Jared Schifrien, Samir Joshi, Jodie Wei
EECS 396 Data Science
June 13, 2018

Data Science Final Project Report

Our Data and Data Cleaning

Our data consists of four datasets: the College Scorecard, the Census dataset, a census Id to zip code map, and a Yelp college dataset.

The College Scorecard dataset is a dataset from the US Department of Education containing the set of all colleges and information they provide. For each college, there are over 1000 columns of potential data, ranging from education and test score statistics, majors offered, diversity statistics, family information including income, earnings after college, graduation rates, death statistics and more. Additionally, this information is provided for colleges from the years 1999 to present. The dataset can be found at https://collegescorecard.ed.gov/data/. While this dataset contains an extremely wide breadth of information, in practice many fields were null for many of the colleges and some columns were analogous to other columns, leaving information split into two columns. As a result, data cleaning and dealing with null values was imperative, removing all colleges with any null values would result in an extremely small usable dataset. Instead, we removed nulls and consolidated columns on as needed basis, either using OpenRefine or using a python script we wrote for cleaning this set. Additionally, columns we did not use were removed entirely, which was necessary due to PostgreSQL row size limits.

The census dataset is a dataset for all census tracks across America, and can be found at https://www.kaggle.com/muonneutrino/us-census-demographic-data/versionhttps://www.theverge.com//1. This dataset contains a variety of diversity and demographic information, in addition to income, transportation, employment, and poverty statistics. In contrast to the college dataset, since all fields are mandatory to be computed, this dataset contains all fields, which no missing entries.

The Census To Zip dataset is a small dataset containing a set of census Ids and the zip code they correspond to. This is necessary because the census is conducted on the county level. Each county has a unique census Id, which may contain/map to various zip codes. While we do not do any analysis directly on this, we use it to join the Census dataset, which contains census Id, and the College Scorecard dataset, which contains the zip code for each college.

The Yelp college dataset is a subset of the Yelp Dataset Challenge, contains locations, business categories, reviews, images and more, split into various datasets. This dataset can be found at https://www.yelp.com/dataset/challenge. From this dataset, we extracted colleges and their locations. We then matched colleges with their reviews from the review dataset within the

Yelp dataset. Through this process, our Yelp college dataset consists of a set of colleges and their corresponding Yelp reviews and ratings. We use this dataset in the machine learning and neural networks components of our analysis.

Insights

Traditional Analytics

Utilizing the college scorecard dataset we were able to aggregate overall statistics to get a big picture view of different college statistics. For example, we were able to aggregate the total number of graduates per year to learn that only about 19% of total college students graduate every year, even among 2-year programs instead of a perfect 25%-50%.

Another aggregation we conducted was the diversity rates of the US versus the diversity of all US colleges. Due to uncleaned data, we originally calculated very shockingly different rates between the US population and college population. To fix this we used OpenRefine to calculate diversity statistics for colleges with complete data and we found much more intuitive results. We learned that while about 9% of the US population is African American, about 12 percent of the college population is African American and while only 1.2% of the US population is Asian, 2.5% of the college population is Asian.

Furthermore, joining the census data with the college scorecard data we were able to learn interesting facts such that colleges located in areas with the highest income per capita are ones located in New York City and that top private schools, medical schools, and law schools have the highest average faculty salaries. Future work could be to remove other contributing factors such as normalizing salaries to account for standard of living or the ratio of degrees awarded so that we are not comparing specialized medical schools in rich urban areas to large state undergraduate schools in middle-income areas.

Workflow Analytics

Utilizing Spark, we were able to calculate a lot of statistical queries on the datasets. For example there does not seem to be much of a correlation between the rate of students without loans and the income per capita of the zip code in which the university is located. This seems counterintuitive that people don't need to take out more loans to go to a university in a higher cost of living area. We can guess that this is because the cost of going to a school fluctuates a lot more based on more impactful factors such as private/public, financial aid of the school and spending habits of students than solely on the cost of living of the general area. A school such as University of Chicago or Yale are in less expensive areas than Northwestern, but generally cost about the same.

Another Spark outcome we found was that as the rate of engineering degrees awarded (pcip24) increases, the average earnings 6-years after graduation (MN_EARN_WNE_P6) increases as

well. On the flipside, we were not able to find a positive or negative correlation between liberal arts degrees awarded (pcip11) and future income. Spark was also able to visualize these rates and we can see that liberal arts degree rates are uniformly distributed across income levels whereas engineering degree rates are very skewed to the right. This makes sense if we consider that most colleges offer liberal arts degrees as the most popular degree awarded, but a college that awards a lot of engineering degrees tends to specifically be a technical college where a majority of students go into high earning technical professions.

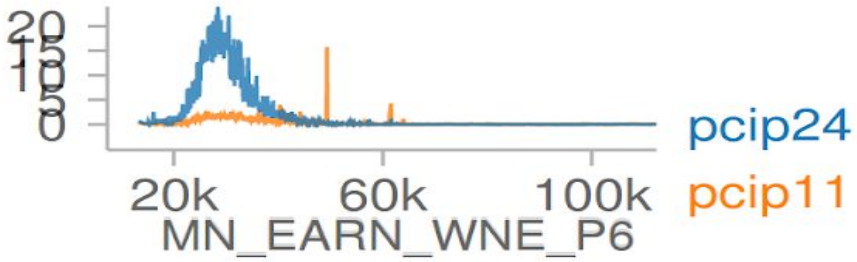

Figure 1: Rates of Engineering Degrees Awarded(pcip24) and Liberal Arts Degrees awarded(pcip11) vs Income 6 years after graduation (MN_EARN_WNE_P6)

Machine Learning and Tensorflow Insights

For machine learning, we used Spark MLlib to investigate if it was possible to predict loan rates given a distribution of majors, post college income given faculty salaries, and admission rate. To investigate loan rate, we constructed a linear regression model, with the percentages of engineering, liberal arts, and business degrees awarded as features. Using these features, the model achieved an RMSE of 0.188 with weights [0.22048810362,-0.275437568969,0.26231465493], which shows that our model was able predict loan rate, which was somewhat surprising.

However, in predicting admission rates, we found no correlation between and were unable to successfully predict admission rate given post college earnings, diversity info and SAT scores, which an near 0 correlation. One explanation for this, beyond no correlation existing in practice, could be the lack of substantial data for this model. In the College Scorecard, the number of colleges that reported each of these columns and admission rate was surprisingly low.

Finally, we found a strong correlation between faculty salary and the earnings of students after college. We were pleased but surprised by this relationship, and the graph of our model and the data can be seen below in Figure 2.

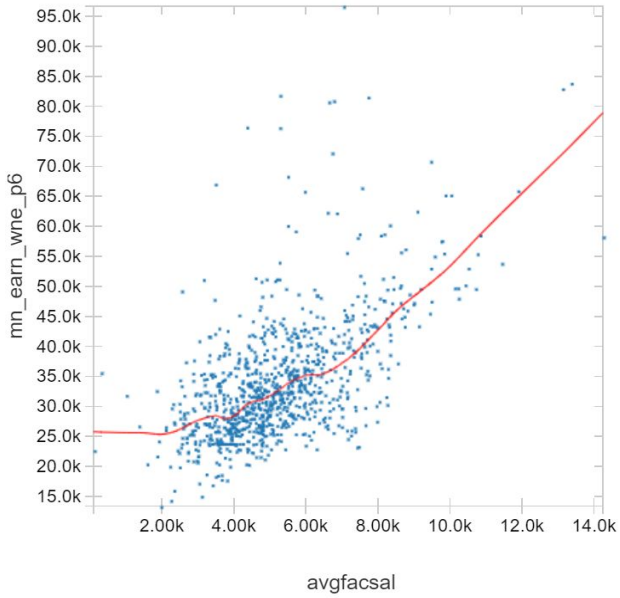

Figure 2: Average Faculty Salary vs Earnings of Students
6 Years After College

Using neural networks, we then created neural networks to address our questions for both structured and unstructured data. First, we used an LSTM model from Keras on Yelp reviews of colleges to determine which colleges were rated most favorably and which were not. We discovered that L'Academie Cafe at the Scottsdale Culinary Institute had a perfect rating, with all of the reviews being positive, while Belmont Abbey College was least favorably viewed with no positive reviews.

Another tensorflow predictive model we created was seeing if we could predict factors such as faculty salary, future income or graduation rates based on solely on the rates of popular majors such as CS, Engineering, Law, Biology, Math, Psychology, Arts, History. Unfortunately, this did not prove to be useful and had a very low accuracy score because the sparseness of the input data. Very few colleges report this data, and when they do they tend to report some either based on schools (engineering, law, arts) or specific majors (Biology, history), but rarely both.

Additionally, using Tensorflow, we were able to perform nearest neighbor searches that match a student with certain statistics to the college that best aligns with those statistics. For example, a student with the [600, 600, 600, 25, 25, 25, None, 4] for the categories (instnm(name),satvr75,satmt75(SAT Math scores),satwr75,actwr75,actmt75,acten75(ACT),costt4_p (Tuition Cost),sch_deg(Degree)) matches best with Pacific College of Oriental Medicine-New York. While there are other factors that a student typically considers that are not possible for this model to capture, such as location

and major preference, this nearest neighbor model could generates a set of baseline matches that a student could use to understand their fit.

Visualization Insights

We used Tableau to visualize different attributes from our dataset against each other to identify trends and also to better understand our data.
Our first focus was on undergraduate student diversity to better understand the racial demographics of the institutions in our dataset. We found that in general, as the undergraduate student body increases in population, representation for all racial groups decreases except for the white demographic, while minority groups are more represented in smaller institutions (fig. 3). This is also supported in fig. 4, where several racial groups are plotted against each other. The majority of colleges have a much higher percentage of students grouped as white; the graphs with this demographic have the majority of points clustered around that axis(on the bottom or on the left). Additionally, there was an unexpectedly low number of colleges in which the percentage of students grouped as Asian were higher than the percentage of any other demographic. This serves to support the fact that higher education remains predominantly white.

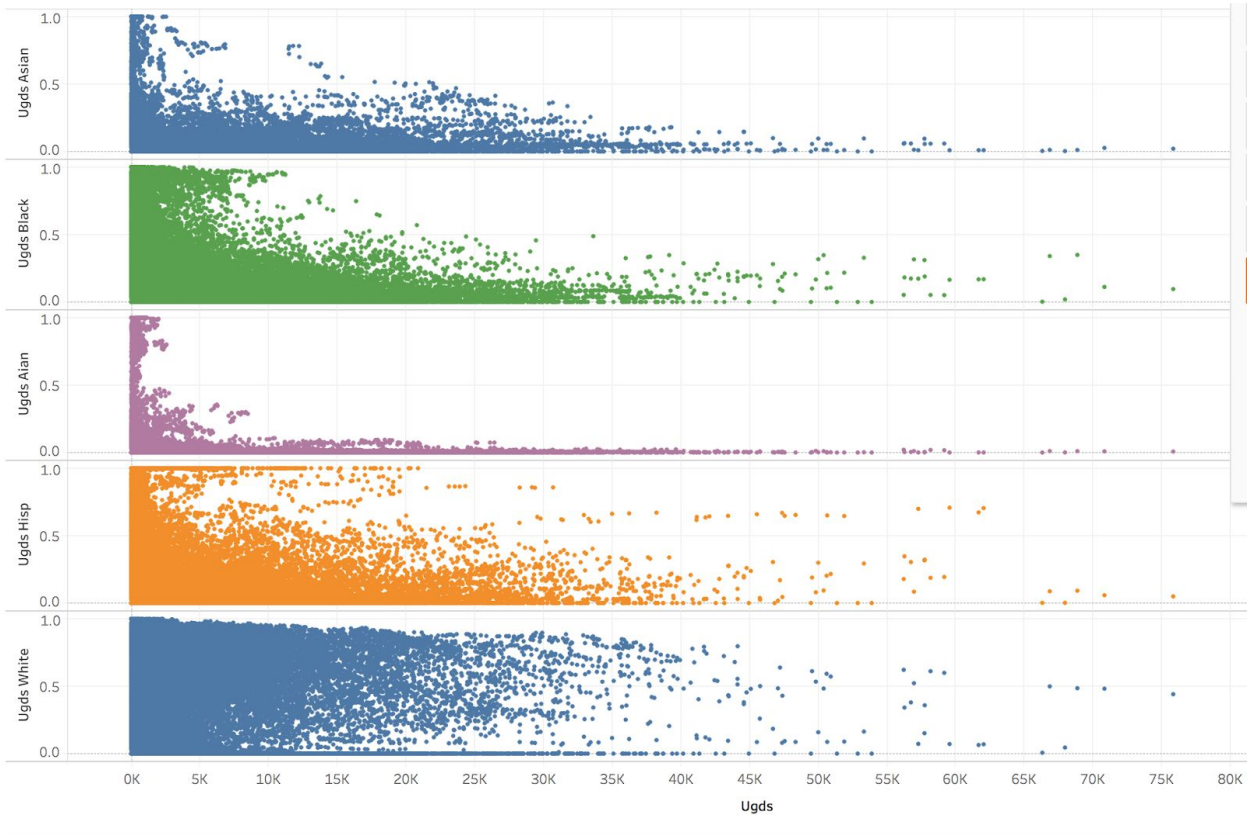

Figure 3 (X axis: undergraduate student population. Y axis: percentage of white, Hispanic, American Indian, black, or Asian undergraduate students)

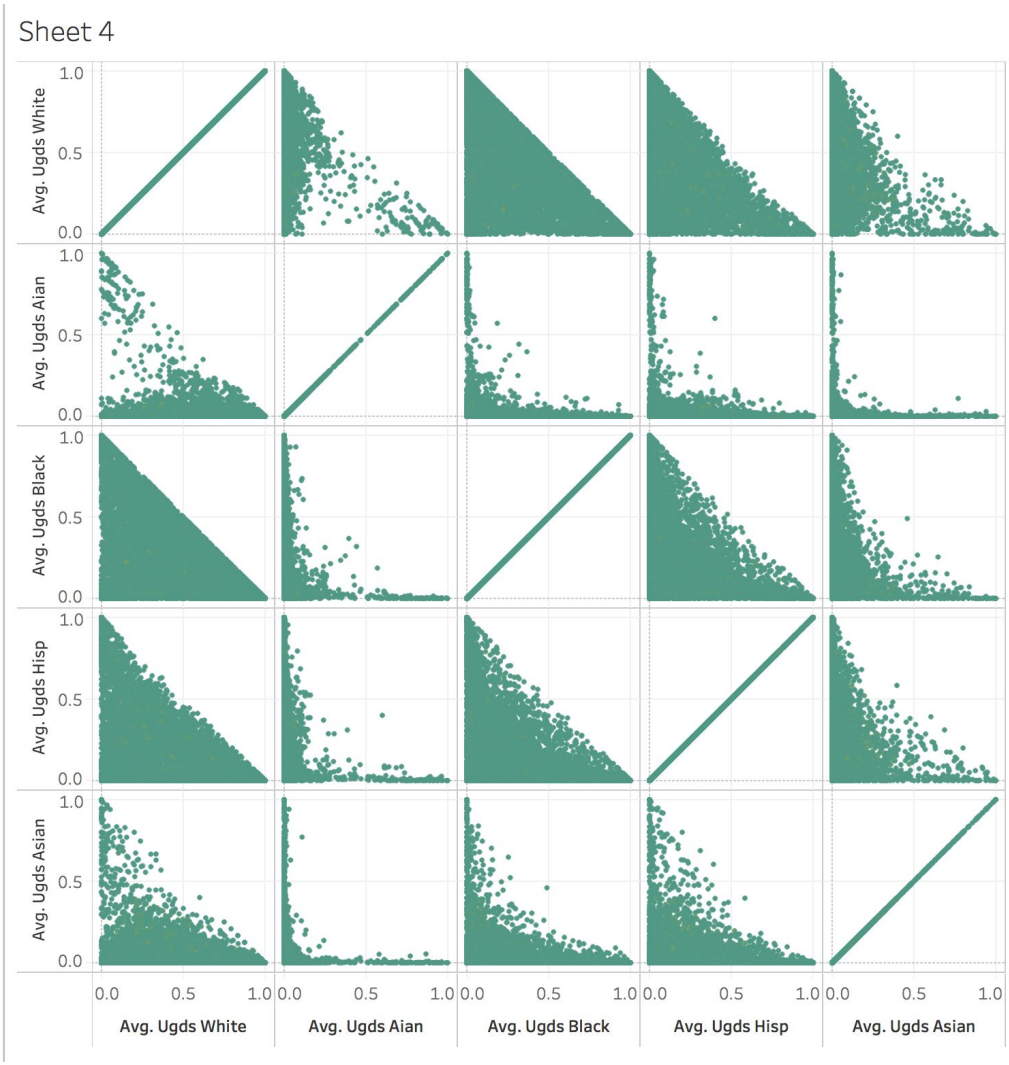

Figure 4 (Axes are the average percentages of racial demographics in the undergraduate student populations)

We wanted to visualize and compare the change in family income and the cost of education over the years. Although the average graduating debt has increased over time, it seems to match up with the increase of family income over time; however it is interesting to note that the increase in family income plateaus while debt steadily increases (fig. 5). We generally expected average debt to increase at a much higher rate over the years, but the data could also be skewed since the data averages from all family wealth levels, so huge debt differences between low- and high- income families could be all averaged together. However, the increasing debt along with plateauing income are concerning observations.

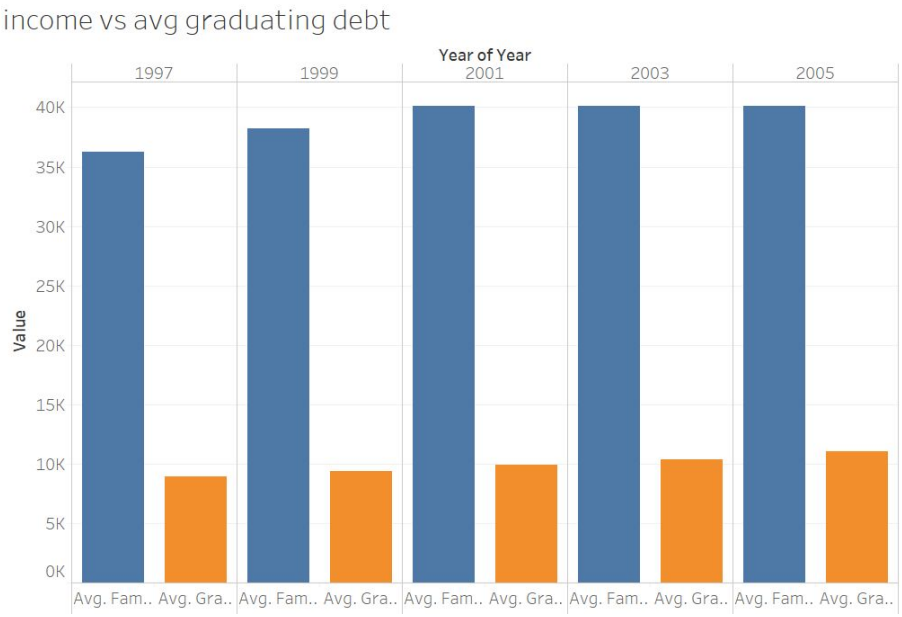

Figure 5 (Blue: average family income. Orange: average graduating debt)

Finally, we wanted to examine average post-graduate incomes, in order to better understand the outcomes of attending higher-education institutions. Surprisingly, they have not increased over time as expected, and the higher-percentile demographics have actually suffered from a greater decrease in income than the lower-percentile demographics (fig. 6). This may be due to the impact of the 2007-2008 financial crisis on certain white-collar jobs that would generally place someone in the upper percentiles, but the most interesting takeaway is that post-graduation earnings are not as optimistic as expected.

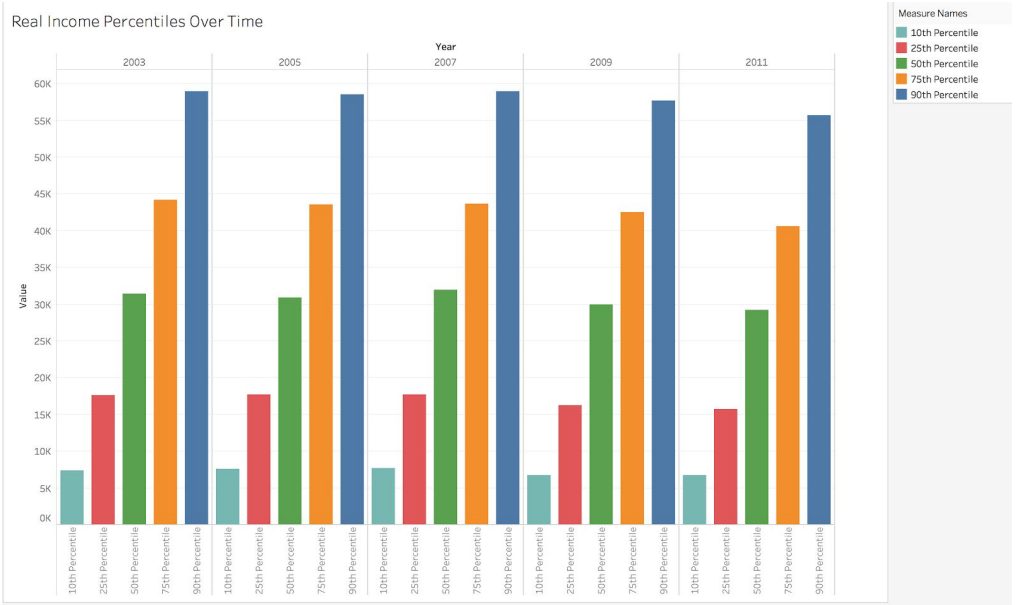

Figure 6 (Teal: 10th percentile, Red: 25th percentile, Green: 50th percentile, Orange: 75th percentile, Blue: 90th percentile)

Outstanding Questions

Based on our analysis over the past quarter, we are still curious about:

1. Post-graduate debt and earning statistics along field of study, race, gender, and income percentile demographics. Since information about debt and earnings along these lines was not provided in the dataset, we were unable to analyze these statistics on a more precise level.

2. While we were able to calculate static diversity statistics, we would like to use per college and per year diversity rates to see both where rates are headed, where in the country the most change is happening, and how these rates correlate to other factors we are looking at such as income and graduation rates.

3. Income at various time periods after leaving college and the factors that influence them. While data exists for certain time periods, such as 2 or 6 years after college, the columns of the dataset included 1, 5, and 10 years, many of which were null. In would be interesting to know as students continue on in the workforce which of the factors about the college was most impactful.

Our Experience with the Tools

Tools that worked well

Of the tools we used this quarter, we found OpenRefine, Spark and Spark MLlib, Tensorflow and Tableau to be the most useful, the more user friendly, and the most powerful. First, OpenRefine was extremely useful for cleaning and consolidating the College Scorecard dataset. Using OpenRefine, we were able to remove rows with empty columns for certain tasks, and filter for values such as greater than 0 with ease. Additionally, OpenRefine made it easy to create new columns based on other columns, such as generating population counts for ethnicities from percentages.

With Spark, loading in and interacting with data was surprisingly straightforward, especially after our experience with Impala (see below in the “Tools that didn’t work well” section). Additionally, due to the use of Dataframes and the Jupyter Notebook style UI, using Spark seemed familiar to us even while it was new. In particular, we found the graph and graph modifications, shown above, to be useful ways to understand the data and understand basic trends.

Tensorflow and Keras are tools that we have used before, and worked as expected and promised. It was straight forwards to get started and allowed us to tweak our models to achieve promising results for tasks such as sentiment analysis using transfer learning and k-means clustering.

Similar to Spark, we found Tableau accessible and quick to load data and get up and running. This allowed us to focus on trying out various visualization to view our data in different ways to better address the questions we had.

Tools that didn't work well

However, we did find that Impala and parts of PostgreSQL gave us expected difficulty. Additionally, we originally had some trouble due to the missing values in our College Scorecard dataset. With Impala, the experience of needing the Cloudera VM (to use/ try it out for free) was difficult and slow. Additionally, even after gaining permissions to all of the necessary files for all users, the command line interface in the VM failed to properly add the data to the VM. However, the local website UI was able to load the data, after which we were able to perform our queries. Once the data was loaded, we were able to perform our queries. However, due to the lack of performance in the VM and the difficulty of loading data, we did not find Impala to be worth the hassle.

With PostgreSQL and the College Dataset, the presence of thousands of columns resulted in problems loading into PostgreSQL, which has a limit on the size of a row in bytes. This required some of the dataset to be removed in order to use PostgreSQL. Additionally, due to the missing values in the College Dataset, queries were more complex than expected.

Conclusion

Overall, we found it fascinating to explore the intersection of the College Scorecard and Census datasets to address many of the questions we had about secondary education in America. In doing so, we learned how to approach data science and use many of common tools.

One of the big drawbacks we had to address was the completeness of the college dataset. Since the information was self-reported by each institution, many of the fields were left unanswered, which made it difficult for us to examine certain attributes across all institutions. Despite missing a lot of data, there were also a huge number of very specific attributes such as "Percent of dependent students with status unknown within 3 years at original institution" that weren't very useful for our purposes, so we had to remove them.

Our key takeaways are the importance of good questions and sufficient data over the importance of technologies Additionally, we learned that patience is required to try various models and play with the data to achieve results that address the questions.