Tanner Coon

June/6/2021

Github project link: <https://github.com/TannerCoon/dst1_final_project>

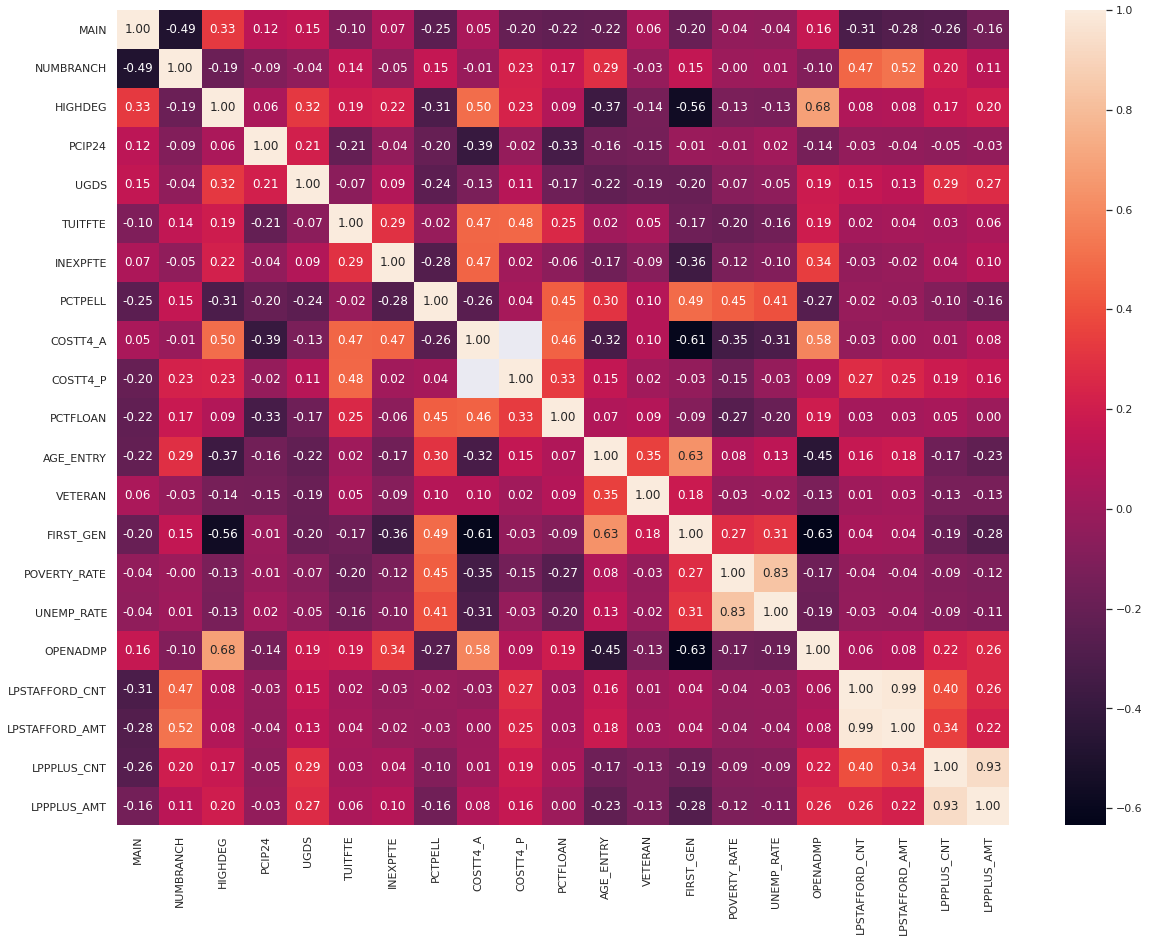
Data Science Tools 1 Final Paper

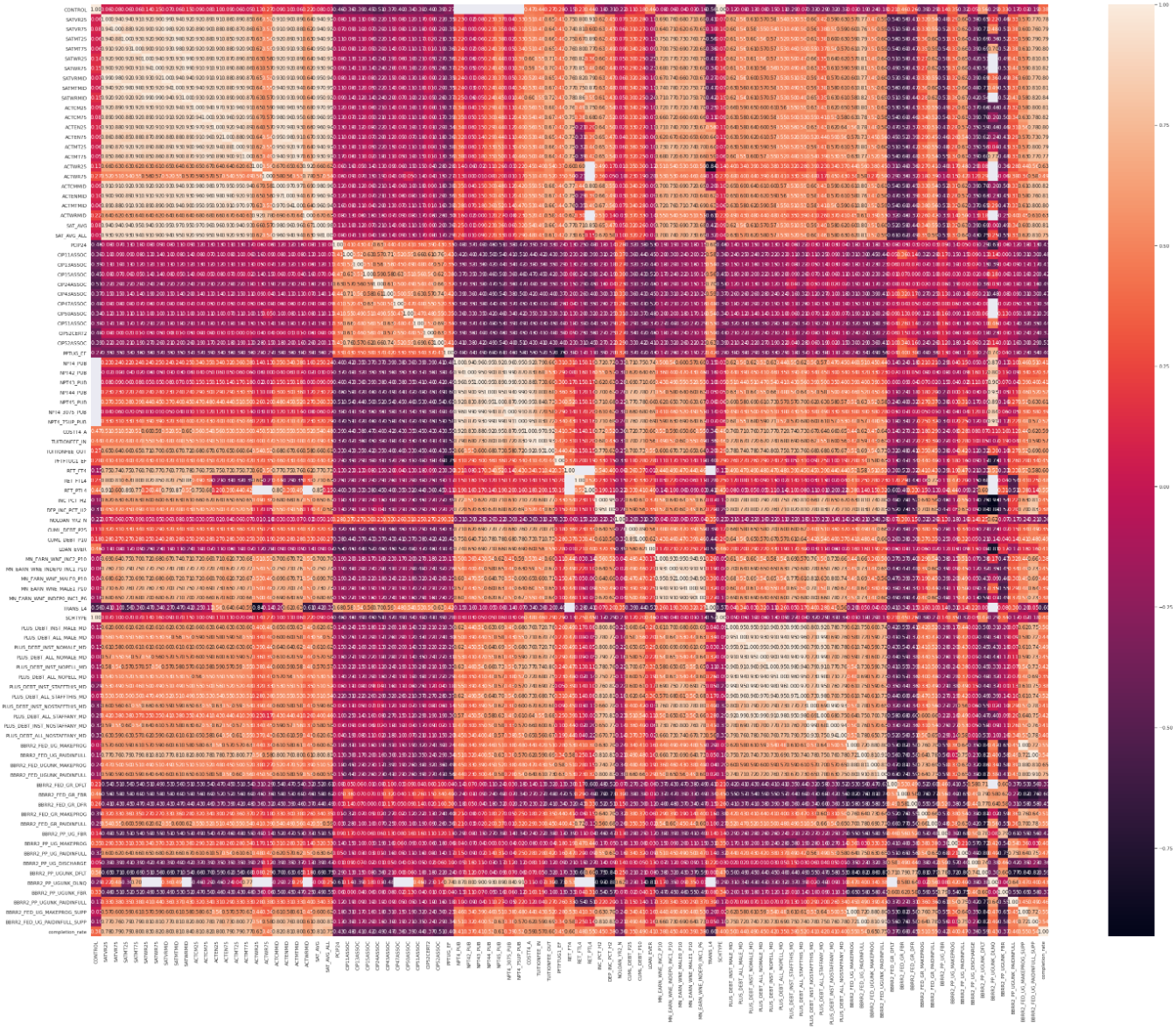
I previously took a similar course to this one when I was an Undergrad. In that class a group and I developed a website that would give useful information on ski resorts. This site displayed several charts and metrics that would help people decide which ski resort they would like to visit. I decided I wanted to do something similar with this project and I came across some interesting US college metrics data. The data is “CollegeScoreCard” data from <https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources>. The data covers everything that colleges get scored on including completion, debt, repayment, and earnings information. The data came in the form of a zip file that contained several csvs. The ones I focused on in my project were “Institutional-level” files. These included the Merged####\_##\_PP.csv files and the Most-Recent-Cohorts-All-Data-Elements.csv file. The files cover college data from 1996 to 2018. When combined the resulting DataFrame of this data had 2,384 columns and 163,331 rows which is a total of 389,381,104 data values. With all this data I hoped to be able to generate some plots that could be useful to people considering different colleges to enroll in.

The metadata in this is quite overwhelming. As stated earlier, there are 2,384 columns and all of them are abbreviated in the data. These abbreviations can be completely unobvious such as the column, **MD\_INC\_WDRAW\_ORIG\_YR2\_RT**, being described as:

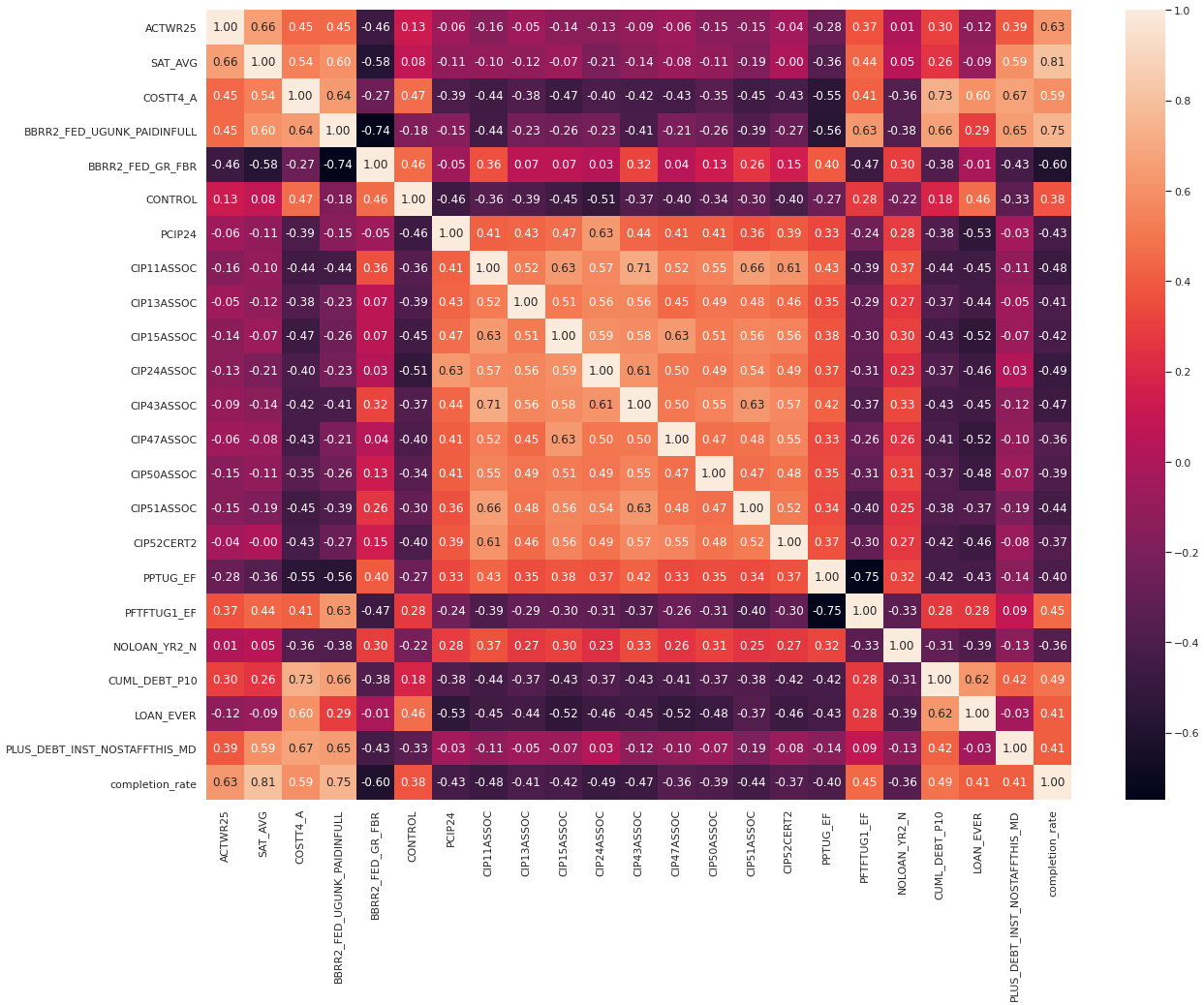
“Percent of middle-income (between $30,000 and $75,000 in nominal family income) students withdrawn from original institution within 2 years”

within the data dictionary file, data.yml. The data I end up using in this project covers everything from loan metrics, cost averages, poverty rates, average age of admission, to the percentages of various degree fields offered at an institution. The data I collected identified based on institution and year for each row.

Can you use a set of features to write an accurate model that guesses at the completion rate for an institution? Before I continue, when I refer to completion rate, this is the joining of two columns, **C150\_4** and **C150\_L4**. Both are described as “Completion rate for first-time, full-time students” with the only difference being whether the institution was a “4-year college” or a “less than 4-year college”. There is are no columns where the two values overlap, and the data could have just as well created a completion rate column and a separate year-range indicator column. Therefore, I feel there is no issue in combining the data into a single variable to be the completion rate for any college present in the data. To answer my research question, I came up with two sets of variables within the data to use for estimating the completion rate using Random Forest Regression. The 1st set of variables where chosen by hand. I went through almost all the row descriptions in the data dictionary and chose ones that I thought would be interesting to use to estimate the completion rate. Here is the correlation matrix of this set of variables and the completion rate for 2018:

To note some of the variables: **UGDS** contains the number of undergrads enrolled, **VERTERAN** contains the percentage of enrolled students who are veterans, and **OPENADMP** is a flag for whether the institution has an open enrollment policy. The next set of variables where those that had an absolute correlation higher than 0.35 and that did not directly imply completion rates for their columns. When I say directly imply completion rates, I mean to describe those variables that are described as being calculated for students who completed their degrees such as **MALE\_COMP\_ORIG\_YR2\_RT** which is described as “Percent of male students who completed within 2 years at original institution”. Here is the original matrix for this set of variables:

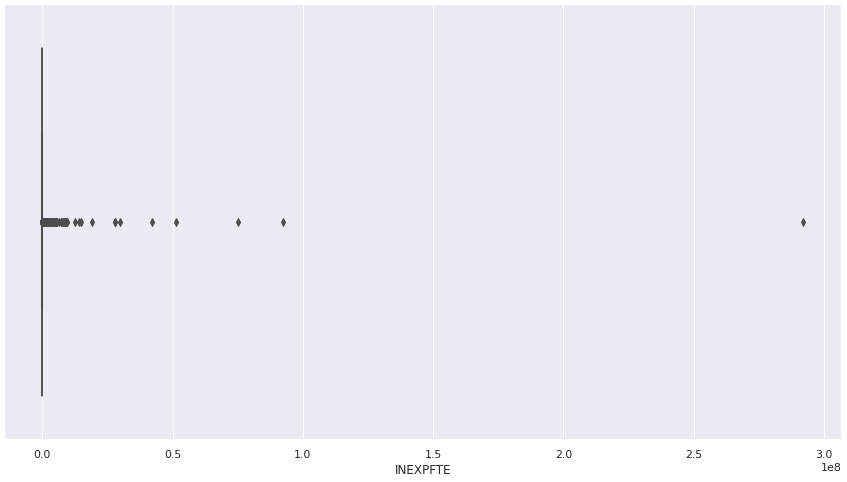
From this correlation matrix, I noticed several clusters variables that seemed to have high correlations between themselves. To avoid collinearity, I selected variables from these clusters that gave the best correlation value to the completion rate without being highly correlated between themselves. This resulted in the following set of variables creating this matrix:

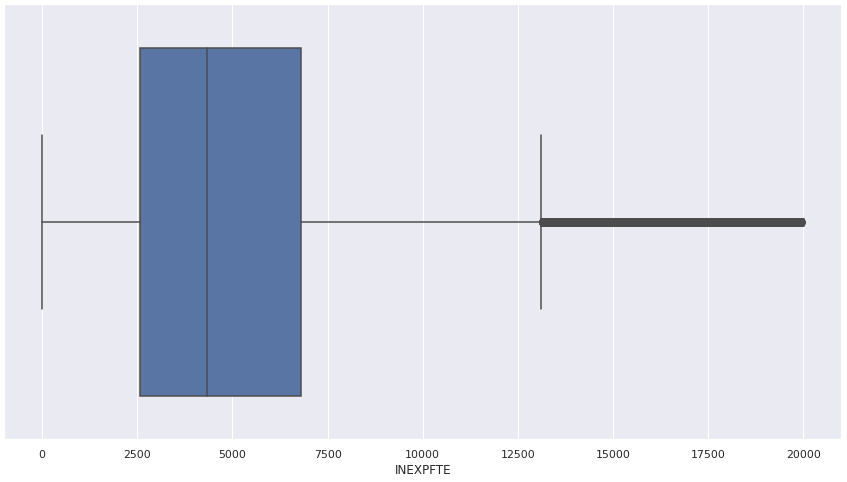
To again note some of the variables: **PCIP24** is the percentage of Liberal Arts degrees awarded that year, and **LOAN\_EVER** is the percentage of students who received a federal loan while in school. These two sets of variables will be inputs to Random Forest Regressor models that will output predicted values for completion rates.

I actually stumbled upon a project made by StackExchange user [Michael T.](https://opendata.stackexchange.com/users/13928/michael-t) that created a tool using this same set of data. This tool takes information from you including what field you want to go into and what region in America you’d like to attend school in and then plots a bar chart for recommended colleges. The tool can be seen at <https://thompsonml.shinyapps.io/BestCollegeApp/>. I also found some statistical case studies that use this data. At <http://jasontdean.com/R/collegeScoreCard.html>, you can find a paper on increasing transparency between the costs of colleges and the potential returns. They plot several variables against the average cost for institutions. They also show visualizations that compare the debts for students who complete their degrees and those who withdraw at different categories of institutions. At <https://anson.ucdavis.edu/~jsharpna/DSBook/unit1/cost_of_uni.html>, you can find a paper that creates several charts that are useful for exploring and understanding the data. The work I’ve done has some overlap with these projects however none of them applied Random Forest Regressor Models to the data and I never saw a chart that directly matched my own.

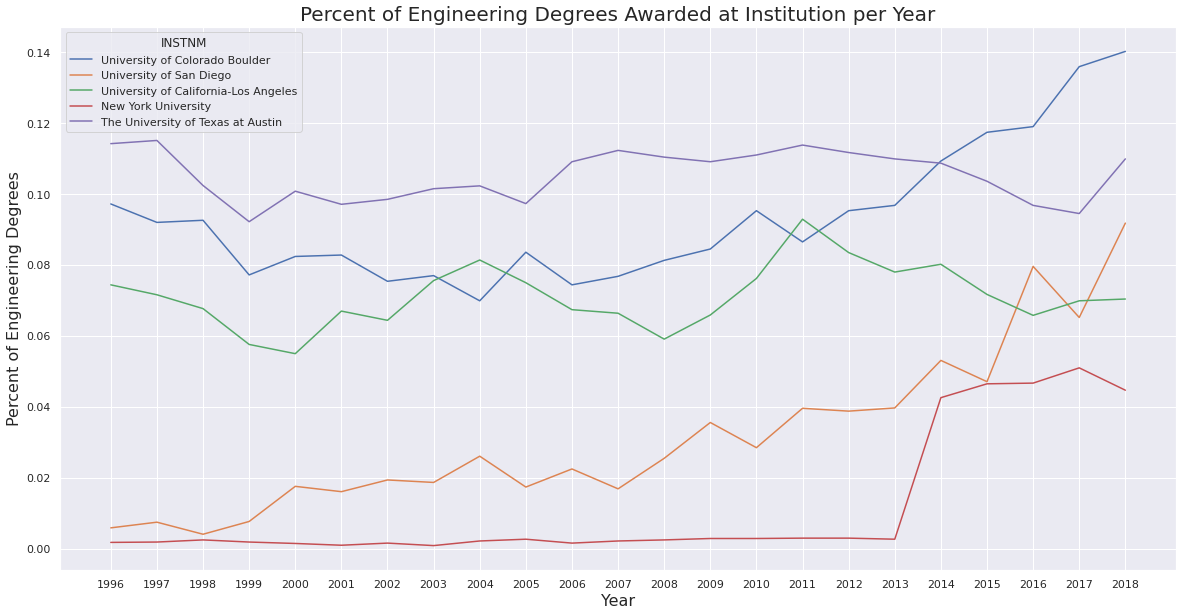
**Data Cleaning:**

The data I collected is relatively clean from a certain point of view. There weren’t any misspelled words that I had to work with nor were there many outliers in the variables I ended up working with. However, this isn’t to say I didn’t need to put in effort into getting the data into a usable state. To start off, I pulled in just the data from 2018 in Most-Recent-Cohorts-All-Data-Elements.csv. I did this because I was running into memory issues when trying to get the columns for the correlated set of variables I will be using later. I did this as a first step because the data is simply too large for my equipment. I tried running through some of the cleaning steps with all the data first and quickly ran out of memory. Therefore, I decided I would collect all variables I would need throughout the rest of the project first and only pull in the data I needed into a DataFrame. After gathering this info, I pulled in the data from all the CSVs and added a year row to indicate which year each row was collected. Then, I replaced the “PrivacySuppressed” values from all rows in the data with NaN values. This appears all over the data and just means the institution suppressed the data due to privacy reasons. From here I ensured there were no columns that only contained NaN values by dropping any that were. I then pulled in the data dictionary and extracted the type information for each column. There was one column I would be using that didn’t have a type listed but I checked the data to figure it out for myself. I then applied these types to the data.

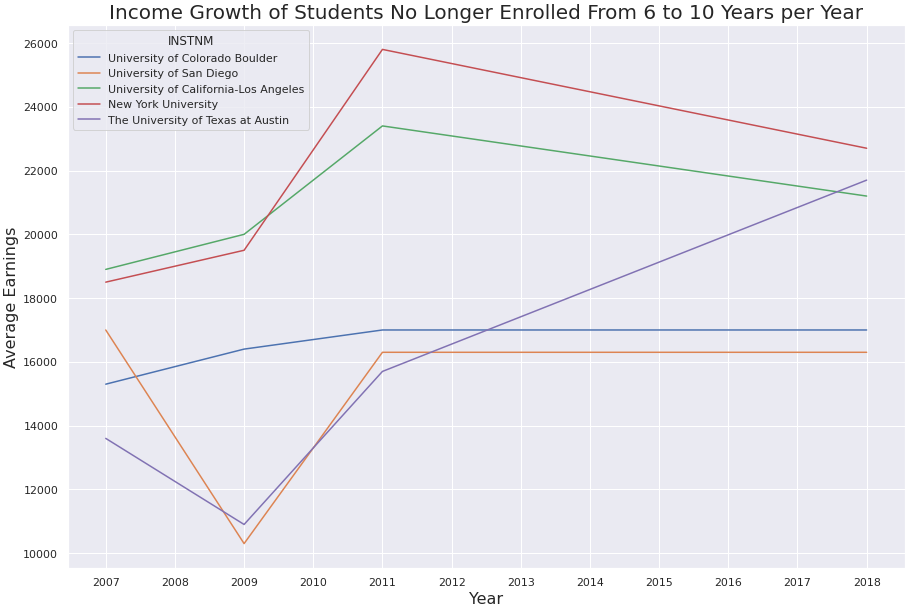
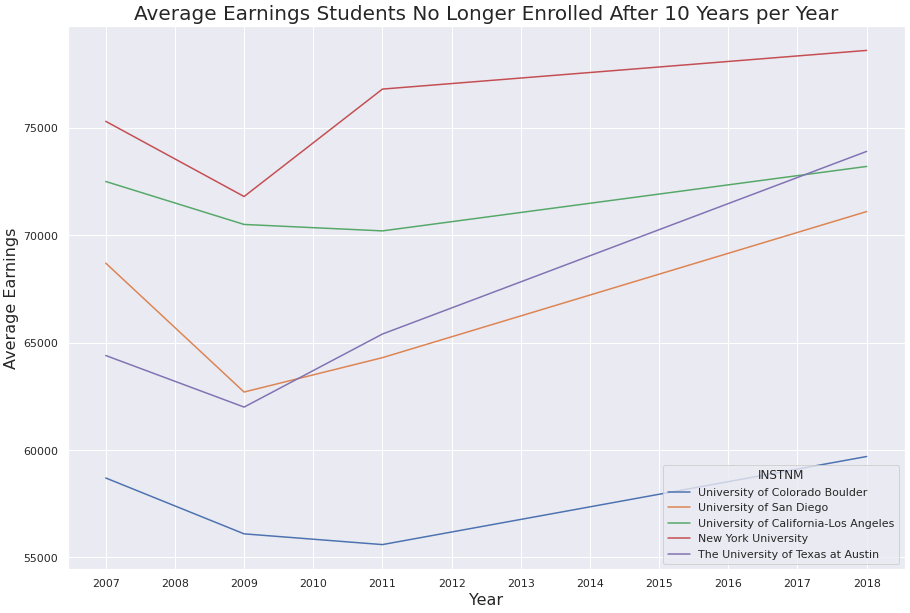
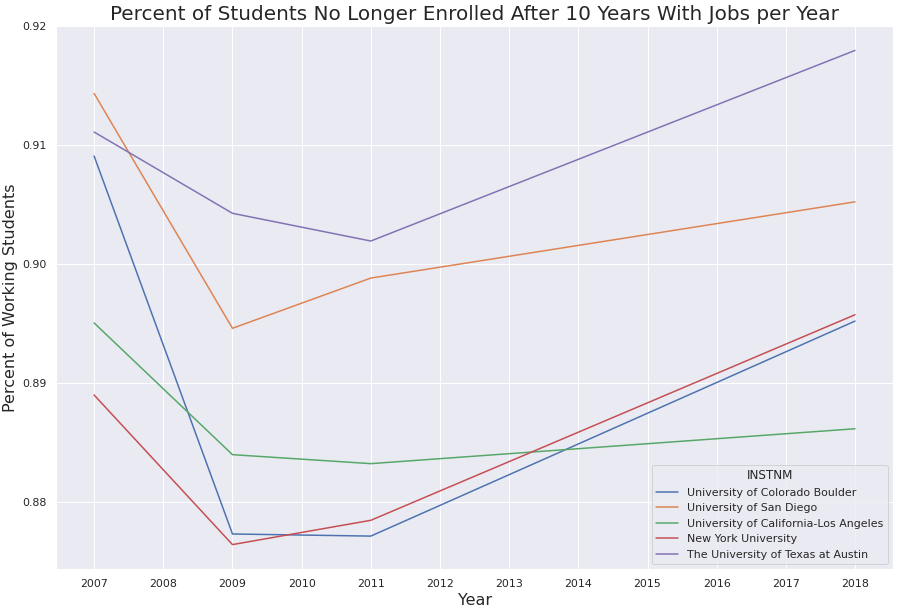
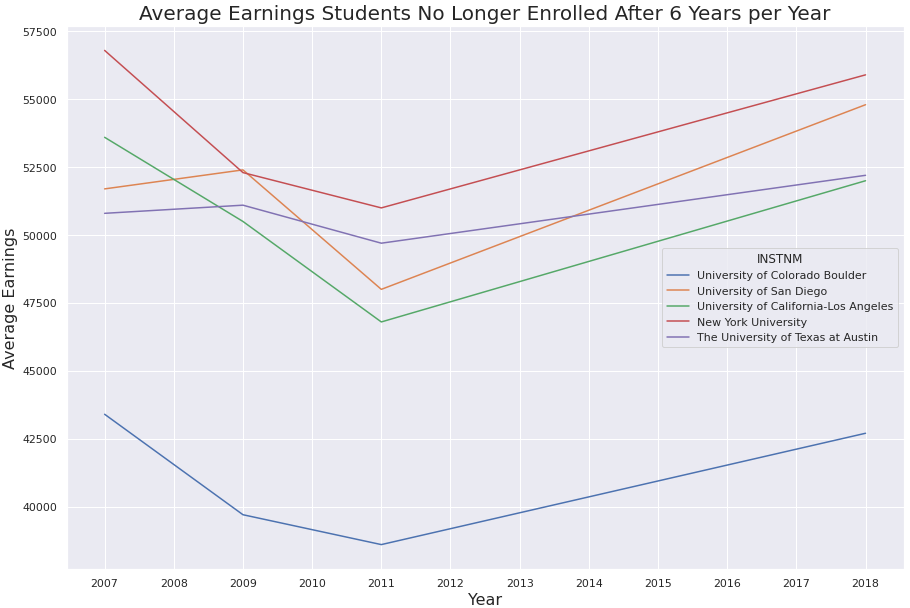
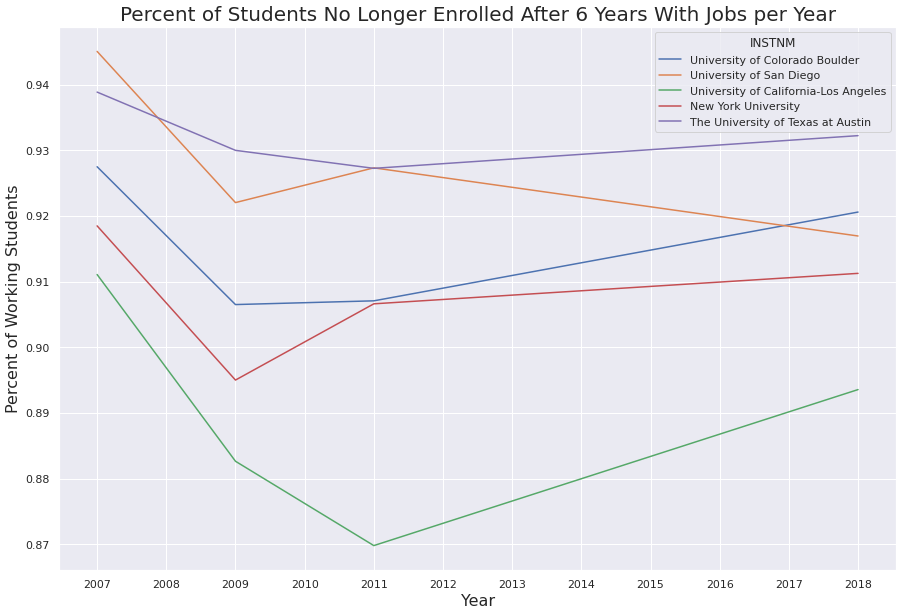
From here I started to explore the distributions for the variables in my data. I plotted boxplots and checked outliers to see if everything made sense. For **NUM\_BRANCH** there were a few with really high values but upon checking the data, these institution branches were in fact present. For **SAT\_AVG**, I found a few outliers. The most outrageous was for the intuition, The New England Conservatory of Music, which claimed to have an average SAT score for students of 1,599 one year and 813 the next and NaN values for every other year. I chose to drop the rows relating to this school because. California Institute of Technology also had high scores, but they were consistent across the years, so I kept this institution’s data in. The only major problem I felt I needed to deal with was **INEXPFTE** which is described as “Instructional expenditures per full-time equivalent student”. I found one case that had an institution report a value of 300,000,000. And comparing to the data present in the rest of that institution’s values, I determined something was off. Here is the boxplot for this variable’s distribution:

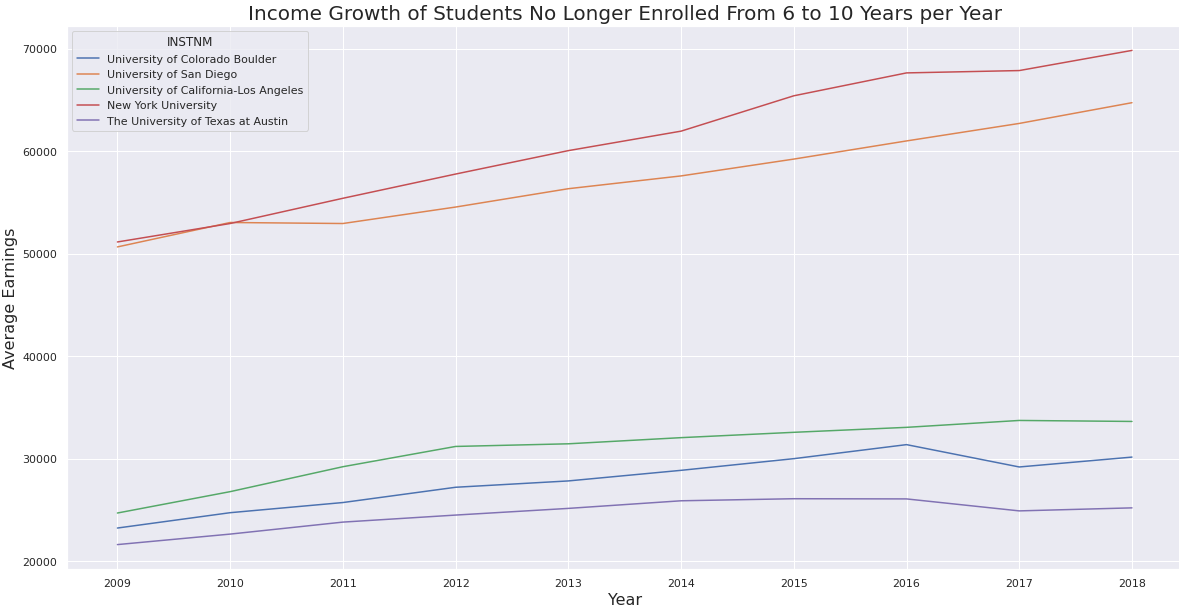
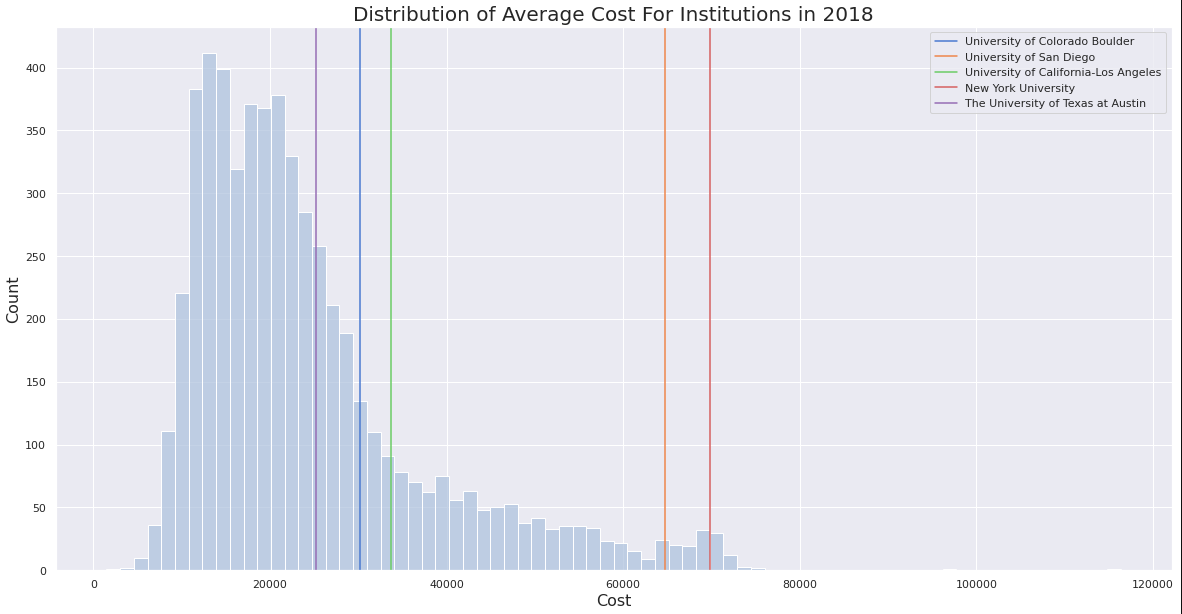
To address this, I wanted to replace each value over 20,000 with the mean of the values for each instituion that were less than 20,000, but I couldn’t figure it out to save my life. Instead, I replaced all values over 20,000 with NaN. I know this isn’t the best but it did mostly fix the outlier problem:

Data Visualization:

I stated that I wanted to create a tool that could display useful charts for students interested in applying to college to help them pick which one to attend. The overall idea here would be that an individual could enter in a couple of institutions they would like to attend and what degree they are interested in pursuing in to a tool such as a webapp. From here the tool would display several charts using the data. My first plot shows the percentage of degrees awarded in Engineering over the years for 5 different colleges:

This chart can show people which instituions focus more in their desired field of study. It also can show whether a university seems to be increasing or decreasing their focus on that field of study over the years. My next set of plots, on the following page, show the percentages of people working who are no longer enrolled after 6 years and 10 years and have jobs for each college. These plots are then side by side with the average earnings a student makes the same years after being enrolled. Below all these plots is one that shows the average earnings growth between students who are no longer enrolled after 6 years and those who are no longer enrolled after 10 years. An interesting note here is you can see the recession hurting a lot of the values between 2007 and 2009. An individual can compare colleges and infer which ones have a good record for students getting jobs after college. Someone can also see which ones seem have graduates getting paid the most. They can also see trends for each of these ideas by looking at the progression over the years.



The next plots show the increase of average cost for each selected college and a distribution plot for average institution cost in 2018 and where each selected college falls within:

Model:

I chose to use create a few DataFrames to generate a Random Forest Regressor model and test for accuracy. The first consisted of the variables chosen correlation to completion rate including without dropping null values. The second consisted of variables I hand picked without dropping null values. The next two were derived by dropping null values for each of the previous DataFrames. The First two couldn’t run with the Regressor Model because of the Null values. However, I got the following results for the two that had dropped null values:

Correlation\_No\_Null:

* Mean Absolute Error (MAE): 0.04137272413793107
* Mean Squared Error (MSE): 0.003246414794241384
* Root Mean Squared Error (RMSE): 0.056977318243678195
* Mean Absolute Percentage Error (MAPE): 0.06463450024470295
* Explained Variance Score: 0.8909981721984553
* Max Error: 0.1312519999999997
* Mean Squared Log Error: 0.0011845133759998251
* Median Absolute Error: 0.025043000000000037
* R^2: 0.880187712610196
* Accuracy: 93.54

HandPicked\_No\_Null:

* Mean Absolute Error (MAE): 0.08772994411764705
* Mean Squared Error (MSE): 0.013273031615126468
* Root Mean Squared Error (RMSE): 0.11520864383858734
* Mean Absolute Percentage Error (MAPE): 0.2814460575108755
* Explained Variance Score: 0.6786552239621412
* Max Error: 0.34266299999999994
* Mean Squared Log Error: 0.006481341713847337
* Median Absolute Error: 0.06609349999999986
* R^2: 0.6719127164655458
* Accuracy: 71.86

The Correlated values generated a much higher accuracy value. I would talk more about the rest of the random forests I generated and how I filled in null values as well as hyper parameter tuning but I have run out of time.

What I would do if I could continue the project:

* I would reach out to experts on the data to help assist me in filling null values.
* I would test different models as well as Random Forest Regressor such as a Polynomial Model
* I would test this on better hardware to run through more hyper parameter tuning variables.

Overall it’s hard to say whether I was successful or not since the best result only guessed across 114 rows out of the original 163,331.