2022

Data Wrangling and Exploratory Data Analysis Project Summary

NATIONAL CENTER FOR EDUCATION STATISTICS DELTA COST PROJECT RAW DATA

JAKE GREENBERG

Data Preparation and Cleaning

Initial Import

For this project, I've opted to work with the Delta Cost Project data from the National Center for Education Statistics. As you will see, I worked with very specific sections of data provided, as the dataset contains 87560 rows, and 974 columns. This results in over 85 million potential elements, which I deemed too large to feasibly work with.

The first thing I noted is that the data contained international data for the entire country. To limit my scope, I subset only the rows that relate to the state of Florida. This limited my rows from 87.5 thousand to only 4.16 thousand. Much more manageable, and faster to compute.

For the next part of preparation, I decided to select specific variables that I wanted to utilize. Many variables are computed numbers based off of other values. To keep the analysis straightforward, I opted to use…

Selected Variables and Data Dictionary

- '**academicyear'**: The Academic Year the row was recorded for. The spreadsheet I used started at 2000, and went until 2012.
- '**instname'**: The name of the Institution the row was recorded for. As mentioned, this subsection contained only institutions in Florida.
- **'nettuition01'**: Net tuition fees and revenue. Subtracts institutional grant aid.
- '**net_student_tuition'**: Net tuition directly from students. Subtracts all grants from nettuition01.
- **'fte_count'**: The number of Full-Time Employees each institution has.
- '**unrestricted_revenue'**: Revenue in unrestricted funds
- **'restricted_revenue':** Revenue in restricted funds
- **'studserv02'**: Salary information for Student Services employees.
- '**instsupp02':** Salary information for Institutional Support employees.
- '**opermain02':** Salary Information for Facilities / Groundskeeping employees.
- '**auxiliary02':** Salary information for Auxiliary Employees
- '**applicantcount':** Amount of applications received by an institution.
- **'admitcount':** Amount of applications admitted by an institution.
- **'enrollftcount'**: Amount of accepted students that enrolled full-time at an institution.

These 15 variables were selected out of all 947 because they had significant response, and categorized areas of interest that I wanted to look at further.

Missing, Incorrect, and Invalid Data

With a dataset this large, it is very important to note that there is a LOT of missing data. University's are not required to submit all of the data that is being requested over the 947 variables, and many of them haven't. For the purposes of numerical graphs, they are shown as blank.

Because we are not doing any predictive analytics, I've decided that outliers should be left as is. If I was opting to create regression models or perform other types of forecasting, this would be a much bigger concern.

Analysis and Visualization

Because this is not a full report, but rather an open ended analysis, most of the visualizations can be found within my Jupyter Notebook. But, for the conclusion and things like that…

Narrow Scope

To keep the scope narrow, and work with as little missing data as possible, I specifically subset the Florida data to **only** the institutions that are apart of the State University System (SUS). This means that most of the data is complete, and there are not a lot of missing values.

Additionally, this means that I do not have to worry about several hundred institutions for the purposes of visuals, and I can only work with the few that were selected. This dataframe was called 'sus_data'.

Interesting Conclusions

Application Count, Acceptance Rate, and Enrollment Rate

I made many graphs relating to different variables in my data. Not all of them made the cut, but a lot of the remaining ones are interesting.

Notably, I found the graph titled "Student Application Acceptance Rate Percentage" highly interesting, as there is a trend from 2000-2012 where the acceptance rate has been steadily decreasing in the SUS system. All SUS schools saw their acceptance rate drop by 10% or more in that span of 12 years.

Additionally, while their acceptance rate went down, some of the institutions saw their enrollment rates decrease as well. (I expect that this is due to a major housing shortage issue in the mid 2000s, but that is a discussion for an entirely different paper.)

After mutating an "is_SUS" variable for the SUS institutions, I tried to check the reported applicant counts for the SUS vs non-SUS schools. There was too much missing data for the non-SUS schools, so I opted to disregard and abandon that little project. I had a similar issue with Net Tuition.

Salaries for University Services

With very few exceptions, I was surprised to see a significant increase in salaries for various university departments. Most departments received major increases. What was odd however, was how some specific departments saw relative decreases at their respective institutions. All of this can be viewed in the "[Department] Salaries by State University by Academic Year" tables in the Jupyter-Notebook.

A Brief Conclusion

Overall, I think the largest issue with this admittedly insane dataset is how little consistency there is across the submissions from universities and other schools.

If all of the data was complete and compiled, I believe that there would be significant strides in the way that universities handle their applications, tuition, finance, etc.

The data that **is** present is useful, but often suffers from the lack of continuity between years. This became more noticeable as I sub sectioned the data more and more.

Maybe worth revisiting, but I'd rather find more complete data on a more specific topic to get more use out of it.