Final Project

1) What is this dataset and where did I get it?

I have decided to select College Scorecard Data for the final project. This is a very famous dataset and has been subjected to numerous studies and exploration. The data set can be found at the following link https://collegescorecard.ed.gov/data/ and has been made public by the US Department of Education to bring transparency and help students make informed choices.

2) Why this dataset

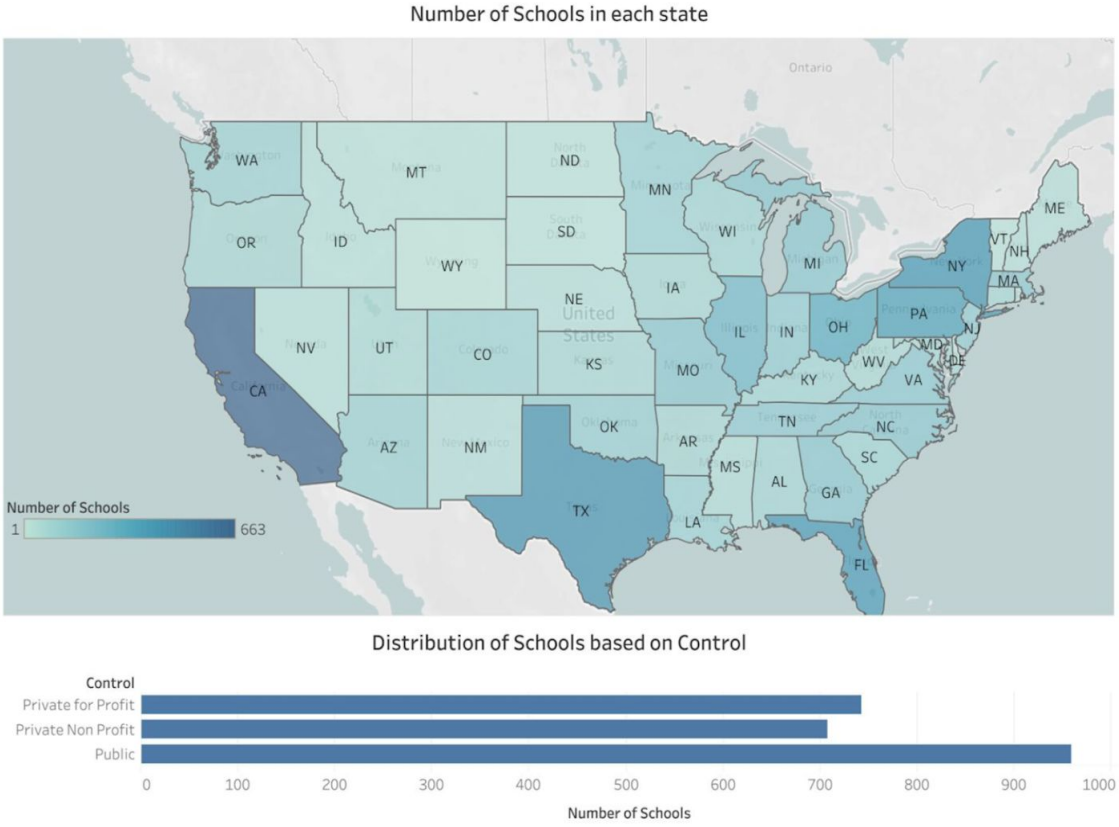
Attending college is a big financial burden for most of the students and everyone wishes to make an informed decision. Knowing the cost of attendance across different schools and states and the return on investment can help students choose wisely and be prepared for future implications. As part of the final project, I wish to explore this data to understand the financial implications associated with different schools across geographies and control structures.

3) What types of questions were you hoping to explore with this data?

I mainly want to compare the cost of attendance and returns across geographies and control structure.

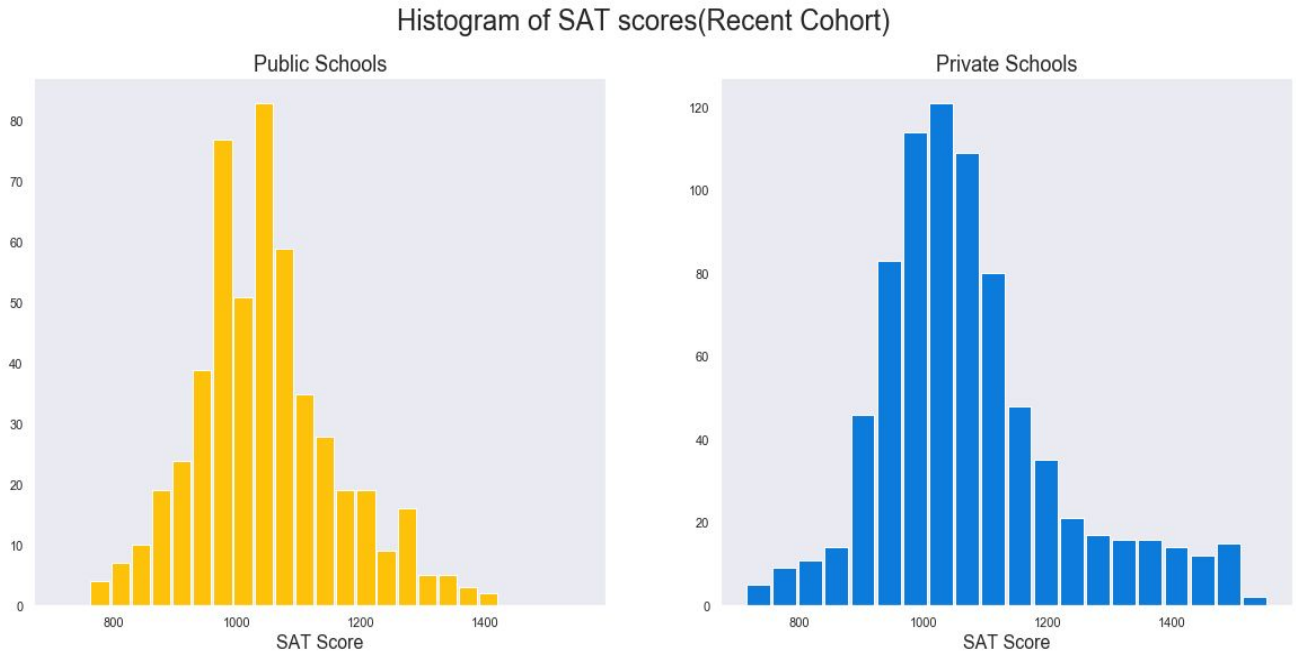
Visualizations

Q) Number of schools across states and their distribution based on control

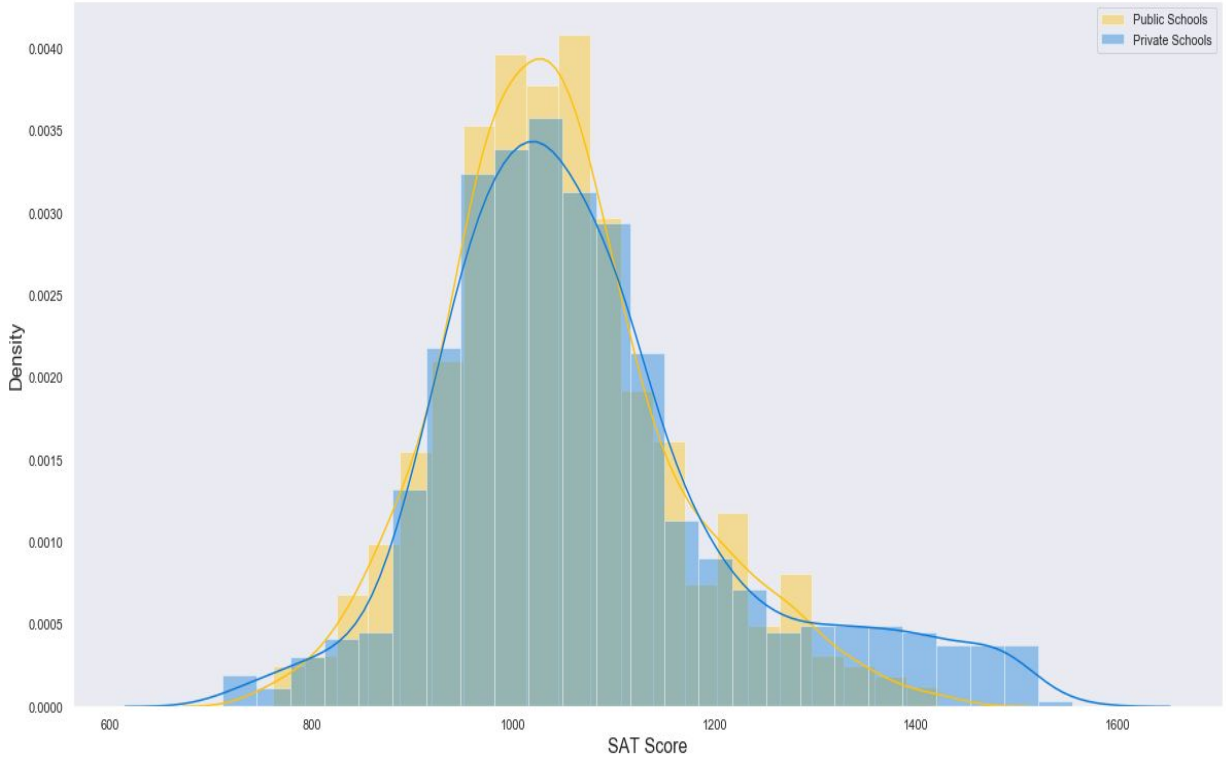


Following is an interactive Tableau dashboard showing the number of schools across each state in US using Choropleth map. The color gradient has been varied based on the number of schools with darker shade corresponds to a larger number of schools in a given state. The color mapping to the number of schools can be found on the bottom left corner of the Choropleth map. The map also shows the state name abbreviation for each state for easy identification. We can see that California(CA) has the maximum number of schools followed by Texas(TX) and New York(NY). These names do not surprise us as these states are renowned for their world class schools and educational infrastructures. Vermont(VT) has the lowest number of schools and is a neighbour to NY. The bar graph below Choropleth map shows the distributions of the schools based on control structure. This information is very useful as Public schools are much more economical to attend compared to Private schools. It also gives us an insight into government spending on schools per state. On an overall level, we can see around 900 public schools and around 700 private for profit and non profit school. Cumulatively, private school numbers are higher compared to public schools. Note that the above dashboard is interactive and give us state wise school distributions based on control if we select a particular state on the plot. I would encourage you to try it on the following link.

Q) Is it easier to get into public school compared to private?



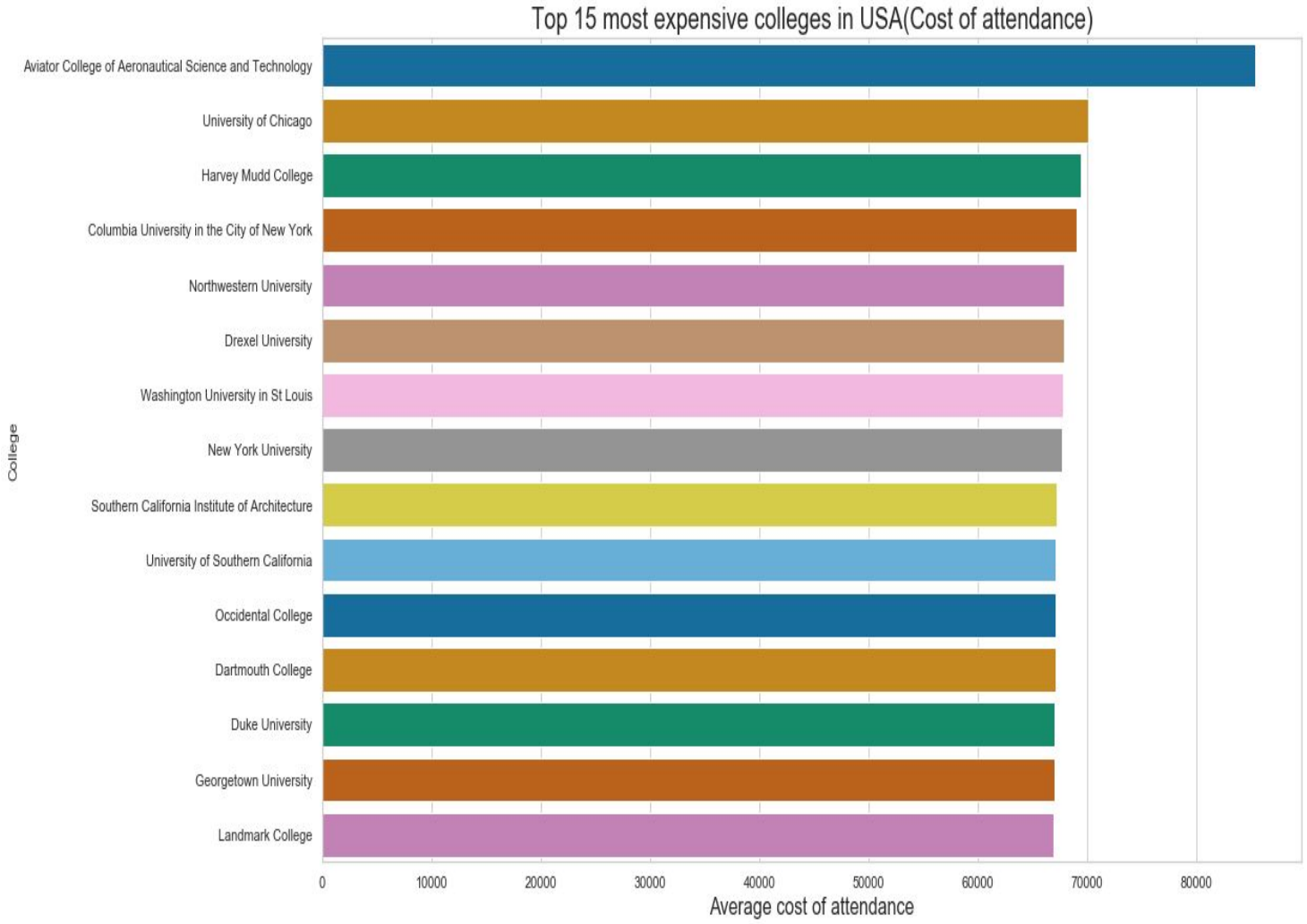
Following histogram plot is used to visualize the distribution of total SAT score of students admitted to public and private school. On a first glance, we donot see a major difference in the distribution of the SAT score. A density plot might be able to reveal better insight.



Following is a density plot of the SAT score of Private vs Public school. The distribution looks slightly more skewed to the right for the Private School suggesting, students with higher SAT score has a high chance of going to Private school compared to Public school.

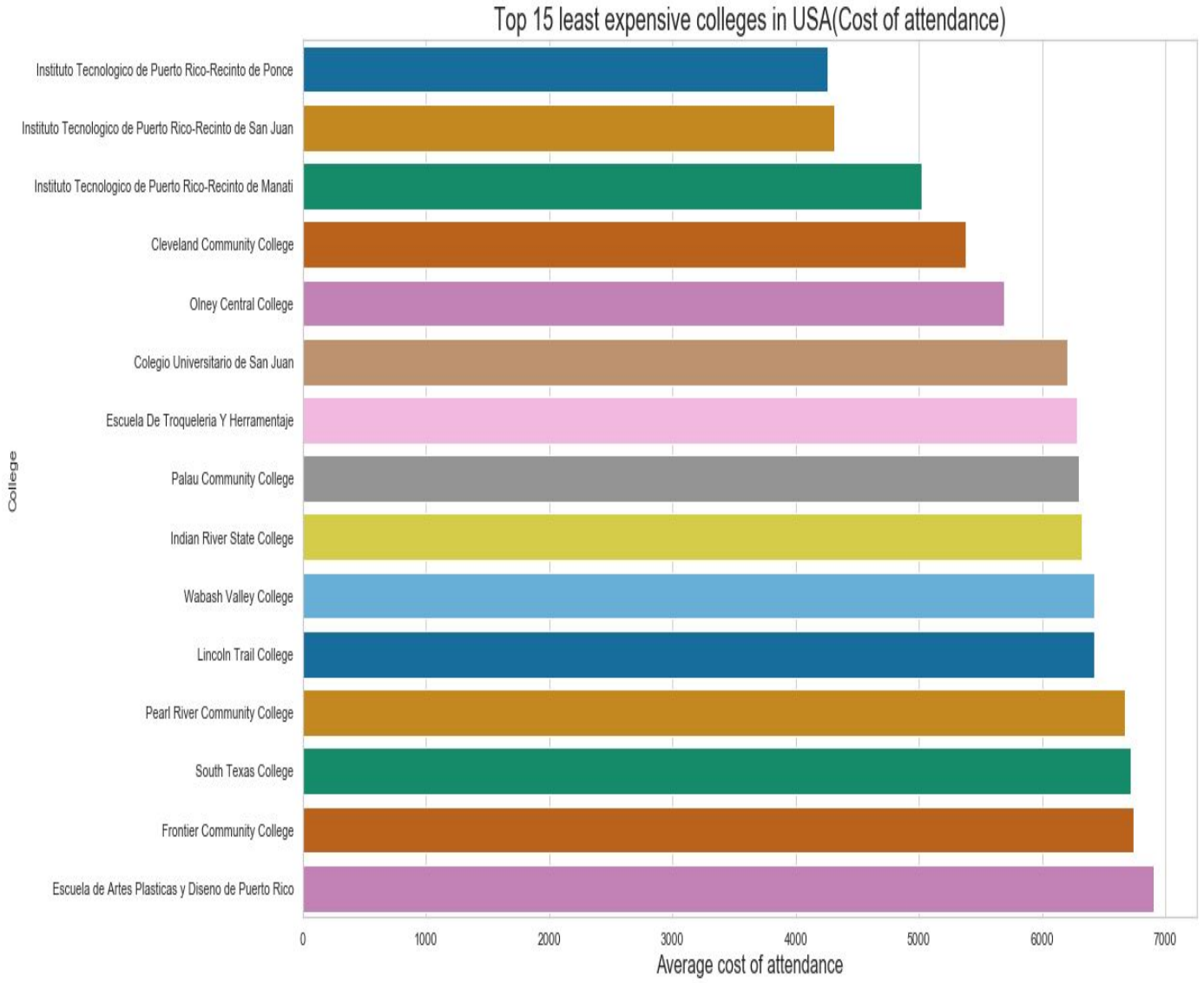
Although the means look quite similar to the median SAT score for Private school is slightly higher compared to Public school.

Q) Which are the top 15 most expensive colleges based on the cost of attendance?



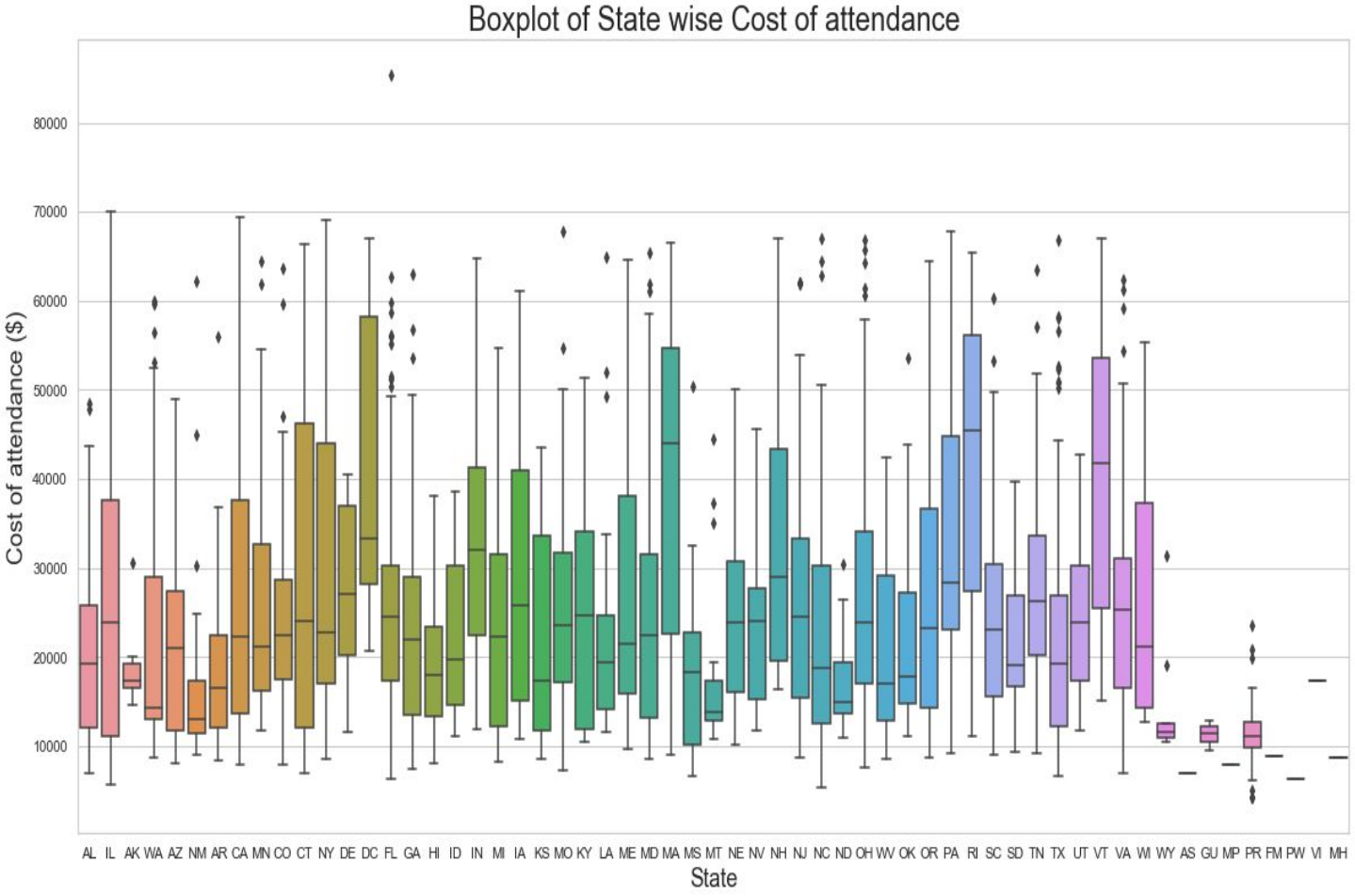
Following bar plot lists the top 15 most expensive colleges in USA. We can see that many of the private and ivy league colleges are a part of this. Also, states like New York and California are known to have high cost of living and also contribute to the high cost of attendance. We also see schools teaching special subjects like aeronautics and aviation tops the list as the course requires expensive equipment and infrastructures.

Q) Which are the top 15 least expensive colleges based on the cost of attendance?



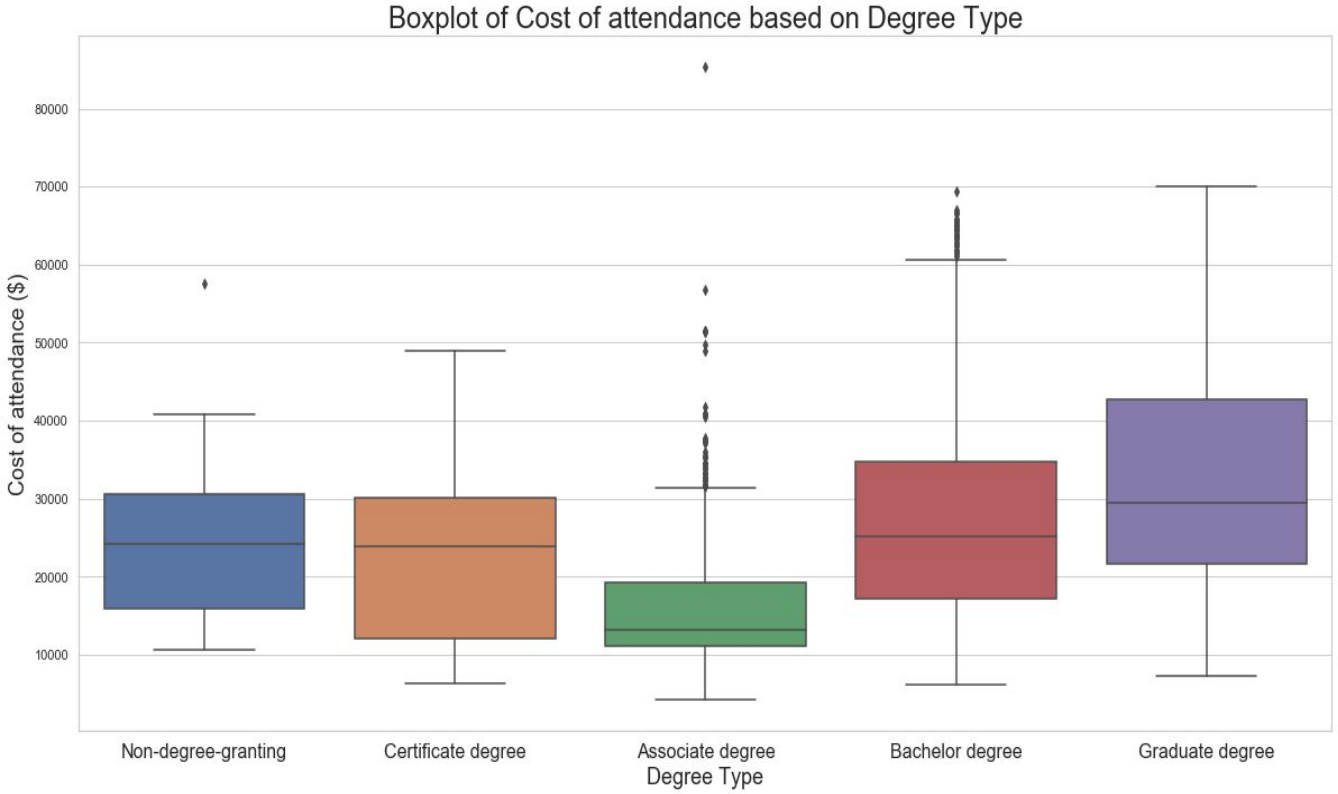
The bar plot above shows the top 15 least expensive colleges in USA. We can see that Puerto Rico has a very less cost of attendance. Also, most of these colleges are Public schools or Community colleges which gives us insight that public schools are cheaper compared to private schools.

Q) What is the state-wise distribution of the cost of attendance?



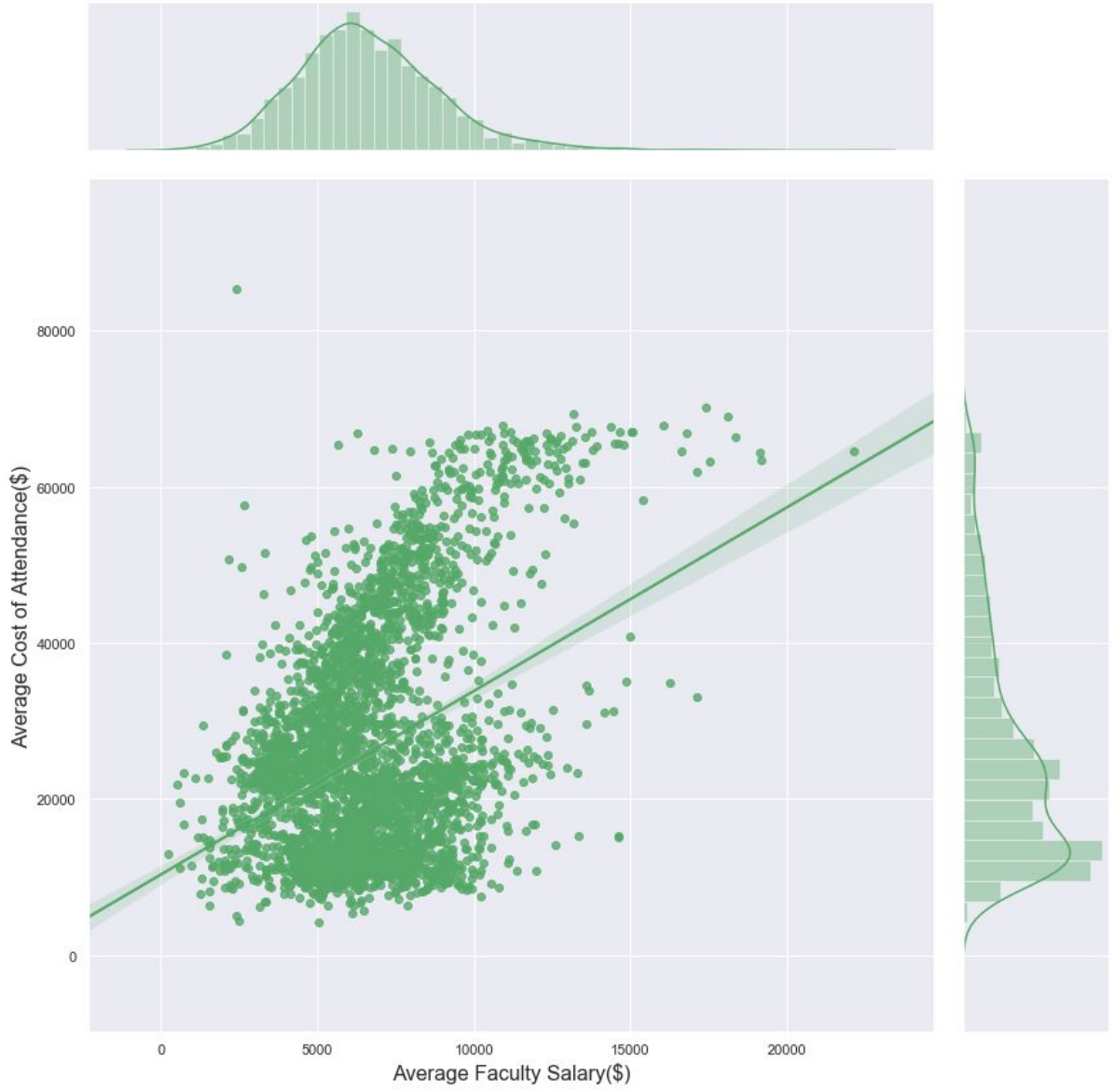
Following box plot shows the distribution of Cost of attendance across states in USA. We can see that the cost of attendance is not uniform across states and few states have very few schools. RI and MA seem to have the highest mean cost of attendance while NM, MT and ND are among the lowest mean cost of attendance.

Q) Distribution of cost of attendance across Degree Type



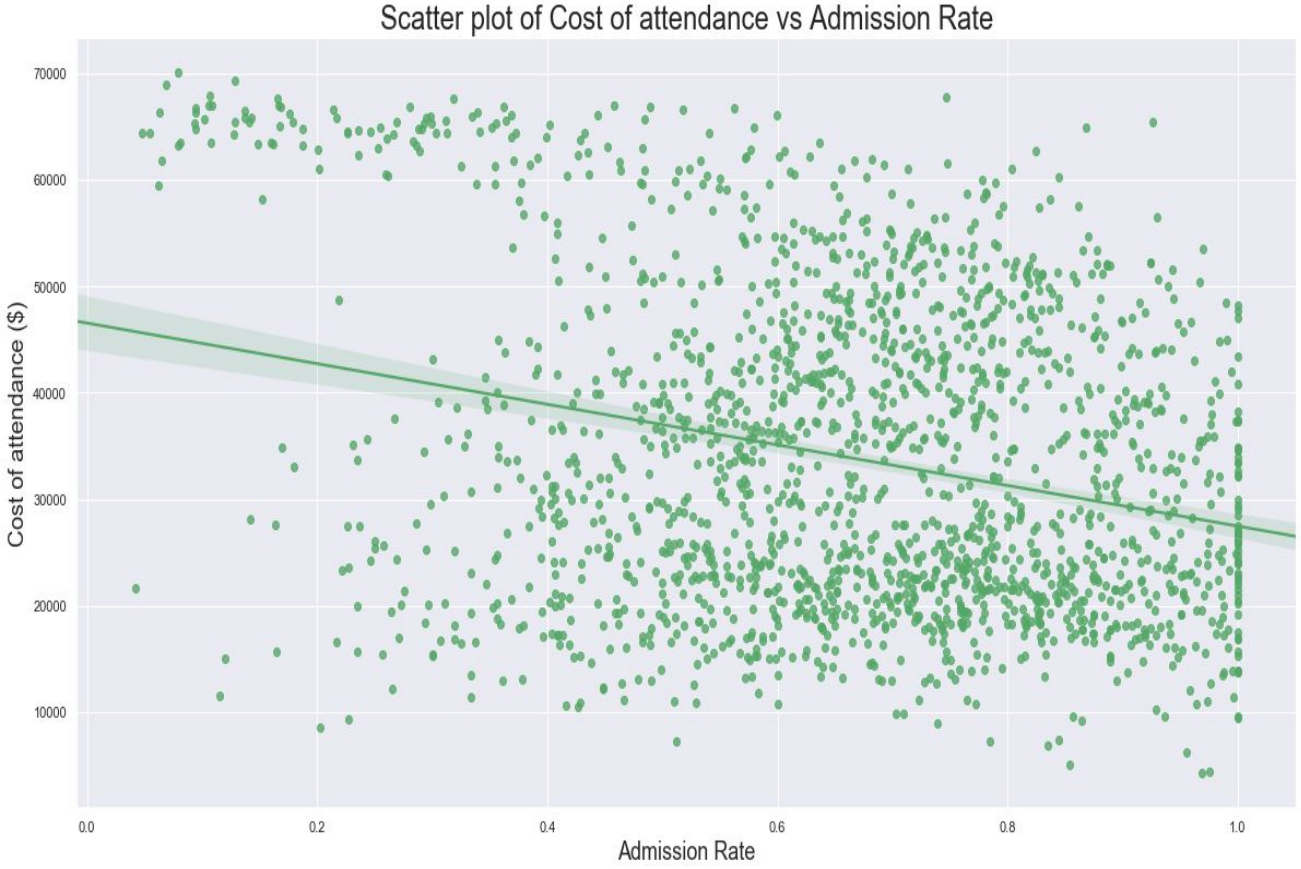
Following Box plot shows the cost of attendance across degree types. As expected, Graduate degree is the most expensive programs followed by Bachelor degree. Associate degree has the lowest mean cost of attendance. This also is one of the reasons for fewer students enrolling in Graduate programs compared to the Bachelor's program.

Q) Do schools with a high cost of attendance also have a high faculty salary?



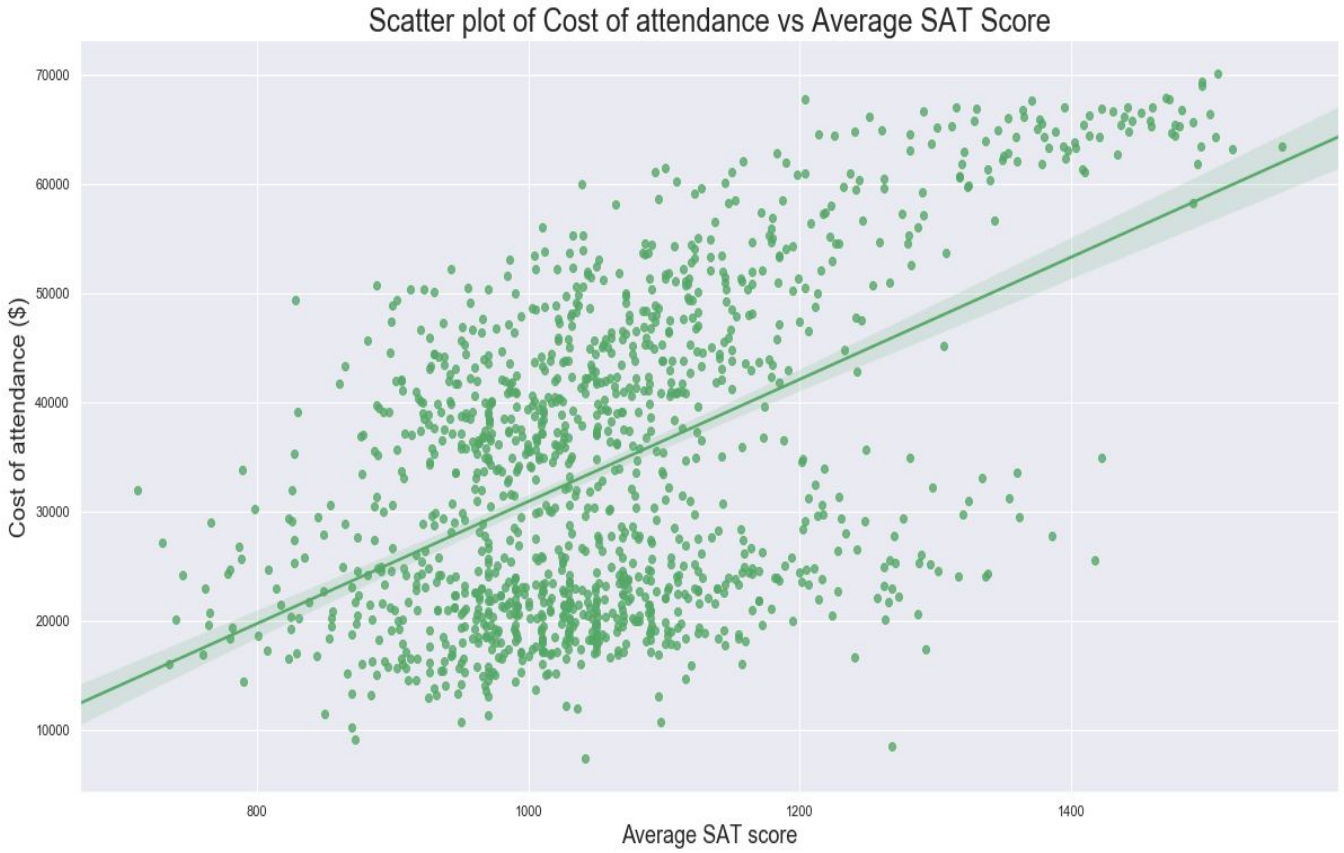
Following scatter plot shows us the relationship between the Cost of attendance vs the mean faculty salary. As clearly shown, there is a positive correlation between the cost of attendance and faculty salary. We also see that the faculty salary is quite normally distributed however the cost of attendance is skewed to the left.

Q) Are schools with a high cost of attendance more selective?



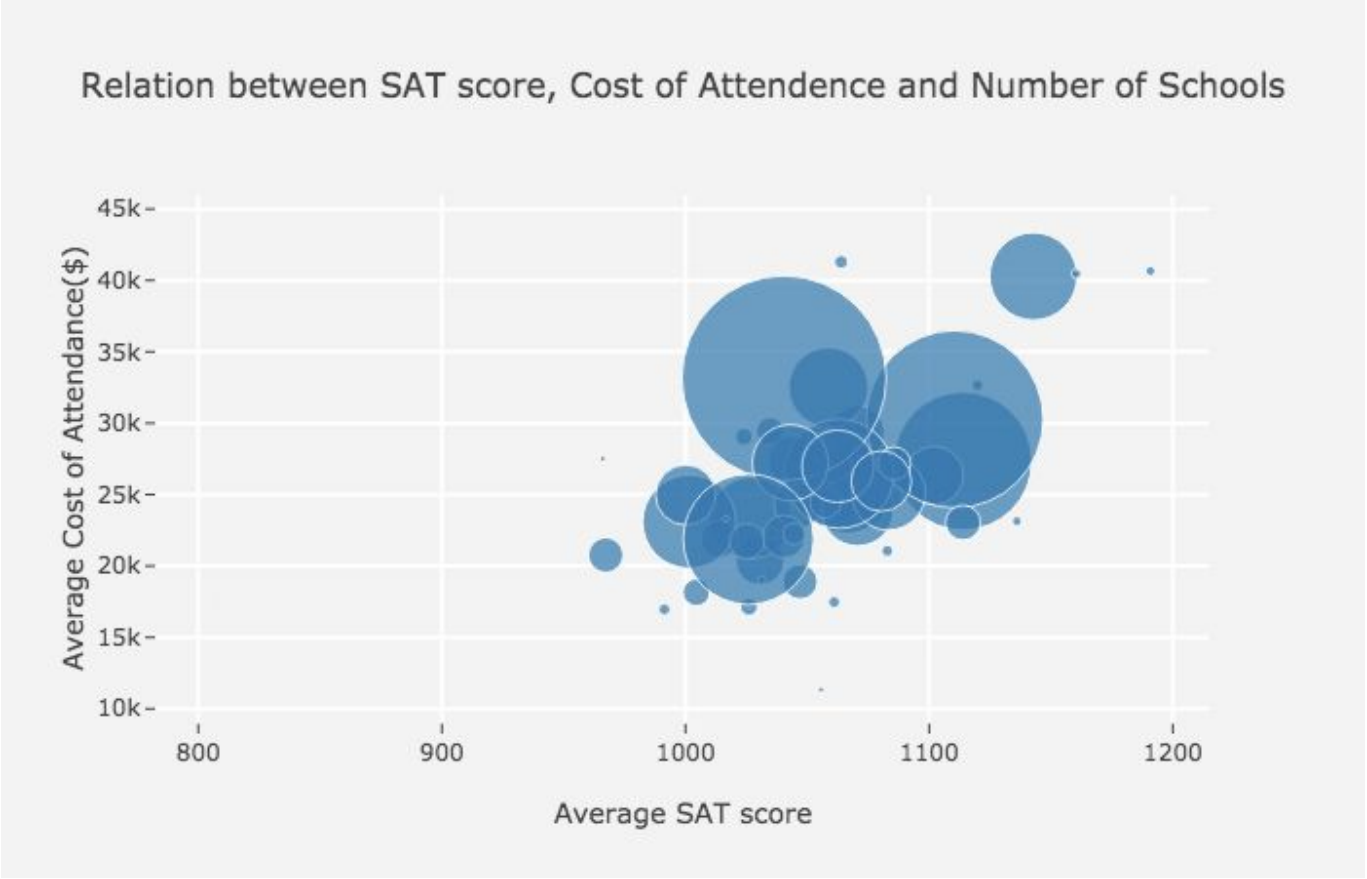
The scatter plot above shows the relationship between the cost of attendance with the admission rate. We can see a negative relation between the cost of attendance and admission rate, suggesting that it is more difficult to get into expensive schools compared to cheaper once. It also hints that the admission rate of Private schools which are generally more expensive is lower compared to the public schools.

Q) Which kind of school higher SAT score student prefers?



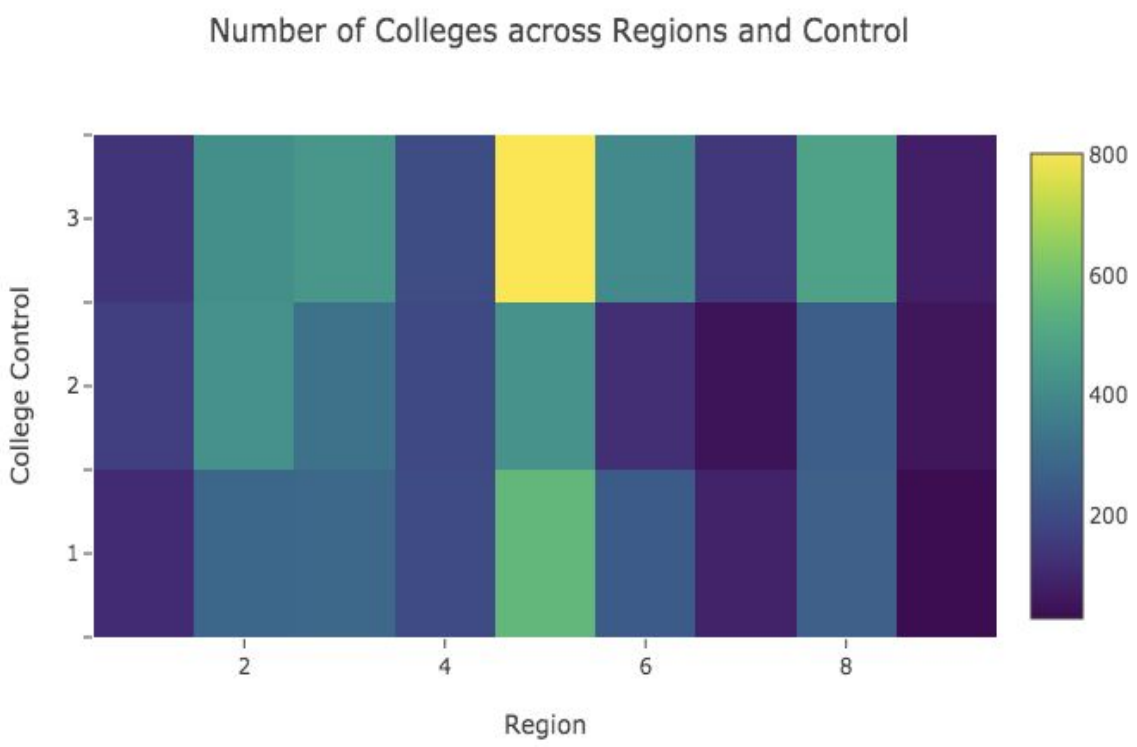
The scatterplot above shows a positive correlation between SAT score and the cost of attendance. This suggests that students with better SAT scores normally prefer to join private schools which are more expensive. This can also be as Private schools being more selective in admitting students compared to public schools.

Q) Does states with a higher number of schools have a lower cost of attendance and SAT score.



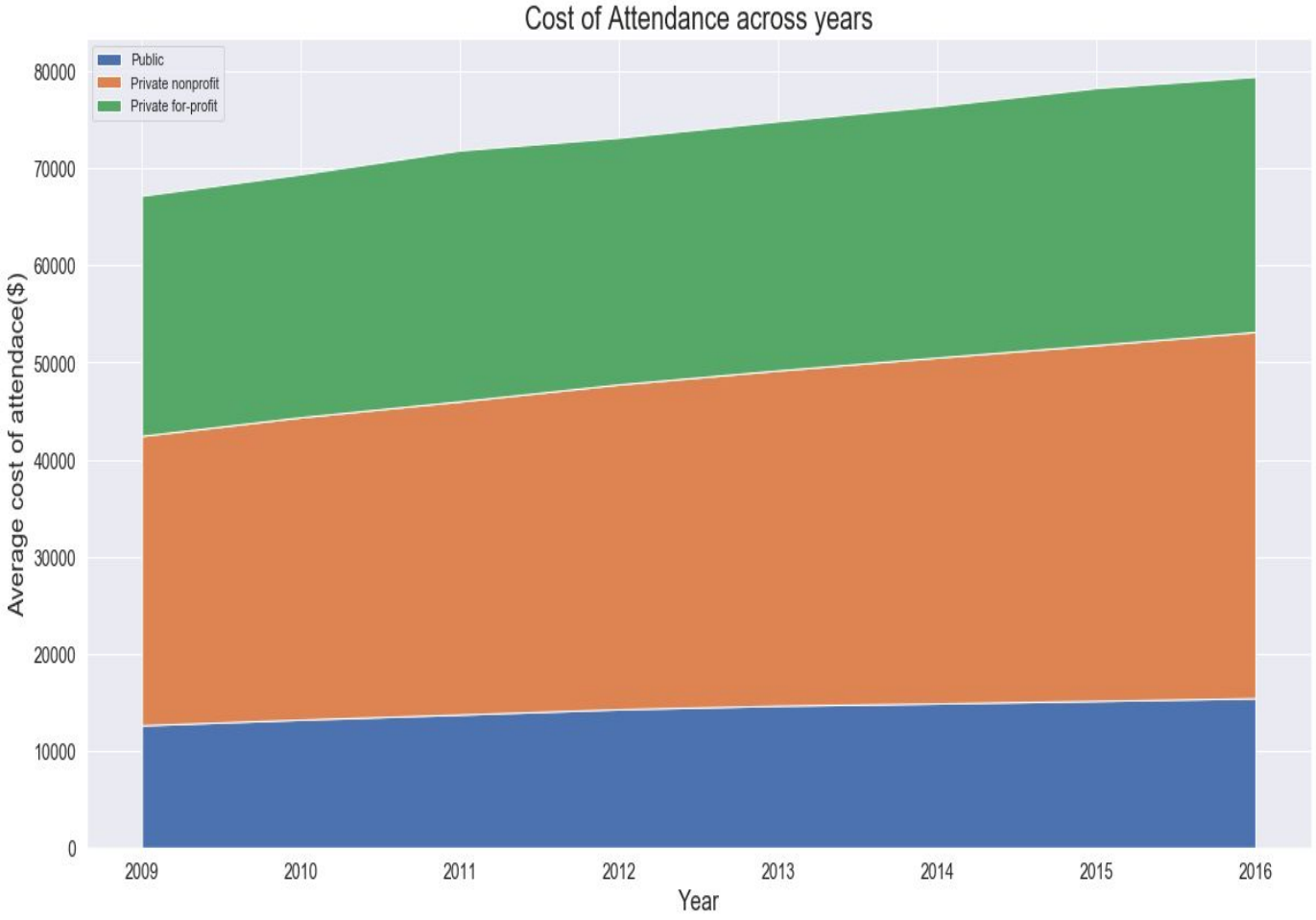
Following bubble plot shows state wise relationship between mean SAT score, cost of attendance and number of schools. We do see a relation between the SAT score and the cost of attendance but the number of schools does not play a huge role in reducing cost. In spite of having a large number of schools in a certain state, the cost of attendance does not go down significantly. MA has a high number of schools and high SAT and still have a very high cost of attendance. **Note that this is an interactive plotly graph that can be better visualized on the jupyter notebook submitted.**

Q) Distribution of Number of colleges across regions and control(school type)



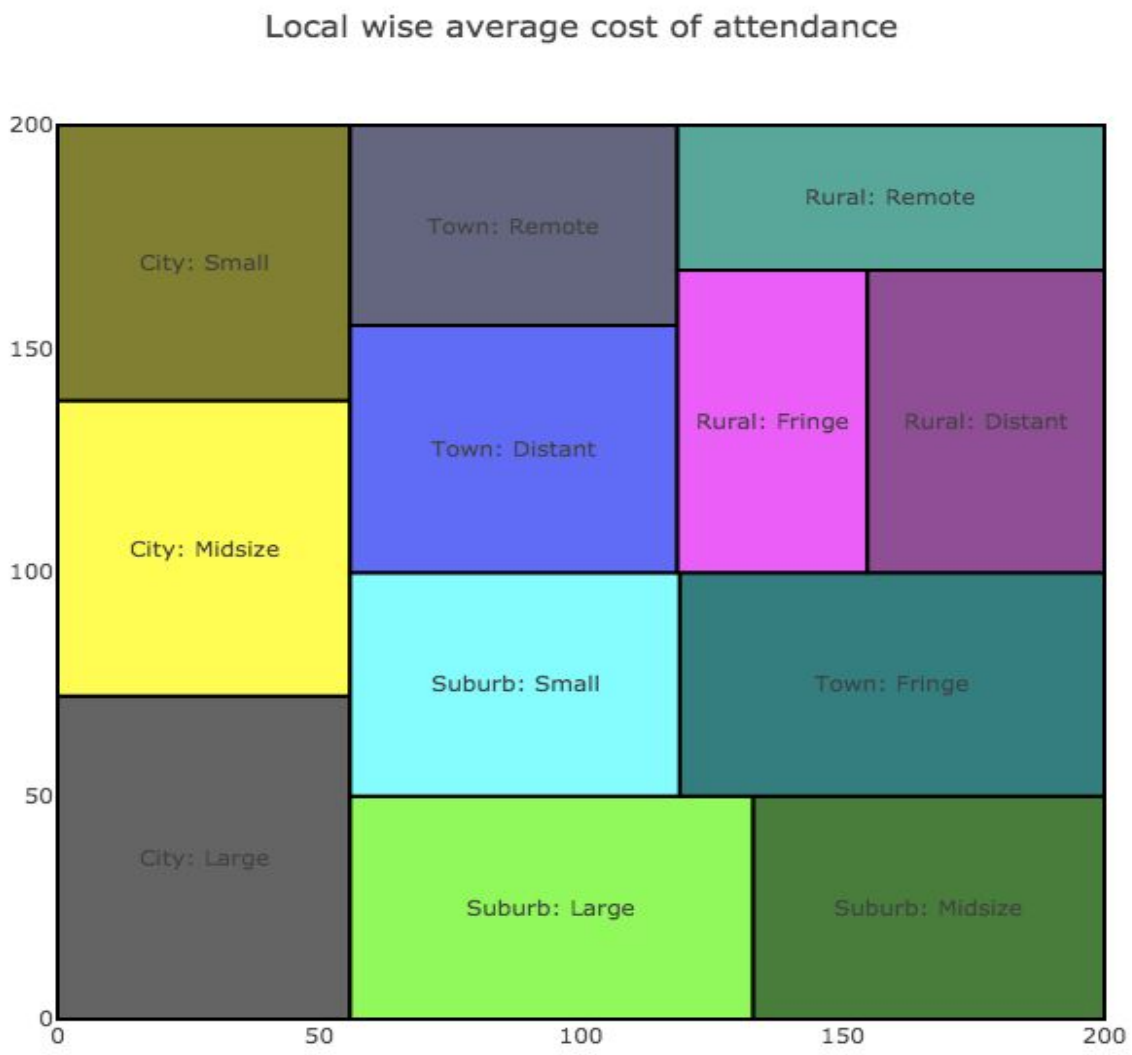
Following is a heat map that shows the distribution of the number of schools across regions and control type. We can see that Region 5 has the highest number of schools under control type 3 while region 9 has the lowest number of schools under control type 1. We also see that certain regions like 7 and 9 have on average less number of schools across all 3 control types while region 5 has a high number of schools across all 3 control types. **Note that this is an interactive plot made on plotly that can be better visualized in the Jupyter notebook shared.**

Q) Has the Cost of attendance increased across years?



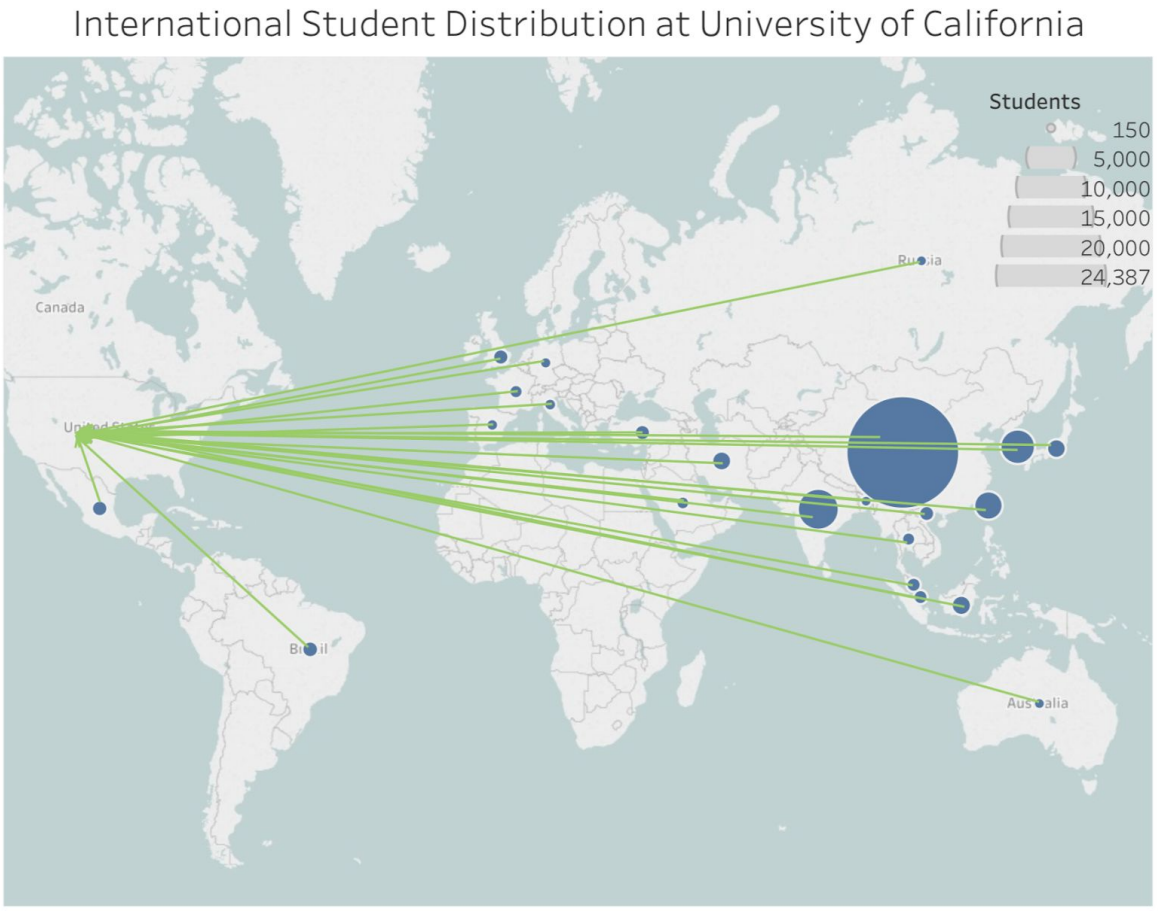
Following is a stacked graphs which compare the cost of attendance across different types of schools since 2009. We can clearly see that the cost increases for all the 3 types of school. We can also see that the increase is slower in public schools compared to the other two school types.

Q) Is the cost of attendance affected by the type of locale?



Following treemap helps us visualize the impact of locality on the cost of attendance. As expected, the cost of attendance is highest in City Large followed by Town Fringe. Rural Fringe has the lowest cost of attendance. **Note that this is an interactive plot made on plotly that can be better visualized in the Jupyter notebook shared.**

Q) Distribution of international students joining the University of California

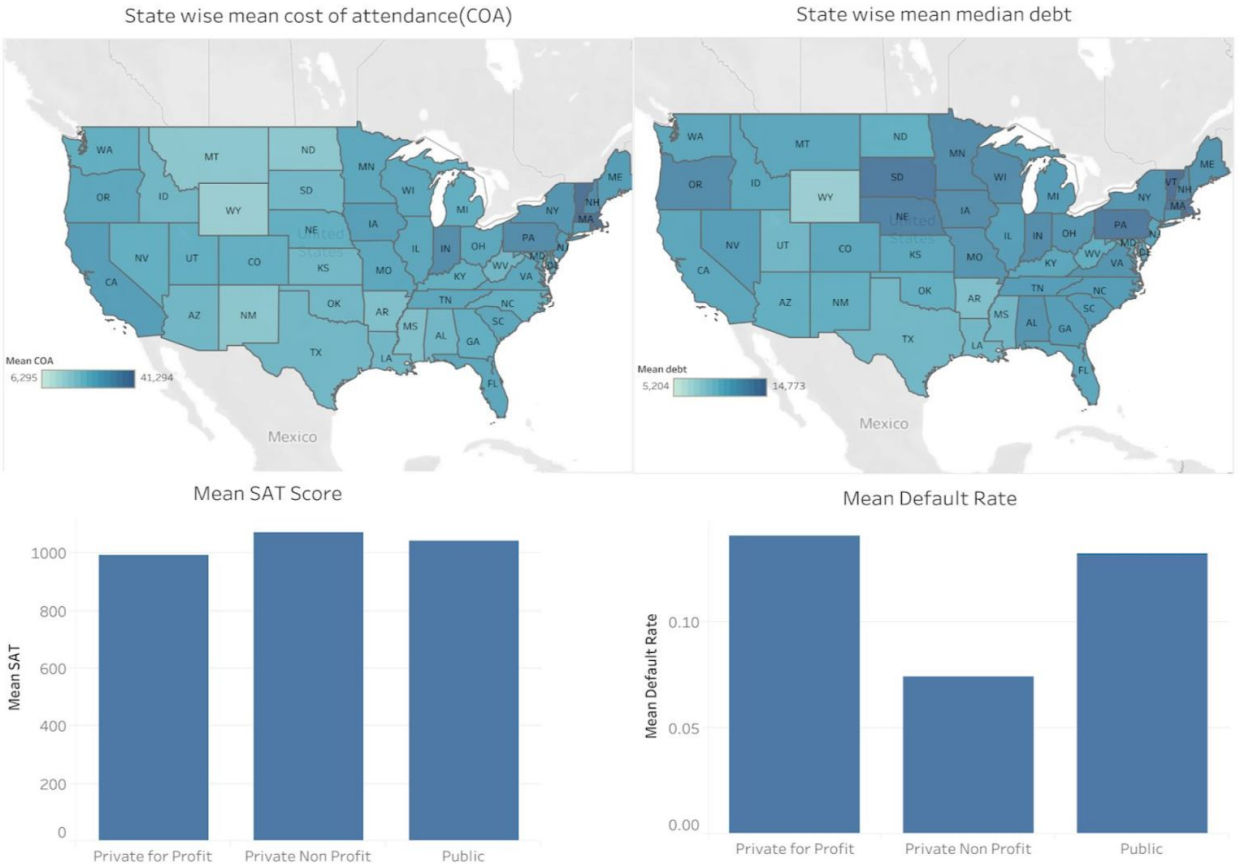


Following is a connection map which shows the number of students coming to university of california across the top 20 countries by the number of students. The bubble represents the number of students and we can see that China is the biggest contributor to the number of students followed by India. In general, Asian contributes the most compared to other continents.

Story Line:

Till here, we have seen a lot of plots which show the relationship between the cost of attendance across regions/locale, states and also its relationship with many factors. In this part, I will use a single graph to answer 4 major questions and convey the key story as follows:

Q) Relation between Cost of attendance, median debt, SAT and default rate



Following is an interactive dashboard created using tableau. The Choropleth plots are click activated and can filter a particular state and the rest of the graphs changes based on that. This helps us capture both global and state wise data distributions. **I would encourage you to try it on the following link**

a) State wise cost of attendance?

To answer this question, i decided to plot a Choropleth plot where color intensity represents the cost of attendance. The relationship between color intensity and cost is shown on the bottom left corner of the plot. I have used single color scale based on previous feedbacks as it is visually more appealing. The reason for using Choropleth plot and not any other plot like

a bar plot is that we have 50 states and the plot will become difficult to fit and read. Vermont(VT) has surprisingly the highest mean Cost of attendance among US states and as shown in HW7, has the lowest number of schools. It is followed by Massachusetts(MA) and Pennsylvania(PA). Surprisingly California(CA) does not have a very high Cost of Attendance. Wyoming(WA) has the lowest COA among all the states.

b) State wise average median debt and relation to the cost of attendance?
To answer this, I have similarly plotted Choropleth plot where color intensity represents

the average median debt. Wyoming(WA) has the lowest average median debt post graduation and can be attributed to the low cost of attendance. Similarly Vermont(VT), Massachusetts(MA) and Pennsylvania(PA) are among the top average median debt states. The reason for plotting these two graphs side by side was also to correlate the cost of attendance to average median debt and easily see a relation. The other way to plot this information could have been to use a scatter plot but the state level granularity would have been lost. On average we see a correlation between the Cost of Attendance and Median debt post graduation. South Dakota(SD) and Nebraska(NE) has a surprisingly high average median debt values compared to the cost of attendance and this could be due to job opportunities and average pay in those states.

3) Average SAT score across different school control type
Another question i was interested to explore was to see which school control type attracts a

better pool of students based on SAT score. I was expecting Public school's to have higher mean SAT score but turns out Private Non Profit schools lead in terms of mean SAT scores. As expected, private for profit schools have the lowest mean SAT score as it is normally the last preferred option. Note that for a few states the mean SAT score is not available in the dataset.

4) Mean default rate across control type and relationship to SAT
I also wanted to compare the mean default rate on loans based on school control type and its relation to SAT score. The trend in the default rate is consistent to SAT score with lower SAT score leading to a higher default rate. However, the drop for Private Non profit is significantly higher which could be due to other hidden factors. We can also check these trends state wise and they mostly hold true for most of the states.

Results/Summary/Conclusion:

1) Private schools are more expensive compared to public schools yet attracts the best talent and has a lower acceptance rate

2) Students with a high SAT score prefer to join expensive Private schools compared to public schools.
3) Expensive schools pay more to its Faculty compared to less expensive ones.
4) China and India contribute to the maximum number of international students joining the University of California.
5) Private schools despite being expensive how lower the default rate may be due to a better student pool.
6) A larger number of schools does not impact the cost of attendance but the location/locale of the school does.

Appendix Containing All Code:

Note that all the code and interactive graphs have been uploaded to canvas and github as jupyter notebook

Link to your github page with this analysis:
https://github.com/jyotipmahes/Final-Project-Dataviz

Citations:
None