Probability and Statistics (IE 6200) - Sec 1 - Final Project Report

Gaurav Raj Chattarki, Jiali Chen and Saumya Vora

Objective:

To find the correlation between published university's overall ranking with its associated factors which have been extracted from government published websites with historical data of 9 years. Calculating each factor's ranking separately and comparing it with the general ranking which is obtained from leading american reporter, US News. The obtained results are then used for predicting the selected university's future rank.

Data Description:

The raw data is obtained from the official american government college scorecard website where we've identified 9 main factors in the complete data set that can be utilized to run our analysis to find the university's rank.

Below we have the table giving the detailed description of the data and how the data has been modified to suit the analysis in R Studio.

Data Overview	**Title**	**CollegeScoreCard Raw Data.zip**
	Creator.Curator	**U.S. DEPARTMENT OF EDUCATION**
	Identifier	**https://collegescorecard.ed.gov/data/ - The identifier here is the .ed.gov top level domain (TLD)**
	Method	**These data are provided through federal reporting from institutions, data on federal financial aid, and tax information. These data provide insights into the performance of institutions that receive federal financial aid dollars, and the outcomes of the students of those institutions.**
	Processing	**The data has been processed in excel, and due to the excessive data, the data set has been filtered and then additional columns have been added against each factor for factor ranking purposes. This is then appended with the data set from US news for verification and validation purposes.**
	Source	**https://collegescorecard.ed.gov/data/ https://publicuniversityhonors.com - US News Historical Data**

Assumption:

The project sticks to real time, authentic government approved data and the only educated assumption we've made is in selecting the factors. Even though it is considered an assumption, we've made sure that the selected factors make sense by weighing out each contributor through web and journal research.

Approach:

Once we gathered the data we laid a clear plan as to what we want to convey with our findings and first, we started with the basic Mean, Standard Deviation and Range for specific variables and then we plotted the scatter plot to find which variable plays a major role in deciding the rank, and after cross validation of correlations, as a result, 'admission rate' became our target variable. In addition to covering concepts learned in the lecture/lab, we utilized the line graph to find deeper insights for a few selected universities and we found that, if not in all but in most universities, the admission rate and the no. of undergraduate students was found to be really close to the overall median, this result is shockingly mirrored in our factor ranking analysis which proves as evidence to our objective of validating the current trend of ranks as well as predicting them.To conclude, we use factor ranking, we first choose five universities and append factor ranks to each respective factor to run our analysis. By sorting each factor in ascending order and redefining the factor definition, we seek to find similar results obtained by descriptive analysis.

Analysis:

Based on the nine year general ranking for each of the one hundred universities, we find mean, standard deviation, and range for their nine year ranking, in order to find university which rank changed a lot (>30) over nine years as well as stay at same level over decades based on standard deviation; we divided top 100 universities into two groups which is 1-50 ranking and 50-100 ranking and choose both case in both groups to avoid errors.

We defined a function to find the correlation between each factor and general rank
as the 'contribution estimator' function,

$$\frac{|General\ Rank - Factor\ Rank|}{100}$$

The above function is used to estimate how each factor contributes to the general rank. Basically, we find the difference value between general rank and the factor rank of the respective universities, then we perform normalization by dividing the difference by 100 to get the contribution estimator value.

The resulted value will lie somewhere between 0 and 1. If the value is very close to 0, that means the factor contributes a lot to the general rank; however, if the value is close to 1, that means the factor does not have much contribution to the general rank.

Then we apply Visualization based on each university and for each university, we make two graphs.

The first one is line chart with x-axis as year from 2007-2016 and y-axis as ranks over the given year. We want to fit all factors rank and general rank into one graph for each university to see the correlation and help future analyze.
The other graph will be “factor matters level” graph which based on the function we made. We will make the summation of contribution estimator over 9 years for each factor based on university. So the range change from 0 to 9, and small result (from 0-3) means the factor contributes a lot to general rank and large resutle(from 6-9) means the factor does not have much contribution to the general rank over 9 years.

For Hypothesis testing, we’re going to test if the mean of admission rate for Massachusetts is lower than the mean of national top 100 universities admission rate.

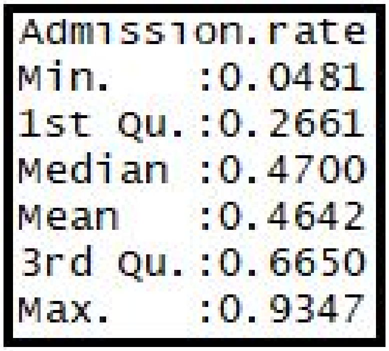
Hypothesis:
1. H0: $x<0.3456$
   H1: $x>0.3456$

Population: the admission rate over nine years for all universities
Sample: the admission rate for universities from Massachusetts over nine years
Since n is greater than 30, we will apply z test in this case

Target variable (Admission Rate) analysis:

Standard Deviation	0.2373
Variance	0.0563
Coefficient of variation	195.5771
Interquartile range	0.398975
Skewness	-0.0271

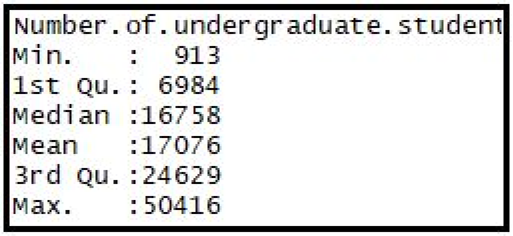
Kurtosis	-1.137078



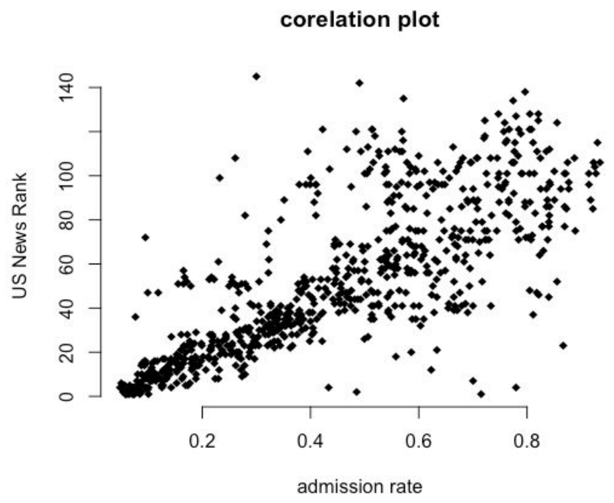
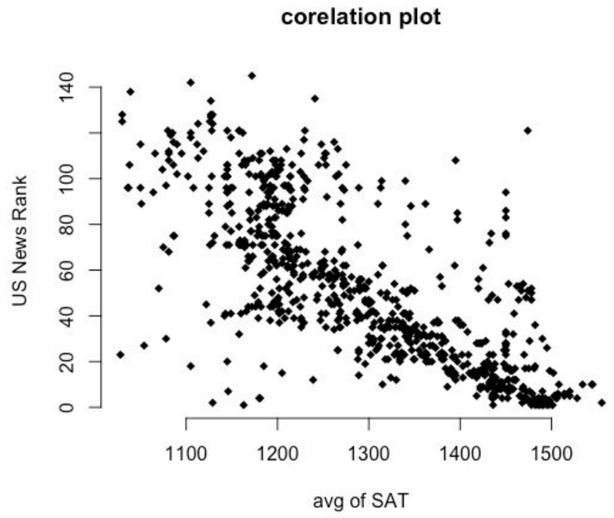
```
Admission.rate
Min.   :0.0481
1st Qu.:0.2661
Median :0.4700
Mean   :0.4642
3rd Qu.:0.6650
Max.   :0.9347
```

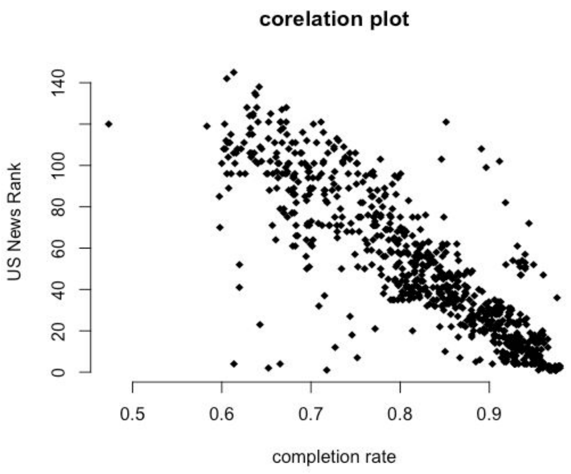
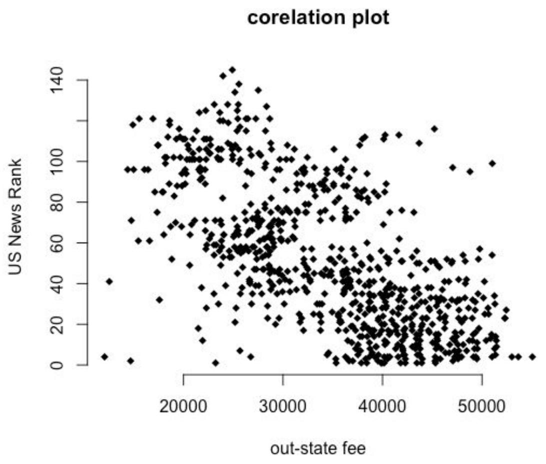
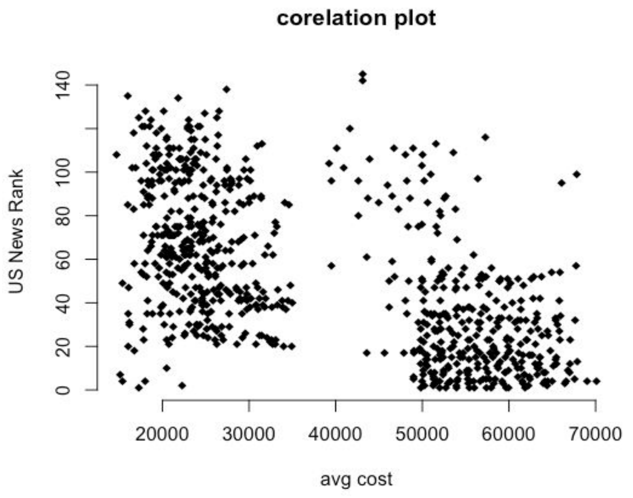
Target variable (No. of undergraduate students) analysis:

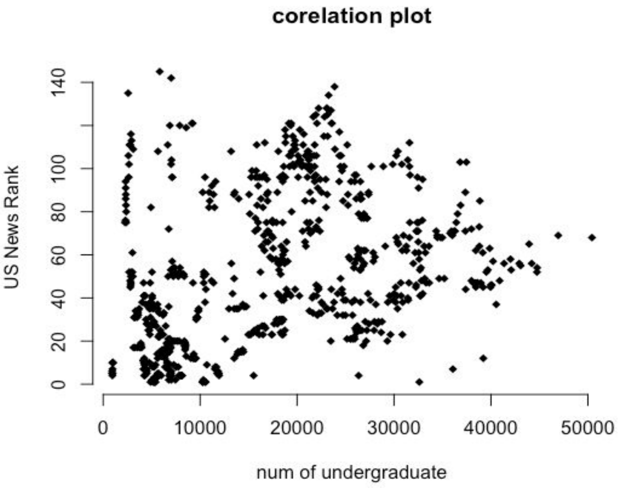
Standard Deviation	10560.7023
Variance	111528435
Coefficient of variation	161.6901
Interquartile range	17644.75
Skewness	0.4564
Kurtosis	-0.6915

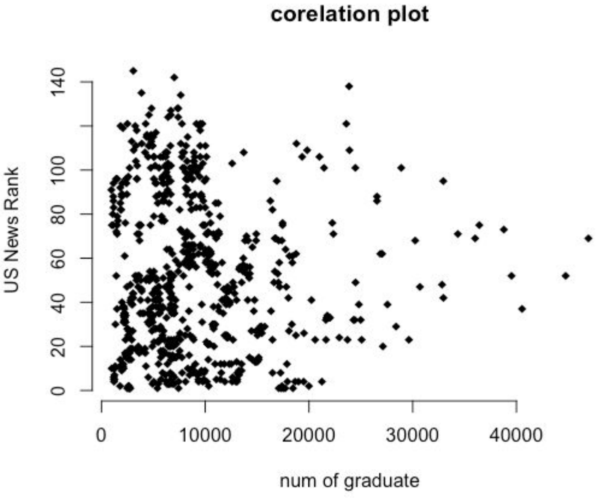


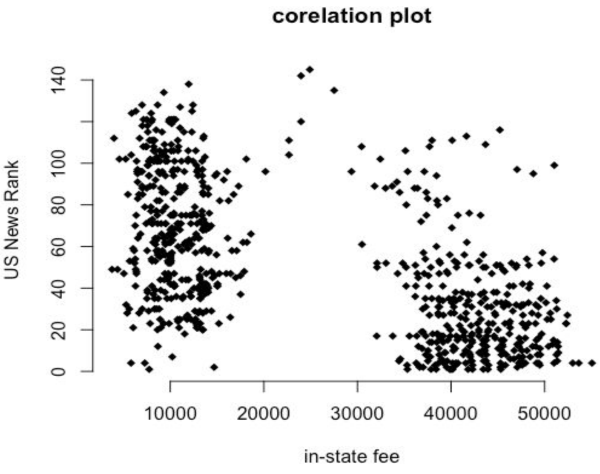
```
Number.of.undergraduate.student
Min.   :  913
1st Qu.: 6984
Median :16758
Mean   :17076
3rd Qu.:24629
Max.   :50416
```

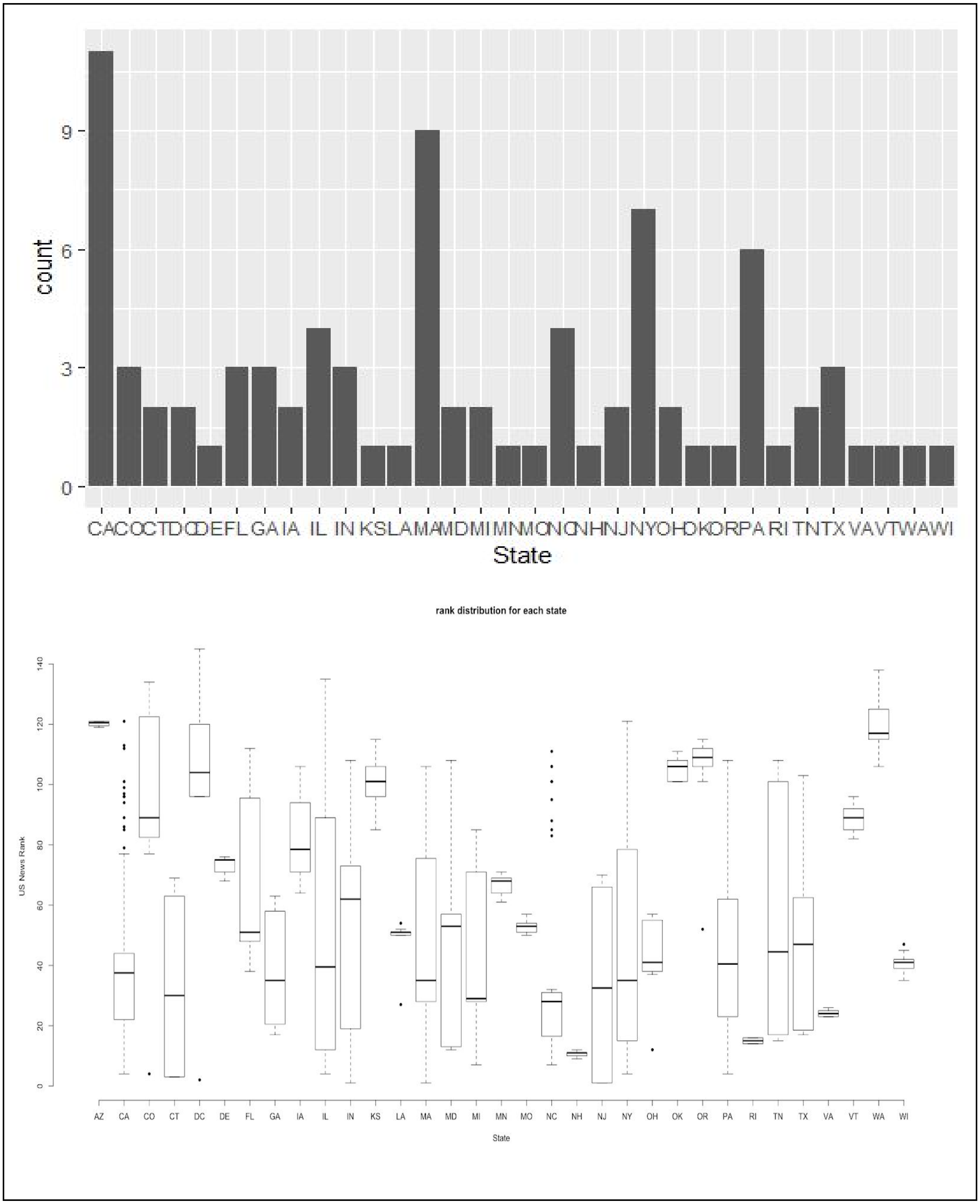
Graph/Plot or Observations	Conclusion
 	**Positive correlation** It refers that with a lower national rank, universities' admission rate is generally higher.
 	**Negative correlation** For the average SAT scores and completion rate, they both have a negative correlation with US News Rank, and out-state-tuition fee also has a weak negative correlation with US News Rank. In our case negative correlation means if a university has a higher rank, they usually has higher SAT score requirement, more expensive tuition fee for out-state student and higher

 	completion rate within required time.
	Points show no significant clustering, so there probably has no correlation.


corelation plot
US News Rank
0
20
40
60
80
100
140
0
10000
20000
30000
40000
50000
num of undergraduate


corelation plot
US News Rank
0
20
40
60
80
100
140
0
10000
20000
30000
40000
num of graduate

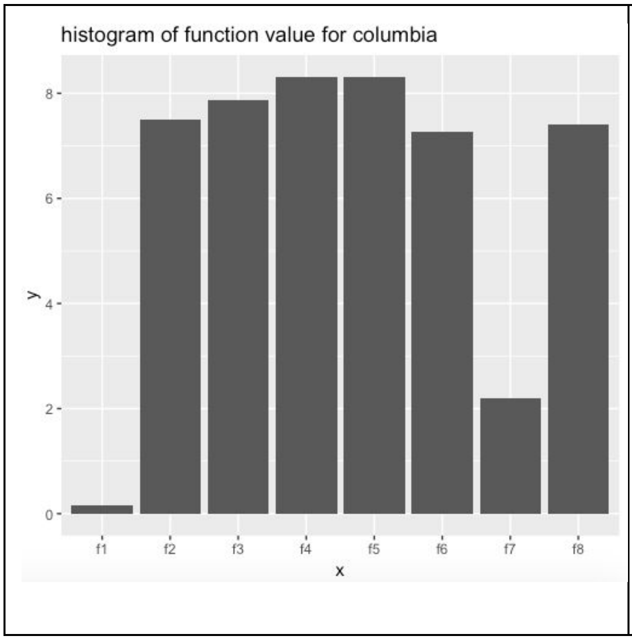

corelation plot
US News Rank
0
20
40
60
80
100
140
10000
20000
30000
40000
50000
in-state fee


count
0
3
6
9
CA CO CT DC DE FL GA IA IL IN KS LA MA MD MI MN MO NC NH NJ NY OH OK OR PA RI TN TX VA VT WA WI
State
rank distribution for each state
US News Rank
0
20
40
60
80
100
120
140
AZ CA CO CT DC DE FL GA IA IL IN KS LA MA MD MI MN MO NC NH NJ NY OH OK OR PA RI TN TX VA VT WA WI
State

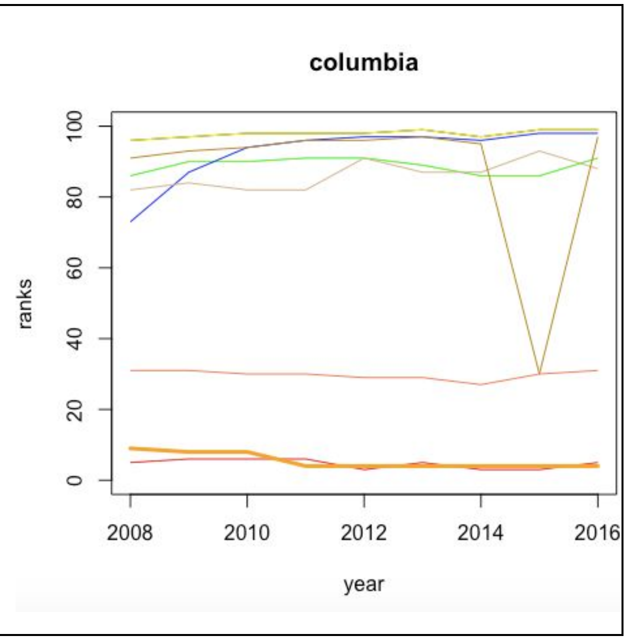
The first graph above shows the histogram defining how many universities from the top 100 list are located in each state.From the histogram we can find that California, massachusetts, New York, and Pennsylvania has more top universities than other state, and from the second graph, the box plot, we can tell that the mean of ranks for those university located around 40. As a result, California, Massachusetts, New York, and Pennsylvania are the states which have better learning atmosphere, interpersonal connection after graduation and employment environment compared with other state. Also we look up the difference between in state student tuition fee and out state student tuition fee, they are almost the same level. This may because the state is developed and they do not need use tuition reduction for local student to attract graduation student. On the other hand, for the state like Delaware, Kansas and Missouri, they have only one or two universities in top 100 ranks and with lower ranking, so they offer a tuition deduction for in state student to attract students, so their in-state-tuition fee is much lower for out-state tuition fee.

The graphs below are the factor ranking based hypothesis results. We've ranked the factors and are now finding the most effective factors that contribute to the general rank of a university.

Please referer the **Appendix** for color codes[1].

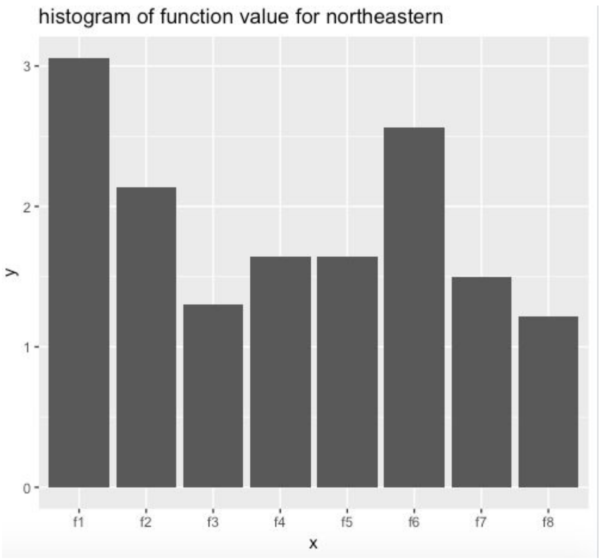


Columbia University Histogram

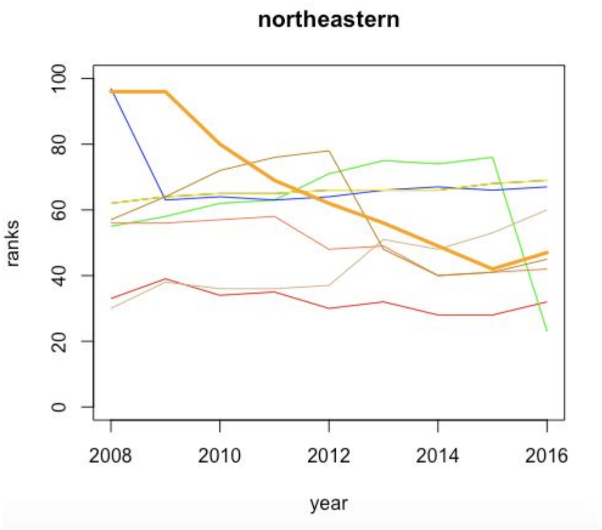


Factor ranking compared to General Rank

The above two graphs act as evidence to proving that, for Columbia University f1 (Admission Rate) and f7 (No. of Number of undergraduate student) play a key role in deciding the rank of the university, with admission rate playing almost a direct role.

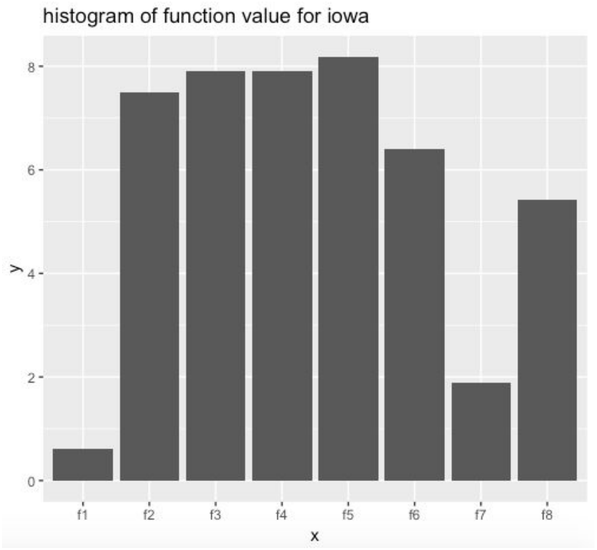


Northeastern University HIstogram (Factor)

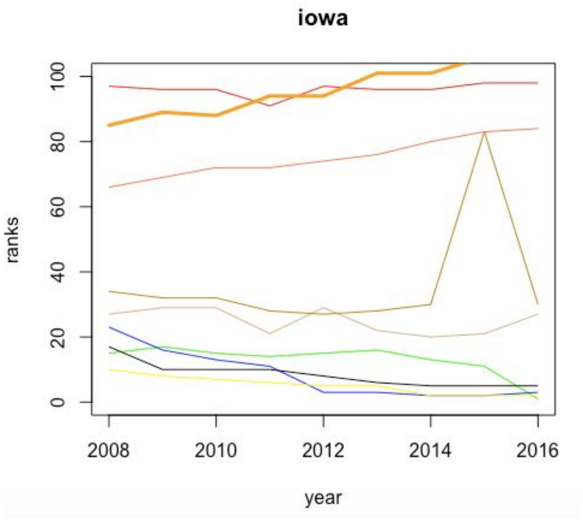


Factor ranking compared to General Rank

Northeastern University’s ranking has increased exponentially from 96th rank to 44th, an incredible 52 place jump. This is largely due to a consistent contribution from all the factors involved and as we can see from the Histogram, it’s an almost equal normalised value (mentioned in the data analysis section). Another shocking discovery is that admission rate has the least contribution to the growth of the university.



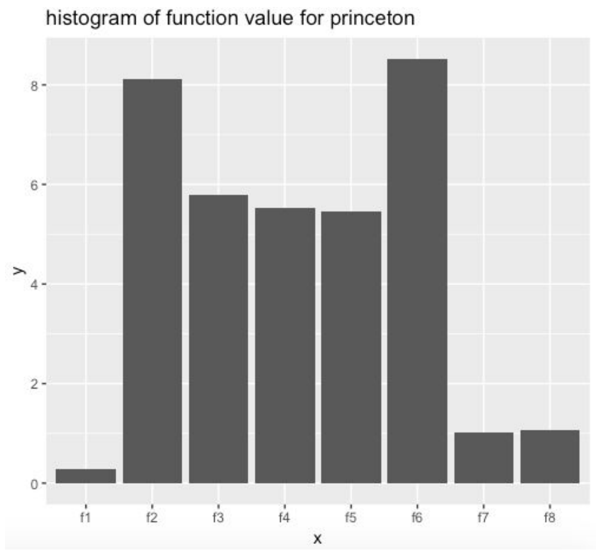
Iowa State University Histogram (Factor)

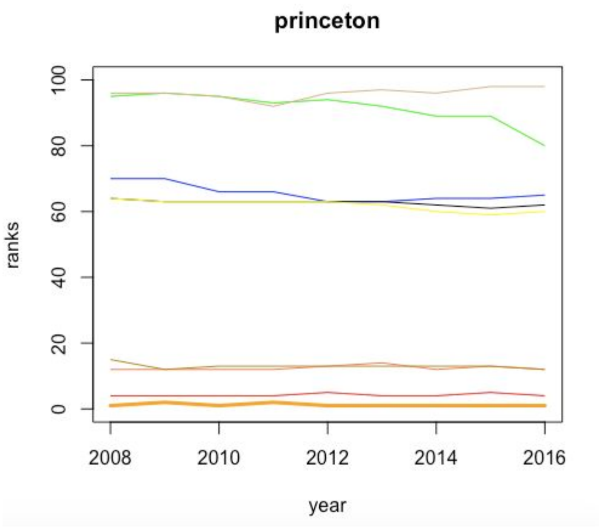


Factor ranking compared to General Rank

The above two graphs act as evidence to proving that, for Iowa State University f1 (Admission

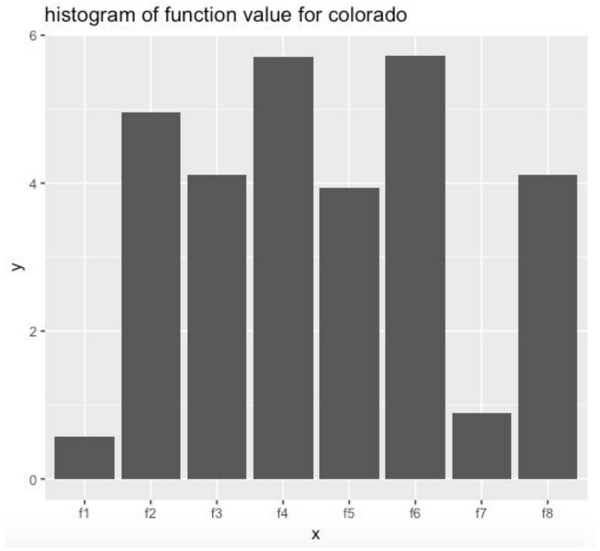
Rate) and f7 (No. of Number of undergraduate student) play a key role in deciding the rank of the university, with admission rate playing almost a direct role.

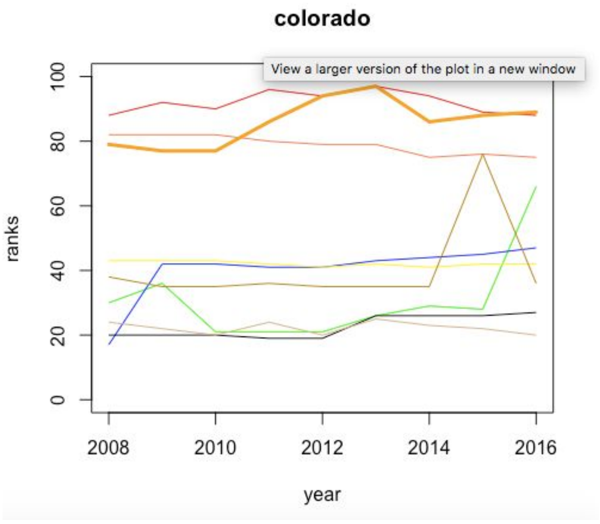




Princeton University Histogram (Factor) Factor ranking compared to General Rank

Princeton University is very consistent when it comes to Ranking and we've found that, in addition to Admission rate and No. of Undergraduate students, Number of graduate students also plays an important role. Because from both the histogram and cross validated by the line graph, we observe that these factors come really close to the general ranking.

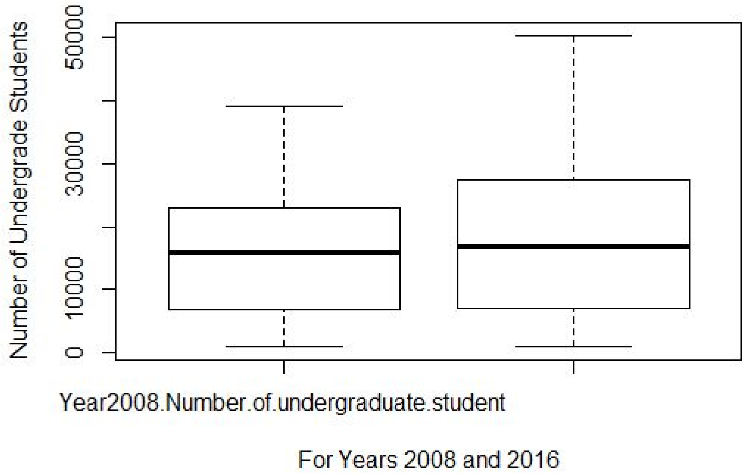


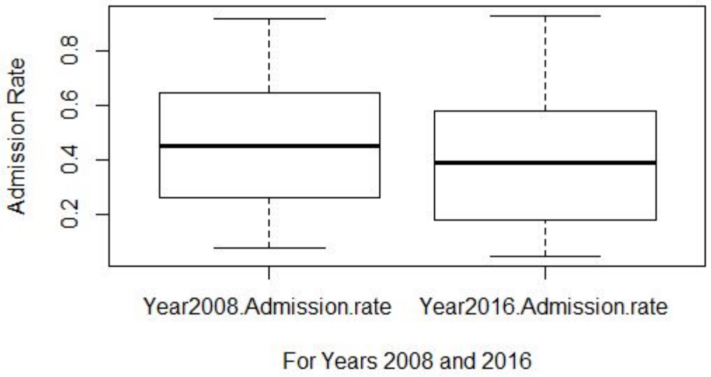


Colorado University Histogram (Factor) Factor ranking compared to General Rank

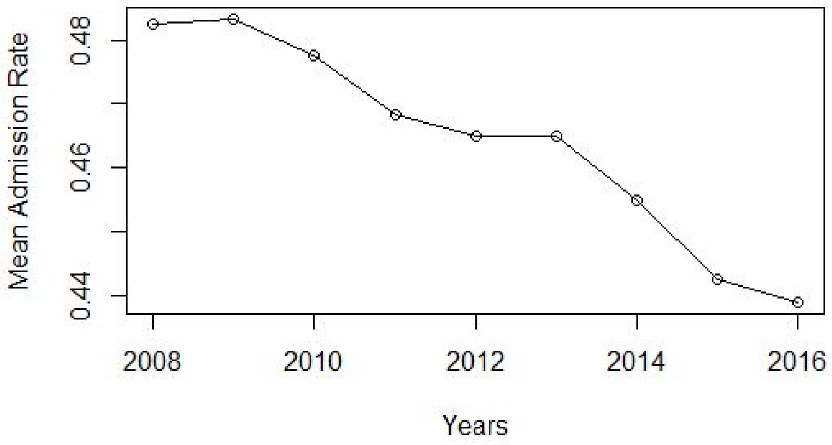
The above two graphs act as evidence to proving that, for Colorado University f1 (Admission

Rate) and f7 (No. of Number of undergraduate student) play a key role in deciding the rank of the university, with admission rate playing almost a direct role.

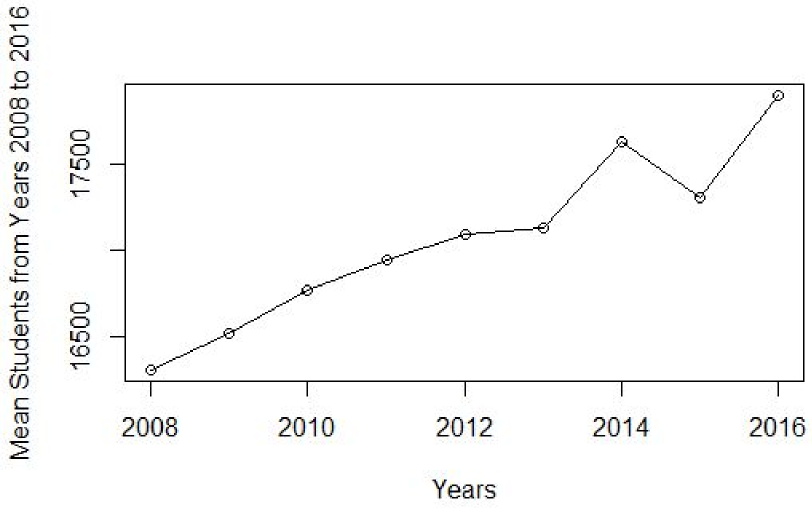


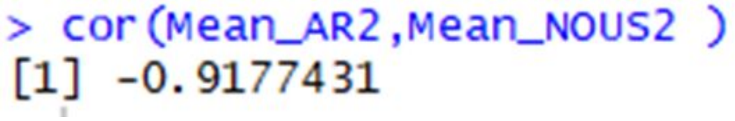


The above two graphs represent a box plot visualizations comparing the years 2008 and 2016, that how they've differed in terms of median shift indicating a negative correlation as mentioned in the table below.



From the above Line graph we observe that the mean admission rate has reduced drastically for universities overall over the past 10 years and when we observe the graph below, we see that the mean incoming students has a drastic increase. The interpretation from this observation is that since the pool of students increased, universities reduced their admission rate due to incompetence in accommodating the spike in applications. The picture below the graph below indicates the correlation, a high negative correlation.





```
> cor(Mean_AR2,Mean_NOUS2 )
[1] -0.9177431
```

Hypothesis test

We want to test if the mean of admission rate for Massachusetts is lower than the mean of national top 100 universities admission rate.

Hypothesis:
H0: x<0.3456
H1: x>0.3456

Population: the admission rate over nine years for all universities
Sample: the admission rate for universities from Massachusetts over nine years
Since n is greater than 30, we will apply z test in this case

Z score = -3.64

Does not fall into Rejection Region, fail to reject the hypothesis. So the mean of admission rate in Massachusetts is lower than the true mean, and with the obtained conclusion from previous project, Universities in Massachusetts has better ranks in the nation.

We will test if the mean of number of undergraduate students in Massachusetts's university is less than mean of national top 100 universities undergraduate student's mean.

Hypothesis:
H0: x<10120.23
H1: x>10120.23

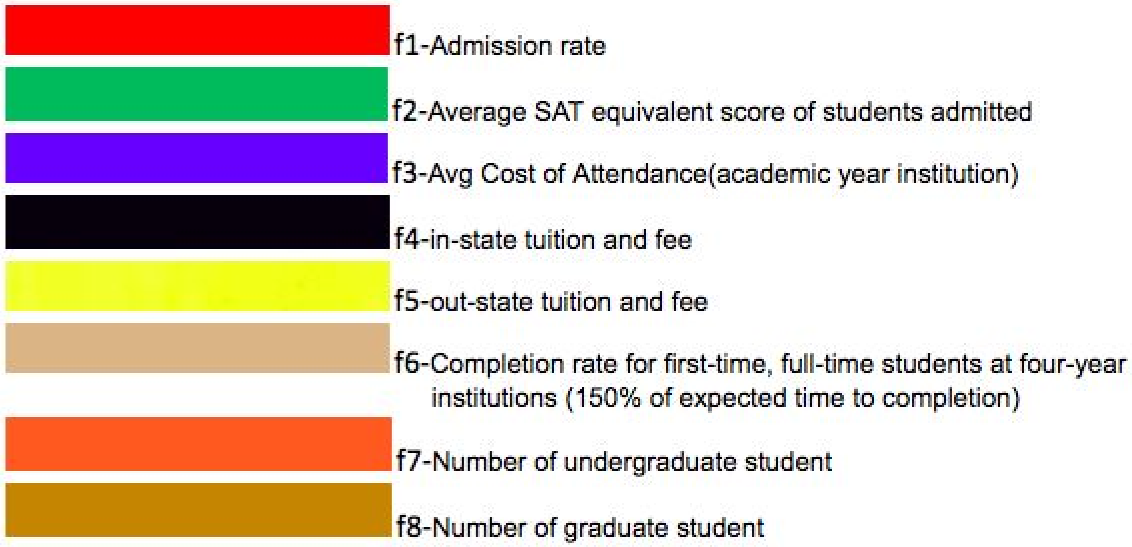
Population: number of undergraduate students over nine years for all universities
Sample: number of undergraduate students for universities from Massachusetts over nine years
Since n is greater than 30, apply z test again.

Z score = -5.76

Does not fall into rejection region, fail to reject the hypothesis. So the mean of number of undergraduate students in Massachusetts is lower than true mean, and followed by the obtained conclusion, larger number of undergraduate student university is trendy to have higher ranks.

Appendix

1. Referal 1- For color codes:



🟥 f1-Admission rate

🟩 f2-Average SAT equivalent score of students admitted

🟪 f3-Avg Cost of Attendance(academic year institution)

⬛ f4-in-state tuition and fee

🟨 f5-out-state tuition and fee

🟫 f6-Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion)

🟧 f7-Number of undergraduate student

🟫 f8-Number of graduate student