Analysis of Student Earnings

Yang Yue, October 2017

1. Executive Summary

This report presents an analysis of data concerning the United States institutions of higher education and their students' earnings. The analysis and income prediction is based on 17107 observations of the institutions' information and the students' earnings at a set interval from when the student first enrolled.

Data exploration with descriptive statistics and data visualization has been done by using Python, Excel and Power BI, revealing the relationships between institution features and student earnings. After that, data cleansing and manipulation was performed in Python, and then a boosted decision tree regression model was trained and tested in Microsoft Azure Machine Learning Studio.

Below conclusions are presented in this report based on analysis and modeling:

- **school__degrees_awarded_predominant_recoded** -- the most important feature in the built regression model. The school with higher predominant degree awarded recoded value, tends to have higher student earnings.
- **school__degrees_awarded_predominant** -- the students from the institutions with entirely graduate-degree granting have apparently higher income.
- **school__degrees_awarded_highest** -- the highest median income and the highest max income both appear in the institutions which can award graduate degree.
- **student__share_firstgeneration** -- the students graduated from the institution with larger share of first-generation students are likely to have lower income.
- **school__faculty_salary** -- the higher faculty salary in a school, the higher income its student will earn, which is especially more significant in the private non-profit schools.
- Health and business marketing are the most popular disciplines.

2. Initial Data Exploration

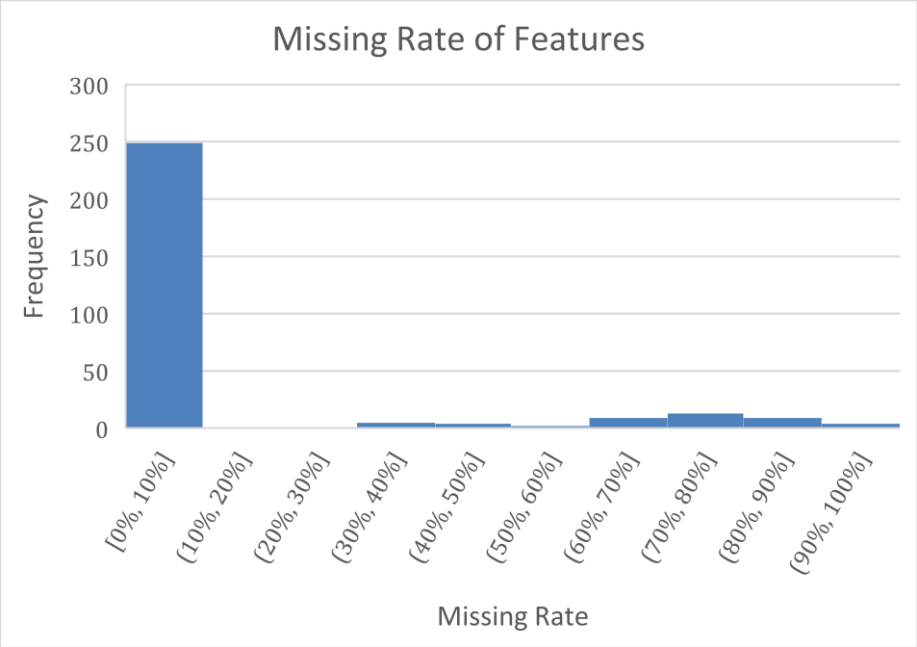
There are 297 features (variables or columns) and 1 label in this dataset. There are total 17107 samples (rows), and each sample represents a set of information from one United States institution of higher education, collected in one of four years denoted in variable **report_year**.

Label **income** is continuous numeric data. And 297 features can be divided into three data types (the number in parentheses indicates the count of relevant features):

- Numeric Features (98)
  - **academics__program_percentage** (38)
  - **admissions__** (25)
  - **completion__** (8)
  - **cost__** (3)
  - **school__faculty_salary** (1)
  - **school__ft_faculty_rate** (1)
  - **school__instructional_expenditure_per_fte** (1)
  - **school__tuition_revenue_per_fte** (1)
  - **student__** (19)
- Integer-type Categorical Features (190)
  - except for the above, other features under **academics__program** (190)
  - **school__degrees_awarded_predominant_recoded** (1)
- String-type Categorical Features (9)
  - **report_year** (1)
  - **school__degrees_awarded_highest** (1)
  - **school__degrees_awarded_predominant** (1)
  - **school__institutional_characteristics_level** (1)
  - **school__main_campus** (1)
  - **school__online_only** (1)
  - **school__ownership** (1)
  - **school__region_id** (1)
  - **school__state** (1)

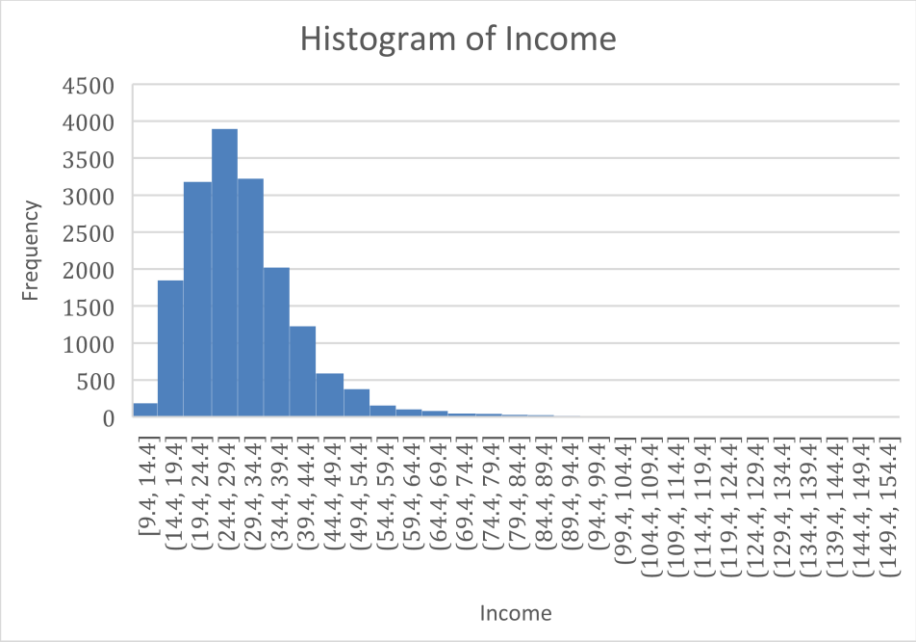
In addition, high missing rate was observed in some features as below histogram. Considering the effectiveness of analysis and modeling, also to avoid the model

deterioration caused by noise, the features with over 45% missing value were removed from the beginning (total 40 features were removed).



2.1 Individual Feature Statistics

For the label **income** which is the target of this project, the histogram shows that student income varied in a large range from 9.4 to 151.5, with significant right-skewed long-tailed distribution. The mean (30.6) and median (28.7) of student earnings is close to each other, showing that the distribution of the majority is still centralized at around 30.



For the numeric features, descriptive summary was done as below table, including min, mean, median, max, standard deviation and percentage of missing value. Noticed that the features with over 45% missing value were not included in this summary as per discussion in last section. And **student__demographics_first_generation** was also removed due to exactly as same as **student__share_firstgeneration**.

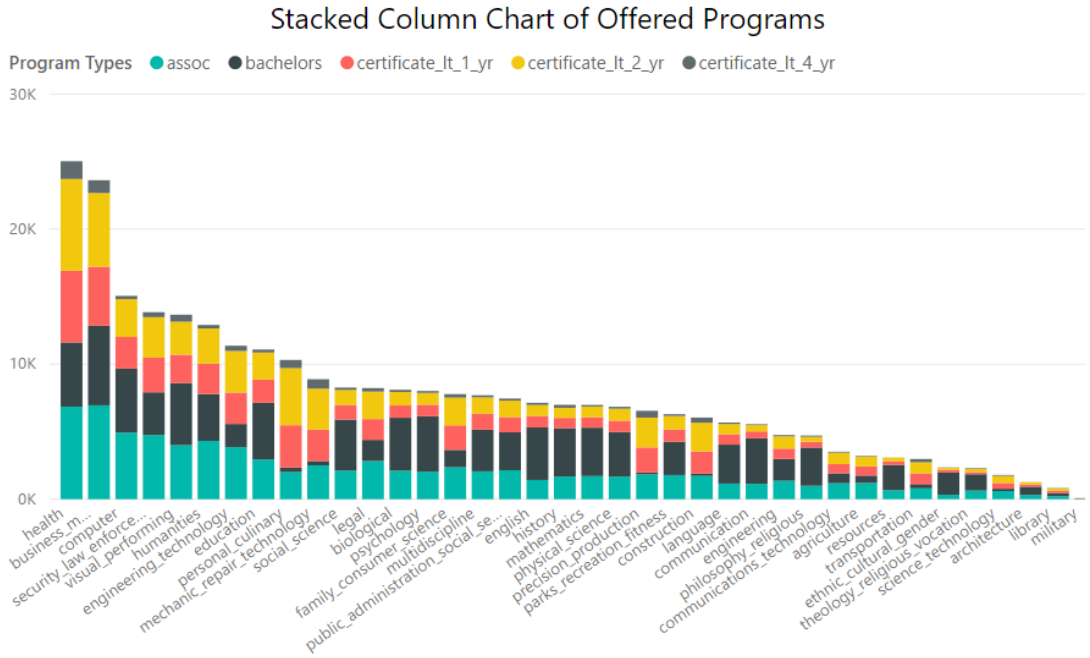
Feature	Min	Mean	Median	Max	Std	Miss %
average of **academics__program** group	0	0.026	0.005	0.858	0.065	5
cost__tuition_in_state	80	12613	11079	49793	9821	33
cost__tuition_out_of_state	80	14974	13305	49793	8926	35
school__degrees_awarded_predominant_recoded	1	2	2	3	0.867	0
school__faculty_salary	153	5798	5575	24892	2050	34
school__ft_faculty_rate	0	0.535	0.494	1	0.29	35
school__instructional_expenditure_per_fte	0	6435	4653	1183028	15484	5
school__tuition_revenue_per_fte	0	9134	7405	1056528	16672	5
student__demographics_age_entry	17.503	26.244	26.214	46.397	3.675	0
student__demographics_dependent	0.01	0.458	0.407	0.993	0.242	2
student__demographics_female_share	0.009	0.644	0.637	0.989	0.184	7
student__demographics_first_generation	0.062	0.48	0.507	0.952	0.126	4
student__demographics_married	0.004	0.186	0.18	0.851	0.1	6
student__demographics_veteran	0.001	0.032	0.023	0.417	0.031	37
student__part_time_share	0	0.229	0.165	1	0.235	5
student__share_25_older	0.001	0.402	0.403	1	0.216	30
student__share_first_time_full_time	0	0.548	0.557	1	0.251	41
student__share_firstgeneration	0.062	0.48	0.507	0.952	0.126	4
student__share_firstgeneration_parents_highschool	0.072	0.442	0.459	0.917	0.093	16
student__share_firstgeneration_parents_somecollege	0.048	0.527	0.501	0.938	0.124	9
student__share_independent_students	0.007	0.542	0.593	0.99	0.242	2
student__size	0	2810	760	249604	6034	5

For the categorical features, the detailed information of each is as below:

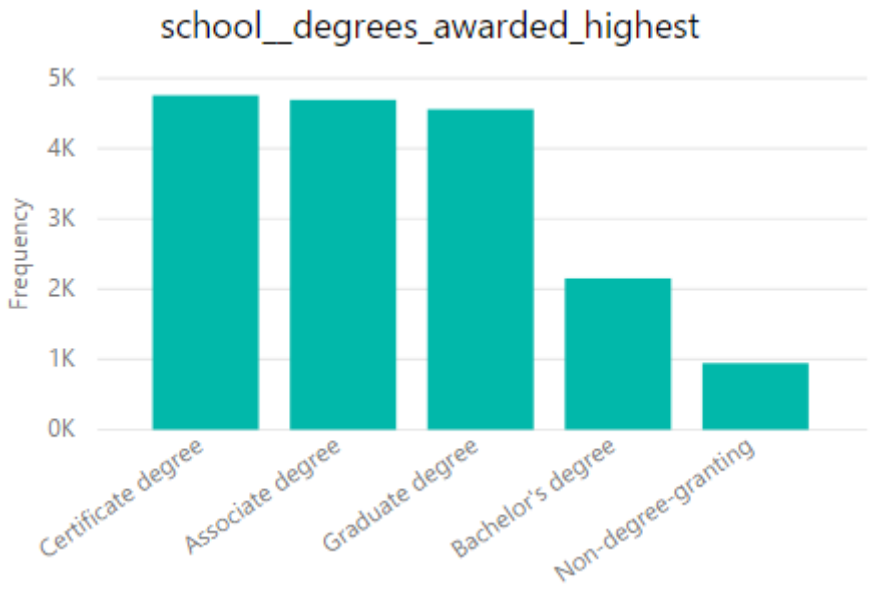
- **academics__program** except for percentage part -- particular program was not offered / offered / offered through and exclusively distance-education program
- **report_year** -- 4 particular years when the information was collected
- **school__degrees_awarded_highest** -- Non-degree-granting / Certificate degree / Associate degree / Bachelor's degree / Graduate degree
- **school__degrees_awarded_predominant** -- Not classified / Predominantly certificate-degree granting / Predominantly associate's-degree granting / Predominantly bachelor's-degree granting / Entirely graduate-degree granting
- **school__degrees_awarded_predominant_recoded** -- recoded 0s and 4s for predominant degree awarded
- **school__institutional_characteristics_level** -- 2-year / 4-year / Less-than-2-year
- **school__main_campus** -- Main campus / Not main campus
- **school__ownership** -- Public / Private for-profit / Private nonprofit
- **school__region_id** -- 9 region in IPEDS system
- **school__state** -- 57 state postcode

Based on the visualization of each above categorical features, following attributes were illustrated:

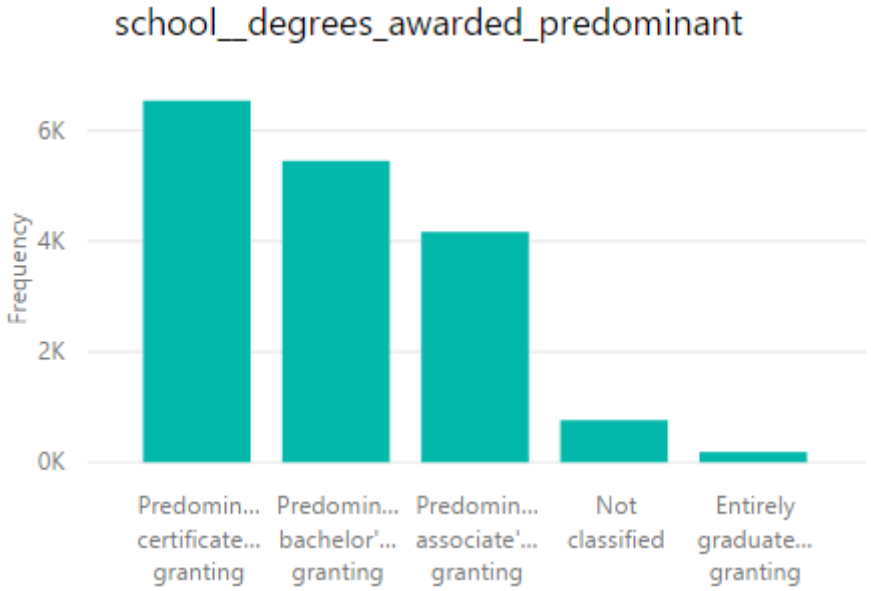
- Health and business marketing are the most popular disciplines (both **program offered** and **program offered exclusively online** were counted).



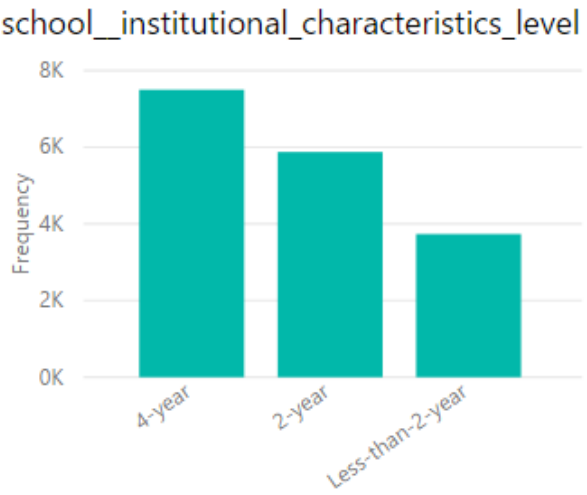
- Non-degree-granting school is relatively uncommon.



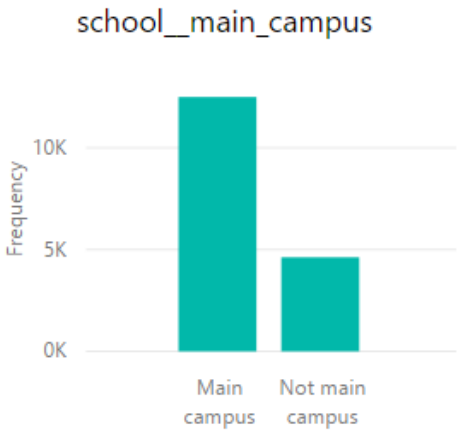
- Most institutions are predominantly granting certificate-degree, while extremely few institutions are entirely granting graduate degree.



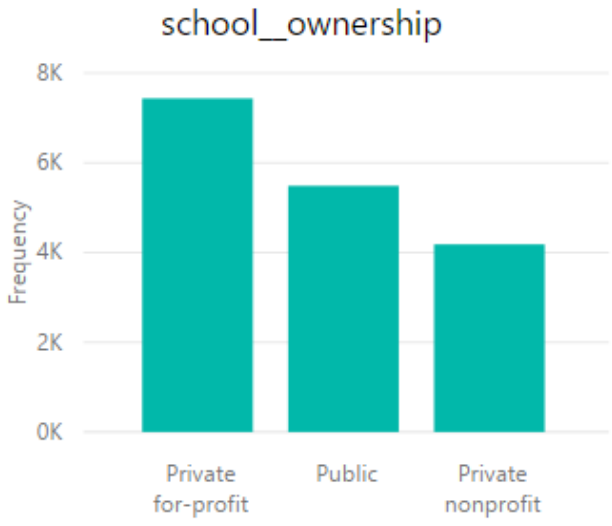
- Schools with 4-year institutional characteristics are more common than 2-year or less-than-2-year ones in this dataset.



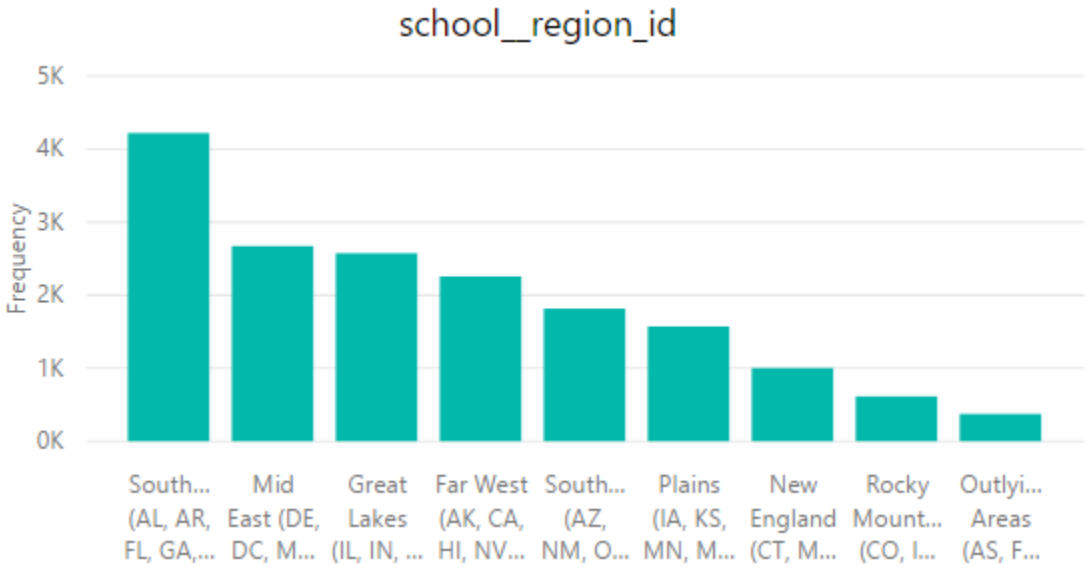
- Most of the information in dataset comes from main campus.



- Private for-profit institutions have more samples in this dataset.



- Around 25% of samples come from southeastern institutions.

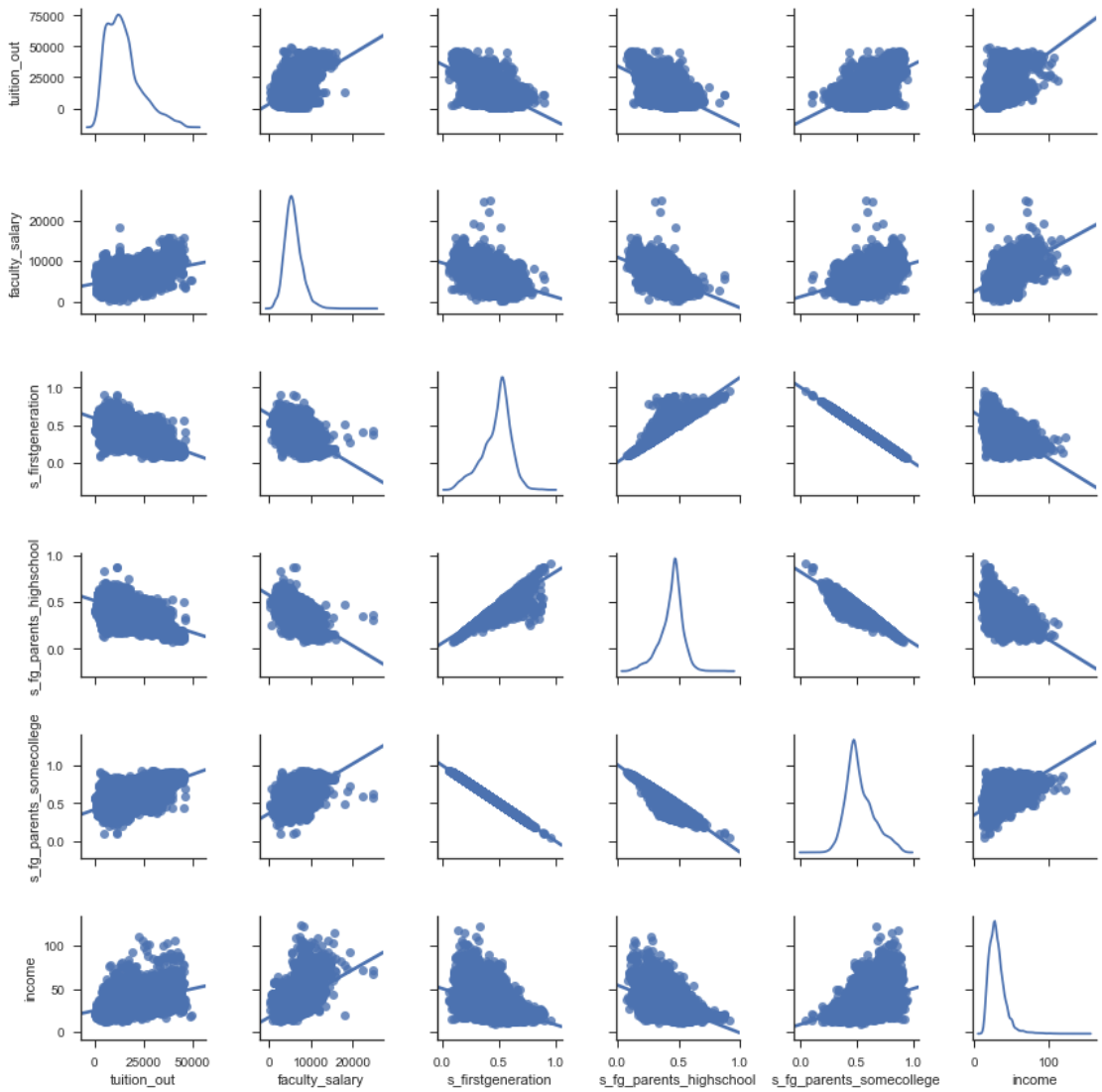


2.2 Correlation and Apparent Relationships

Correlation between features and label income was investigated and illustrated in this section.

2.2.1 Numeric Relationships

Pairwise scatter-plot matrix was created to understand the relationship between numeric features and income. And due to considerable features, the features with less than 0.45 correlation were not shown here.



From the plots in the bottom row and right-most column of above matrix, students from institution with higher out-of-state tuition or higher faculty salary tend to earn more. But the students graduated from the institution with larger share of first-generation students are more likely to have lower income.

Furthermore, if higher percent of the first-generation students whose parents have achieved a postsecondary educational level, the graduates of this school tend to have higher income. On the contrary, if higher percent of the first-generation students whose parents have only completed high school, then the student income tend to be lower.

However, above correlations between features and income are not particularly apparent, and their linearity are not so strong neither.

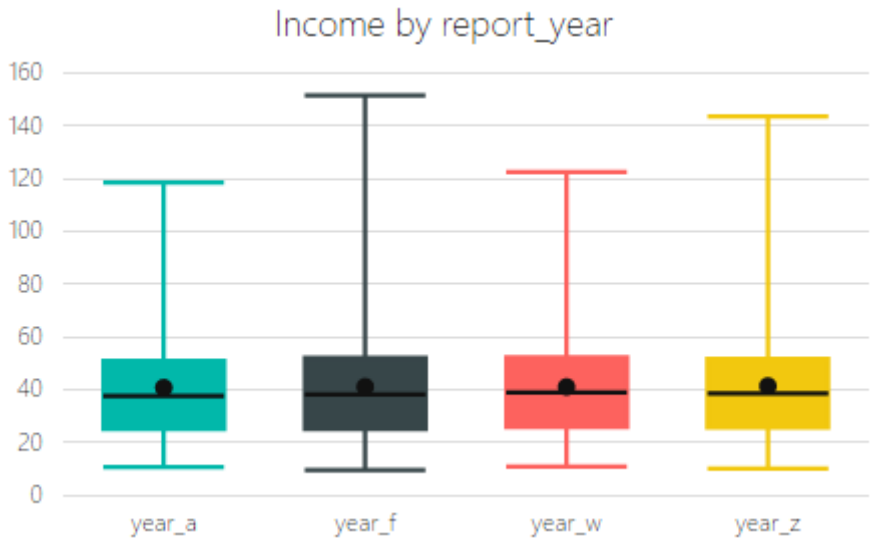
The correlation value matrix was attached as below. And the absolute values of last row and last column are around 0.5, which also prove the moderate correlation between features and income.

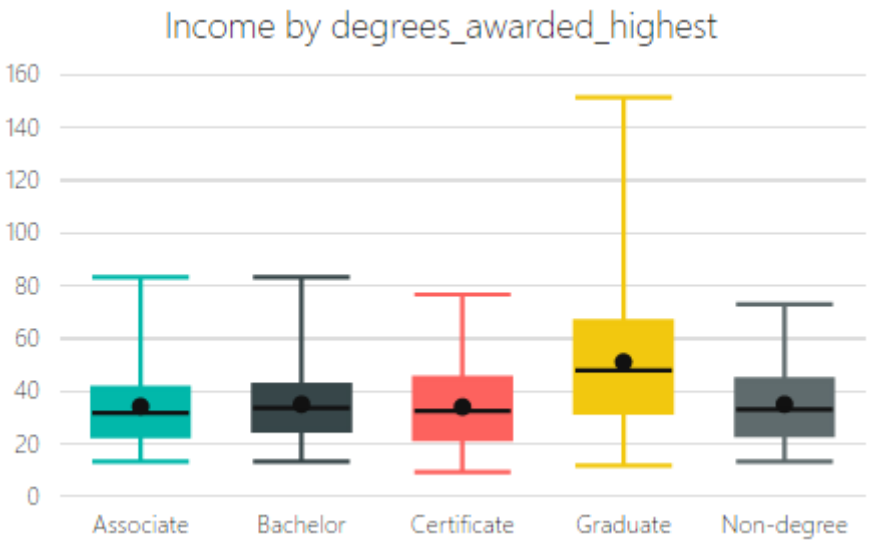
	cost__tuition_out_of_state	school__faculty_salary	student__share_firstgeneration	student__share_firstgeneration_parents_highschool	student__share_firstgeneration_parents_somecollege	income
cost__tuition_out_of_state	1	0.44	-0.65	-0.57	0.65	0.47
school__faculty_salary	0.44	1	-0.53	-0.58	0.52	0.53
student__share_firstgeneration	-0.65	-0.53	1	0.92	-1	-0.5
student__share_firstgeneration_parents_highschool	-0.57	-0.58	0.92	1	-0.94	-0.52
student__share_firstgeneration_parents_somecollege	0.65	0.52	-1	-0.94	1	0.49
income	0.47	0.53	-0.5	-0.52	0.49	1

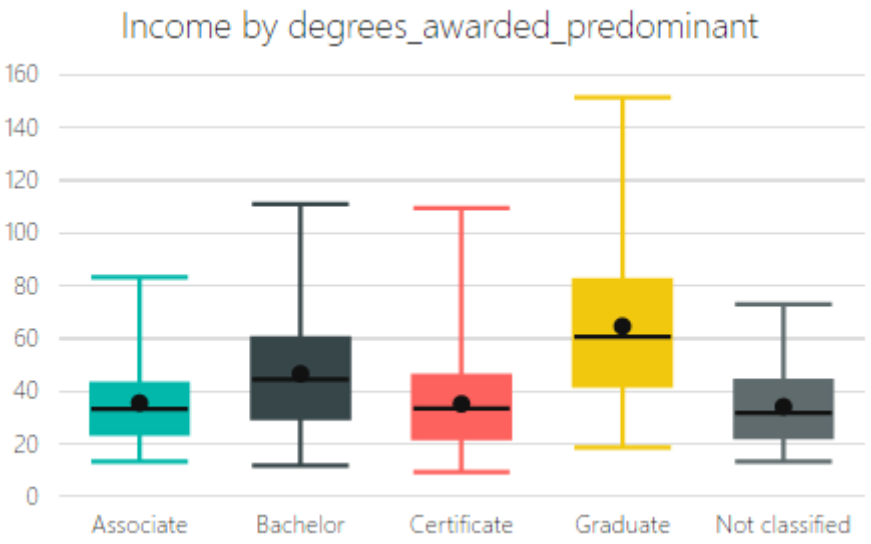
In addition, **student__share_firstgeneration**, **student__share_firstgeneration_parents_highschool** and **student__share_firstgeneration_parents_somecollege** are highly correlated to each other, meaning they are dependent variables, which can deteriorate some models which assume input features independent.

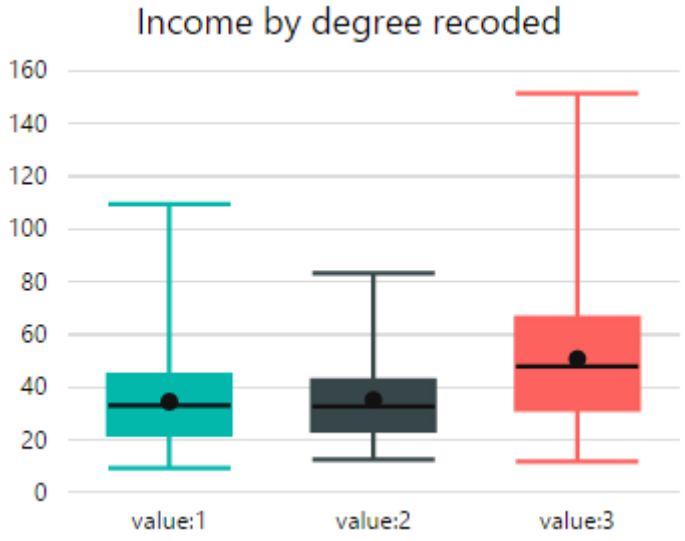
2.2.2 Categorical Relationships

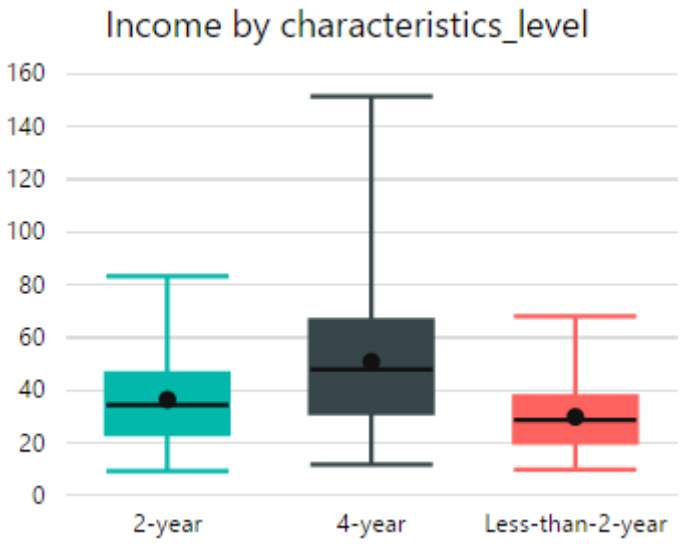
To discern the relationships between categorical features and income, following box plots were exhibited:

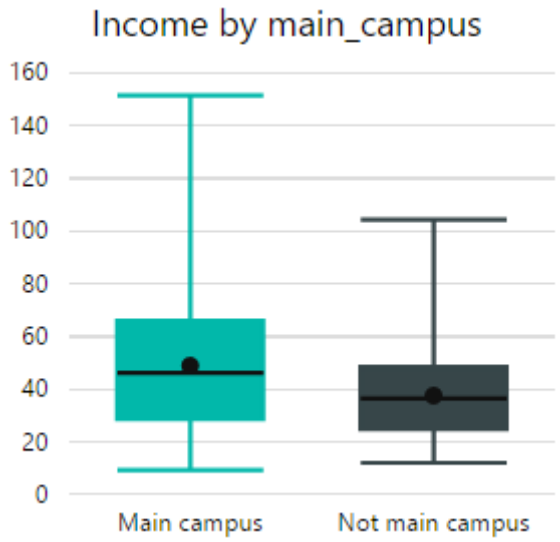


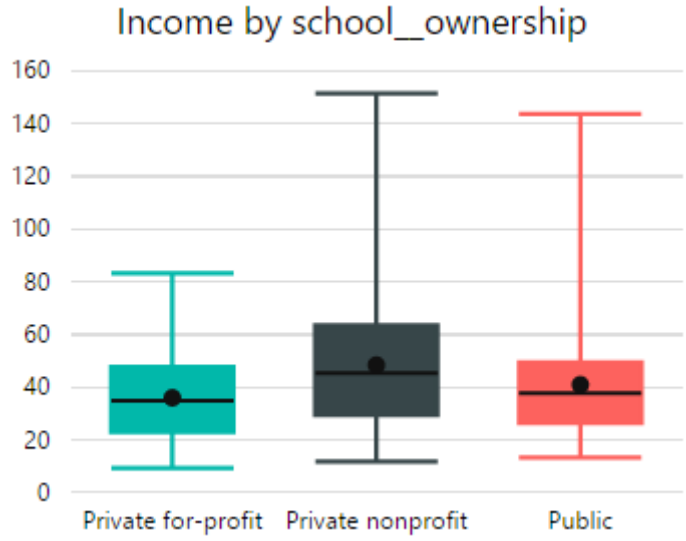

Income by degrees_awarded_highest
160
140
120
100
80
60
40
20
0
Associate
Bachelor
Certificate
Graduate
Non-degree

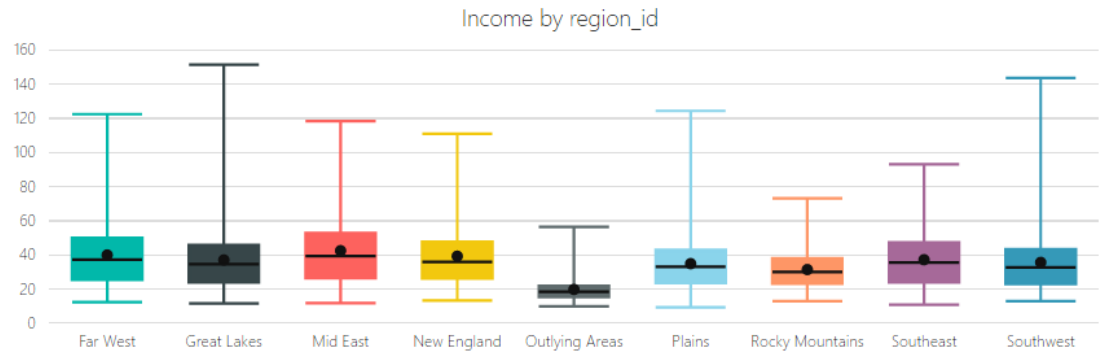

Income by degrees_awarded_predominant
160
140
120
100
80
60
40
20
0
Associate
Bachelor
Certificate
Graduate
Not classified

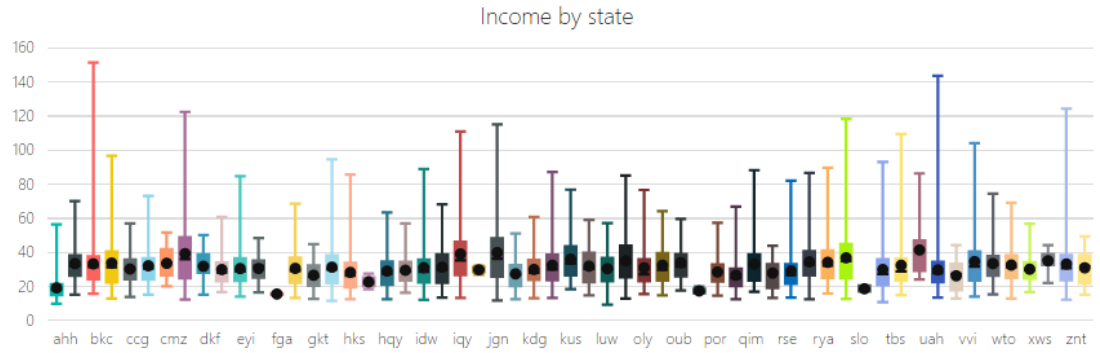

Income by degree recoded
160
140
120
100
80
60
40
20
0
value:1
value:2
value:3


Income by characteristics_level
160
140
120
100
80
60
40
20
0
2-year
4-year
Less-than-2-year


Income by main_campus
160
140
120
100
80
60
40
20
0
Main campus
Not main campus


Income by school__ownership
160
140
120
100
80
60
40
20
0
Private for-profit
Private nonprofit
Public





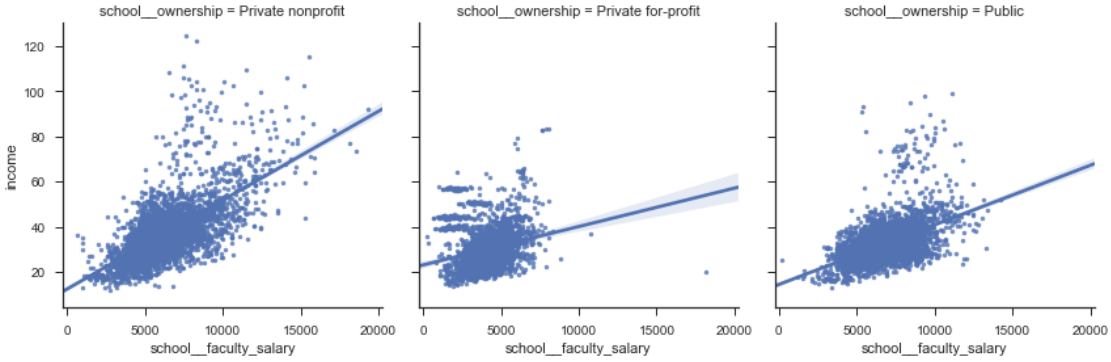
From above box plots, some conclusions about the relationships between categorical features and income were drawn:

- The income information collected in 4 particular years has no significant difference.
- Considering highest degree awarded, the highest median income and the highest max income are both appeared in the institutions which can award graduate degree.
- The students from the institutions with entirely graduate-degree granting have apparently higher income.
- The school with higher predominant degree awarded recoded value, tends to have higher student earnings.
- The institutions with 4-year characteristics level or private nonprofit ownership have higher student income.
- No remarkable income difference are found considering region and state of the institution.

2.2.3 Multi-faceted Relationships

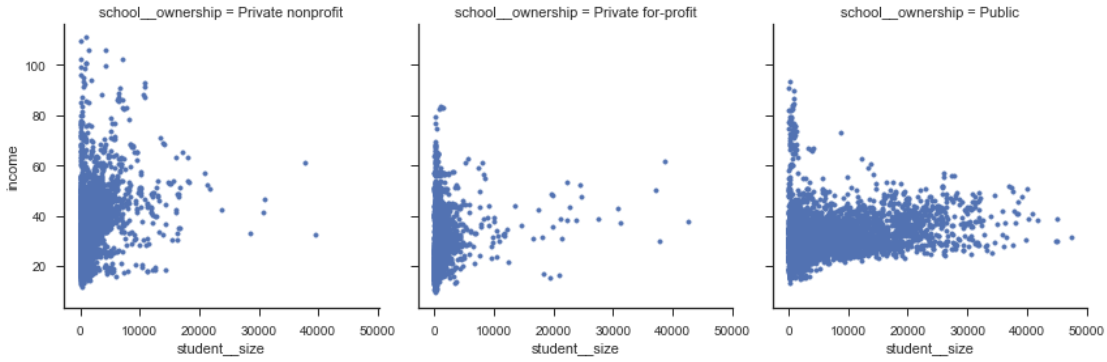
Faceted grid plot is a useful technique to analyze the same relationship conditioned on different levels of some variable. To be clear, only the most significant and meaningful relationships were shown below.

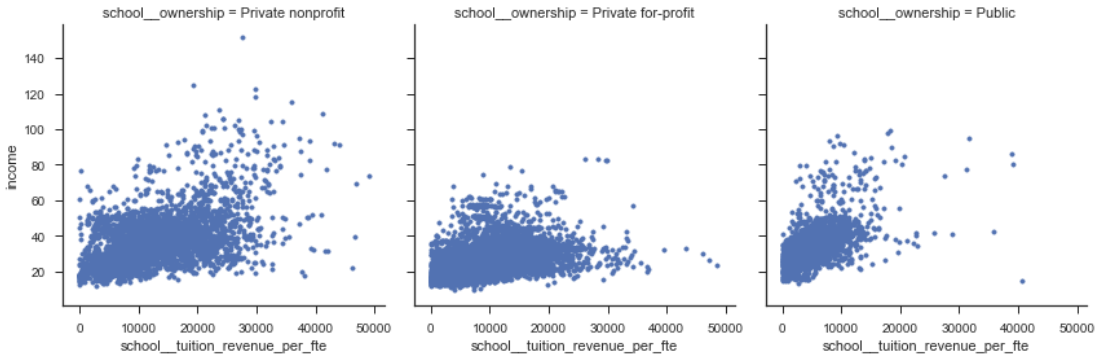
From previous discussion, private non-profit institutions have 20% higher median student income than public ones. And this can be explained by their different correlations between their faculty salary and income as below faceted plots.



Although the distribution of faculty salary is similar for both private non-profit and public institutions, the private non-profit institutions have stronger correlation (0.652) between faculty salary and student income, compared to the public ones (0.537). Intuitively speaking, if a private non-profit school pays more to its faculties, normally its graduated students will have higher income, while this is less significant in a public school.

Moreover, at an identical level of student income less than 60k, the public institutions cover large range of student size, while the most of private non-profit ones keep small size. However, also at a same income level, the most of private non-profit schools have higher net tuition revenue per full-time equivalent student than the public schools.





3. Regression Model of Student Earnings

3.1 Data Manipulation

Before feeding data into the machine learning model, following manipulation of raw data was performed using Python to improve model predictive power. And if not specified, the column operations were applied on both train values and test values dataset, while the row operations were applied on both train values and train labels dataset.

1. Remove less value features and samples
    - remove columns with more than 45% missing values in train values dataset, then remove same features in test values dataset
    - remove rows with more than 85% missing values
2. Remove features highly correlated to/dependent on **student__share_firstgeneration** (absolute correlation value over 0.9) due to using decision tree based model
    - **student__demographics_first_generation**
    - **student__share_firstgeneration_parents_somecollege**
    - **student__share_firstgeneration_parents_highschool**
3. Transform string-type categorical features into dummy/indicator variables
4. Fill missing value NaN
    - fill mode value for NaN in integer-type categorical features
    - fill median value for NaN in numeric features

The prepared train values dataset had 335 features and 16340 samples.

3.2 Modeling and Testing

Because the dataset has both numeric and categorical variables, to predict the student's income, tree based regression model is more suitable compared to other regression models. And after comparing multiple tree based regression models, boosted decision tree regression model in Azure Machine Learning Studio was chosen.

Following experiment was established for model training and tuning:

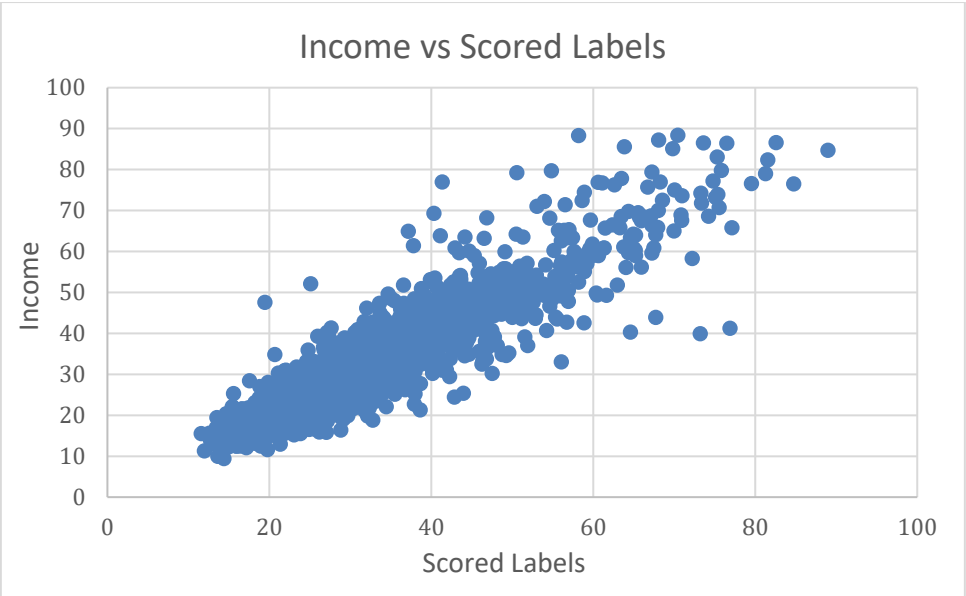
1. Left join train values and train label dataset
2. Remove **row_id**
3. Define categorical features and label (Edit Metadata module)
4. Split data into 70% training set and 30% testing set
5. Tune model by sweeping hyperparameters with cross validation using training set
6. Retrain model using the training set
7. Validate model using the testing set

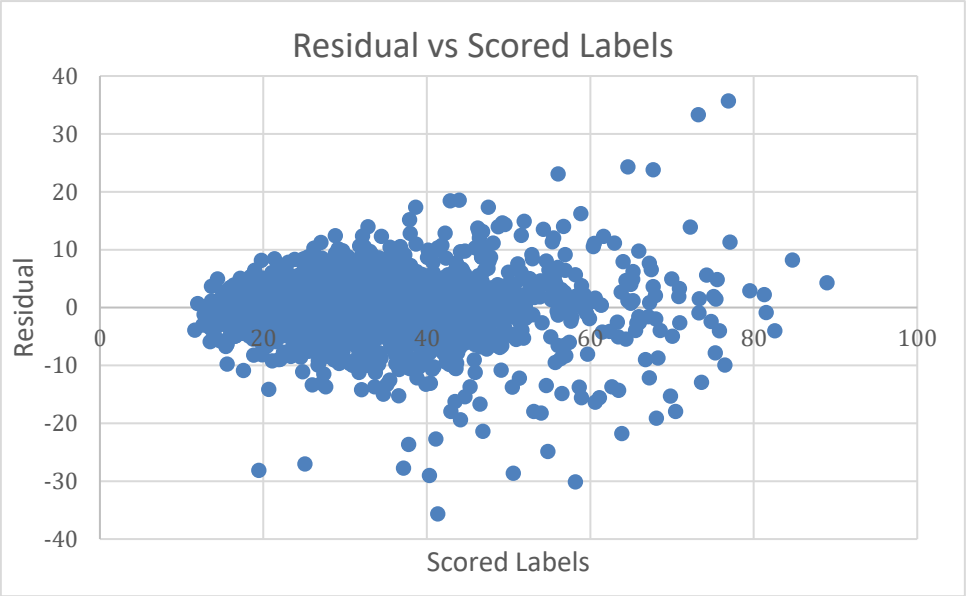
3.3 Model Evaluation

The selected trained model had RMSE (root-mean-squared error) result as below:

- mean of 10-fold cross validation: 3.5908
- the 30% testing set: 3.7221
- leaderboard score: 3.8298

By using the 30% testing set to validate the model, the scatter plots of income by scored labels (predicted values) and residual by scored labels were created to visualize the performance of model.

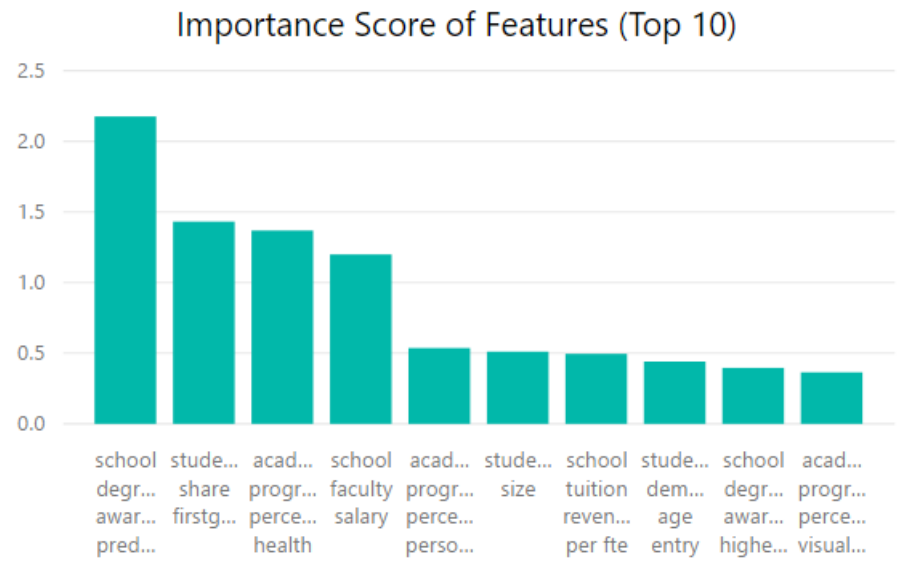




From the income plot, a notable linear correlation between real income value and predicted value is shown, which illustrates that the regression model is working well. To be more detailed, the model has strong predictive ability at the income range from 10k to 50k, but it is less efficient for the income over 50k cases which is explained by the second plot.

The residual plot shows the relationship between residual (the difference between real income value and predicted value) and scored labels. It is obvious that when predicting the income over 50k, the variance of residual increases and not as stable as the cases with 10k-50k income. However, the residual is still randomly distributed around the zero-middle line which is preferred.

In addition, features importance was also evaluated by Permutation Feature Importance module, and the top 10 important features were as shown below.



Combining with the findings in *Correlation and Apparent Relationships* chapter, the most important feature **school__degrees_awarded_predominant_recoded** shows that the higher recoded value the higher income. And both 2nd and 4th important features, **student__share_firstgeneration** and **school__faculty_salary**, have around 0.5 correlation with income.

4. Conclusion

Based on the analysis and modeling result of this project, the student income can be well predicted by current dataset with necessary manipulation. The most important features of prediction (importance score over 1) are **school__degrees_awarded_predominant_recoded**, **student__share_firstgeneration**, **academics__program_percentage_health** and **school__faculty_salary**. And categorical feature **school__ownership** provides additional grouping information of student income distribution.