Capstone 1 | Milestone Report

Problem Statement

There are a lot of reasons students seek a college education. One of these is to earn a living afterward. How can a prospective student choose the school that has the highest average graduate earnings after school?

Intended Audience

Based on the findings of this analysis, students should be able to see which types of schools have the greatest chances of increasing their after school earnings. All other factors being equal, a student will be able to rank their top choices by earnings potential.

Dataset

College Scorecard - *https://collegescorecard.ed.gov/data/*
This dataset tracks 800 variables for every secondary school in the United States. For this study, we I will be focusing on the 4 most recent years of school data (2013 - 16)

Data Wrangling

I collected the dataset by calling the data.gov api. This involved querying the api using a rate limit and parsing the resulting json. I specifically queried for the latest scorecard data for all schools in the United States.

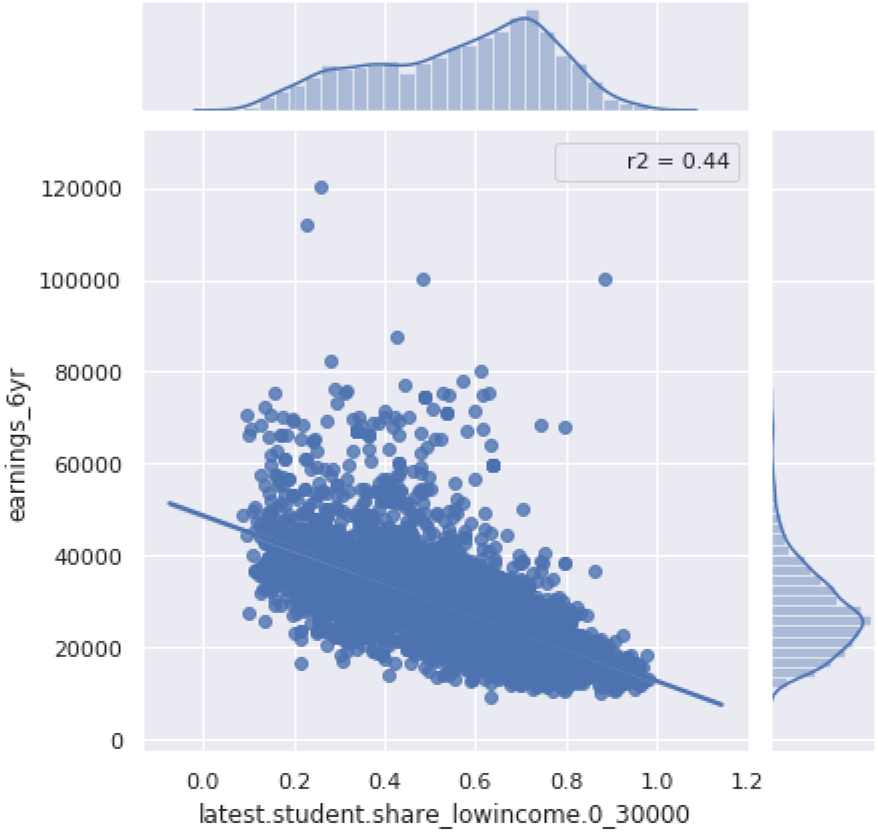
Once I had parsed the json file, I saved as a csv. I then plotted my target variable (median income of 6 yr graduates) on a histogram to view its distribution. I also plotted all the other features available to identify any outliers or trends in the data. I then calculated each independent variable's mean and imputed to any null values.

Initial Findings

Summary

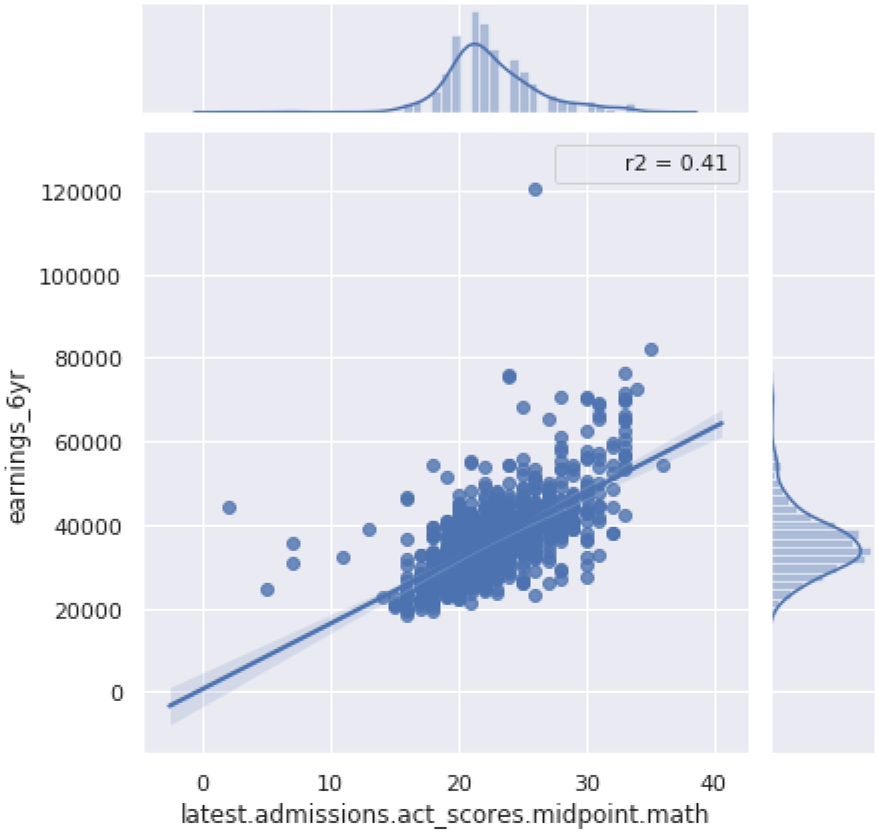
Among all the variables, the most strongly correlated with our target variable (6yr Earnings) were:

1. **Latest.Student.Share_lowincome.0_30000** (r^2 = 0.44):



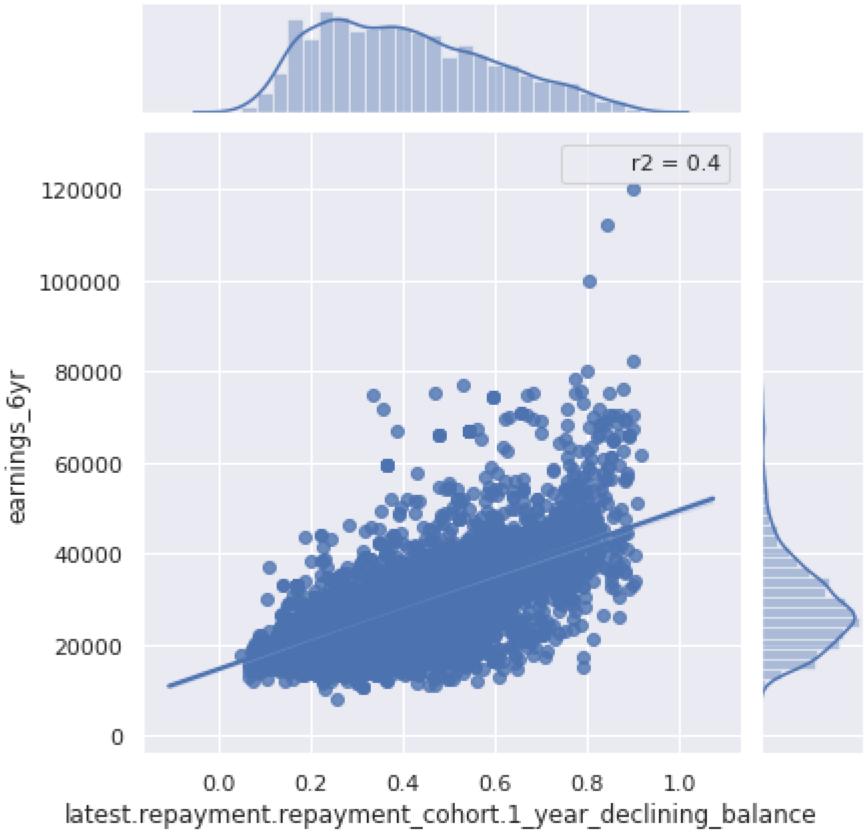
This feature was negatively correlated with our target variable, suggesting that as schools' percent of low income students are higher, graduates' 6 yr earnings were lower.

2. **Latest.Admissions.act_scores.Midpoint.Math** (r^2 = 0.41):



This feature was positively correlated with target varirable. This suggests that as schools' ACT Math scores were higher, graduates' 6 yr earnings were higher as well.

3. **Latest.Repayment.Repayment_cohort.1_year_declining_balance** (r^2 = 0.40):



This feature was positively correlated. This suggests that as schools' percent of graduates with declining loan balances 1yr after graduation is higher, graduates' 6 yr earnings are also higher.

Each of these relatioinships were statistically significant (resulting pvalues were all < 0.01%).

After conducting aPCA, it became apparent that many of the features were duplicative. I used a correlation matrix to determine which variables were strongly correlated with each other and removed duplicative columns.