Colleges that make the American Dream a reality

The American Dream is founded on the ideal that children can attain a higher standard of living than their parents. For many young adults, a college education is the path to the American Dream but not all colleges are equal. Every year journalists and economists (USA today, the Wall Street journal, Forbes etc.) rank the best undergraduate colleges in the USA. While the top colleges are associated with good employment outcomes, they are expensive, highly selective and largely exclude lower-income students. With over 2000 accredited institutions, many students may not know what their options are, especially students from lower income brackets. Further, many colleges lack the opportunities and resources that can help lower-income students succeed. The primary goal of this initiative is to first, explore which institutions promote upward mobility and then develop a predictive model that can help low-income students decide where to go to college to increase their earning potential. I plan to use machine learning to create an application that matches institutions (top 10) with students based on their SAT scores, family income and location.

The data

Data for this project will be acquired from a combination of 1) the College Scorecard data, made freely available by U.S. Department of Education, and 2) the Equal Opportunity Project. The College Scorecard database contains data on all higher educational institutions in the US, including (a) the characteristics of the institutions, (b) academic offerings, (c) the median and mean values of student SAT scores, family income, earnings amongst other metrics, and (d) characteristics of the neighborhood where the institution is located from the census data. The Equal Opportunity Project has dataset on the economic mobility per college including a mobility ranking. Their metrics were calculated from tax records indicating individual income levels and the institutions attended.

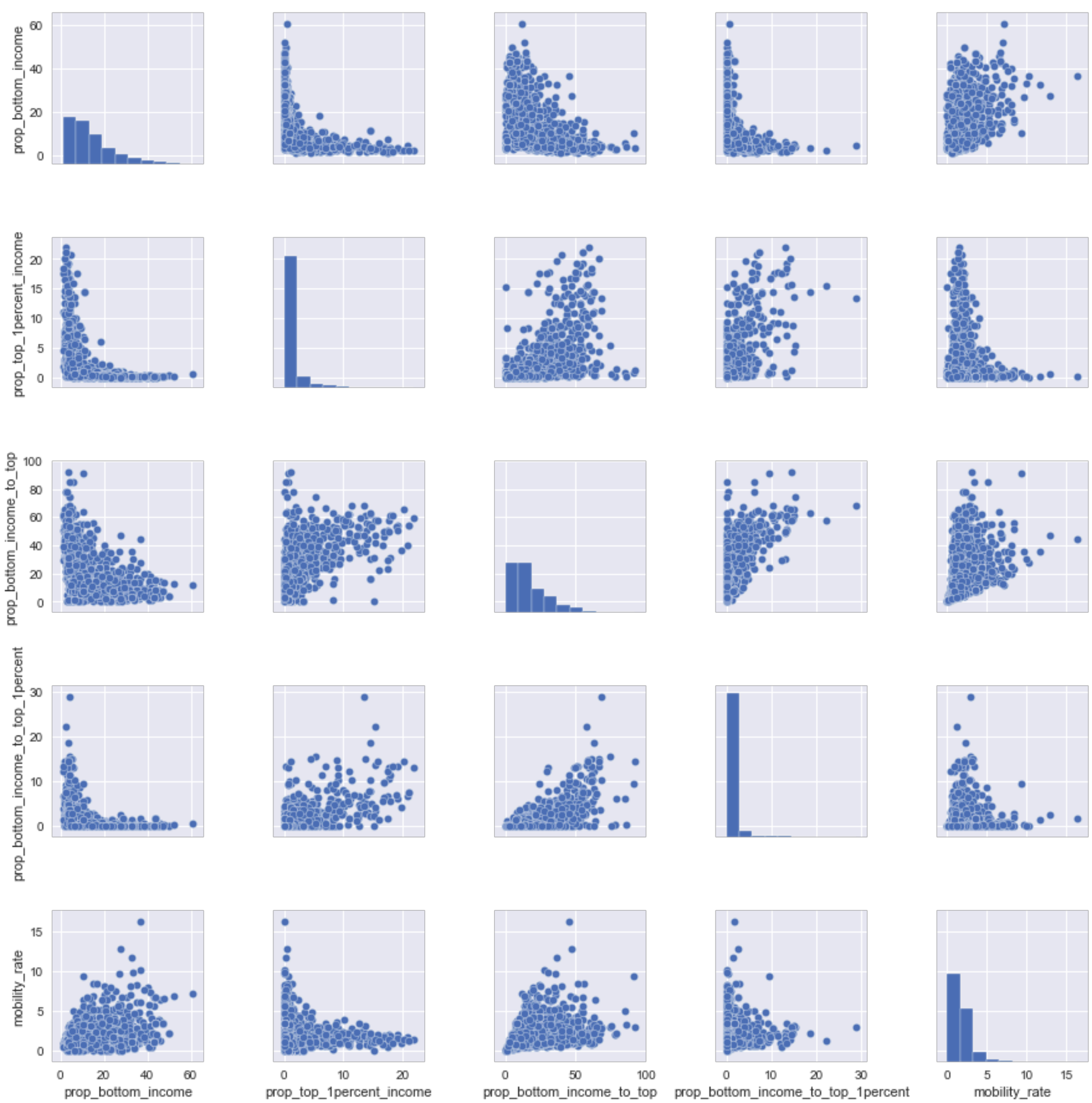
Data Wrangling

- Imported 120 variables using College Scorecard card's API.
- Used the College Scorecard dictionary (https://collegescorecard.ed.gov/data/documentation/) to generate the API.
- Variables of interested in a .csv file 'allVariables.csv'
- Made two API requests, 1) for institutional characteristics for all years 2) student data for 15 years.
- Plotted missing data on a heat map to look for chunks of missing data
- Focused my analysis on data from 2013, because it is the most complete, recent dataset.
- Neighborhood census data is missing for 2013 and imputed from 2005.
- Replaced integers of categorical variables with meaningful strings from the data dictionary
- Imported the Equal Opportunity Project mobility ranking data from the website. These data are clean so I did not change anything.

- Concatenated the College Scorecard data and the Equal Opportunity Project data using the OPE_ID, which is a unique institution ID in both datasets. However, some multi-location institutions (e.g. all the Universities of Illinois) share the same ID. The Equal Opportunity Project assigns a single value for these institutions so the main dataset has repeated values across institutions at multiple locations. Other metrics from the College Scorecard data are not repeated across institutions at multiple locations. Not sure how to address the non-independence of data here.
- Replaced column names with strings that are more intuitive, shorter and pythonic.

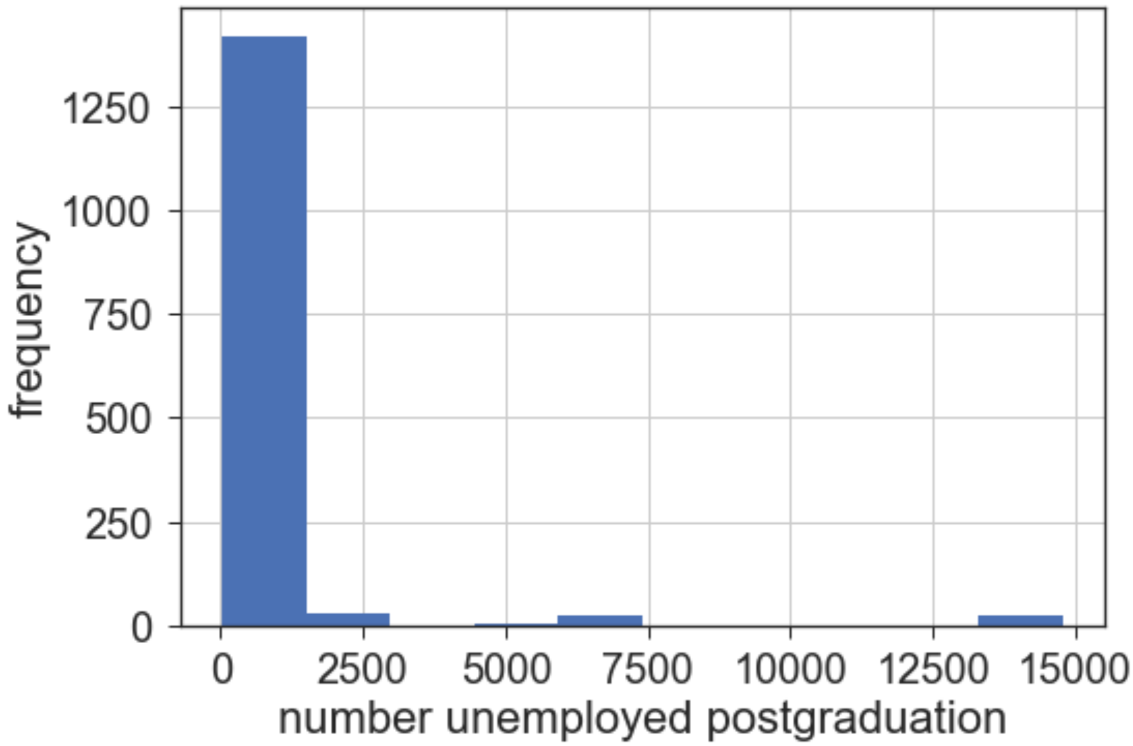
Exploratory Data Analyses

- Pair plots to explore the data and look for any major problems, outliers.

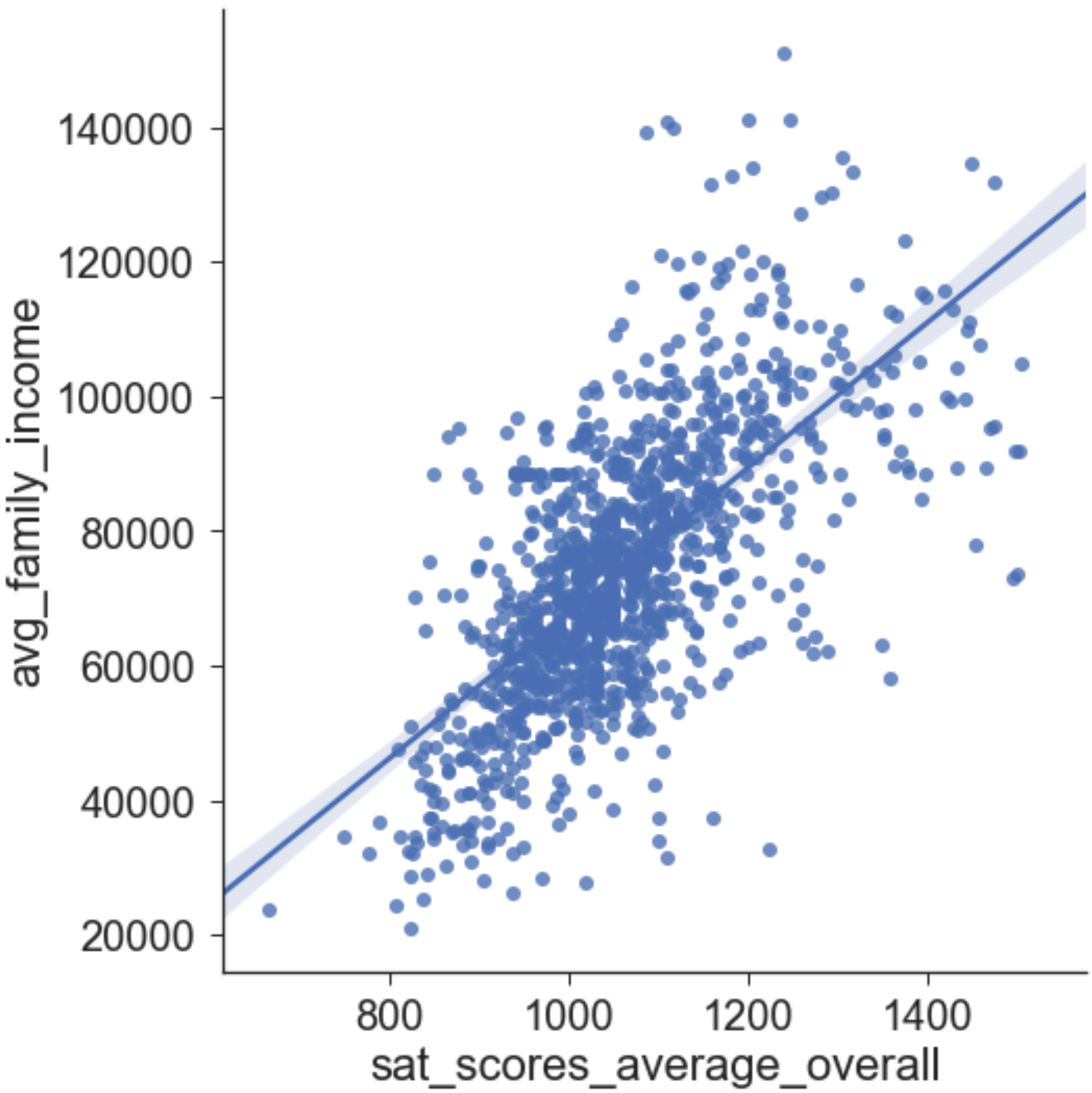


- Many of the metrics are proportions and are non-normal (sqrt transformed these for inferential statistics)

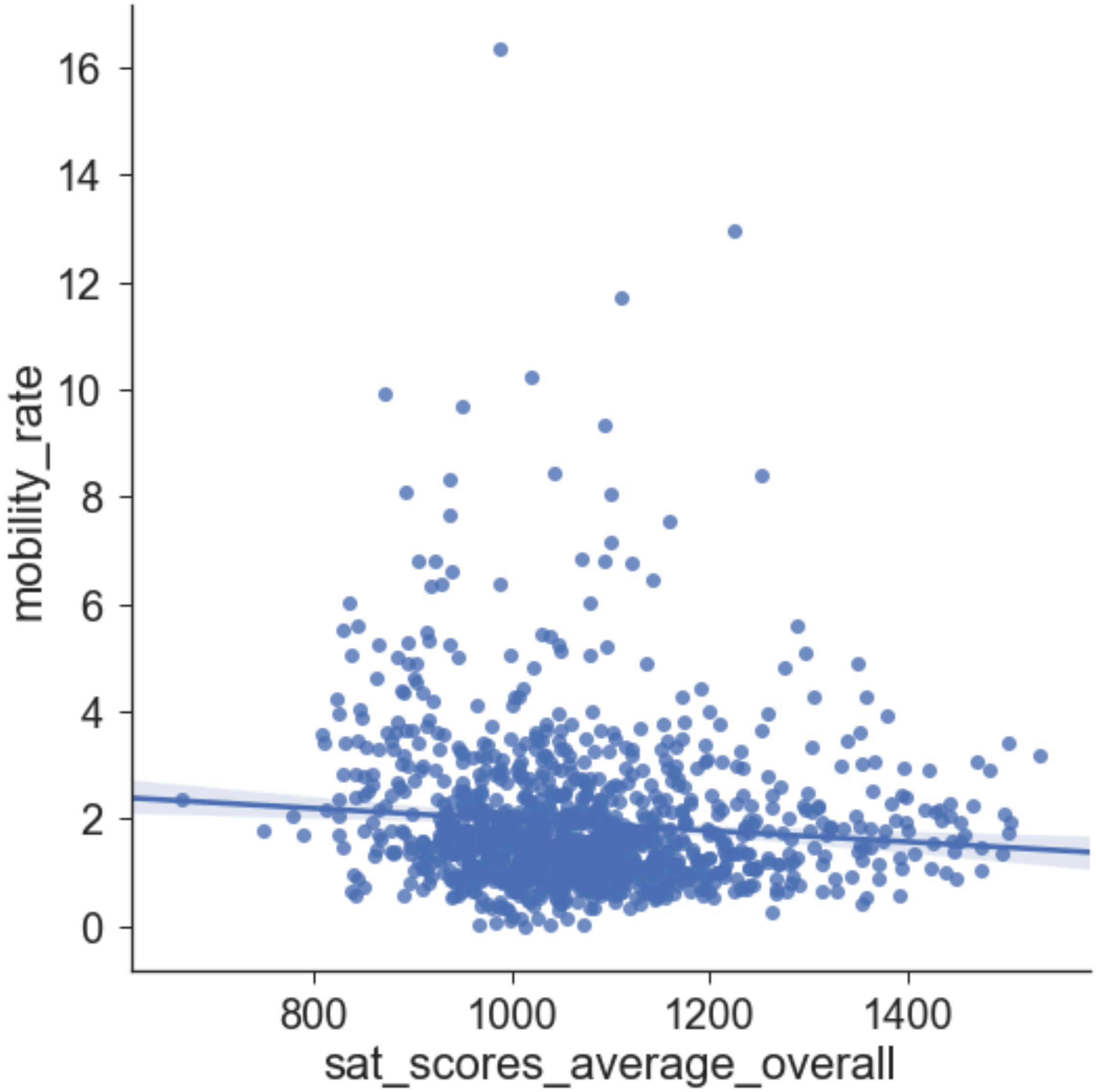
- The unemployment data show that some universities (Phoenix University, DeVry University) have really high unemployment, skewing the data.



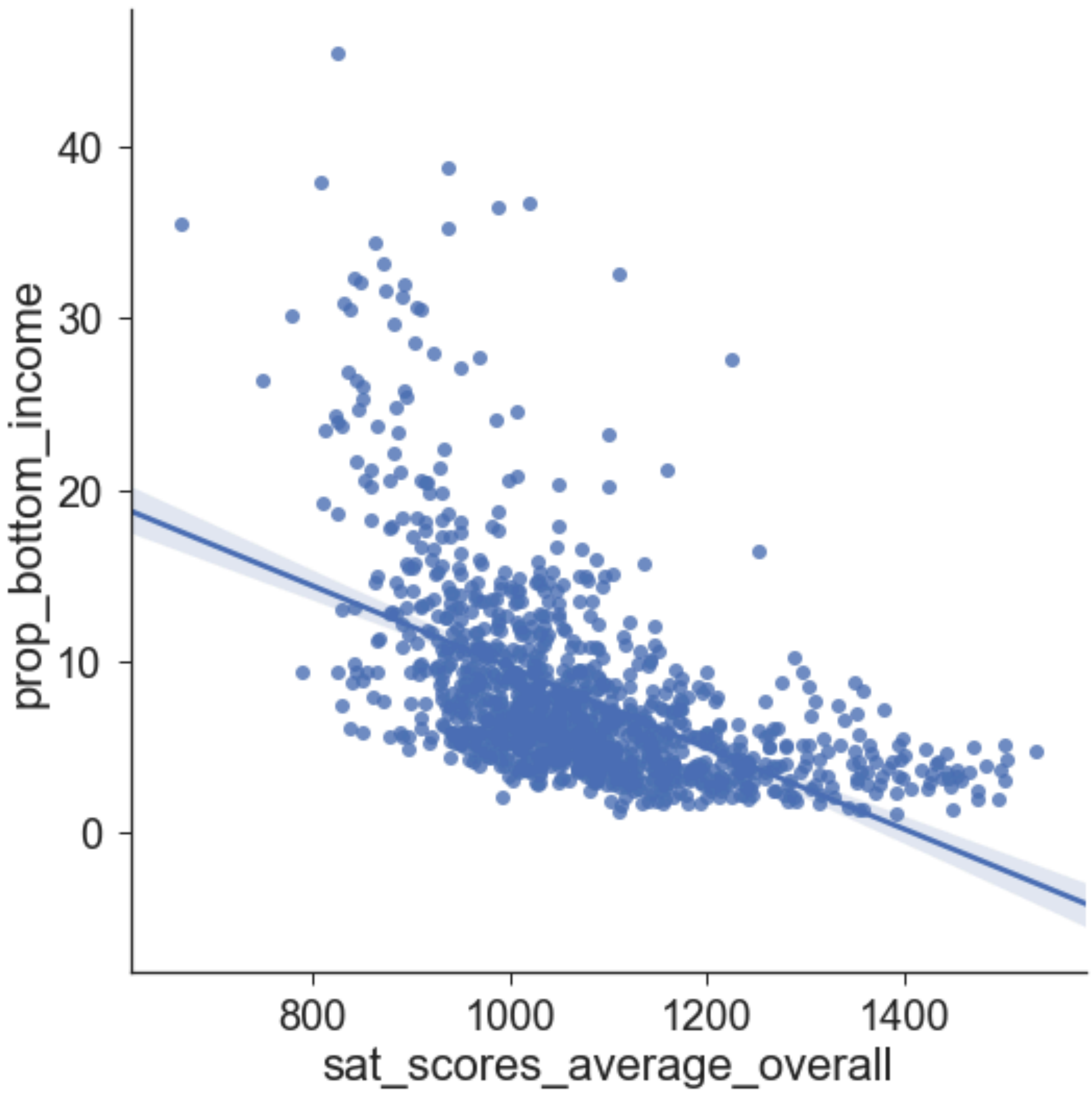
- SAT scores predict college selectivity and is important in determining college admission. I first examined the relationship between SAT scores and family income.
- Rich students go to more selective schools. There is evidence from other data that that kids from poorer families tend do worse on the SATs (https://economix.blogs.nytimes.com/2009/08/27/sat-scores-and-family-income/). What this means for upward mobility is that top tier, selective colleges, are not well-suited to increase the upward mobility of lower income families.



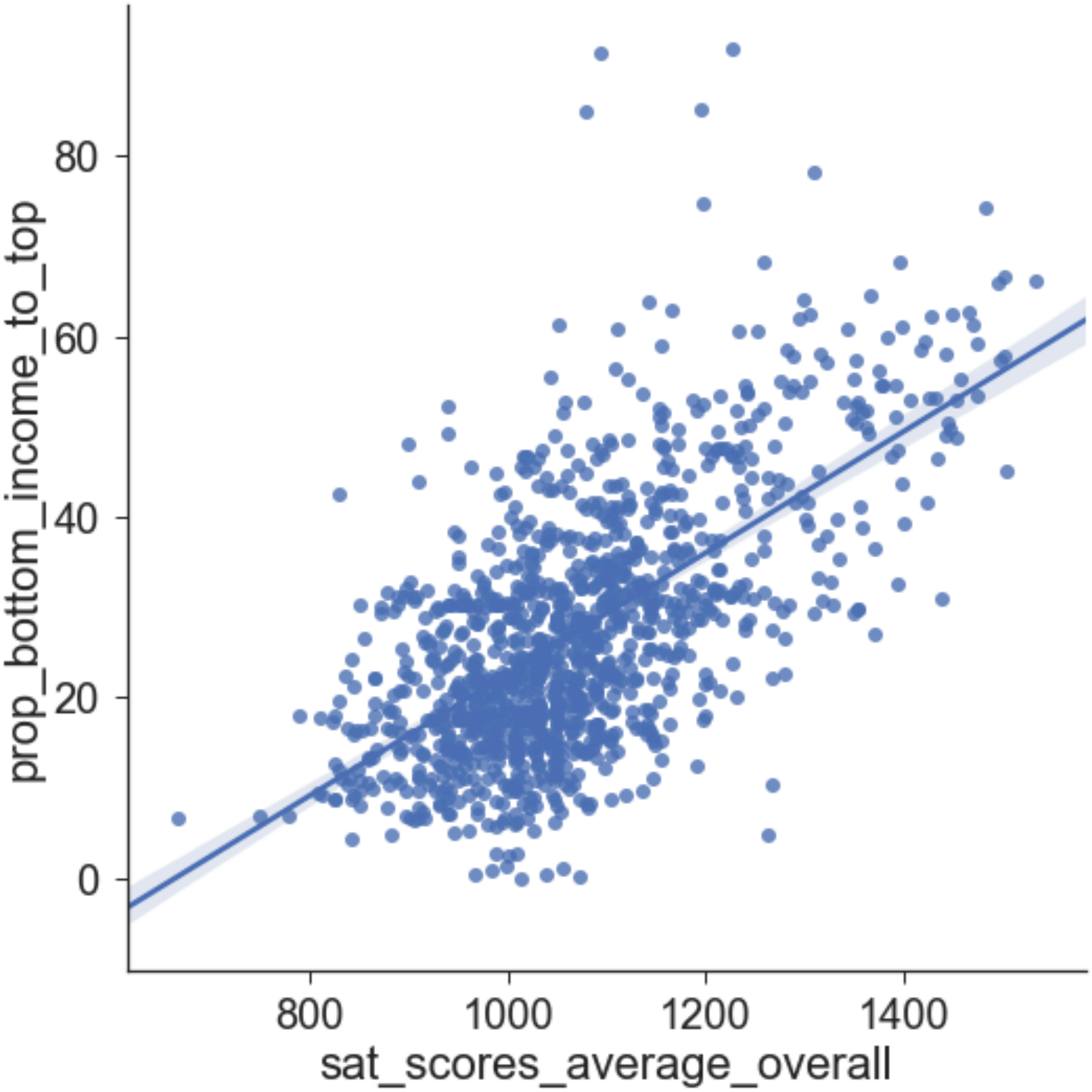
- College selectivity (SAT scores) has a small positive effect on mobility rate. i.e. more selective colleges have marginally higher mobility.



- The small effect size is largely because selective colleges do not take in low income students

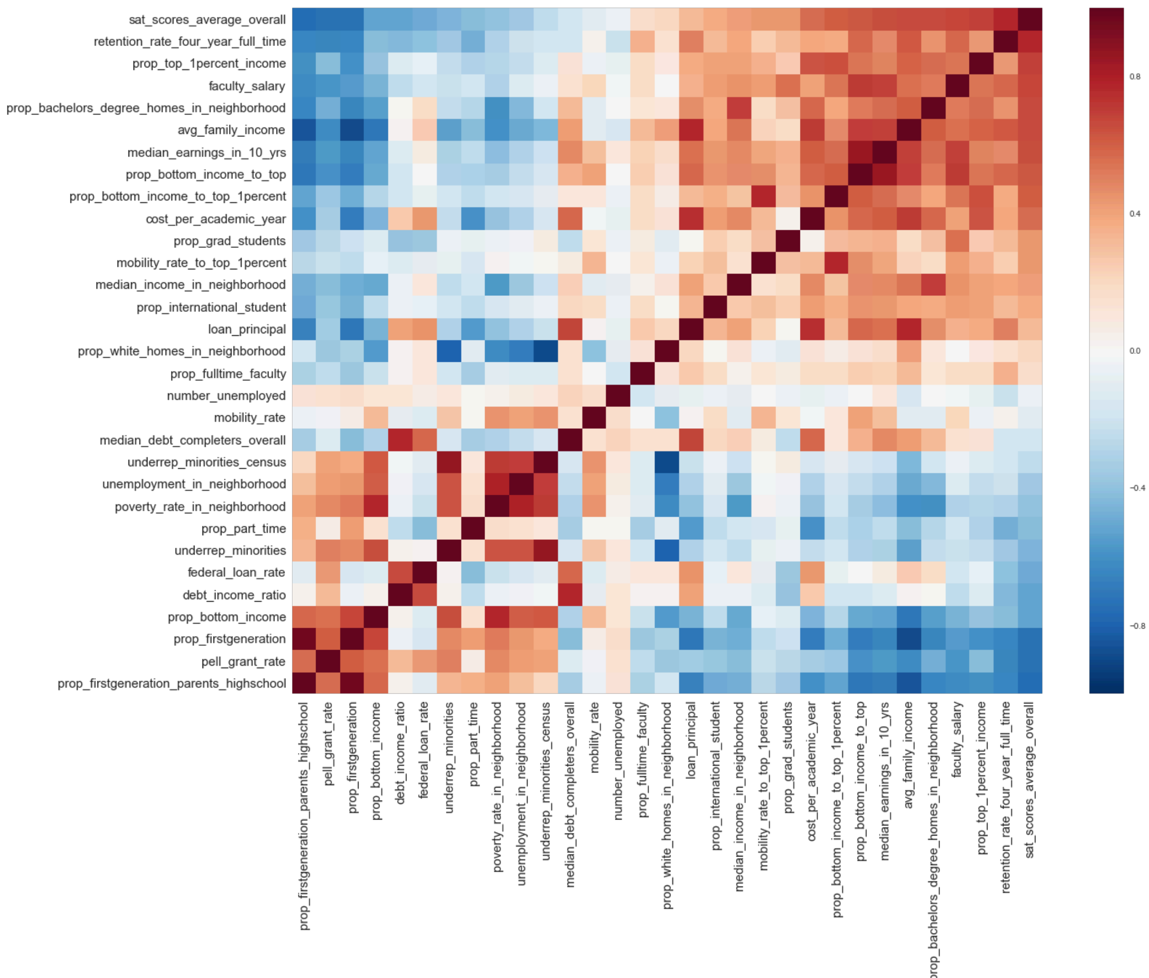


However, for low-income students who can get into more selective colleges, mobility is much higher.

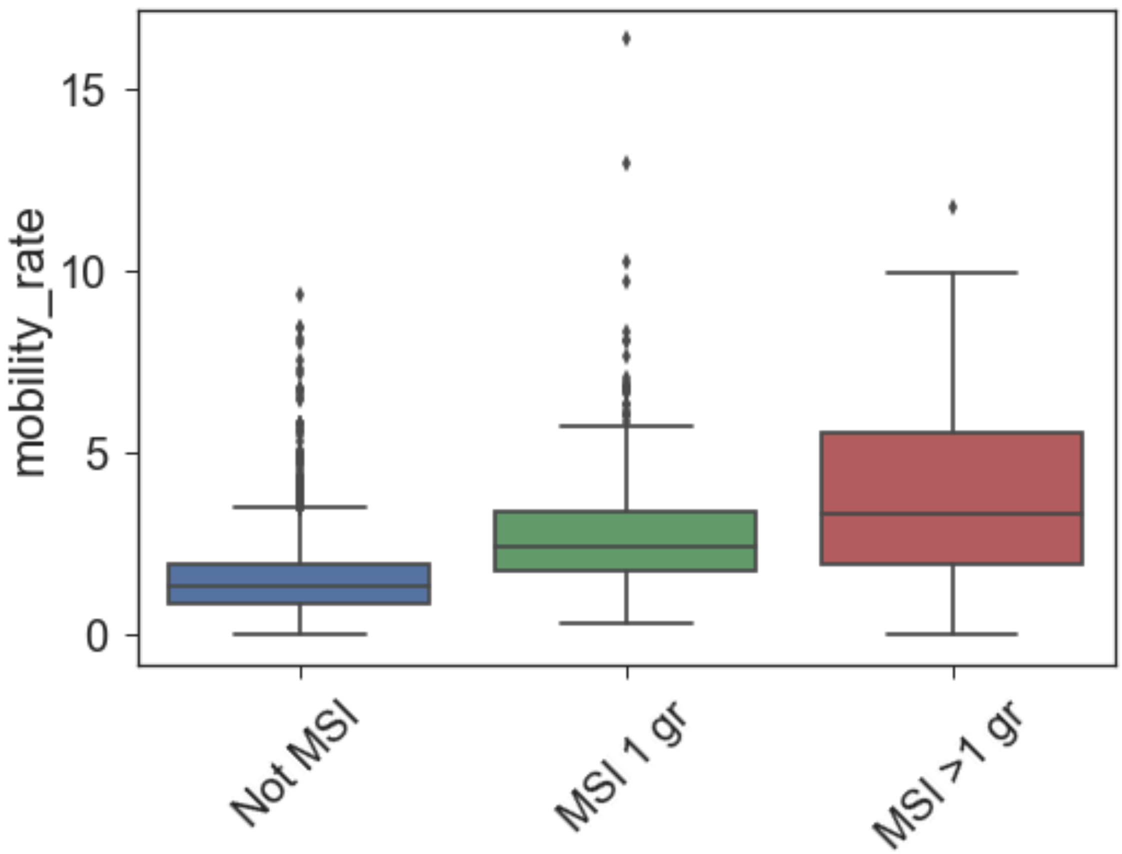


- To look for overall relationships in all predictors, I plotted a correlational heat map. I ordered the heat map by SAT scores because this is such an important determinant of admission. The heat map shows strong positive associations between average college SAT scores (selectivity), students from top income brackets, high tuition and affluent white neighborhoods. Mobility is high for low-income students who do make it to top selective and expensive universities (prop_bottom_income_to_top). However, these colleges take in very few students from the bottom income brackets (prop_bottom_income).
- Elite universities are doing a poor job of facilitating upward mobility. These universities don't contribute to the American dream.
- Upward mobility is associated with diversity at the institution and where the institution is located. It is also interestingly associated with colleges that have

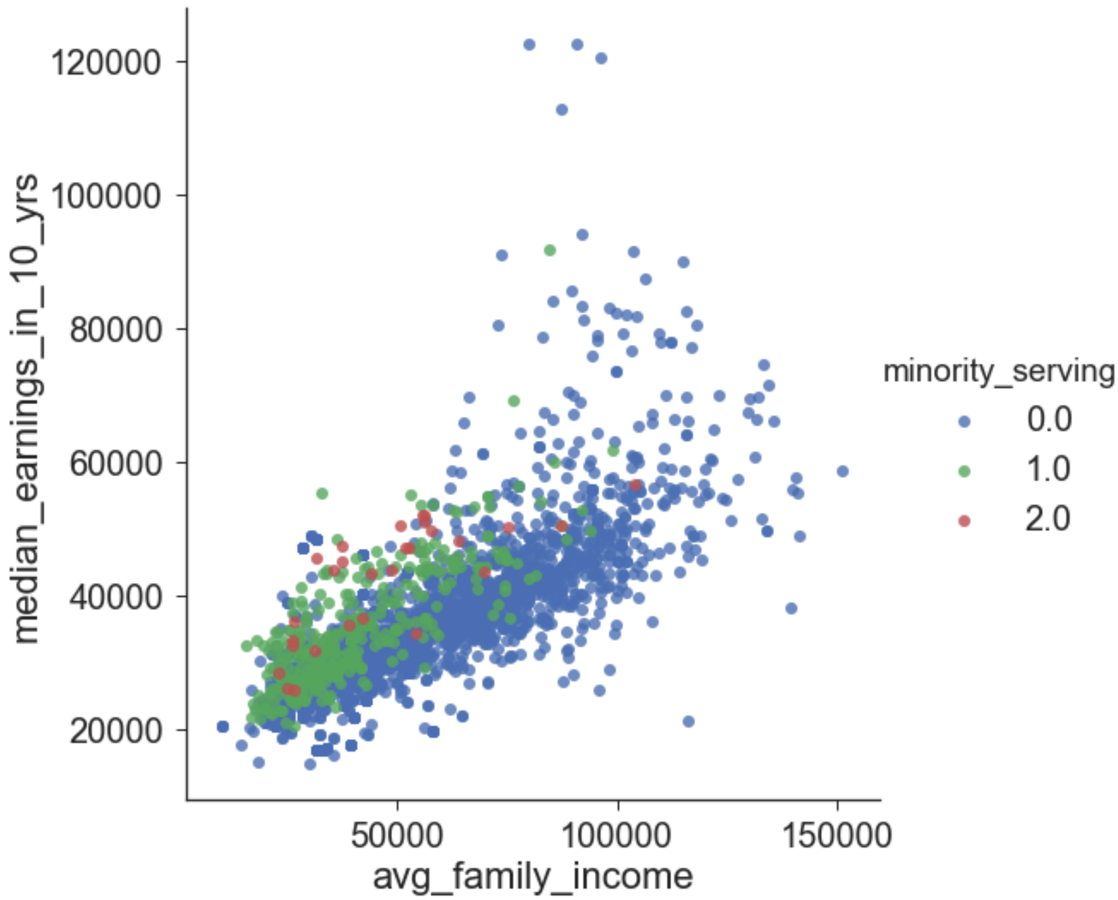
more graduate students, international students, full time faculty and faculty that get paid more (i.e. likely bigger more diverse universities).



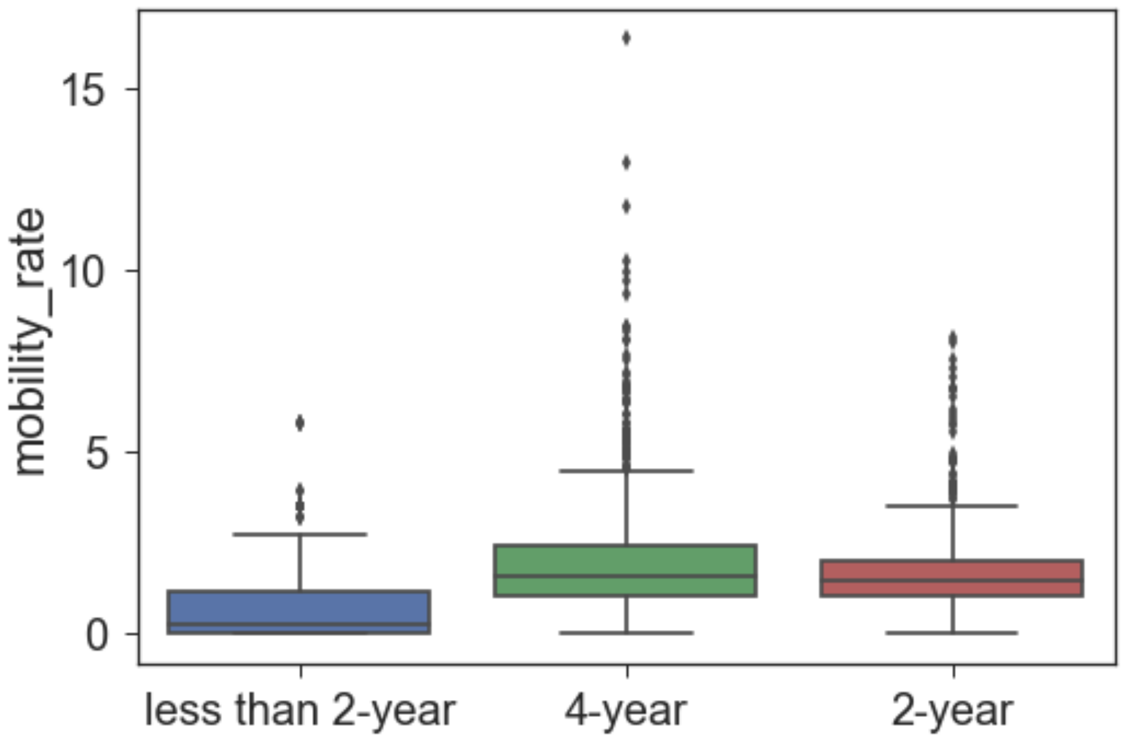
- Diversity matters for upward mobility. Minority serving institutions (MSIs) enroll a large proportion of minority students and have developed strategies to help often-underprepared students succeed in college. These institutions make a difference for helping students at lower income levels rise up the economic ladder.



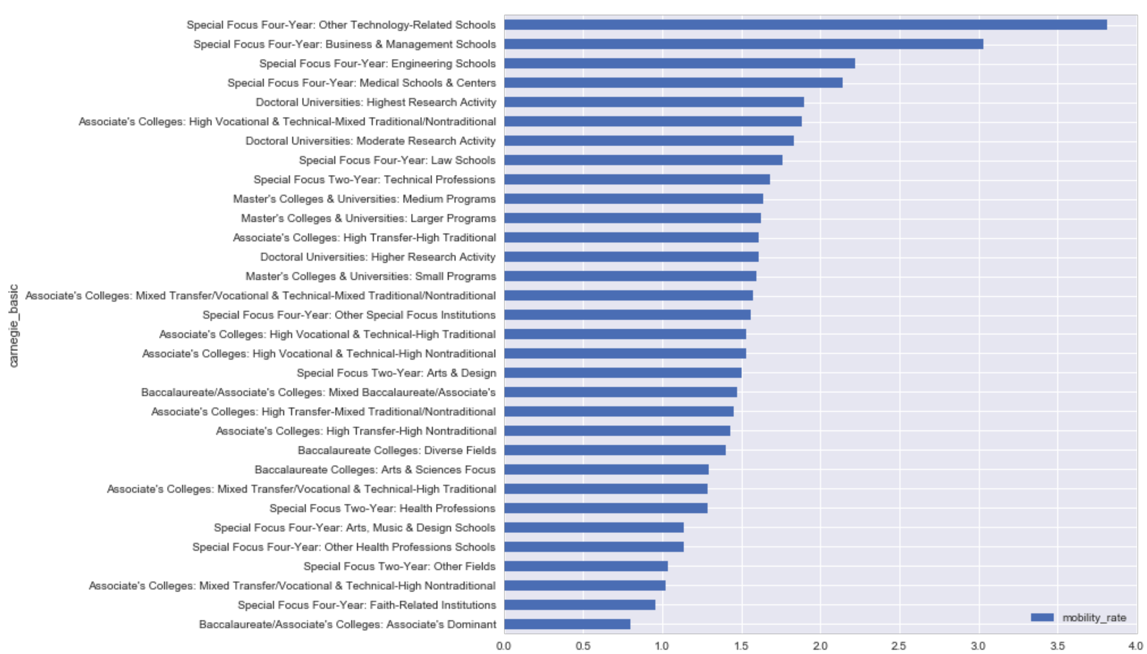
- Students with the same family income tend to earn more after attending a minority-serving institution. MSIs also tend to be cheaper



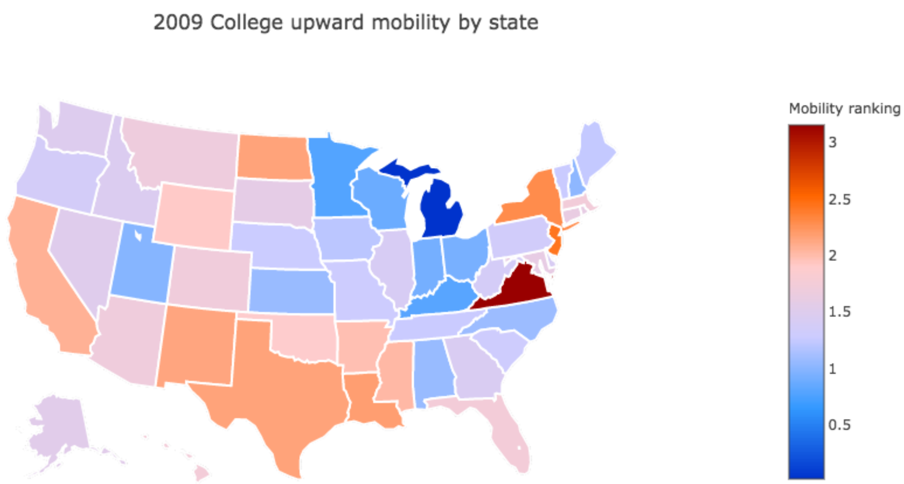
- Traditional 4-year institutions engender the highest mobility.



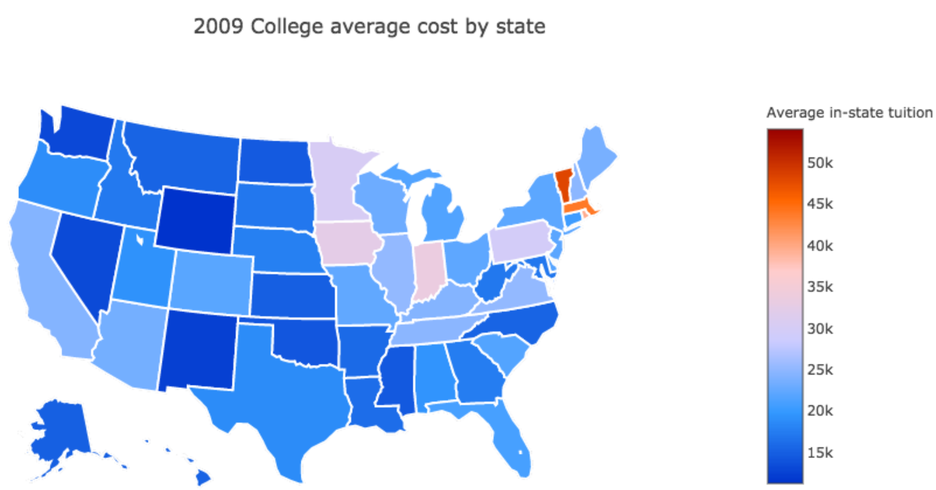
- Applied schools (Buisness, Management, Tech related) tend to be better at increasing economic mobility that those focused on the arts and religion



- California, Texas, New Mexico, North Dakota, New York and Virginia and have much higher mobility.



- This does not correspond to the cost of tuition by state!



Schools with top Economic Mobility from bottom 20% to top 20% income levels

	name	mobility_rate
1374	Vaughn College Of Aeronautics And Technology	16.357975
555	CUNY Bernard M. Baruch College	12.938586
2239	City College Of New York - CUNY	11.723747
1025	CUNY Lehman College	10.235138
3044	California State University, Los Angeles	9.918455
2234	CUNY John Jay College Of Criminal Justice	9.691438
2139	MCPHS University	9.343507
580	Pace University	8.432647
994	State University Of New York At Stony Brook	8.412747
2241	New York City College Of Technology Of The Cit...	8.334076

Schools with Economic Mobility from bottom 20% to top 1% income levels

	name	mobility_rate_to_top_1percent
286	Claremont Mckenna College	1.249444
2139	MCPHS University	0.963851
3058	Kiamichi Technology Center	0.798517
1432	Huntingdon College	0.778688
243	University Of California, Berkeley	0.763982
586	Columbia University In The City Of New York	0.750328
703	University Of Texas Of The Permian Basin	0.737258
1412	California Institute Of Technology	0.723216
555	CUNY Bernard M. Baruch College	0.706509
1728	Maine Maritime Academy	0.693706

Inferential Statistics

To look for statistically significant associations, I ran general linear models on most of the exploratory relationships.

1) A 10,000 dollar increase in family income increases SAT scores by on average 13.6 points and this relationship is statistically significant

```
                                OLS Regression Results
==============================================================================
Dep. Variable:     sat_scores_average_overall   R-squared:                       0.958
Model:                                    OLS   Adj. R-squared:                  0.958
Method:                         Least Squares   F-statistic:                 2.607e+04
Date:                        Thu, 16 Nov 2017   Prob (F-statistic):               0.00
Time:                                08:58:08   Log-Likelihood:                -7876.9
No. Observations:                        1156   AIC:                         1.576e+04
Df Residuals:                            1155   BIC:                         1.576e+04
Df Model:                                   1
Covariance Type:                    nonrobust
==============================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
avg_family_income   0.0136    8.4e-05    161.451      0.000         0.013     0.014
==============================================================================
Omnibus:                       45.480   Durbin-Watson:                   1.446
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               65.894
Skew:                          -0.362   Prob(JB):                     4.91e-15
Kurtosis:                       3.918   Cond. No.                         1.00
==============================================================================

Warnings:
```

2) College selectivity (SAT scores) have a small positive effect on mobility rate. i.e more selective colleges have marginally higher mobility. The response variable is sqrt transformed to meet assumptions of normality so a little difficult to interpret the slope parameter.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          mobility_rate   R-squared:                       0.623
Model:                            OLS   Adj. R-squared:                  0.622
Method:                 Least Squares   F-statistic:                     1963.
Date:                Tue, 14 Nov 2017   Prob (F-statistic):          4.58e-254
Time:                        14:31:27   Log-Likelihood:                -2145.0
No. Observations:                1191   AIC:                             4292.
Df Residuals:                    1190   BIC:                             4297.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
sat_scores_average_overall     0.0017   3.94e-05     44.308      0.000         0.002     0.002
==============================================================================
Omnibus:                      849.720   Durbin-Watson:                   1.417
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15751.745
Skew:                           3.100   Prob(JB):                         0.00
Kurtosis:                      19.702   Cond. No.                         1.00
==============================================================================
```

3) College selectivity (SAT scores) have a small positive effect on mobility rate. i.e more selective colleges have marginally higher mobility. However, for low-income students who can get into more selective colleges, mobility is much higher. The response variable (mobility_rate) is sqrt transformed to meet assumptions of normality so a little difficult to interpret the slope parameter.

```
                               OLS Regression Results
==============================================================================
Dep. Variable:     prop_bottom_income_to_top   R-squared:                       0.957
Model:                                   OLS   Adj. R-squared:                  0.957
Method:                        Least Squares   F-statistic:                 2.652e+04
Date:                       Tue, 14 Nov 2017   Prob (F-statistic):               0.00
Time:                               14:35:05   Log-Likelihood:                -1783.4
No. Observations:                       1191   AIC:                             3569.
Df Residuals:                           1190   BIC:                             3574.
Df Model:                                  1
Covariance Type:                   nonrobust
=============================================================================================
                               coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------------
sat_scores_average_overall   0.0047   2.91e-05    162.857      0.000         0.005     0.005
==============================================================================
Omnibus:                       36.615   Durbin-Watson:                   1.766
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               73.858
Skew:                          -0.177   Prob(JB):                     9.16e-17
Kurtosis:                       4.168   Cond. No.                         1.00
==============================================================================
```

4) Which factors significantly predict mobility rate? I ran a multiple regression but there are very likely problems with multicollinearity with these data (many of the predictors are very correlated with each other e.g. underrep_minorities and underrep_minorities_census). The solution is dimension reduction (pca) followed by a regression of pca loadings against the response. Nevertheless, most terms are significant at alpha = 0.05. The effect sizes are large for debt: income ratio, the number of first generation students and the number of international students. Schools that support a lot of first generation students increase mobility by almost 1%!

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          mobility_rate   R-squared:                       0.953
Model:                            OLS   Adj. R-squared:                  0.952
Method:                 Least Squares   F-statistic:                     2141.
Date:                Tue, 14 Nov 2017   Prob (F-statistic):               0.00
Time:                        16:57:52   Log-Likelihood:                -256.37
No. Observations:                1178   AIC:                             534.7
Df Residuals:                    1167   BIC:                             590.5
Df Model:                          11
Covariance Type:            nonrobust
=======================================================================================================
                                         coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------------------------
debt_income_ratio                     -0.3049      0.081     -3.782      0.000        -0.463    -0.147
median_earnings_in_10_yrs           8.717e-06    1.4e-06      6.240      0.000      5.98e-06  1.15e-05
loan_principal                      2.125e-05   2.91e-06      7.312      0.000      1.55e-05  2.69e-05
prop_firstgeneration                   1.0403      0.119      8.778      0.000         0.808     1.273
underrep_minorities                   -0.1342      0.102     -1.318      0.188        -0.334     0.066
underrep_minorities_census             0.0054      0.002      3.182      0.002         0.002     0.009
poverty_rate_in_neighborhood           0.0557      0.004     12.899      0.000         0.047     0.064
prop_white_homes_in_neighborhood      -0.0052      0.001     -5.516      0.000        -0.007    -0.003
prop_international_student             0.9024      0.222      4.069      0.000         0.467     1.337
prop_grad_students                 -6.007e-06   3.81e-06     -1.576      0.115     -1.35e-05  1.47e-06
faculty_salary                      3.458e-05   6.32e-06      5.468      0.000      2.22e-05   4.7e-05
==============================================================================
Omnibus:                       60.374   Durbin-Watson:                   1.814
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              211.094
Skew:                           0.018   Prob(JB):                     1.45e-46
Kurtosis:                       5.073   Cond. No.                     1.23e+06
==============================================================================
```

5) Minority serving institutions for a single underrepresented minority increases mobility by 0.2 percent. Those that serve two or more minorities increase mobility by 1%

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          mobility_rate   R-squared:                       0.147
Model:                            OLS   Adj. R-squared:                  0.146
Method:                 Least Squares   F-statistic:                     249.2
Date:                Thu, 16 Nov 2017   Prob (F-statistic):          1.43e-100
Time:                        08:42:54   Log-Likelihood:                -4623.9
No. Observations:                2895   AIC:                             9254.
Df Residuals:                    2892   BIC:                             9272.
Df Model:                           2
Covariance Type:            nonrobust
====================================================================================================
                                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------------------
Intercept                           2.8809      0.065     44.620      0.000         2.754     3.008
C(minority_serving)[T.MSI >1 gr]    1.1278      0.231      4.877      0.000         0.674     1.581
C(minority_serving)[T.Not MSI]     -1.3589      0.069    -19.747      0.000        -1.494    -1.224
==============================================================================
Omnibus:                     1569.253   Durbin-Watson:                   1.555
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            20716.297
Skew:                           2.279   Prob(JB):                         0.00
Kurtosis:                      15.287   Cond. No.                         14.1
==============================================================================
```

6) Two-year, junior colleges decrease mobility by 2.12% and colleges with shorter programs decrease mobility by almost 2.9%

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          mobility_rate   R-squared:                       0.048
Model:                            OLS   Adj. R-squared:                  0.047
Method:                 Least Squares   F-statistic:                     77.85
Date:                Tue, 14 Nov 2017   Prob (F-statistic):           1.04e-33
Time:                        17:19:16   Log-Likelihood:                -5081.4
No. Observations:                3087   AIC:                         1.017e+04
Df Residuals:                    3084   BIC:                         1.019e+04
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================================================
                                                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------------------------------
Intercept                                      1.8959      0.029     65.730      0.000         1.839     1.952
C(institutional_characteristics_level)[T.2]   -0.2381      0.050     -4.746      0.000        -0.337    -0.140
C(institutional_characteristics_level)[T.3]   -1.0227      0.083    -12.253      0.000        -1.186    -0.859
==============================================================================
Omnibus:                     1850.285   Durbin-Watson:                   1.407
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            27398.804
Skew:                           2.575   Prob(JB):                         0.00
Kurtosis:                      16.656   Cond. No.                         3.99
==============================================================================
```

Initial findings

The most striking findings from this preliminary analysis is that diversity in the school and in the neighborhood facilitates economic mobility for low-income students. At a time when Affirmative Action is under the radar and we consider the importance of these measures, it is important to know that diversity plays a big role in making the American dream a possibility. Diverse schools with a high proportion of first generation students, likely have the resources and the inclusivity to retain low-income students and see them graduation. Many of these schools are also more inclusive and affordable which is especially important since low-income students, have poorer SAT outcomes and cannot make it to more elite schools. However, for the low-income students who do make it to the Ivy Leagues, the probability of improved economic mobility is much higher.

Additional exploration

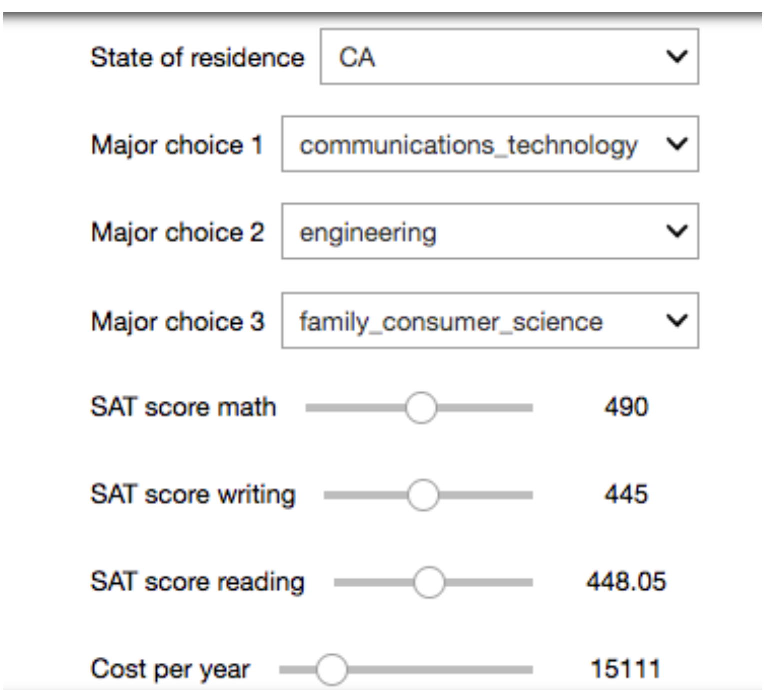
- Data on school academics. The college Scorecard database has information on the academic programs and the proportion of degrees granted in each program by institution. I would like to explore these data as metric for economic mobility
- Plot these data on a map at the county level (still figuring this out – I need to get a dataset that matches zip codes to counties)

Machine Learning: College Recommender

Develop an algorithm to match students (based on SAT scores, cost of tuition, academic interests) to 10-20 schools based on state and display matches with mobility rankings. That is, develop an algorithm that outputs which in-state schools are an option and which ones are most likely to increase the students upward mobility.

Pipeline:

- K means clustering to produce even clusters of similar colleges on:
    - SAT scores
    - Cost of attending
    - Programs offered
- Run model for each state, splitting colleges into approximately 2-20 clusters per state
- Train a Random Forest classifier on these clusters to match student data to colleges
- Enter data using an ipywidget



- Display the options with the mobility opportunities that the college provides.