University of California, Berkeley  
School of Information  
W200, Introduction to Data Science Programming

Thursday 630pm, Spring 2022, Tracy Huang

Project 2 Proposal

**Team member names:**

* Tatianna Martinez, Nicholas Lee, Che Guan, Judge Hiciano

**Name of GitHub Repository:**

* UC-Berkeley-I-School/Project2\_Martinez\_Lee\_Guan\_Hiciano
* <https://github.com/UC-Berkeley-I-School/Project2_Martinez_Lee_Guan_Hiciano.git>

**Primary Dataset to Analyze:**

* This data set contains information about and aggregated data for the institutions across the United States from the years 1996 to 2020. Information includes, but is not limited to, graduate rates, degrees granted, loans disbursed, enrollment rates by institution.
* We will be using the “MERGED\*” csv files, excluding any “FIELDOFSTUDY\*” csv’s. The field of study csv’s contain information about student income 1 year post graduation and cumulative student debt at graduation that are not of interest at this time. Adding the field of study datasets would also make this field extremely large and difficult to work with on our current equipment.
* The data dictionary in the “data.yaml” file that comes with the zip folder provides a dictionary for descriptions for the encoded columns.
* A technical document provides additional information about columns such as “CONTROL” which refers to the type of institution, public or private.
* These documents have helped us to narrow down the data set from 2990 columns to ~206 columns based on patterns in the column names.
* Link to dataset:<https://collegescorecard.ed.gov/data/>

**Supplemental datasets, if any, to complement your primary dataset - this means links, columns that you'll join on, etc.: None**

* None.

**What you plan to cover in the final report and how you plan to organize it:**

In the final report, we plan to discuss the findings of our analysis and what we learned. Our goal is to understand the factors that contribute to graduation rates at different colleges. We are interested in analyzing growing popular degree programs and completion rates of those popular programs. Also, we would like to compare completion rates to various different factors and find possible disparities and correlations. In addition, our team is interested in building a predictive model to determine graduation rates for certain schools based on various factors. Due to time constraints and data limitations, we are not sure if this will be possible, but we hope to answer this question.

Our report will be split into the following sections:

* Introduction
* Initial exploratory analysis
* Analysis
* Final Findings
* Lessons Learned

**Research Questions:**

1. What are the top growing degree programs over the last 5 years?
2. What are the associated costs?
3. How does the graduation rate change based on?
   1. Race/Ethnicity
   2. First Generation students
   3. Public/Private Schools
4. What is the predicted graduation rate for a particular school given certain factors?

**Initial Exploration:**

Shape of merged dataset: (170026, 2990)

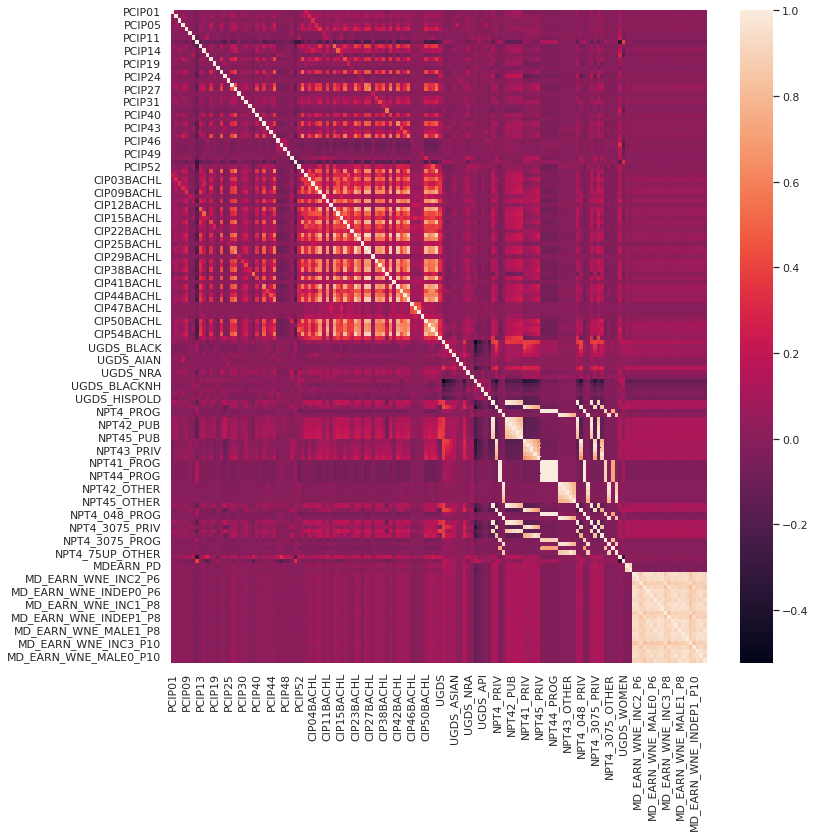
Refined dataset shape: (170026, 206)

Performing a descriptive statistical analysis using df.describe() for count, mean, standard deviation, min, quartiles, and max showed that a large portion of the dataset is missing values. This exercise also helped to identify which variables were categorical and continuous.

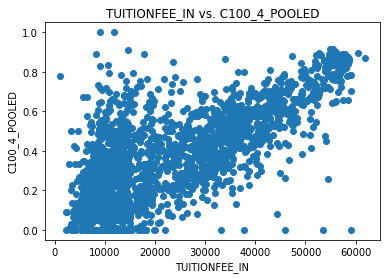


Snapshot of a correlation table from refined dataset. Correlations less than -0.7 and greater than 0.7 and were looked at to see if any strong correlations existed between variables. This was helpful for exploring correlations between columns that had a stronger relationship. Additional row were excluded to keep the proposal short.

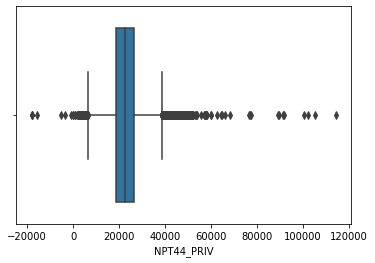
To visualize our correlations, a heatmap was created to further explore any linear correlations. There may be some potential variables to explore. Most of the correlations that are on some extreme end have related variables. Such as NPT4 variables all look at price for an institution based on a grouping variable like public school or family income.



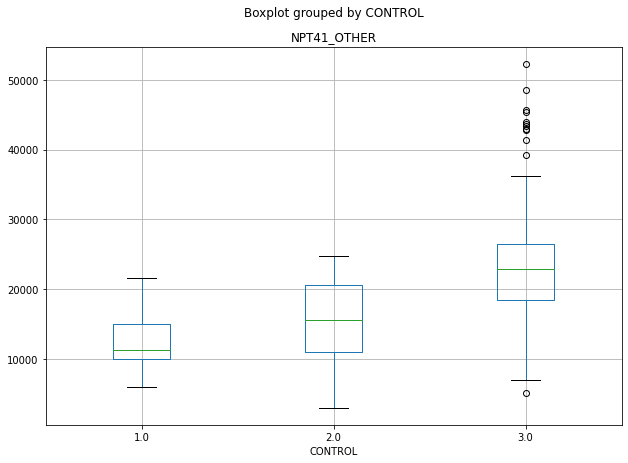
Upon further analysis some columns correlations were skewed by the data.



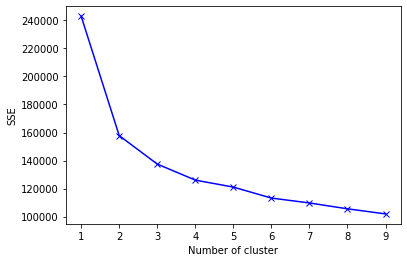
In this image above, the correlation was > 0.9 between pooled completion of a 4 year degree and in-state tuition. The pattern does not appear to be completely linear and is impacted by the quantity of data we have. Linear correlations will have to be investigated further before reporting with confidence there is a linear relationship.



To look at distribution of the variables in our refined dataset we looked at the boxplots. Example of a boxplot created is above. This boxplot refers to the average net price for $75,001-$110,000 family income (private for-profit and nonprofit institutions). The data related to price, cost or family income are well documented in this data set. Categorical variables such as the type of institution were identified to further explore the distribution of the cost of tuition for colleges.



Example of differences between average net price for the largest program at the institution for schools on "other" academic year calendars based on institution type. 1.0 is public, 2.0 is private non-profit, 3.0 is private for-profit. There appears to be an effect of program price based on institution, which makes sense based on these categories.



SSE is defined as the sum of the squared distance between centroid and each member of the cluster. As shown in the figure, as K increases SSE decreases. The key idea is to choose the value K at which the curve decreases abruptly. Performing k-means clustering on our refined dataset shows that we may be able to split our groups into 4-6 groups. We will use this information to explore the features that created these separate groupings.

**Some of the variables (column names) you intend to explore and what kind of insights you expect to glean:**

Potential columns to explore are below. Columns denoted with wildcard “\*” symbol denote that text comes either before or after the string and all columns with this pattern will be explored.

* Completion rate
  + columns:
    - C100\_4 : Completion rate at 4 year
  + Insights:
    - How many students completed their 4 year program
* Institution features:
  + Columns:
    - The columns for degrees are separated by degree name
    - UGDS: Enrollment of undergraduate certificate/degree-seeking students
    - HIGHDEG: Highest degree offered
    - RELAFFIL: Religious affiliation
    - ADM\_RATE: Admissions rates
    - ADM\_RATE\_ALL: Admission rate for all campuses rolled up to the 6-digit OPE ID
      * Identification number used by the U.S. Department of Education's Office of Postsecondary Education (OPE) to identify schools that have Program Participation Agreements (PPA) so that its students are eligible to participate in Federal Student Financial Assistance programs under Title IV regulations. This is a 6-digit number followed by a 2-digit suffix used to identify branches, additional locations, and other entities that are part of the eligible institution. [<https://ceds.ed.gov/CEDSElementDetails.aspx?TermId=15203>]
    - CURROPER: Flag for currently operating institution, 0=closed, 1=operating
  + Insights:
    - Information associated with the institution that will help to draw out features for admission rates
* Finances:
  + Columns:
    - COSTT4\_A: Average cost of attendance (academic year institutions)
    - COSTT4\_P: Average cost of attendance (program-year institutions)
    - TUITIONFEE\_IN: In state tuition
    - PCTFLOAN: Percent of all undergraduate students receiving a federal student loan
    - Family income are grouped
    - Might be able to indirectly measure income based on repayment rate of loans and disbursement amounts
    - \*DEBT\_MDN: Median debt for different demographic groups
    - \*\_EARN\_\* : Earnings of students working and not enrolled 10 years after entry
  + Insights:
    - The average cost of tuition for the institutions. This is likely to influence the number of students taking out federal loans. Family income may also influence the schools students choose to attend and how much debt they enter into during their program.
* Demographics:
  + Columns:
    - “UGDS\_\*” where the \* is the ethnicity
    - UGDS: Enrollment of undergraduate certificate/degree-seeking students
    - UG: Enrollment of all undergraduate students
    - PAR\_ED\* Parents education middle school, high-school, college
  + Insights:
    - Total enrollment for undergraduate programs at colleges and universities. Adding parent education as a contributing factor to graduation rates because it may be an indirect indicator of family attitudes towards school.
* First time Students:
  + Columns:
    - FIRSTGEN\* denotes a lot of the column names
    - NOT1STGEN\* to contrast first gen
    - PFTFTUG1\_EF: Share of entering undergraduate students who are first-time, full-time degree-/certificate-seeking undergraduate students
  + Insights:
    - Describe the factors that contribute to first generation students graduating their programs compared to non first gen students. These columns include information such as house hold income, the number of students that are first generation. It would might be helpful to explore the difference between full-time versus part-time.