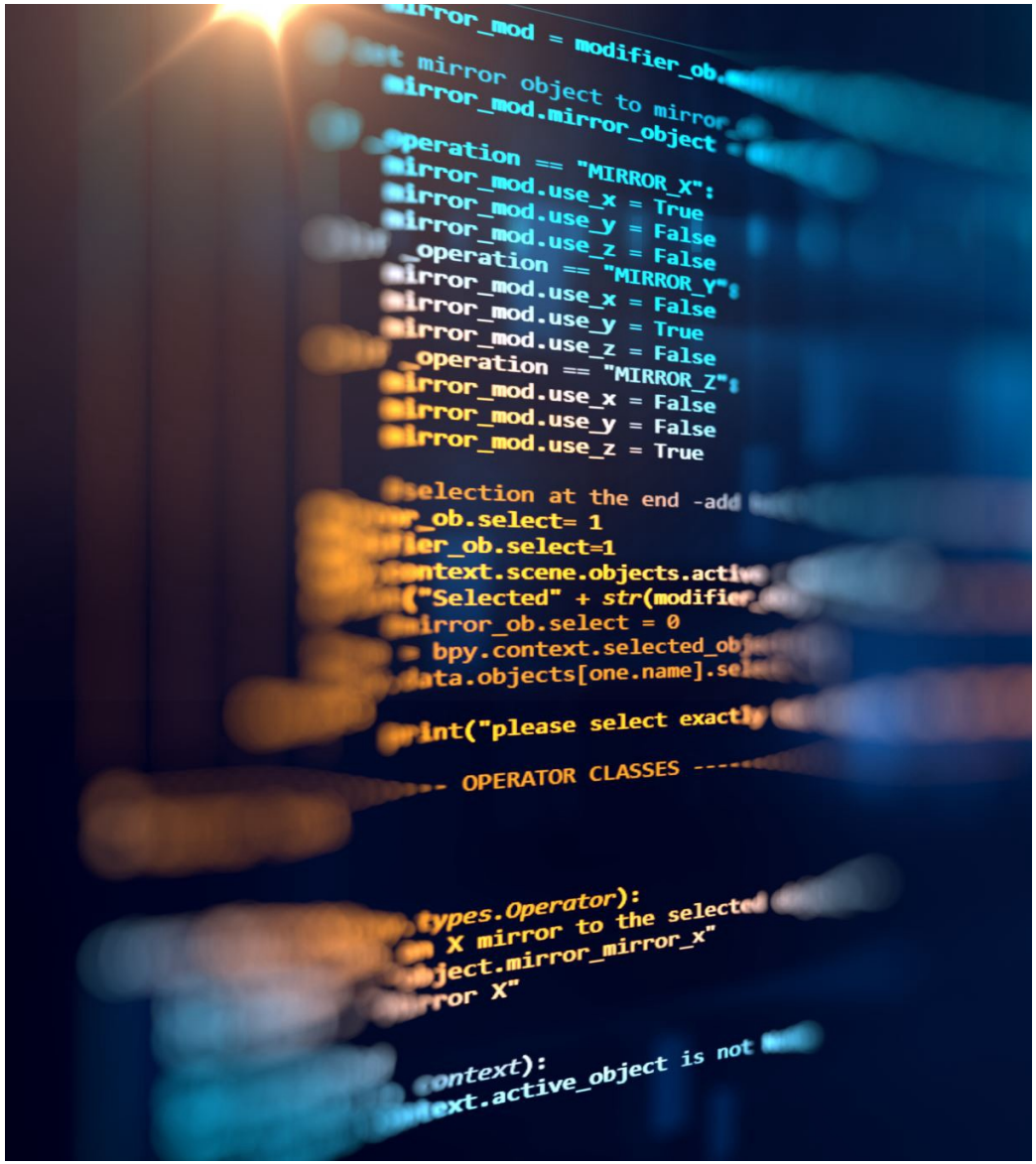
https://www.linkedin.com/in/thomasseeber

Technologies Used for Analysis

seaborn

matplotlib

Geopandas



- Complete Source Code and Project files Available at
- https://github.com/tomseeber/college_data_analysis

Exploratory Data Analysis (EDA): New Tools Implemented

▸ 1000465	99 date	security_id	int_val	Col0	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11
0	2018-09-13	100000	13700556066	1.65	1.68	-0.09	0.68	2.04	1.40	-0.74	-0.15	-0.40	-0.07	0.16	0.90
1	2018-09-13	100001	32370834703	0.81	-0.68	-2.29	-0.58	1.05	0.88	-0.01	0.40	-0.63	-0.34	-1.62	1.99
2	2018-09-	12	74359132292	0.27	1.39			-1.1	27	-1.4	0.28	0.04			0.61
3	2018-09-13		3555690688	0.24			-0.31		.95	-(	0.95	-1.2	21	-1.74	-0.48
4	2018-09-13	00004	105740597	-0.64	-1.38		1.0		0.44		-1.81			2.02	1.05
5	2018-09-13	100005		00	0.7	1.73				6	-0.85		1.18	0.41	1.03
6	2018-09-		30414943406	0.09		0.9		-(	0	0.39	-1.16			1.76	1.11
7	20	00007	94115021503	0.07		0.8	0.60						0.16	-1.00	-1.23
8	2018-09-	100008	7207483836	-0.03	-0.68	-0.06	-0.58	-(		-1.13	0.3	0.09	-1.13	0.07	-0.07
9	2018-09-13	100009	24054692457	0.78	0.28	0.66	0.88	0.7	-0.57	-0.51	-0.29	-0.71	0.48	-1.07	0.41
10	2018-09-13	100010	55729553089	-0.24	-0.04	-0.29	1.63	-2.35	-0.18	0.35	-1.19	1.96	-1.02	0.43	1.40
11	2018-09-13	100011	79679700690	-0.41	0.34	0.27	0.07	0.94	-0.35	-1.54	-0.74	-0.81	-1.05	0.57	-0.20
12	2018-09-13	100012	35404502206	0.26	0.93	0.09	-0.11	0.77	-0.41	0.25	-1.92	1.05	-1.74	-0.58	-0.36
13	2018-09-13	100013	54366289347	1.60	0.55	-1.83	-0.74	1.90	-1.42	0.52	-0.86	1.19	0.47	1.69	-0.17
14	2018-09-13	100014	29512587692	0.74	-1.22	-0.79	0.40	0.20	-0.61	-1.58	1.87	1.11	-0.48	0.21	0.12
15	2018-09-13	100015	59921981555	1.70	0.39	-0.51	-0.77	0.99	-0.69	0.04	-0.87	1.60	0.31	1.39	-0.09

D-TALE

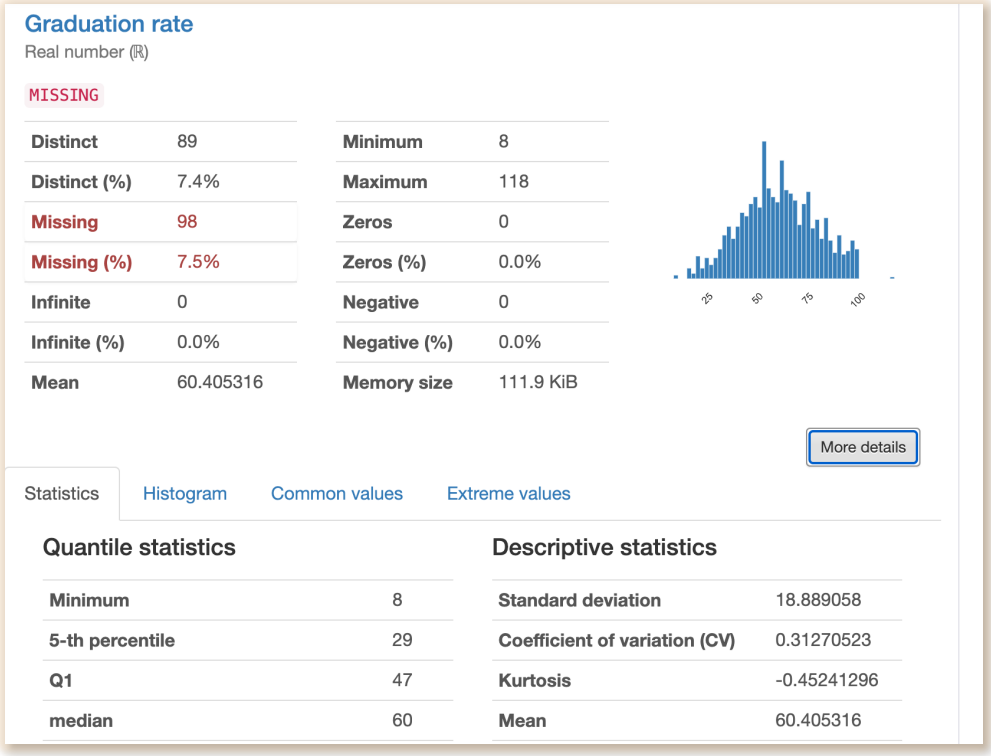
Exploratory

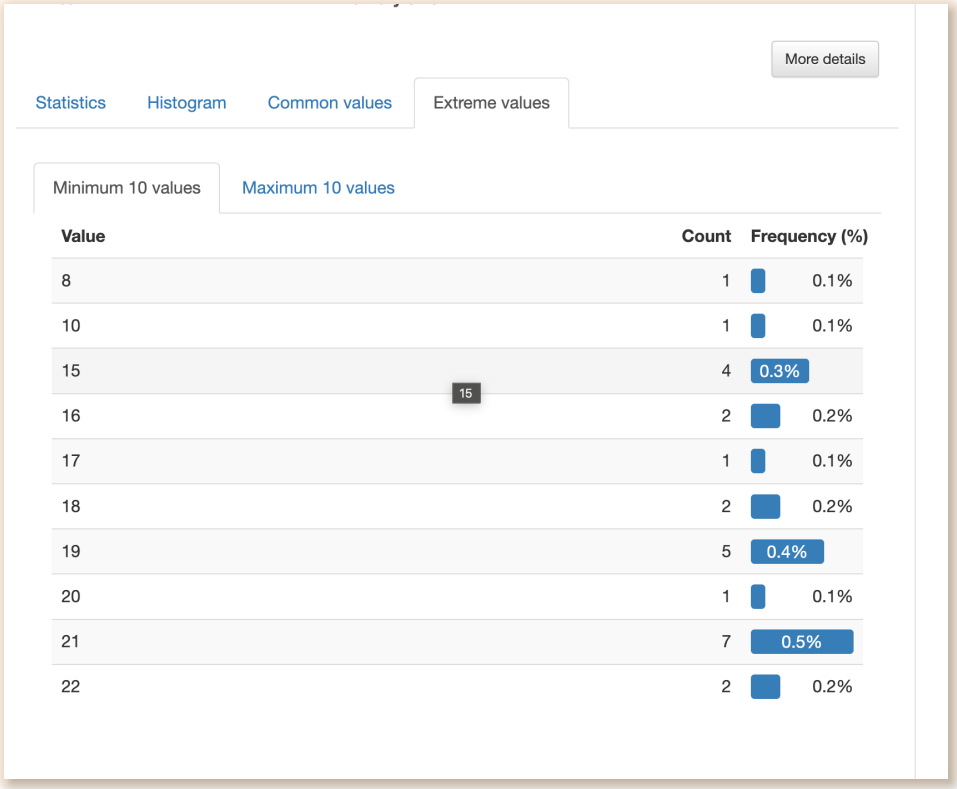
Some basic EDA was created using Automated tools. Jupyter notebooks are provided.

https://github.com/tomseeber/college_data_analysis/tree/main/notebooks

EDA: Panda Profiling (New Tools)

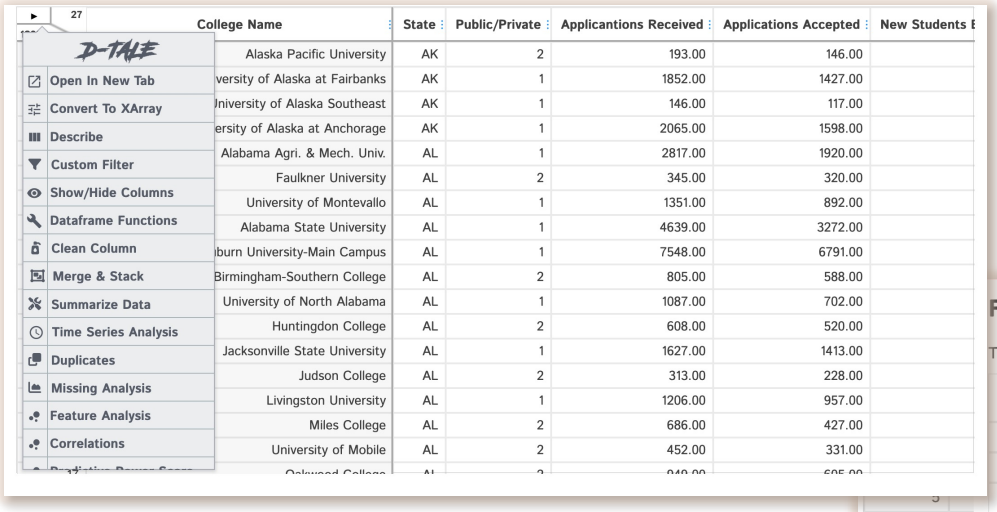
Panda Profiling lets you derive all basic statistics, outliers, correlations, and basic EDA information in one interface.





https://pypi.org/project/pandas-profiling/

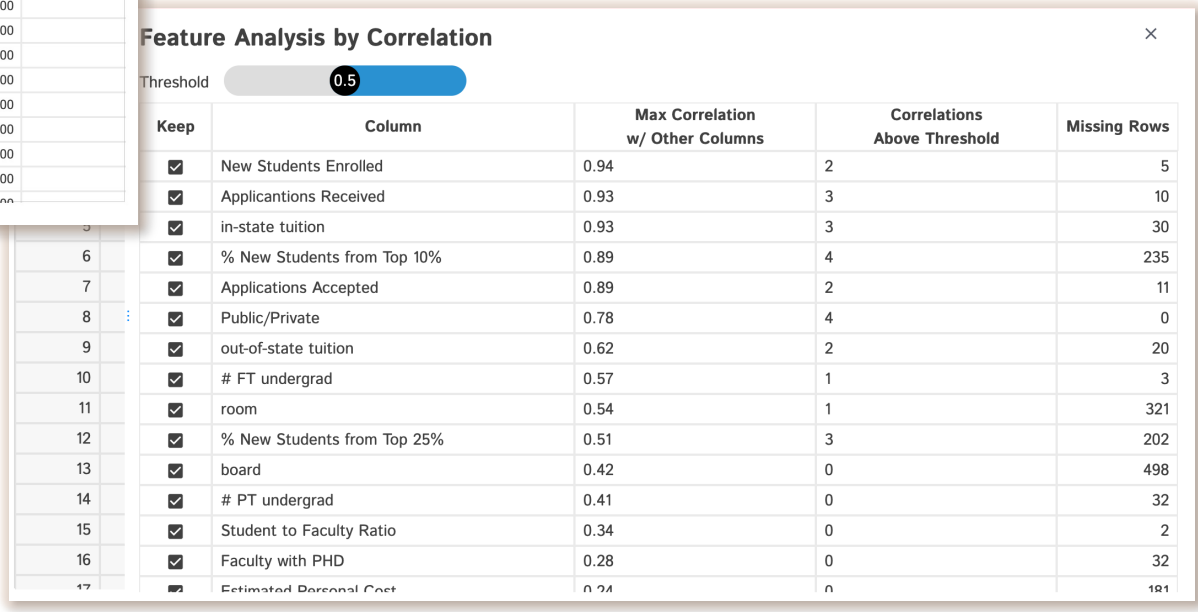
EDA: D-Tale (New Tools)



	College Name	State	Public/Private	Applicantions Received	Applications Accepted	New Students E
1	Alaska Pacific University	AK	2	193.00	146.00	
2	versity of Alaska at Fairbanks	AK	1	1852.00	1427.00	
3	Iniversity of Alaska Southeast	AK	1	146.00	117.00	
4	ersity of Alaska at Anchorage	AK	1	2065.00	1598.00	
5	Alabama Agri. & Mech. Univ.	AL	1	2817.00	1920.00	
6	Faulkner University	AL	2	345.00	320.00	
7	University of Montevallo	AL	1	1351.00	892.00	
8	Alabama State University	AL	1	4639.00	3272.00	
9	burn University-Main Campus	AL	1	7548.00	6791.00	
10	Birmingham-Southern College	AL	2	805.00	588.00	
11	University of North Alabama	AL	1	1087.00	702.00	
12	Huntingdon College	AL	2	608.00	520.00	
13	Jacksonville State University	AL	1	1627.00	1413.00	
14	Judson College	AL	2	313.00	228.00	
15	Livingston University	AL	1	1206.00	957.00	
16	Miles College	AL	2	686.00	427.00	
17	University of Mobile	AL	2	452.00	331.00	

https://pypi.org/project/dtale/

D-Tale lets you derive all basic statistics, outliers, correlations, and basic EDA information in one interface as well, but also does feature analysis, and on the fly graphing with GUI



Feature Analysis by Correlation

Threshold 0.5

Keep	Column	Max Correlation w/ Other Columns	Correlations Above Threshold	Missing Rows
☑	New Students Enrolled	0.94	2	5
☑	Applicantions Received	0.93	3	10
☑	in-state tuition	0.93	3	30
☑	% New Students from Top 10%	0.89	4	235
☑	Applications Accepted	0.89	2	11
☑	Public/Private	0.78	4	0
☑	out-of-state tuition	0.62	2	20
☑	# FT undergrad	0.57	1	3
☑	room	0.54	1	321
☑	% New Students from Top 25%	0.51	3	202
☑	board	0.42	0	498
☑	# PT undergrad	0.41	0	32
☑	Student to Faculty Ratio	0.34	0	2
☑	Faculty with PHD	0.28	0	32
☑	Estimated Personal Cost	0.24	0	181

ETL Processes: Adding In Additional Data - Geocoding

- Added Latitude and Longitude to Address Data (Geocoding) through the Geofy
  - Allows for direct mapping and geospatial analysis
  - Maps
  - Connection to US Census and demographic data

```
def geocode_dataframe(college_info_df):
    lat = []
    lon = []
    county = []
    address1 = []
    address2 = []
    zip_code = []

    for index, row in college_info_df.iterrows():
        target = f'{row["College Name"]},{row["State"]},USA'
        params = {"text": target, "apiKey": geoapify_key, "filter":"countrycode:us"}
        response = requests.get(base_url, params=params).json()
        print(index)
```

ETL Processes: Adding In Additional Data - US Reports Regions

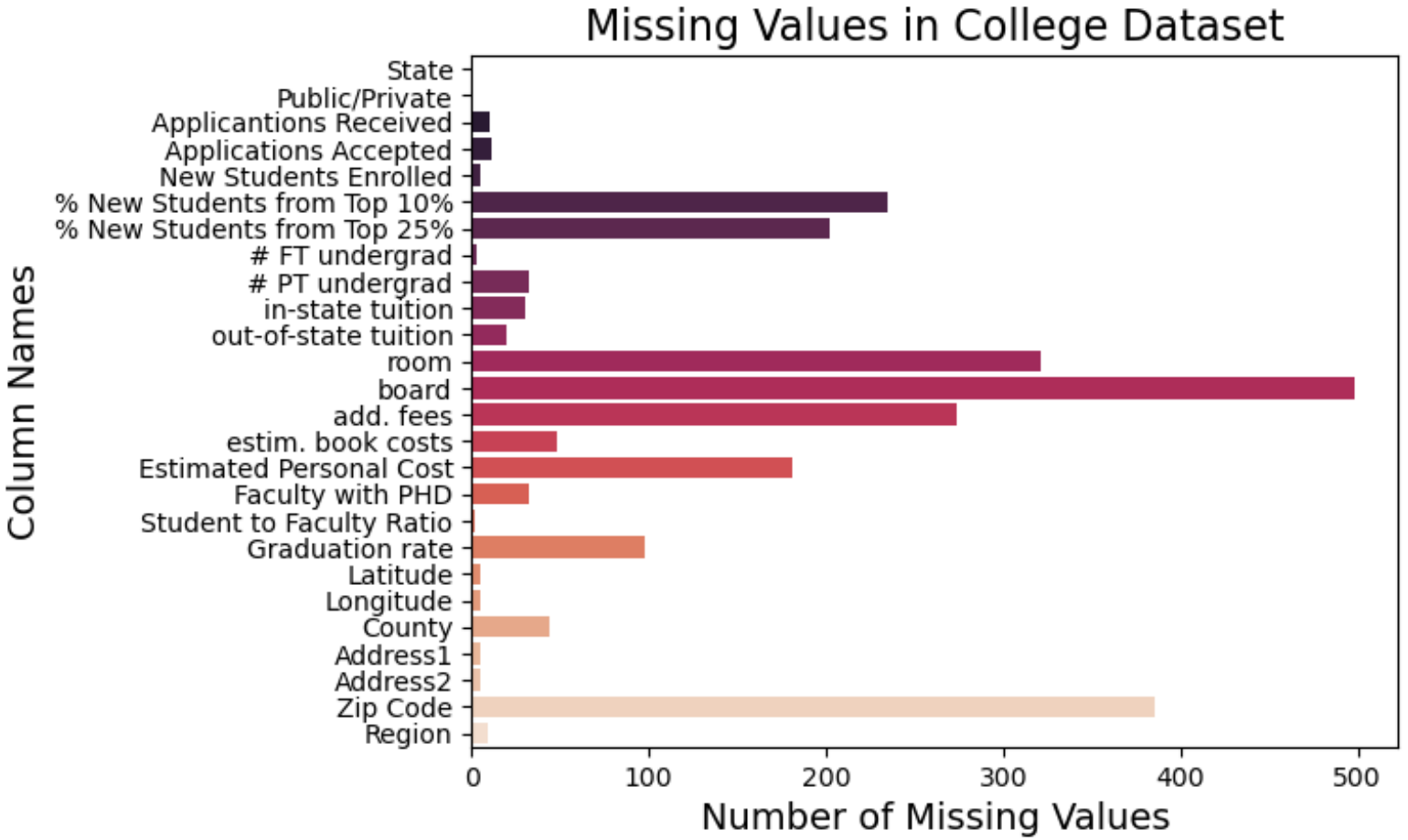
- Added Regional Data from the US and News and Reports
  - Original States per Region in US News and Reporting
  - Extracted from 2016 Information from ChatGPT
  - Cross-Region comparisons

```
regions_df = pd.read_csv('data/us_news_and_reports_regional_label.csv')
college_info_df=college_info_df.merge(regions_df, on='State', how='left')
```

Adding Regional Information for DC, since DC was not included in the State extraction

```
#fixing DC region to North
college_df.loc[college_df['State'] == 'DC','Region']='North'
```

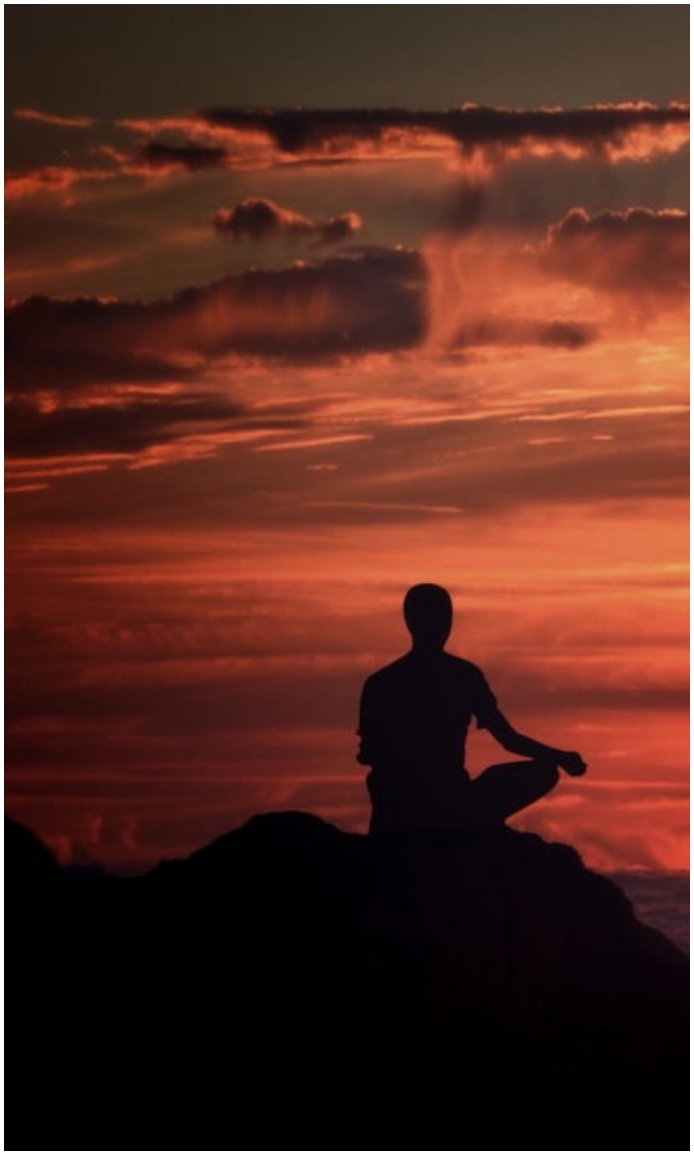
Data Cleaning: Missing Data: No drop_na



- Using the Pandas Method Drop_na takes our record count from 1302 to 471 with a 63.82% reduction in data.
- We can't use such a blunt tool.
- Case-by-case dropping using dropna with a subset on specific columns

Data Cleaning: Correcting Types

- Converted string columns to numeric (Float or INT) in Pandas
- Corrected any typos or mistakes that would prevent this conversion via Pandas and Loc.
- Kept data in Excel using Pandas Excel functionality conversions to avoid unneeded string-to-number conversions.
- No Data Time functionality was needed.



Outliers: Detection

The IQR method uses the quartiles +- 1.5 (Code is on the right).

Easy to deal with outliers:

- 120 for the Graduation Rate is impossible; manually fixed this with correct data. (Fixing these)
- Case Western Reserve University does not have a 2.9 Student to Faculty Ratio, and most lower than Q1-IQR*1.5 are probably incorrect data

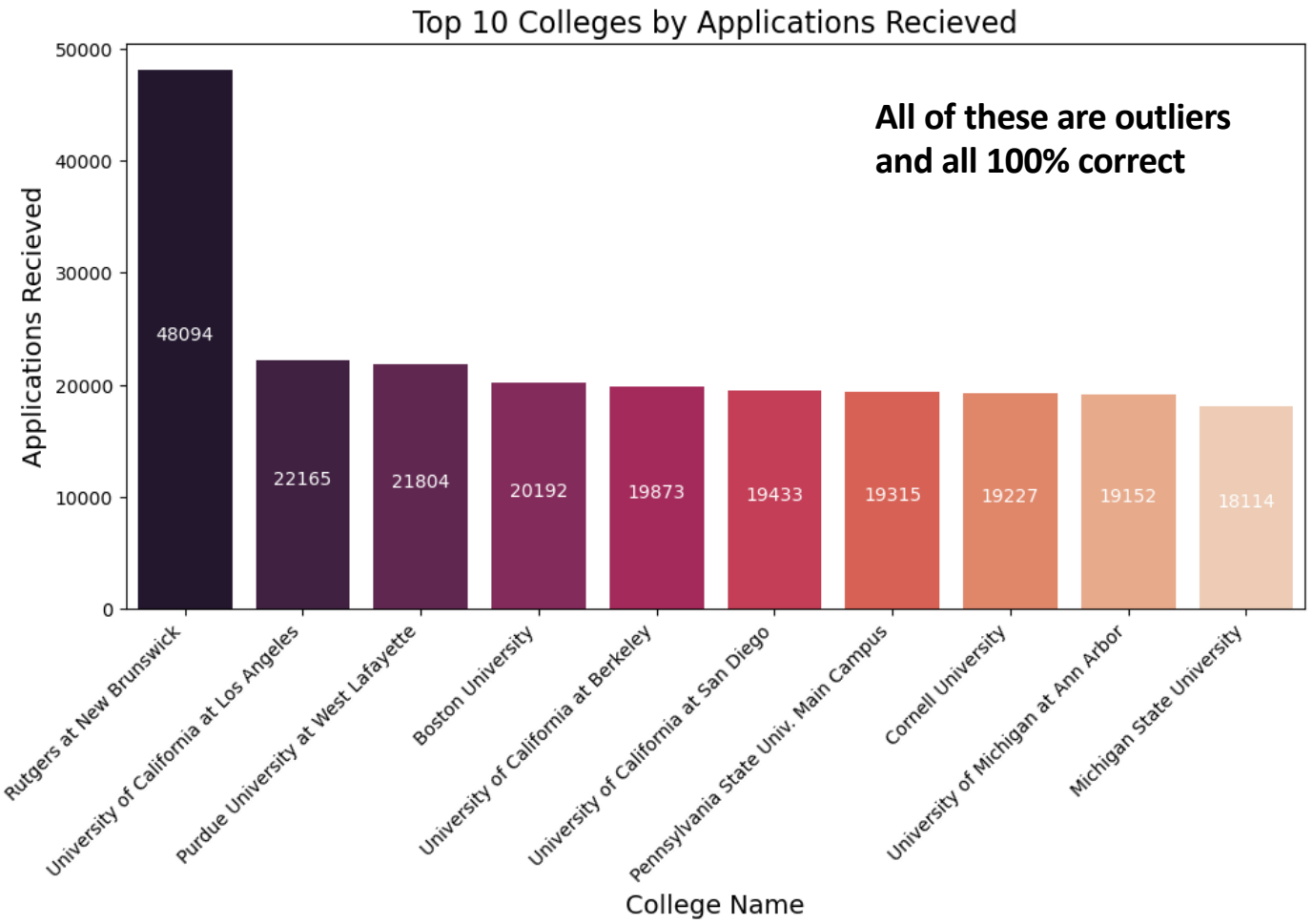
Harder to Dismiss

- ~48000 applications for Rutgers University in 2016 are correct.
- Similarly, in and out-of-state tuitions, some schools are correctly expensive outliers. Can we drop MIT from a college report?

```python
def IQR_outlier_report(column_list):

    '''
    This function takes list and returns a report
    of the number of outliers in the column based on the IQR method.
    '''
    q1 = column_list.quantile(.25)
    q3 = column_list.quantile(.75)
    # print(column)
    # print(f'Q1: {q1}')
    # print(f'Q3: {q3}')
    q2 = column_list.quantile(.5)
    iqr = q3 - q1
    lower_bound = q1 - (1.5 * iqr)
    upper_bound = q3 + (1.5 * iqr)
    return lower_bound, upper_bound, q1, q3, q2
```

✓ 0.0s

Data Cleaning: Most Outliers



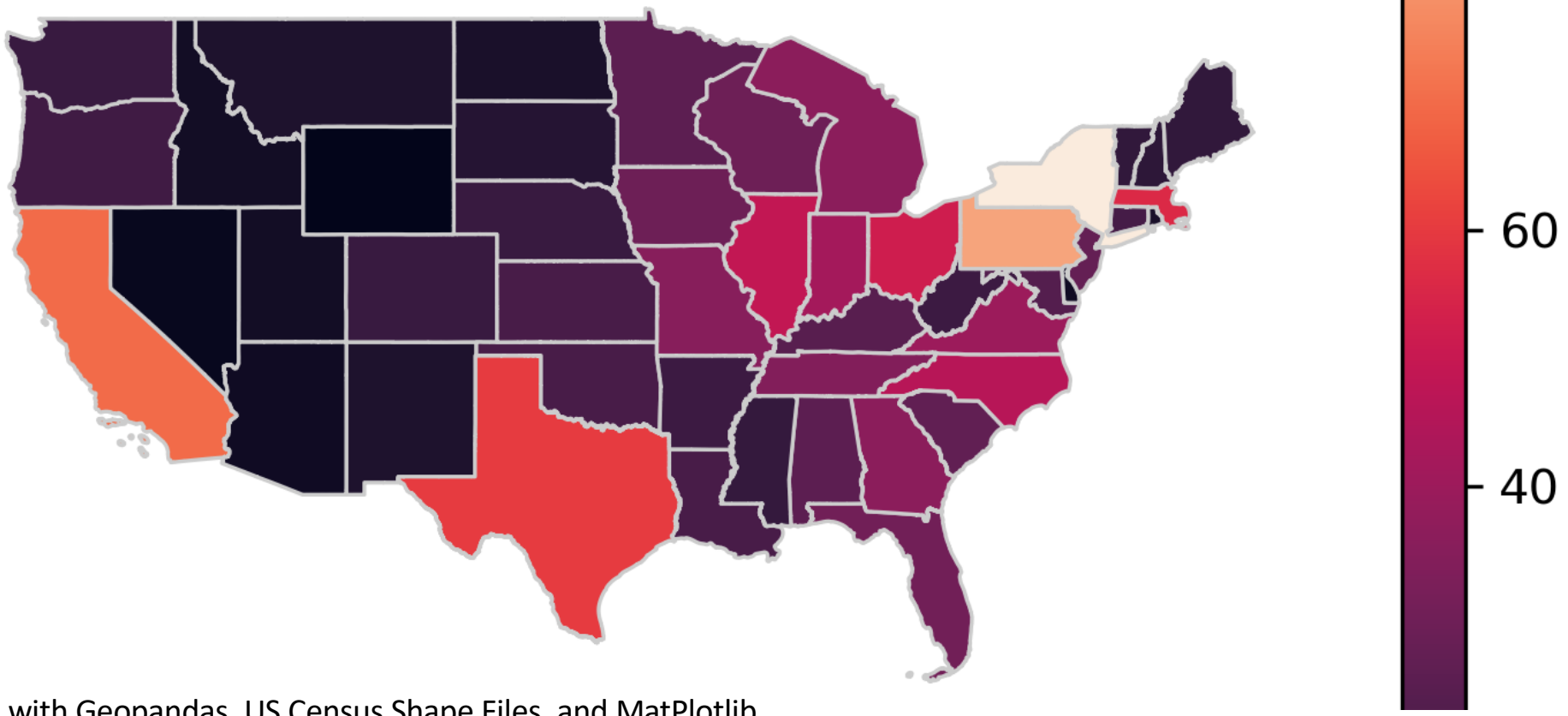
Most columns in the data have higher variability.

Outliers tend to look like the top 10 colleges by Applications Received.

All items are outliers in this chart, and all are relevant.

Outliers in this data can not be summarily dismissed.

Number of Colleges by State



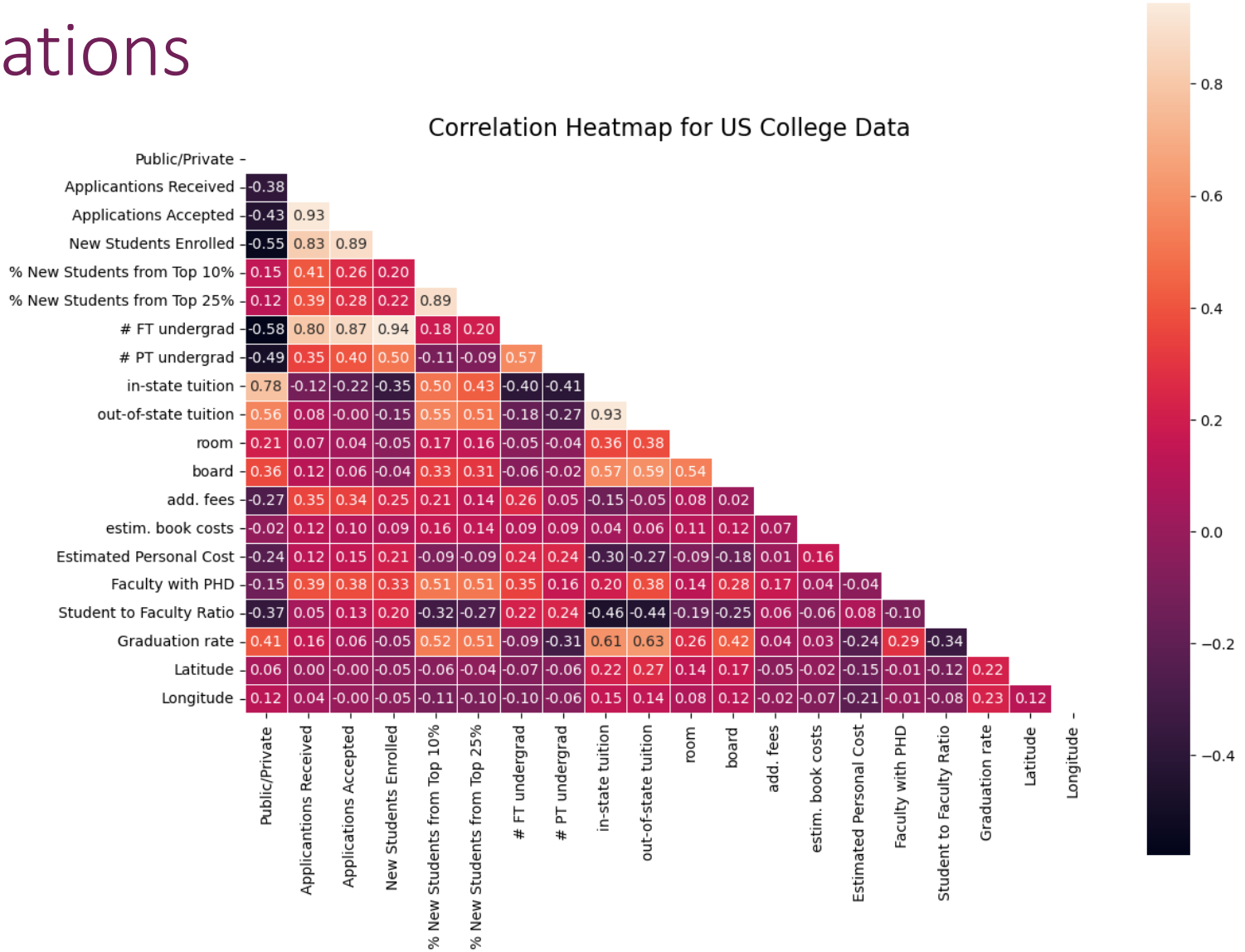
Created with Geopandas, US Census Shape Files, and MatPlotlib

General Correlations

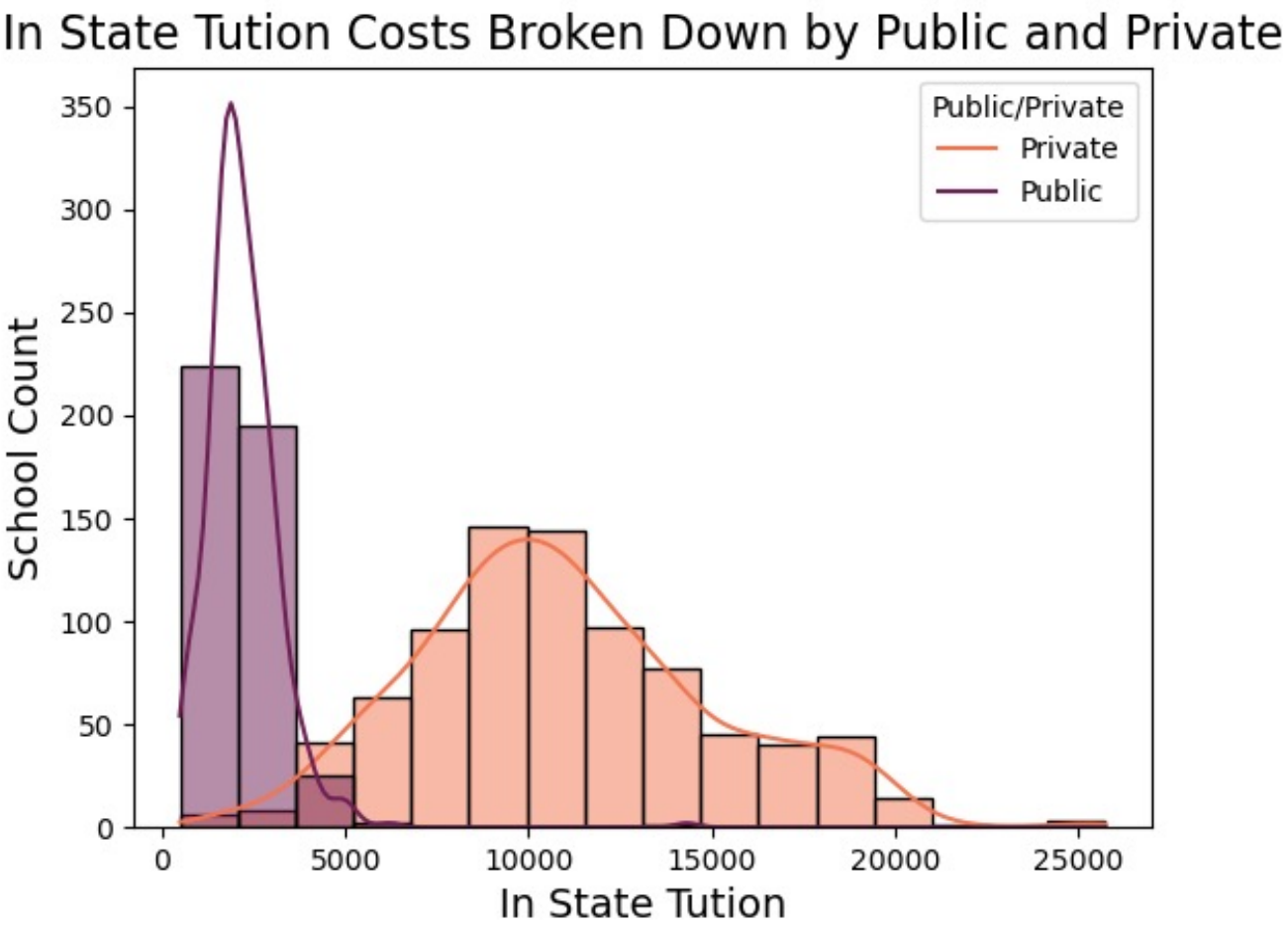
Applications Received, Applications Accepted, and New Students Received are all highly correlated. More enrollment can take more students and get more applications.

Part-time student status moderately negatively relates to Graduation Rates
(Correlation: - 0.31)

Additional Fees and Estimated Personal costs negatively correlate to Public/Private schools. At the same time, room and board are moderately correlated with public/private, meaning **Private schools tend to bundle more components into the room and board**, and **public schools tend to itemize costs and pass them to students.**



Cost: In-State Tuition



2208

Mean In-State Tuition for Public.

11008

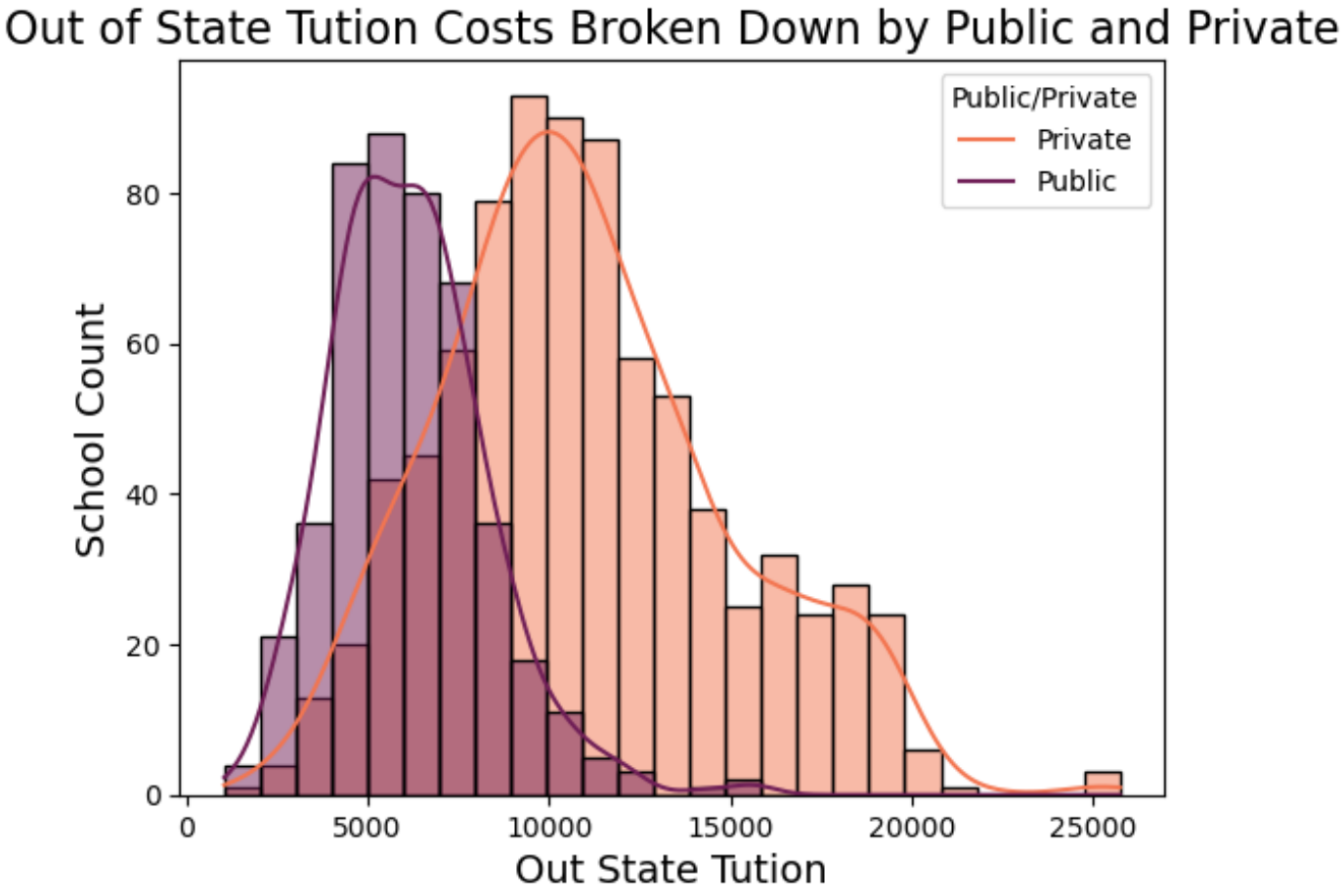
Mean for In-State Tuition for Private Schools

(2-tailed ttest p-value = 0.0, Statistic=58.6)

Schools over 20K a year in order of highest to lowest

Middlebury College
Bates College
Franklin and Marshall College Bennington College
Hampshire College
Massachusetts Institute of Technology

Cost: Out-State Tuition



6152

Mean Out-of-State Tuition for Public.

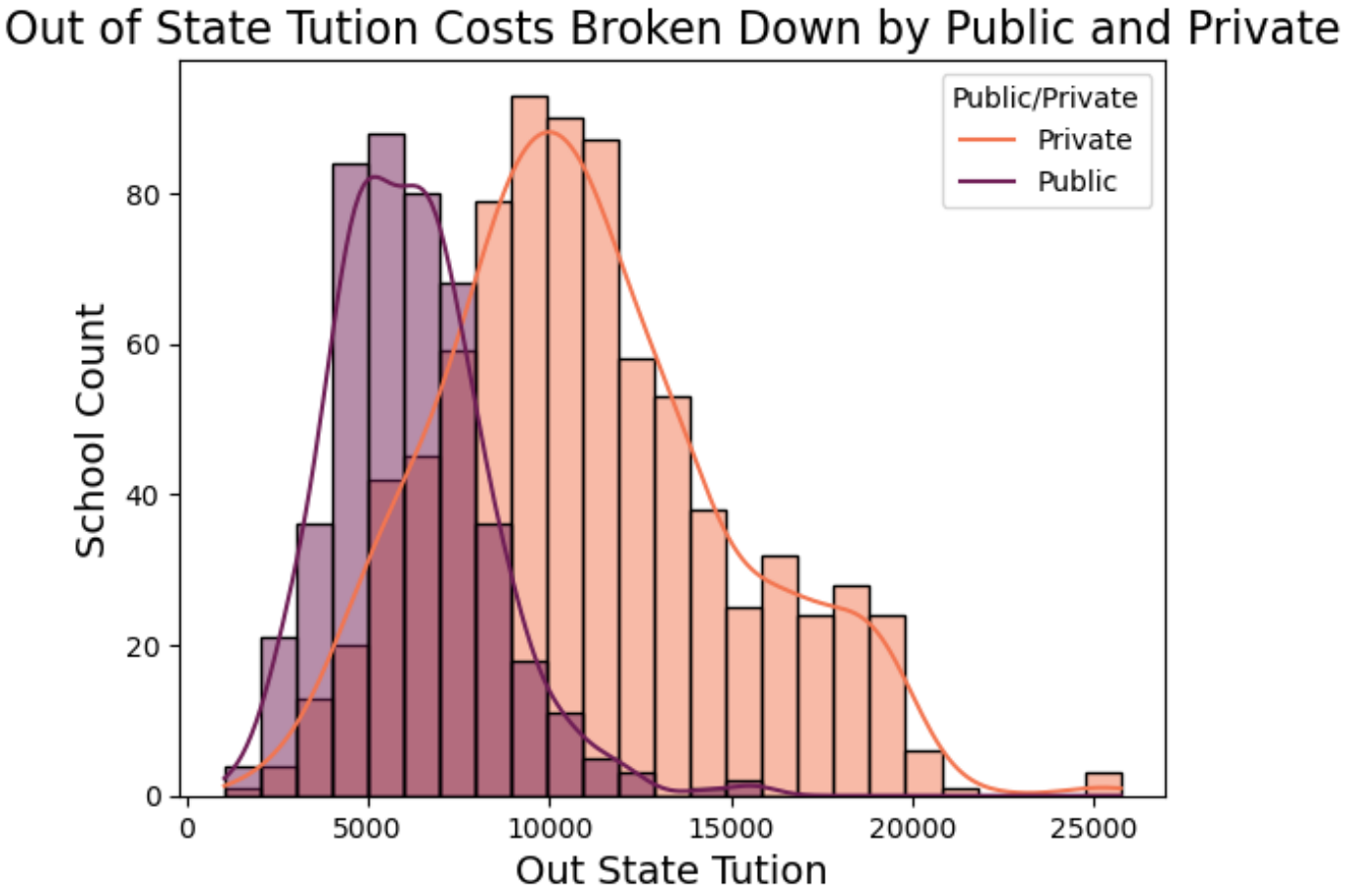
11008

Mean for Out-of-State Tuition for Private Schools
(2-tailed ttest p-value = 0.0, Statistic=58.6)

Lowest Cost for Out-of-State Colleges

College Name	out-of-state tuition
Grambling State University	1044.0
Adams State College	1672.0
Livingston University	1740.0
Brigham Young University - Hawaii	1780.0
University of Central Oklahoma	1928.0
Langston University	2064.0
Boise State University	2093.0
Texas Woman's University	2279.0
Northeastern State University	2280.0
Southern Arkansas University	2340.0

Cost: Out-State Tuition



6152

Mean Out-of-State Tuition for Public.

11008

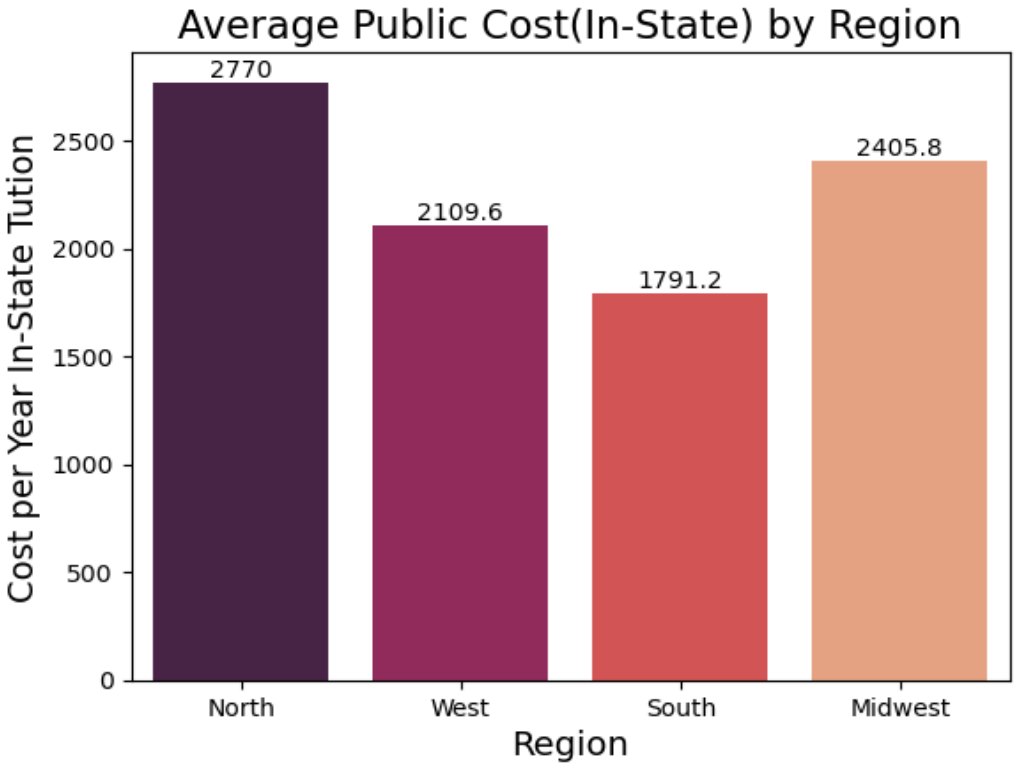
Mean for Out-of-State Tuition for Private Schools
(2-tailed ttest p-value = 0.0, Statistic=58.6)

Lowest Cost for Out-of-State Colleges

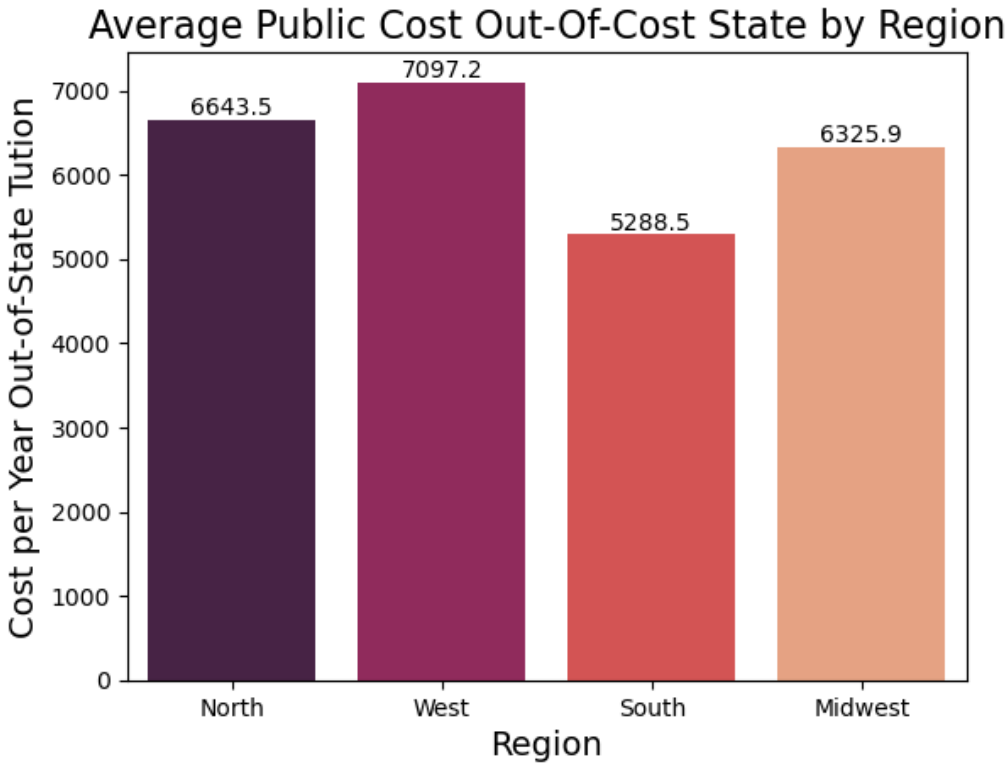
College Name	out-of-state tuition
Grambling State University	1044.0
Adams State College	1672.0
Livingston University	1740.0
Brigham Young University - Hawaii	1780.0
University of Central Oklahoma	1928.0
Langston University	2064.0
Boise State University	2093.0
Texas Woman's University	2279.0
Northeastern State University	2280.0
Southern Arkansas University	2340.0

Public to Public

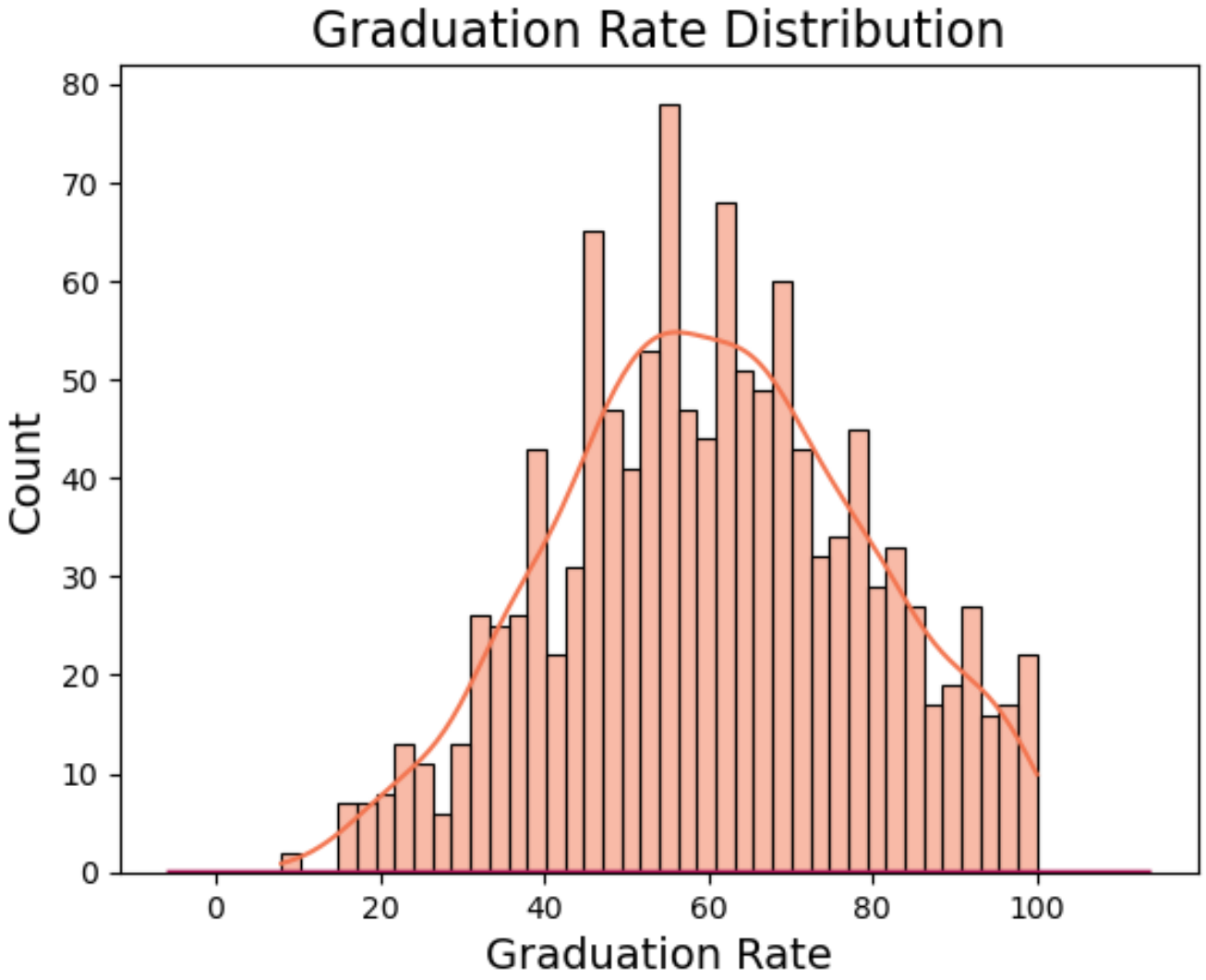
Regions Charge Differently for In and Out of State



Significant Differences Between the Average Cost In State For Different Regions (ANOVA fvalue=22.97054536966959, pvalue=7.732162944526283e-14)



Significant Differences Between the Average Cost In State For Different Regions (AMOVA fvalue=18.14586886782207, pvalue=3.897390493224823e-11)



mean	**60.343023**
std	**18.822515**
min	**8.000000**
Max	**100**
25%	**47.000000**
50%	**60.000000**
75%	**74.000000**

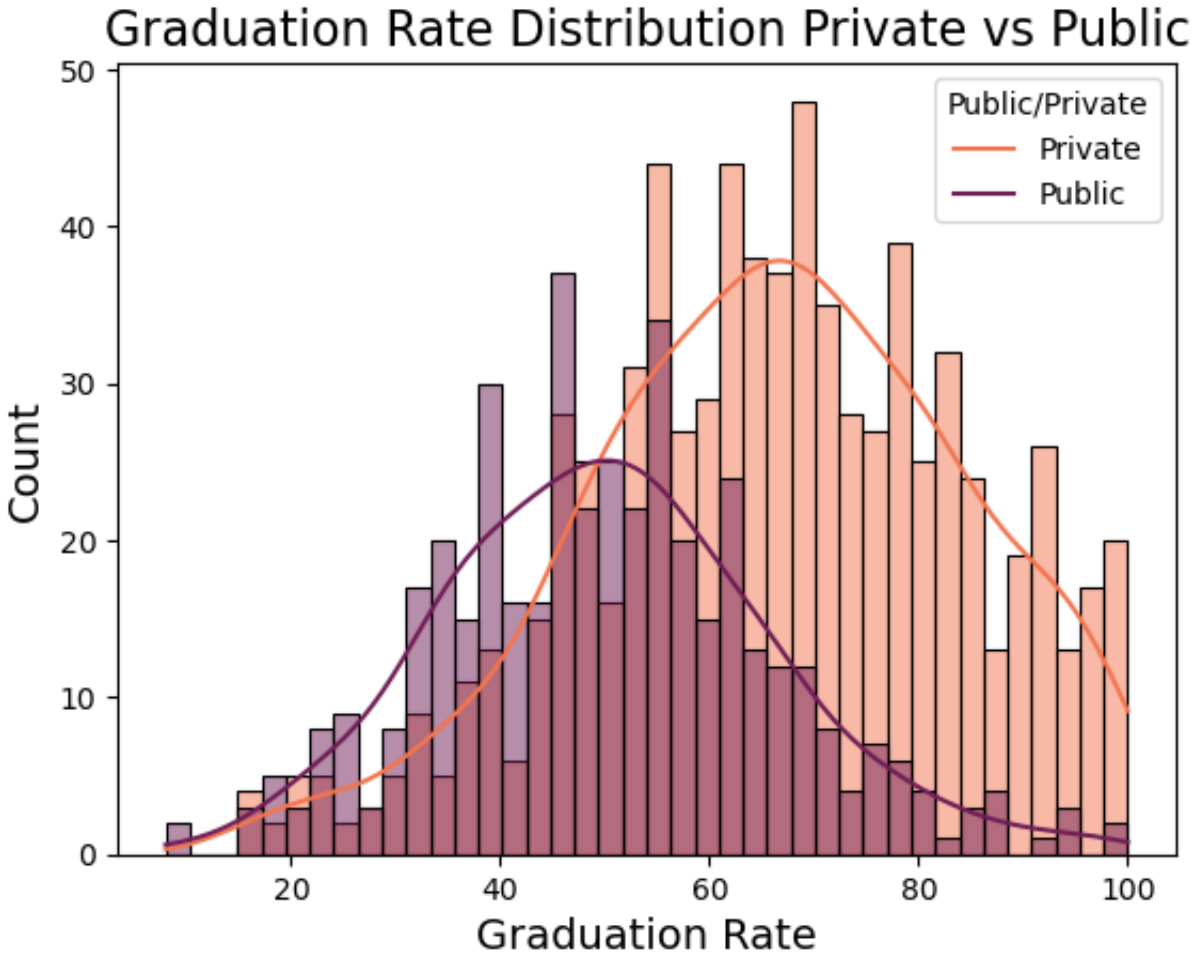
THE BEST AT 100% GRADUATION RATE

- Harvey Mudd College
- Santa Clara University
- Amherst College
- Harvard University
- Lindenwood College
- Missouri Southern State College
- Siena College College of Mount St. Joseph Grove City College
- University of Richmond Goddard College
- Heritage College

THE WORST AT 8% GRADUATION RATE

UNIVERSITY OF HOUSTON DOWNTOWN

Graduation Rates: Private Schools Do Better



Private Average Graduation Rate

66.07

Public Average Graduation Rate

50.18

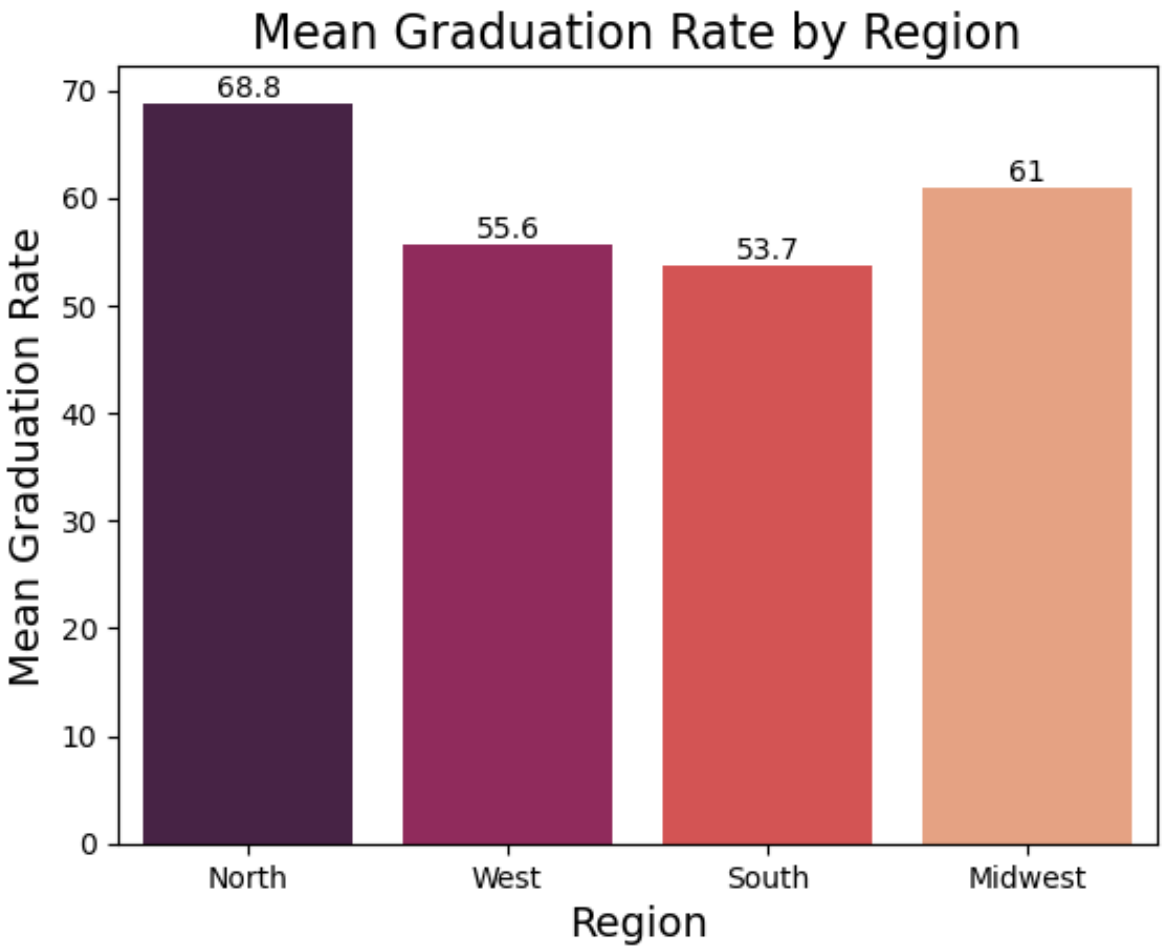
2-tailed t-test (p-value: 0.0, Statistic = 15.03*)

Statistically Significant Difference

We must reject the null hypnosis that private and public school graduation rates are the same.

* p-values are exceedingly small

Graduation Rates: Regional Differences Exist



Different US News and World Report Best College Regions 2016 have different graduation rates

(ANOVA p-value= 0.0, fvalue=47.13*)

The North Region does significantly better then even the next highest Midwest Region

(t-test p-value=0.0, statistic=5.8 *)

The West and South Region are not significantly better than each other

(t-test p-value=0.29, statistic=1.05 *)

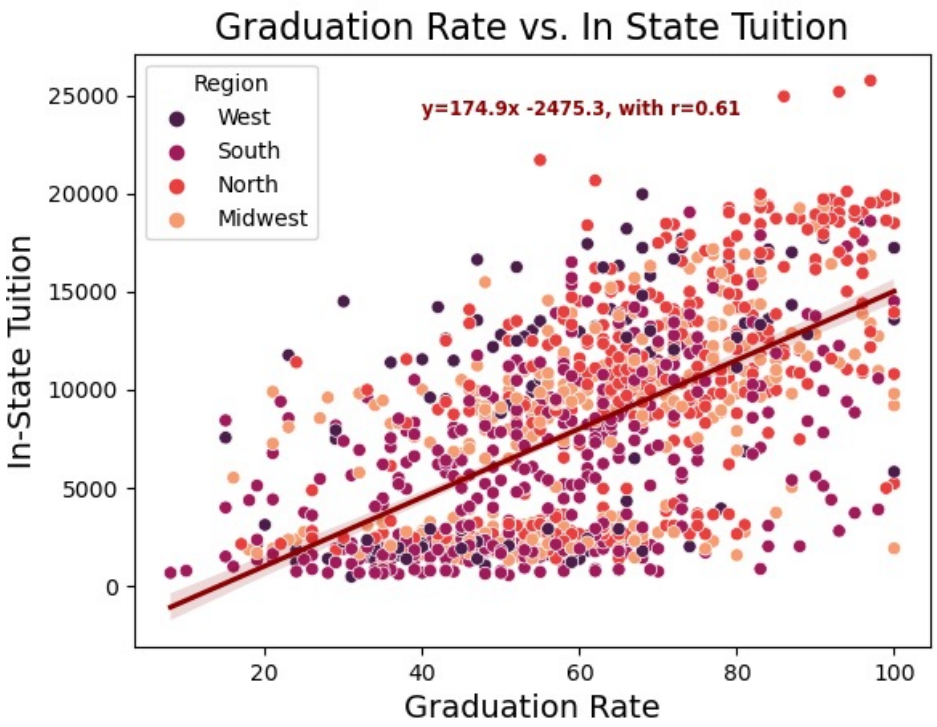
The Midwest does better than the West and South

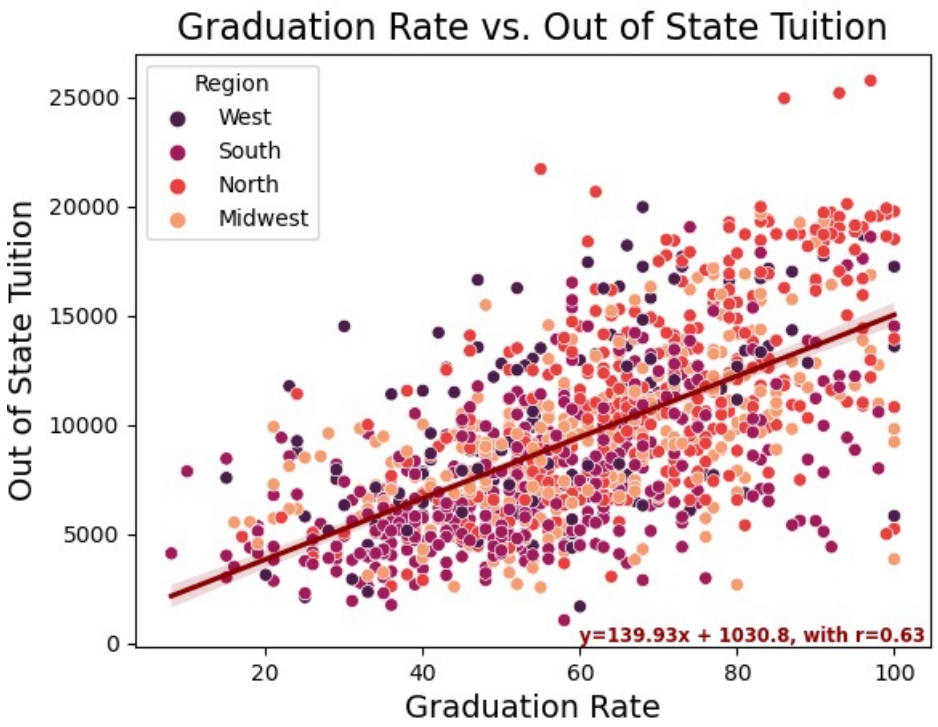
(t-test p-value=0.002, statistic=3.02 *)

If you must graduate, attend North Region schools.

* p-values are exceedingly small

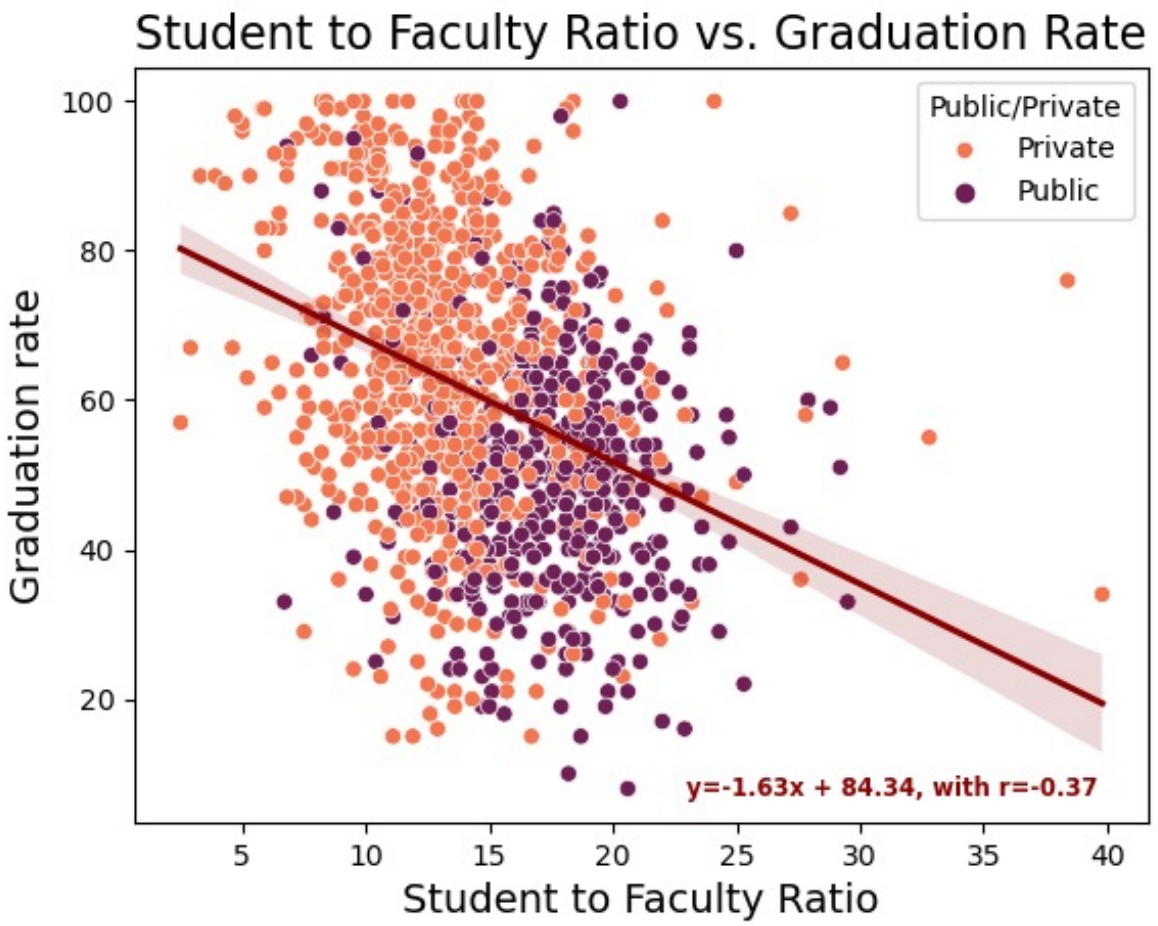
Graduation Rates: Tuition Up, Graduation Rate Up





Spend More to get more? Graduation rates suggest yes.

Graduation vs. Faculty to Student Ratio



A lower student-to-faculty ratio does help with higher graduation rates, but it is only moderate.
(Correlation: -0.34)

Most Schools have a ratio of student to faculty ratio between 14-15
(mean: 14.9, median: 14.3)

Private Schools have statistically significantly smaller Student to Faculty Ratios.
(2 tailed t-test p-value = 0.0*., Public student to faculty ratio = 17.4, Private student to faculty ratio = 13.1)

Next Steps In Analysis

- Separate Public and Private fully and analyze each separately. In many of the graphs and analyses, the two categories acted like they came from different populations.
- Much of the outlier (but valuable data) problem will be helped with that split.
- More Geopandas with Regional Tests and Regional Visualizations.
- Clean Github and Clean-Up Jupyter notebooks
- Would like to do an Unsupervised K-Means as I suspect there is a public-private and small-medium-large clustering in the data.
- Would like to test the PCA reductions against Random Forest.

PCA Dimension Reduction

The step before machine learning.

Preparation Work

Dropped Columns that were too unique

- Address1, Address2, Latitude, Longitude, Zip Code, County. Dropping Region as it was not part of the original data.
- Reduction from over 3000 columns hot-encoded to just 68

Hot-Encoding data VIA Pandas pd.get_dummies.

Imputing missing data via Median Strategy.

Scaling via Standard Scalar Modules

```
college_df = pd.read_excel(io='data/geocode_college_with_address.xlsx', index_col=0)
college_df.value_counts()

# Droping columns high uniques values.
college_df.drop(['Address1', 'Address2', 'Latitude', 'Longitude', 'Zip Code', 'Region', 'County'], axis=1, inplace=True)
print(college_df.shape)
```

✓ 0.3s

```
(1302, 19)
```

```
### Hot encoding
import pandas as pd

# One-Hot Encoding with Drop-First
df_encoded = pd.get_dummies(college_df, drop_first=True)
print(df_encoded.shape)
```

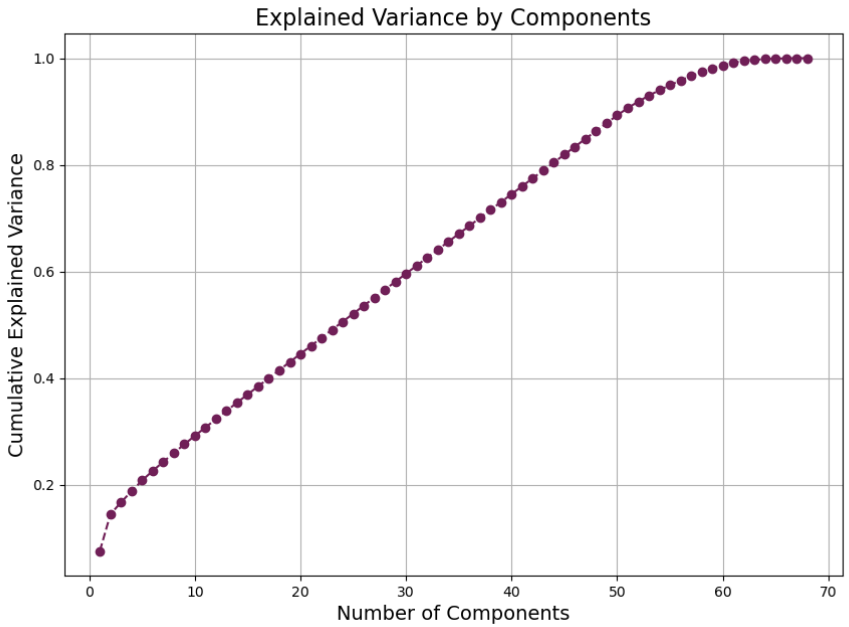
✓ 0.0s

```
(1302, 68)
```

```
# IMPUTE data, Median Imputation
college_imputed_df= df_encoded.apply(lambda x: \
    x.fillna(x.median()), axis=0)
print(college_imputed_df.shape)

```

✓ 0.0s

Scree Analysis for PCA

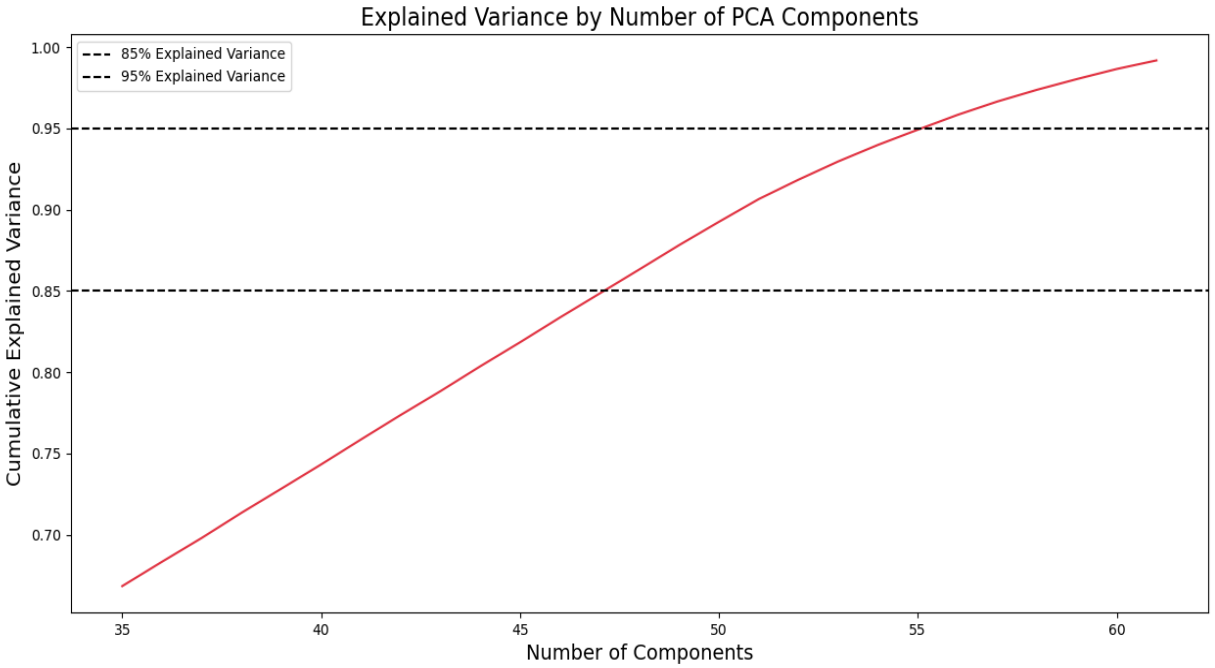


Scree graph stables around 58, checking for the best coverage at the lowest level of n_components.

```
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split

pca_df = pd.DataFrame()
pca_df['Covered Variance'] = ""
pca_df['n_components'] = ""

totalvariance = []
n_components = []
for i in range(35, 62):
    X_scaled = scaler.fit_transform(college_imputed_df)
    pca = PCA(n_components=i)
    ccinfo_pca = pca.fit_transform(X_scaled)
    variance = pca.explained_variance_ratio_.sum()
    rowcount = i
    totalvariance.append(variance)
    n_components.append(i)

pca_df['Covered Variance'] = totalvariance
pca_df['n_components'] = n_components

```

Based on Scree, Ran PCA with n-Components between 35-62 to capture 75% -> 95% of the variability

Outcome of Multiple PCA Dimension Reduction



55 PCA components to have 95% of the summed variance covered.

47 PCA components to have 85% of summed variance covered.