Analysis of Students' Income

Aram Tovmasyan, October 2017

Executive Summary

This document presents an analysis of data concerning students and their income. The analysis is based on about 17107 observations of US higher institutions (schools), each containing specific characteristics of some institution and students' earnings. After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between higher educational institution characteristics and students' income were identified. After exploring the data, predictive model to classify students into two income categories was created. Another model of unsupervised learning to find out the number of possible domains of schools was created. Predictive model to classify schools into five future income categories was created, and finally a regression model to predict a student income from its school features was created.

After performing the analysis, the author presents the following conclusions:

While many factors can help indicate the students' income, significant features found in this analysis were:

- **school__degrees_awarded_predominant** - Predominant undergraduate degree awarded
- **school__degrees_awarded_highest** - Highest degree awarded
- **school__institutional_characteristics_level** -Level of institution (Possible Values: 2-year, 4-year, Less-than-2-year
- **school__faculty_salary** -Average faculty salary
- **school__ownership** - Control of institution (Possible Values: Public, Private for-profit, Private nonprofit)
- **student__retention_rate_lt_four_year_full_time** - student retention rate at less-than-four-year institutions First-time, full-time
- **student__retention_rate_four_year_full_time** - Student retention rate at four-year institutions First-time, full-time
- **student__demographics_first_generation** - Share of first-generation students
- **student__share_firstgeneration_parents_highschool** - Percent of students whose parents' highest educational level is high school
- **student__share_firstgeneration_parents_somecollege** - Percent of students whose parents' highest educational level was is some form of postsecondary education
- **cost__tuition_out_of_state** - In-state tuition and fees
- **cost__tuition_in_state** -Out-of-state tuition and fees

- **admissions__sat_scores_average_overall**- Average SAT equivalent score of students admitted

Initial Data Exploration

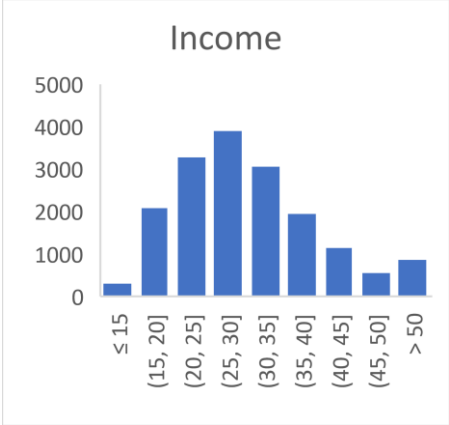
The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 17107 observations are shown here:

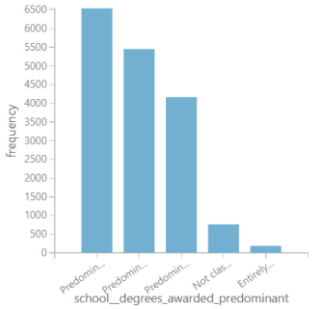
	school__faculty_salary	retention_rate_lt_four_year_full_time	student__demographics_ first_generation	student__retention_rate_four_year_full_time	student_share_firstgeneration_parents_highschool	academics__program_percentage_personal_culinary	cost__tuition_out_of_state	cost__tuition_in_state	admissions__sat_scores_average_overall	student__share_firstgeneration_parents_somecollege
Mean	5797.64	0.66	0.48	0.69	0.44	0.15	14973.92	12612.9	1052.18	0.53
Standard Error	19.27	0.00	0.00	0.00	0.00	0.00	84.65	91.79	2.01	0.00
Median	5575.00	0.66	0.51	0.72	0.46	0.00	13305	11079	1035	0.50
Mode	5120.00	1.00	0.52	0.50	0.47	0.00	18048	18048	1010	0.48
Standard Deviation	2050.42	0.17	0.13	0.19	0.09	0.34	8925.56	9821.37	127.67	0.12
Minimum	153.00	0.00	0.06	0.00	0.07	0.00	80.00	80.00	660	0.05
Maximum	24892	1.00	0.95	1.00	0.92	1.00	49793	49793	1520	0.94
Count	11323	8684	16421	6204	14300	16210	11117	11448	4025	15532

Since Income is of interest in this analysis, it was noted that the mean and median of this value are not significantly different and that the comparatively small standard deviation indicates that there is no considerable variance in the future earnings of the students. A histogram of the Income column shows that the Income values are normally distributed – in other words, most students' future earnings are in the middle of income range, as shown here



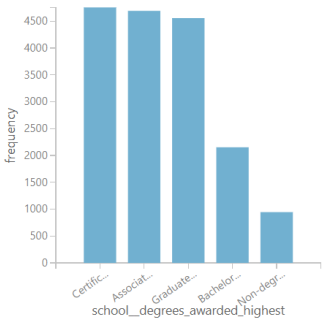
In addition to the numeric values, the schools observations were included categorical features, such as

- school__degrees_awarded_predominant



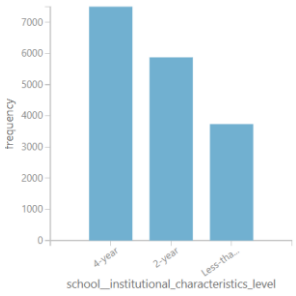
Predominant undergraduate degree awarded
0 Not classified
1 Predominantly certificate-degree granting
2 Predominantly associate's-degree granting
3 Predominantly bachelor's-degree granting
4 Entirely graduate-degree granting

- school__degrees_awarded_highest



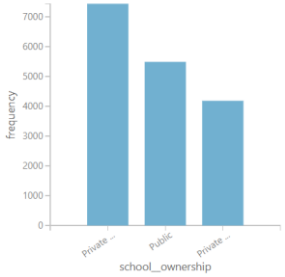
Highest degree awarded
0 Non-degree granting
1 Certificate degree
2 Associate degree
3 Bachelor's degree
4 Graduate degree

- school__institutional_characteristics_level



Level of institution
2-year
4-year
Less-than-2-year

- school__ownership



Control of institution
Public
Private for-profit
Private nonprofit

Bar charts created to show frequency of these features and indicate the following:
• Most schools are awarding predominantly certificate degree (38%), less -bachelor's degree (32%) and associate's-degree (28%)
• Highest degrees awarded by most schools are Certificate degree, Associate degree and Graduate degree (about 27% each)

- Most schools are 4 years (44%), less are 2 years (34%) and Less than 2 years
- Most schools are private (43%), less are public (32%) and private non-profit (24%)

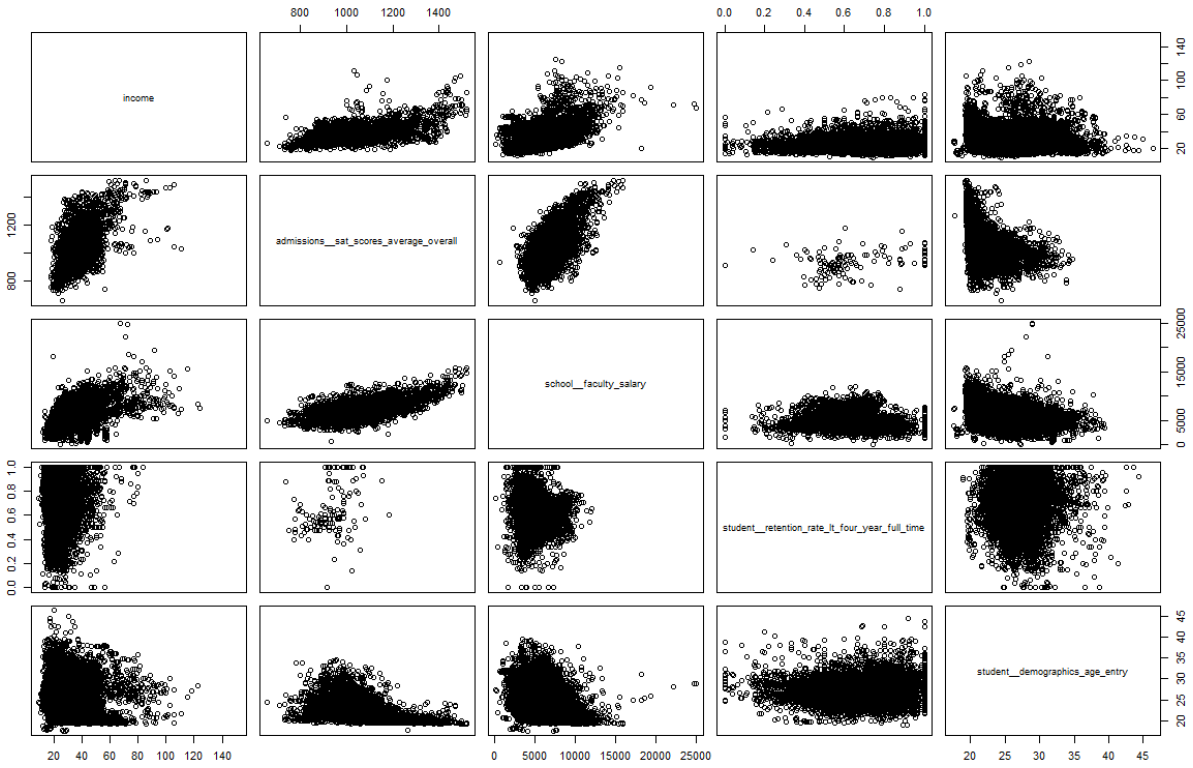
Correlation and Apparent Relationships

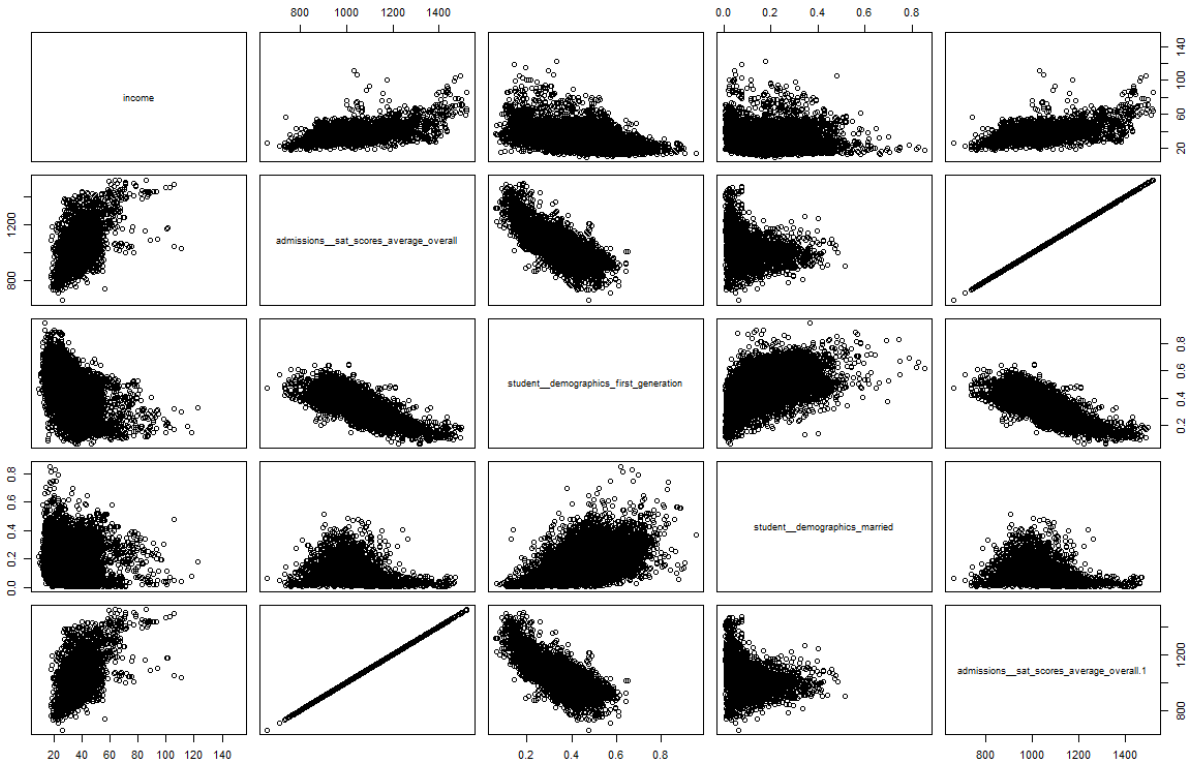
After exploring the individual features, an attempt was made to identify relationships between income and the other features.

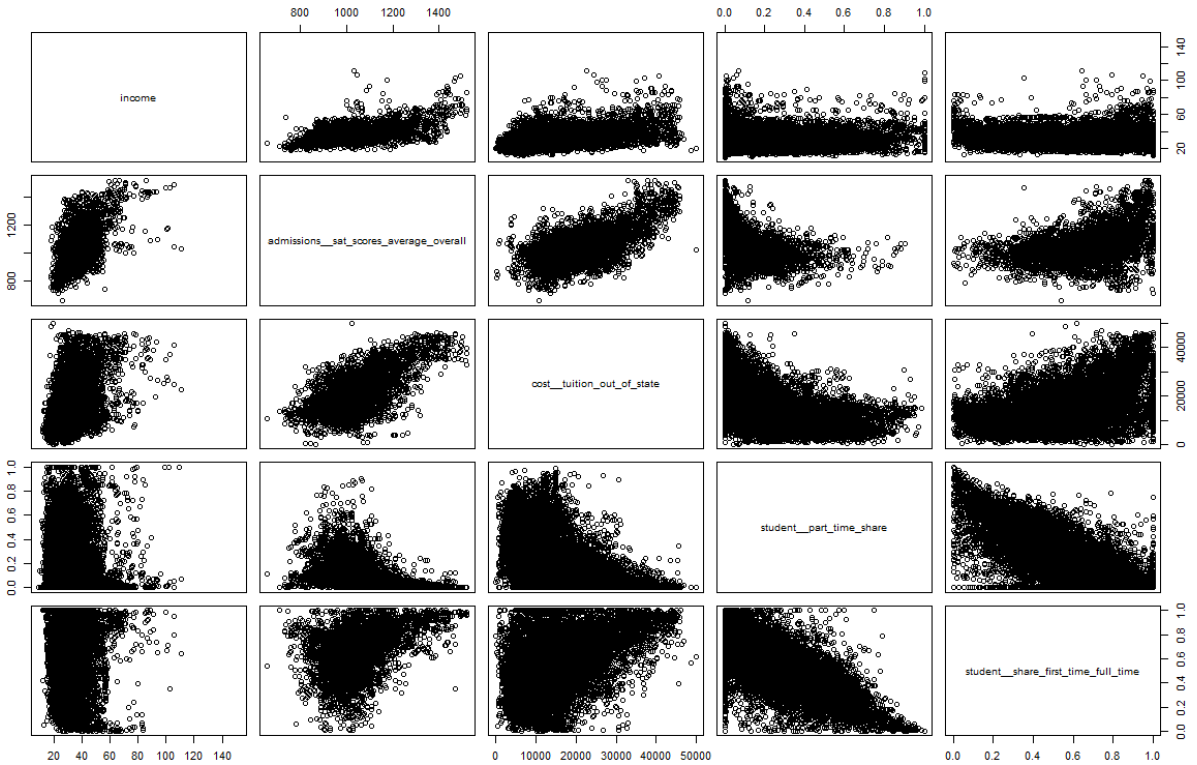
Feature	Correlation	Description
admissions__act_scores_25th_percentile_math	0.5876	25th percentile of the ACT math score
admissions__act_scores_midpoint_math	0.5869	Midpoint of the ACT math score
admissions__sat_scores_25th_percentile_math	0.5706	25th percentile of SAT scores at the institution (math)
admissions__sat_scores_midpoint_math	0.5657	Midpoint of SAT scores at the institution(math)
admissions__act_scores_75th_percentile_math	0.5554	75th percentile of the ACT math score
school__degrees_awarded_predominant_recoded	0.5423	Predominant degree awarded (recoded 0s and 4s)
admissions__sat_scores_75th_percentile_math	0.5383	75th percentile of SAT scores at the institution (math)
admissions__act_scores_25th_percentile_cumulative	0.5376	25th percentile of the ACT cumulative score
admissions__sat_scores_average_overall	0.5297	Average SAT equivalent score of students admitted
school__faculty_salary	0.5280	Average faculty salary
admissions__sat_scores_25th_percentile_writing	0.5280	25th percentile of SAT scores at the institution (writing)
admissions__act_scores_midpoint_cumulative	0.5255	Midpoint of SAT cumulative scores at the institution
student__share_firstgeneration_parents_highschool	-0.5221	Percent of students whose parents' highest educational level is high school
admissions__sat_scores_midpoint_writing	0.5175	Midpoint of SAT scores at the institution(writing)
student__demographics_first_generation	-0.5044	Share of first-generation students
student__share_firstgeneration	-0.5044	Percentage first-generation students
admissions__act_scores_25th_percentile_english	0.5024	25th percentile of the ACT English score
student__share_firstgeneration_parents_somecollege	0.4946	Percent of students whose parents' highest educational level was is some form of postsecondary education
admissions__act_scores_75th_percentile_cumulative	0.4934	75th percentile of the cumulative ACT score
admissions__sat_scores_75th_percentile_writing	0.4892	75th percentile of the SAT scores at the institution (writing)
admissions__sat_scores_25th_percentile_critical_reading	0.4884	75th percentile of the SAT scores at the institution (critical reading)
admissions__act_scores_midpoint_english	0.4775	Midpoint of ACT scores at the institution (English)
cost__tuition_out_of_state	0.4712	Out-of-state tuition and fees

Numeric Relationships

The following scatter-plot matrix was generated initially to compare numeric features with one another. The key features in this matrix are shown here:

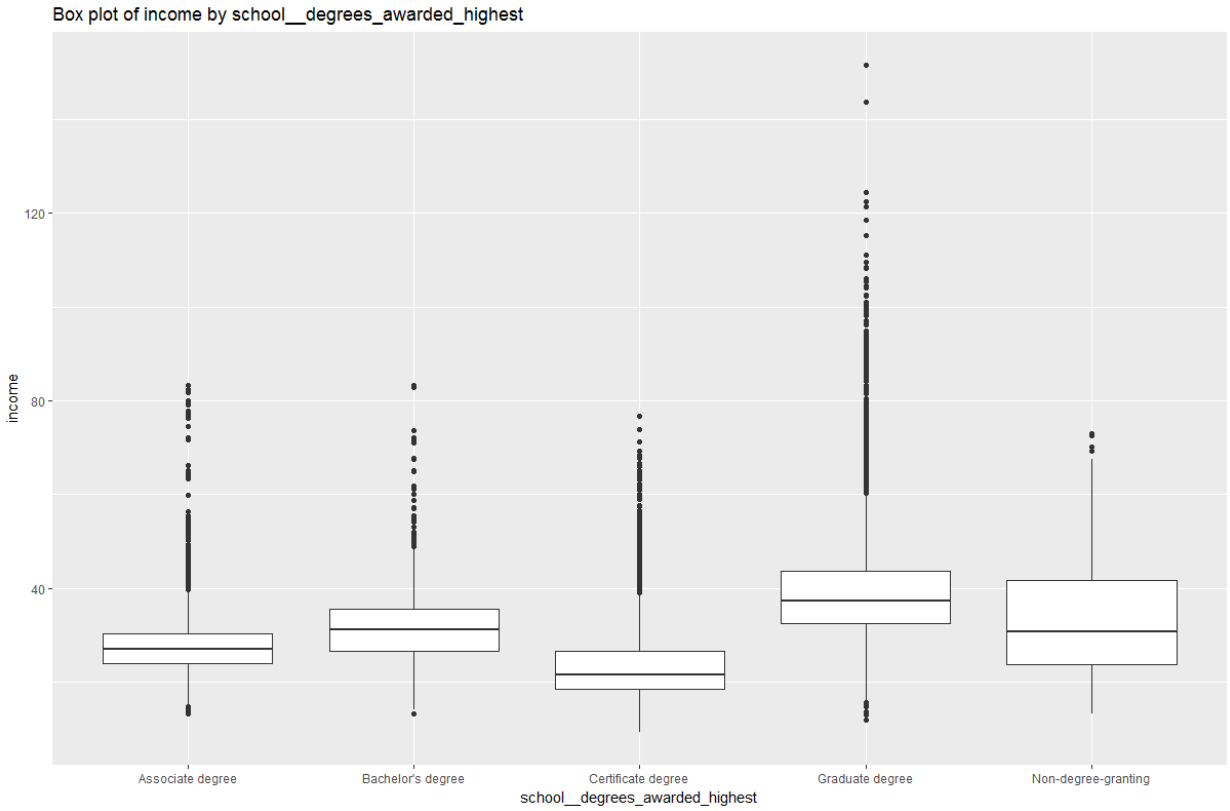


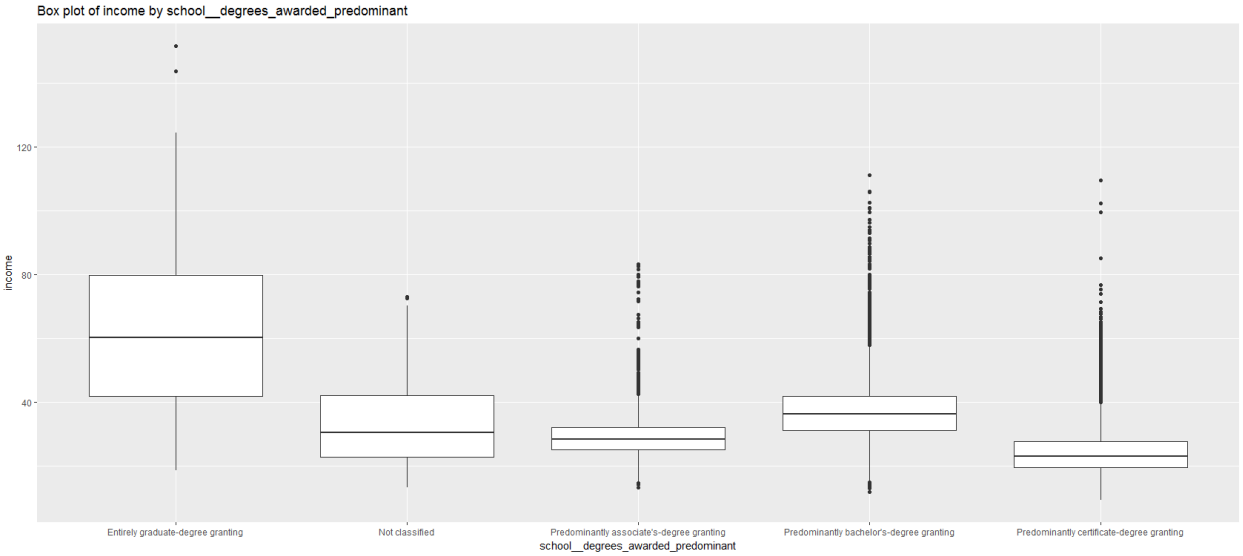

income
admissions__sat_scores_average_overall
student__demographics_first_generation
student__demographics_married
admissions__sat_scores_average_overall.1

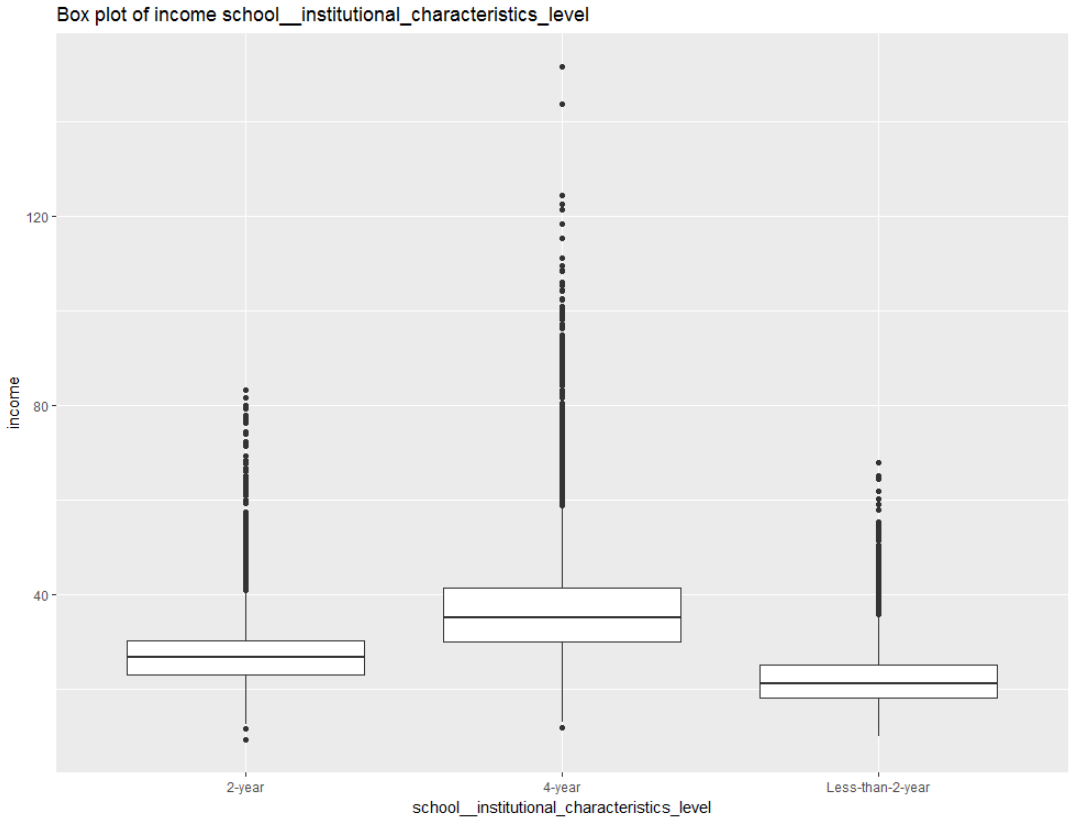

income
admissions__sat_scores_average_overall
cost__tuition_out_of_state
student__part_time_share
student__share_first_time_full_time

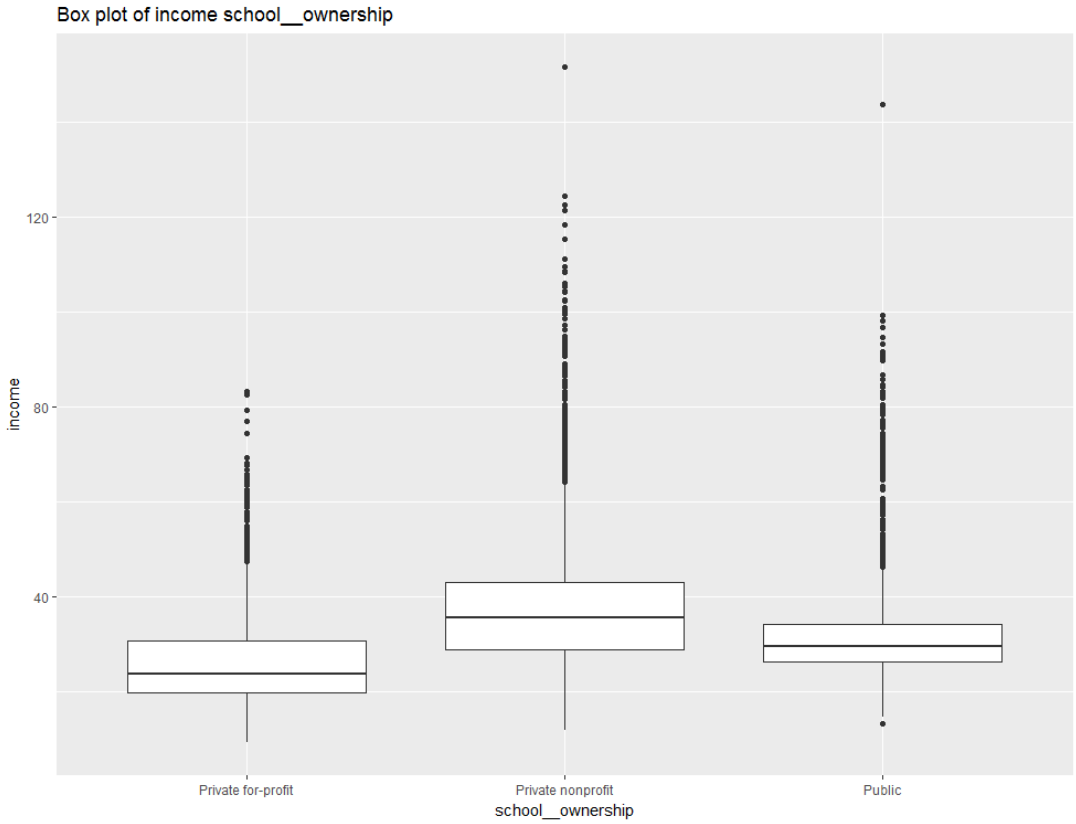
Categorical Relationships

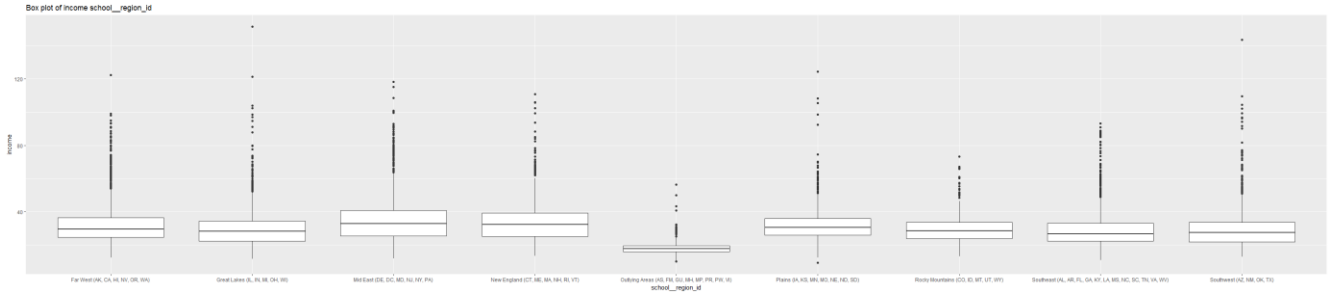
Having explored the relationship between Income and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and income. The following boxplots show the categorical columns that seem to exhibit a relationship with the Income:






Box plot of income school__institutional_characteristics_level
income
120
80
40
2-year
4-year
Less-than-2-year
school__institutional_characteristics_level


Box plot of income school__ownership
income
120
80
40
Private for-profit
Private nonprofit
Public
school__ownership



The box plots show some clear differences in terms of the median and range of Income for different categorical features. For example:

• Students from schools awarding highest degree “Graduate degree” have higher Income

• Though the small percentage, there is a wider range of incomes for schools awarding predominantly “Entirely graduate degree”, and the median income is significantly higher than for other types of schools.

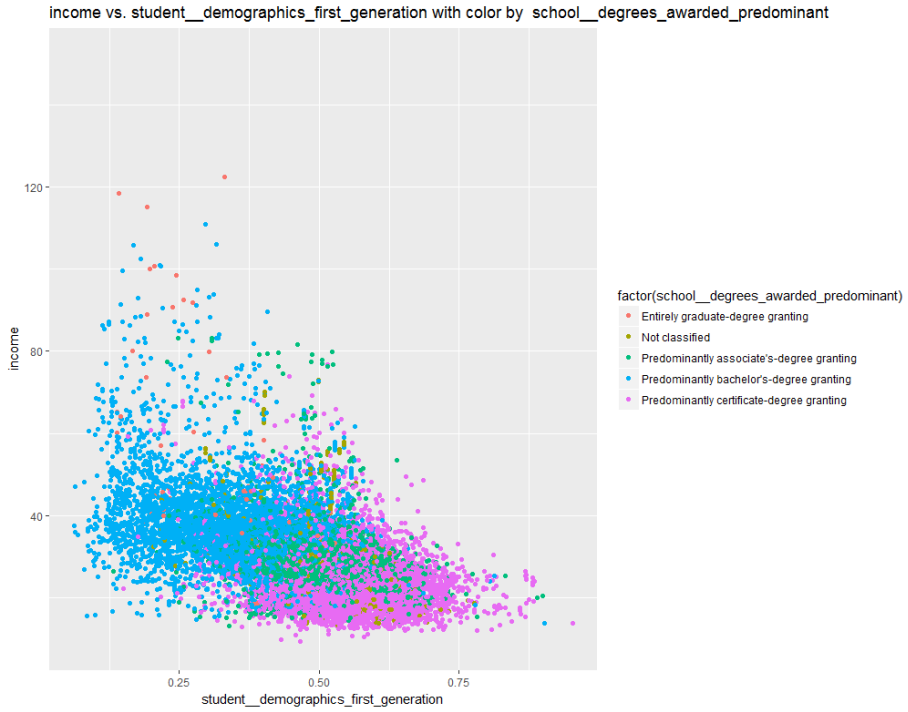
• 4-year schools’ student income is higher than 2-year schools and 2-year schools’ students’ income is higher than Less-than-2-year schools

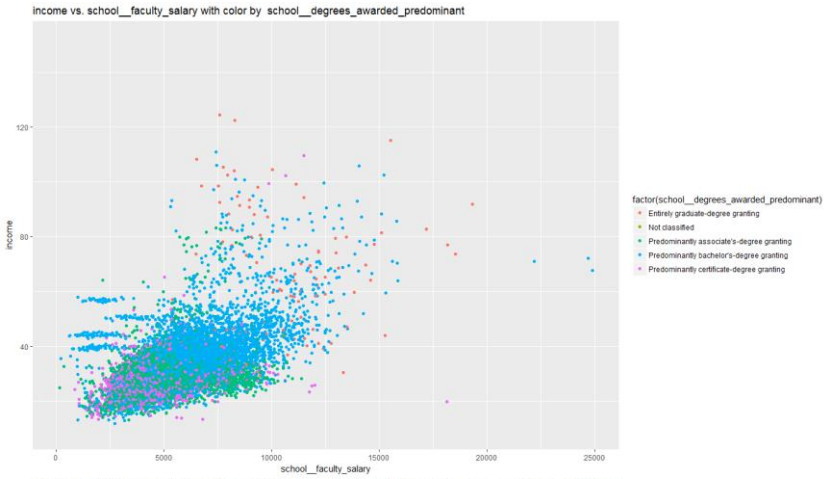
• “Private non-profit” schools’ students’ income is higher than “Public” schools and “Public” schools’ students’ income is higher than “Private for-profit” schools

• “Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)” schools’ students’ income range is very small and median income is significantly less than of other regions schools which are not very different from each other.

Multi-faceted Relationships

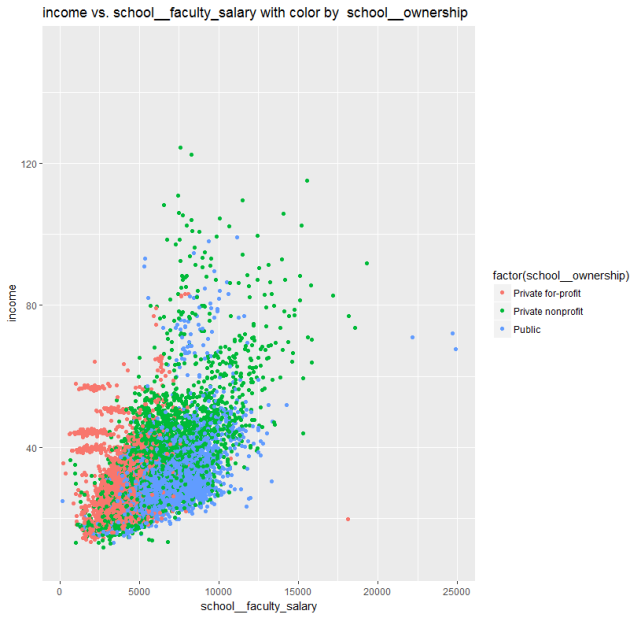
Apparent relationships between income and individual features are helpful in determining predictive heuristics. However, relationships are often more complex, and may only become apparent when multiple features are considered in combination with one another. To help identify these more complex relationships, some faceted plots were created. The following plots show some interesting aspects of school__degrees_awarded_predominant. It can be seen from these plots that school__degrees_awarded_predominant can be indicative of student__demographics_first_generation , school__faculty_salary, student__share_firstgeneration_parents_somecollege which are typically predictive of Income.

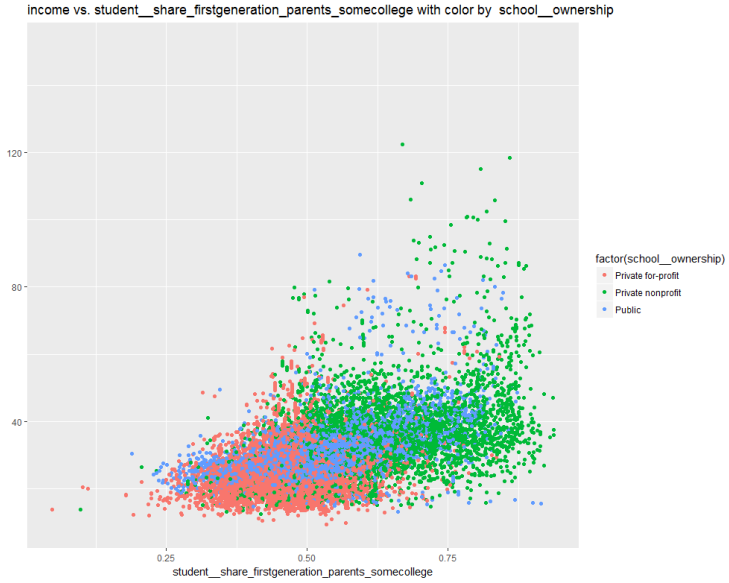

income vs. student__demographics_first_generation with color by school__degrees_awarded_predominant
income
120
80
40
0.25
0.50
0.75
student__demographics_first_generation
factor(school__degrees_awarded_predominant)
Entirely graduate-degree granting
Not classified
Predominantly associate's-degree granting
Predominantly bachelor's-degree granting
Predominantly certificate-degree granting


income vs. school__faculty_salary with color by school__degrees_awarded_predominant
income
120
80
40
0
5000
10000
15000
20000
25000
school__faculty_salary
factor(school__degrees_awarded_predominant)
Entirely graduate-degree granting
Not classified
Predominantly associate's-degree granting
Predominantly bachelor's-degree granting
Predominantly certificate-degree granting


income vs. student__share_firstgeneration_parents_somecollege with color by school__degrees_awarded_predominant
income
120
80
40
0.25
0.50
0.75
student__share_firstgeneration_parents_somecollege
factor(school__degrees_awarded_predominant)
Entirely graduate-degree granting
Not classified
Predominantly associate's-degree granting
Predominantly bachelor's-degree granting
Predominantly certificate-degree granting

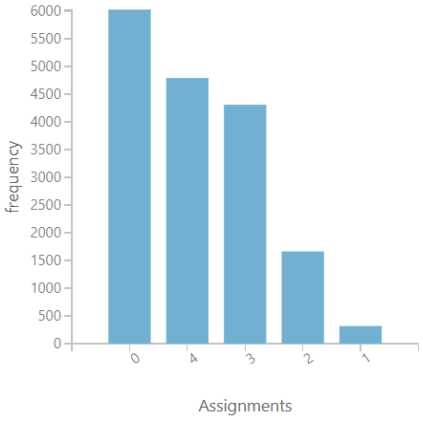
Similarly, school_ownership can be indicative of school__faculty_salary and student__share_firstgeneration_parents_somecollege.

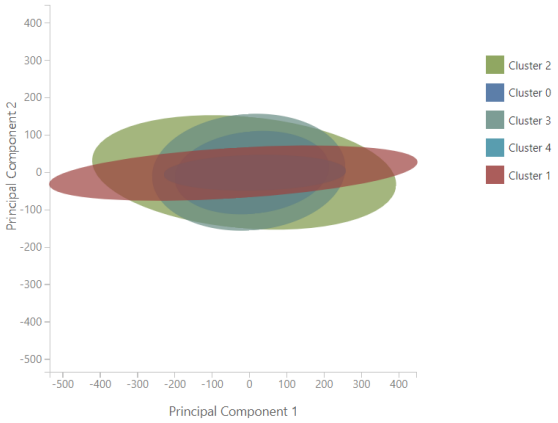




Classification of Schools based on Unsupervised Clustering

Unsupervised predictive model to classify Schools was created, using K-Means Clustering algorithm. Schools data is classified into 5 clusters. Below are provided visualizations of clusters





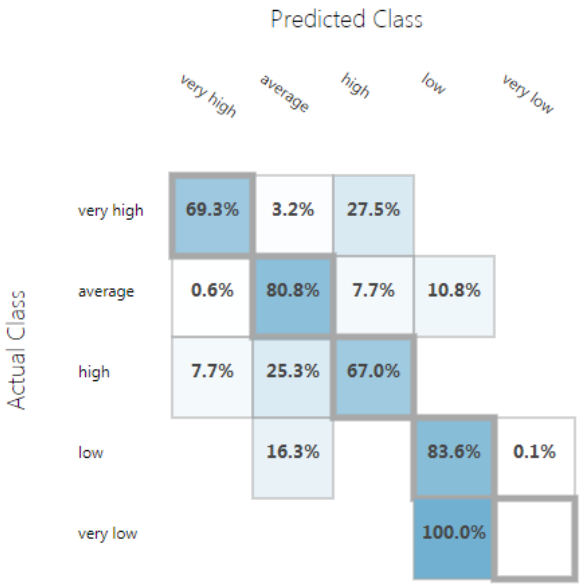
Multivalued Classification of Schools based on Students Income

Based on the analysis of the Schools, a predictive model to classify Schools into 5 Income categories was created. Following ranges were predefined

- Very low - Income <12K
- Low - Income is between 12k and 24k
- Average - Income is between 24k and 36k
- High - Income is between 36k and 48 k
- Very high – Income is higher than 48k

The model was created using the using Multiclass Logistic Regression algorithm and trained with 80%of the data. Testing the model with the remaining 20% of the data yielded the following results:

- Overall accuracy 0.78
- Average accuracy 0.91
- Micro-averaged precision 0.78
- Macro-averaged precision 0.61
- Micro-averaged recall 0.78
- Macro-averaged recall 0.60



Confusion matrix of the model is provided in the right.

Classification of Schools based on Students Income

Based on the analysis of the students' future income, two predictive models to classify Schools into two Income categories: Low (Income < 30,000$) and High (Income > 30,000$) were created, based on Two-Class Boosted Decision Trees and Two-Class Neural Network. Both algorithms were trained with 60% of the data. Testing the model with the remaining 40% of the data yielded.

Following results were received for the Two-Class Boosted Decision Tree:

• True Positives: 3549
• True Negatives: 2682
• False Positives: 339
• False Negatives: 273

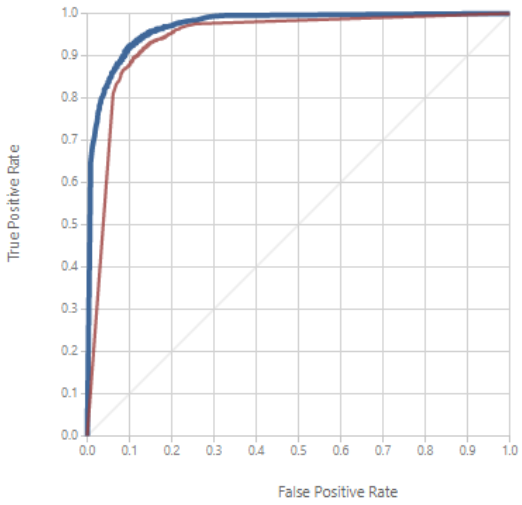
This translates in to the following standard performance metrics for classification:

• Accuracy: 91.1%
• Precision: 91.3%

• Recall: 92.9%
• F1 Score: 92.1%

For the Two-Class Neural Network were received:

• True Positives: 3523
• True Negatives: 2597
• False Positives: 424
• False Negatives: 299



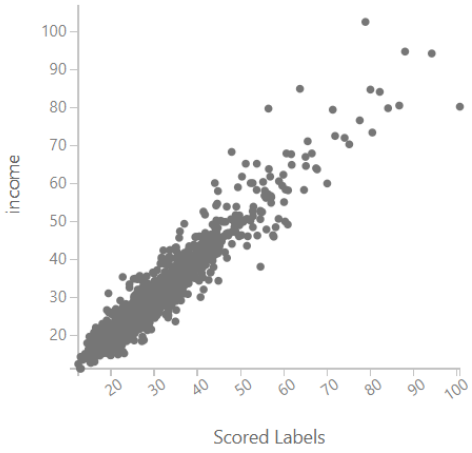
This translates in to the following standard performance metrics for classification:

• Accuracy: 89.4%
• Precision: 89.3%
• Recall: 92.2%
• F1 Score: 90.7%

The Received Operator Characteristic (ROC) curves for the models is shown here, with the blue line for the Two-Class Boosted Decision Tree and the red line for the Two-Class Neural Network. As the blue line is further then diagonal line showing the expected results of a random guess, it was concluded that Two-Class Boosted Decision Tree algorithm was showing slightly better results with selected parameters.

Regression

After creating a classification model to predict Income categories, a regression model to predict the future income of students was created. Based on the apparent relationships identified when analyzing the data, a boosted decision tree regression model was created to predict the value for the Income. The model was trained with 80% of the data, and tested with the remaining 20%. A scatter plot showing the predicted Income values and the actual Income values was shown in the picture.



This plot shows a clear linear relationship between predicted and actual values in the test dataset. The Root Mean Square Error (RMSE) for the test results was **3.19**. The standard deviation of Income was 10.62, which was higher than the variance, indicating that the model performs reasonably well.

Conclusion

This analysis has shown that the income of students can be confidently predicted from the higher educational institution characteristics.