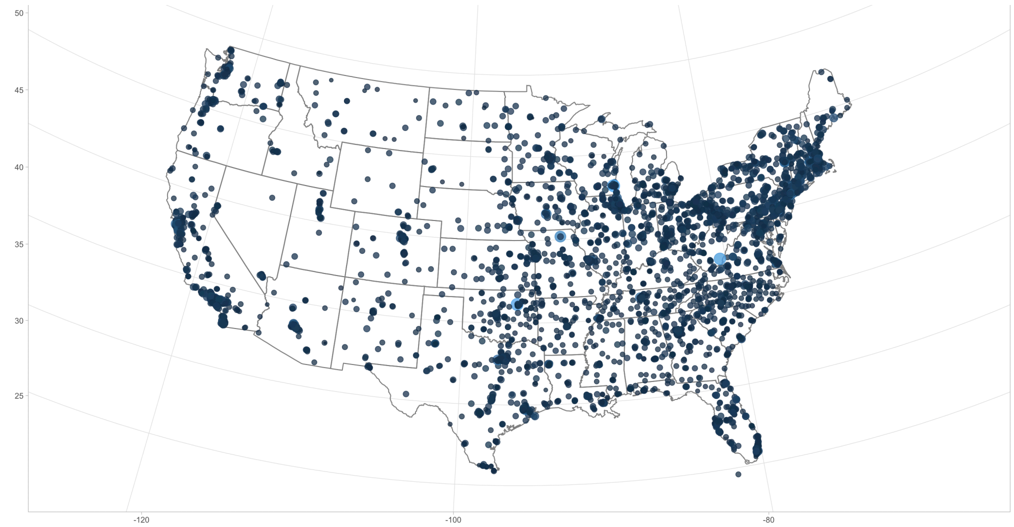
College Student in the US
Exploratory Data Analysis with R



Introduction

College student data plays an important role in helping educators and governments to reflect on the current status different aspects in student life before, during, and after college then take effective actions towards improving the educational systems. In this report, we carry out a exploratory analysis a large student college data consisting of more than 7,000 colleges/universities across US. In particular, we look at three important aspects of a college student including the standardized tests, student debts remaining, and earnings after graduation.

Dataset

The dataset in this project is collected form US Department of Education's College Scorecard [1]. Starting from 1996, Department of Education conducted survey at undergraduate degree-granting institutions of higher education in order to learn more about student program completion, debt and repayment, earnings, and more. Each annual data is presented in a large csv file (∼100MB) containing more than 2000 columns (variables). Nonetheless, most of the variables are not interesting for general audiences therefore we carefully select a smaller subset of variables that we think appealing to the public. Next, we discuss the set of selected variables used for data analysis.

Variable Selection

College Scorecard publishes a data dictionary in which they give brief descriptions for all variables in the dataset. Table 1 shows all the variables used in this work along with their brief descriptions.

[1] https://collegescorecard.ed.gov/data/

Table 1: College Student Variables Used

No	Variable	Description	Variable Type
1	UNITID	Unit ID for institution	Institution Data
2	CITY	Institution city	Institution Data
3	STABBR	State abbreviation	Institution Data
4	STATENAME	State name	Institution Data
5	INSTNM	Institution name	Institution Data
6	ADM_RATE	Admission rate	Institution Data
7	GRAD_DEBT_MDN	Median of student debt	Student Debt
8	SAT_AVG	Mean scores of SAT	Standardized Test
9	ACTCMMID	Mean scores of ACT	Standardized Test
10	UNEMP_RATE	Unemployment rate	Student Status
11	LATITUDE	Institution latitude	Geographical Data
12	LONGITUDE	Institution longitude	Geographical Data
13	REGION	Institution region	Geographical Data
14	DIVISION	Institution division	Geographical Data

Standardized Test Scores

Typically, admission to a US college requires either Scholastic Aptitude Test (SAT) or American College Testing (ACT) or some require both. These two tests are important factor in determining if students should be accepted into a college or not. In this report, we look at two tests using two different means of representing data distribution, box plot for SAT and density plot for ACT.

Figure 1 shows the distributions of SAT score within US divisions. Note that these scores are median SAT scores collected from all colleges in the survey. In general, the median of median SAT scores are relatively similar of around 1,050. The **Pacific** and **New England** divisions have the highest interquartile ranges where their 75th percentile scores are greater than 1,200. For the rest of divisions, their upper outliers fall into 1,300-1,500 range and lower outliers are less than 900 with West North Central, East North Central, and East South Central.

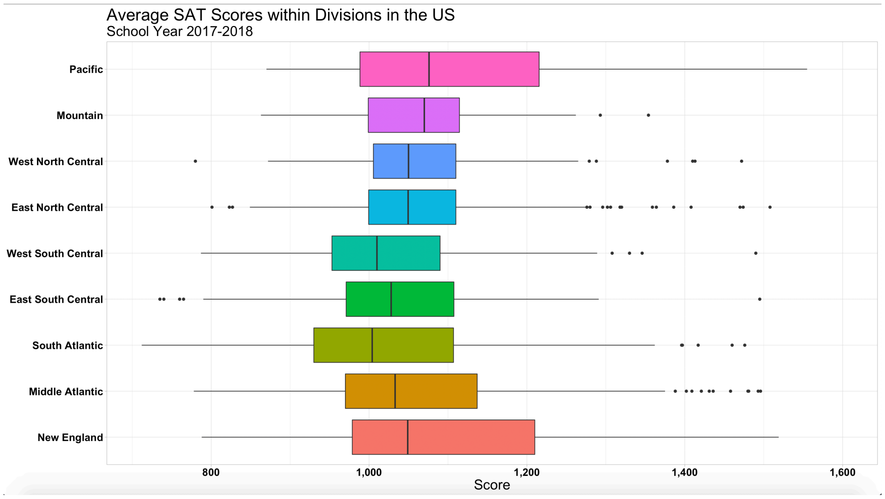


Figure 1: SAT Score across Divisions

Figure 2 shows the density plot of ACT scores over US regions. Most of the scores fall into 20-25 score range with North Central has highest density in this range. In addition, North Central also has plenty of students with ACT scores of less than 10, which is comparatively low compared to other regions in the nation. Looking at the right side of the plot, student applying to Northeast universities tend to perform better where they have highest portions of students with ACT scores of 30 or more. One plausible explanation might be the universities in Northeast are generally more competitive than others therefore they can attract some of the best students in nation to apply to.

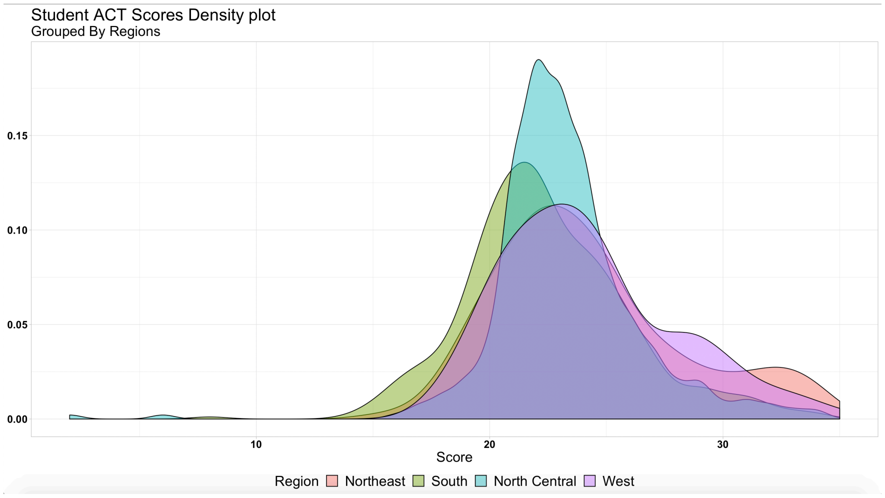


Figure 2: ACT Density Plot

Student Debt

Student Debt Trend between 2007 and 2017

Figure 3 shows the debt remaining in 2007 and 2017. Note that the debt in each surveying year reflects the median debt of the students 10 years after their graduation. We can see from the chart that there is a increasing trend in the debt from all states. Students from **South Dakota** secure the highest debts in both 2007 and 2017 with $17,000 and $25,000, respectively. On the other side, Wyoming students remain lowest debts after 10 years of graduation of in 2007 and 2017 with $6,500 and $8,500, respectively.

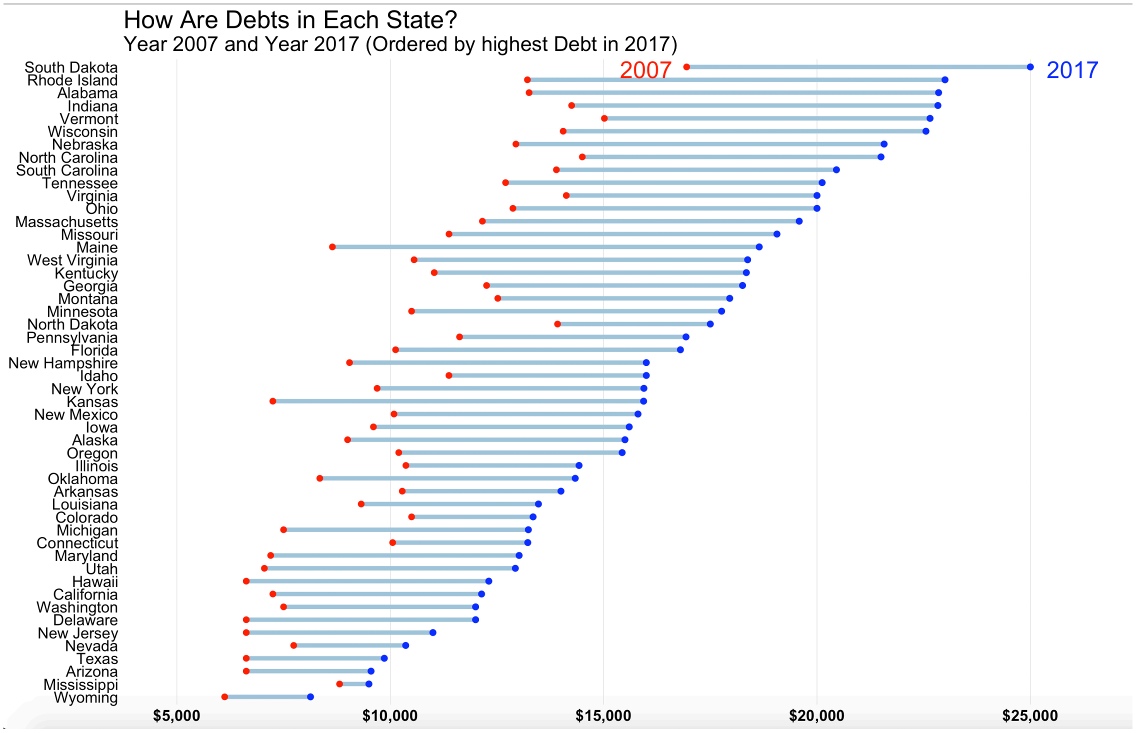


Figure 3: Student Debt in 2007 and 2017

Least Debt States

If we are more interested in a specific aspect of the debt, e.g., top 10 states with least debts and what regions are they from, Figure 3 may become to be over-complicated and also lack of needed information. Instead, we analyze the data and observe two things, 1) **North Central** states are absent therefore the students come from this region tend to own more money than those from other areas. This can be explained by the facts that the tuition fees and costs of living in North Central are significantly higher than other areas in the US and 2) **Western** states dominate the list with 5 appearances. Following Western region, students from the **South** also pay well their debts when they have 4 representatives in the list with Mississippi, Texas, Delaware, and Louisiana. Northwest have only one state in the list, New Jersey. As of 2017, these states have the debts of less than $12,000$, which can hopefully turn to zero in the following years of hard working.

Student Debts across US Regions

In the previous section, we have discussed the general trend in student debts across states. However, it is also captivating to explore insights among groups where the states in the same group are believed to share similar economic and social status. The US is divided into 4 geographical regions, West, Northeast, North Central, and South. The Figure 4 illustrates sorted state's debts grouped by each region. Wyoming and Mississippi, as we have noticed earlier, have lowest student debts in their regions. Vermont has the highest student debt in **Northeast** followed by Rhode Island and Massachusetts. In the North Central, South Dakota tops the charts with roughly $25,000 debts. On the other hand, its neighboring state – North Dakota has lowest student debt of around $12,000, which is pretty interesting.

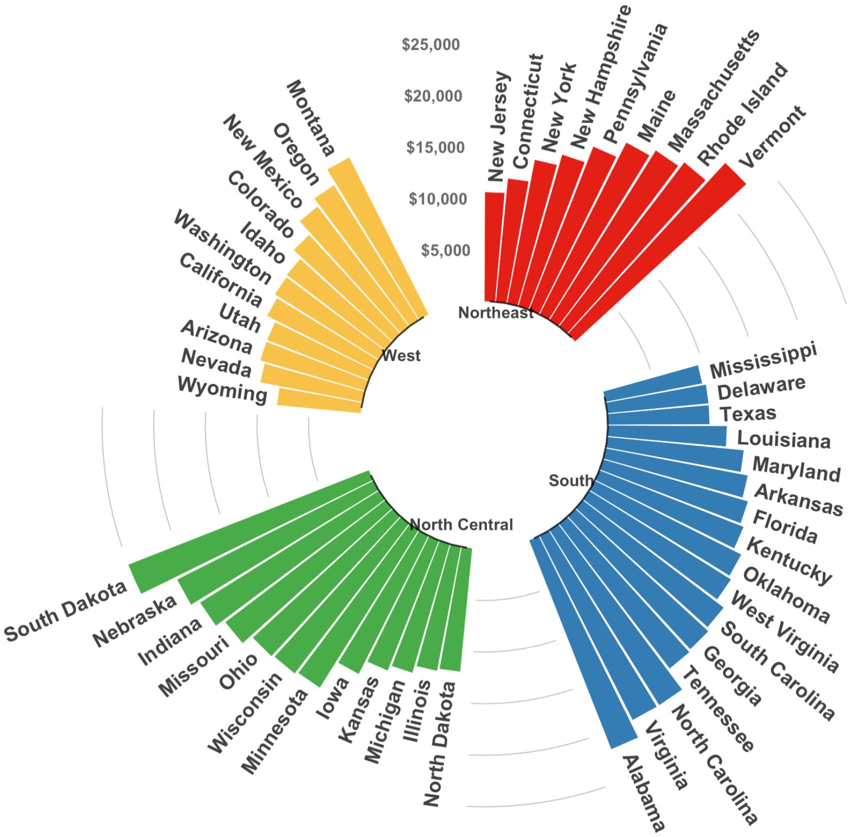
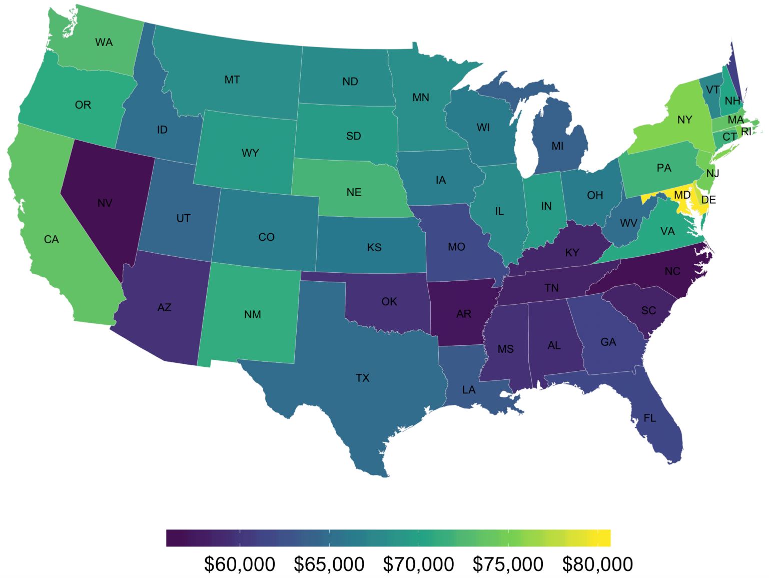


Figure 4: Student Debts Across US Regions

Earnings after Graduations

The dataset contains student salaries after their graduation. We calculate the median salaries of each state then use map data to plot the Figure 5. The median salaries of US states range approximately from $50,000 to $90,000. Maryland has the highest median earnings of students with more than $80,000 per year followed by New York, Massachusetts, and California with more than $75,000 in median earnings. This result is not surprising because these states have some of the most selective universities in the US therefore the student can supposedly get highly-paid jobs after graduation. In addition, these states have some of the places with highest cost of living. As the result, employers often pay more to compensate for the costly expenses. In contrast, Nevada, Arkansas, and North Carolina are at the bottom with median salaries of less than $60,000. The reasons maybe that the majority of the students from these state pursue low-demanding jobs, which possibly leads to more difficult job hunting. Again, cost of livings in these states may be complement factor when it comes to deciding the salary rates.



Data collected in 2018

Figure 5: Student Earnings 10 Years after Graduation

Summary

In this report, we have explored various aspects of US students during and after colleges. We showed that there is an increasing trend in student debts over years at all states. We also looked at SAT and ACT scores and compare the performances of students from different divisions and regions. Finally, we investigated the student earnings across the states. We observed that students from selective universities make more money than the ones from less competitive areas partly due to the difference in cost of livings in those areas. Still, the question of how to choose the right colleges remains unanswered. There need to be more thorough data analysis with much more factors being taken into account in order to draw comprehensive conclusions.