Predict the future income of students

Sandhya Anand Kumar

Predict the future income of students

Sandhya Anand Kumar

1. Introduction

Student loan debt is an important concern right now. As tuition costs rise, more and more students are taking out loans to cover the cost of their educations. In fact, student loan debt in the US has nearly tripled over the last decade to more than $1.25 trillion. The goal of this capstone project is to predict the future earnings of these students a set number of years after they initially enrolled.

2. Summary

After understanding the data using Summary statistics, exploratory analysis the next step is feature selection. There were 298 features across 17,107 samples. Out of these 298, more than 50 features were holding binary data (0/1). These features were converted into categorical features. In total there were 198 categorical features, and 100 numerical features that needed were evaluated.

Prepping the data: Handling NAs. More than 80% of the data were NAs. There wasn't even one complete case in the train data set. PMM data imputation technique was used for imputing numerical features. A new NA factor level was added to the categorical features.

Feature Selection: The Following approach was used for feature selection.

1. Filter out most inter correlated (>0.90) numerical features from the feature set
2. Filter out numerical features that are not significantly correlated with prediction variable "income."
3. Filter out categorical features by chi-squared test that's not significant

List of features used to predict income.

1. admissions__sat_scores_average_by_ope_id
2. school__faculty_salary
3. student__share_firstgeneration_parents_highschool
4. report_year
5. school__institutional_characteristics_level
6. school__main_campus
7. school__ownership
8. school__region_id
9. school__degrees_awarded_highest
10. school__degrees_awarded_predominant
11. school__degrees_awarded_predominant_recoded

After feature selection, various models were built. Linear, Radom and boosted random models were built. Test data was used to predict the income variable. The model with the least **RMSE minimum Root Mean Square Error** was selected. This shows the predictability of student income based on the above feature sets.

3. Data Exploration

Three different data files were provided.

Train_data: which contains the training data set (17,017 samples 298 features)

Train_labels: This contains the income for all rows in Train_data set (17,017 samples, two features)

Test_data: This includes the dataset to test your model before submission. (9192 samples, 298 features)

4. Explore Train Data

As stated earlier the data file consists of 17,017 samples 298 features. 198 of them were converted as categorical variables (as they were binary features 0/1) and 100 were numerical variables. There wasn't even one complete case.

Here is the summary statistics of the data before imputation.

Descriptive Statistics	*admissions__sat_scores_average_by_ope_id*	*school__faculty_salary*	*student__share_firstgeneration_parents_highschool*
Mean	1055.509103	5797.635609	0.441961452
Standard Error	1.898073416	19.26917509	0.000779689
Median	1044	5575	0.458539778
Mode	1010	5120	0.46852687
Standard Deviation	125.0281576	2050.424944	0.093237303
Sample Variance	15632.0402	4204242.453	0.008693195
Kurtosis	0.922106666	3.136692288	1.250706126
Skewness	0.668894708	0.928928228	-0.64381374
Range	860	24739	0.844733461
Minimum	660	153	0.072131148
Maximum	1520	24892	0.916864608
Sum	4579854	65646628	6320.048759
Count	4339	11323	14300

4.1 Income:

The feature of interest **income** was further analyzed.

Mean(Income) = 30.59 and

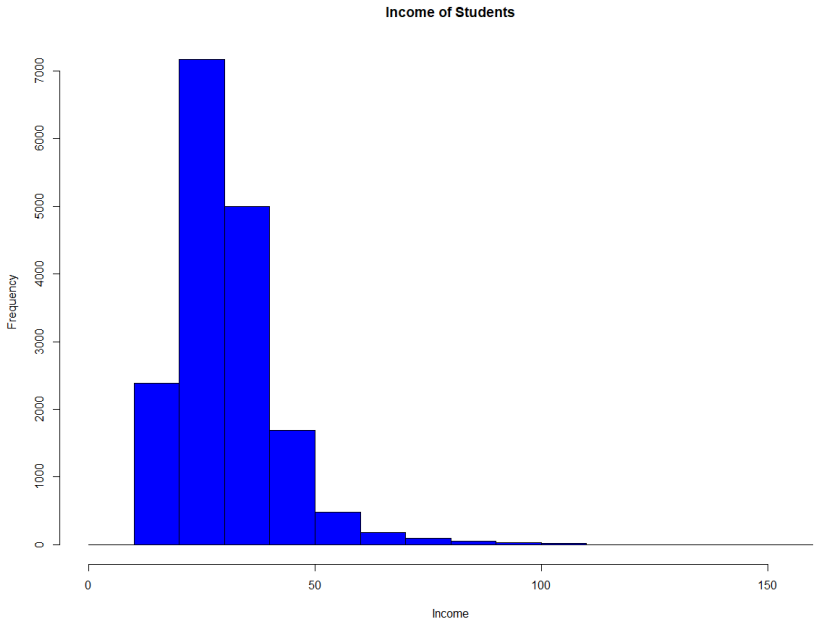
Median(Income) = 28.70.

Skeweness = 1.95594

Kurtosis = 7.971642

Note the **skewness** is 1.95594 which means the curve is not perfectly belled shape. **Kurtosis measures** the relative size of two tails. For a completely normal distribution, this value is 3. In our case, the value is 7.9 meaning the tails are not equal.

The following histogram of income shows the skewness and kurtosis.

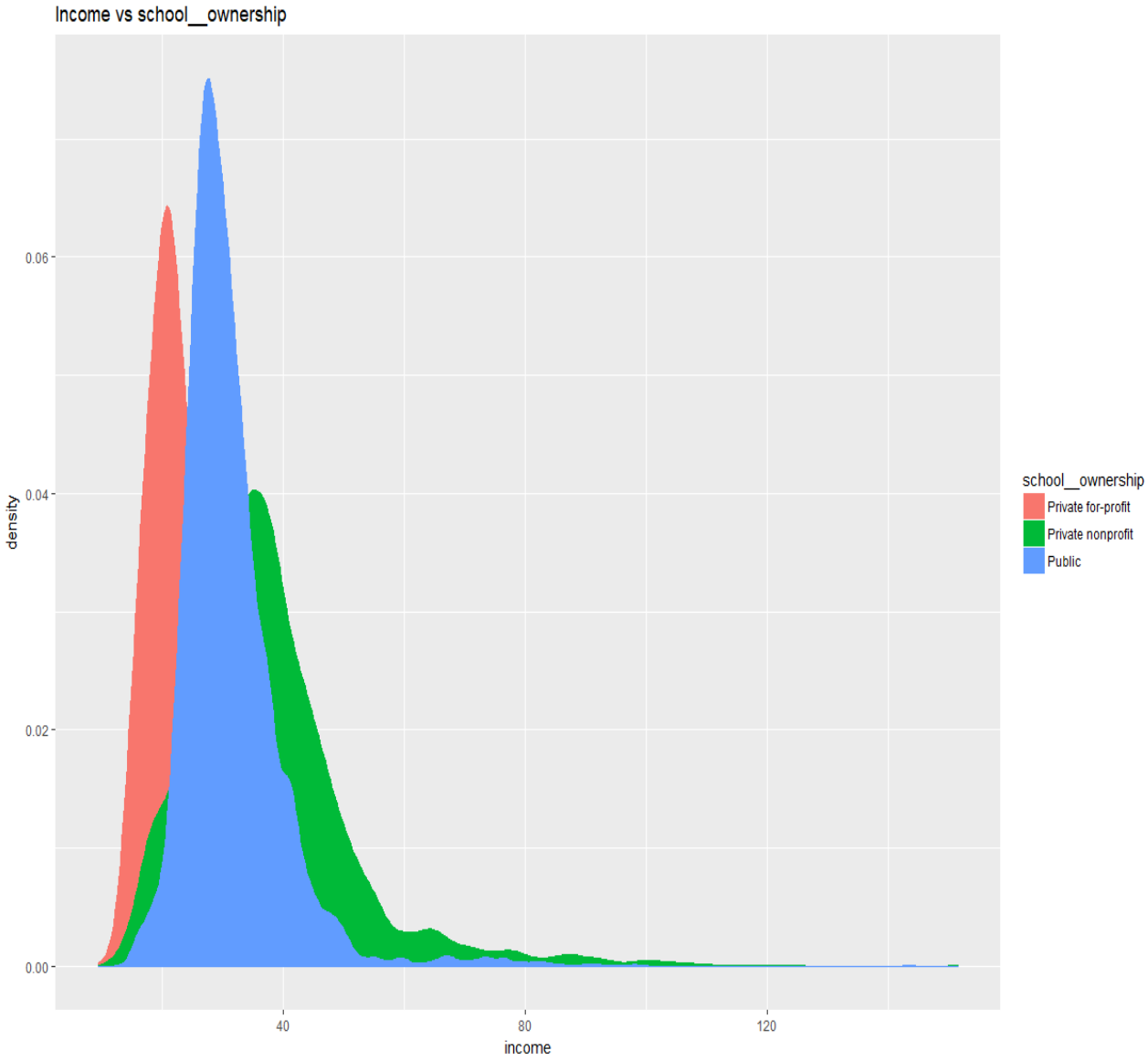


4.2 Exploratory Analysis

The following section shows the exploratory analysis between few key features that are used for modeling and income.

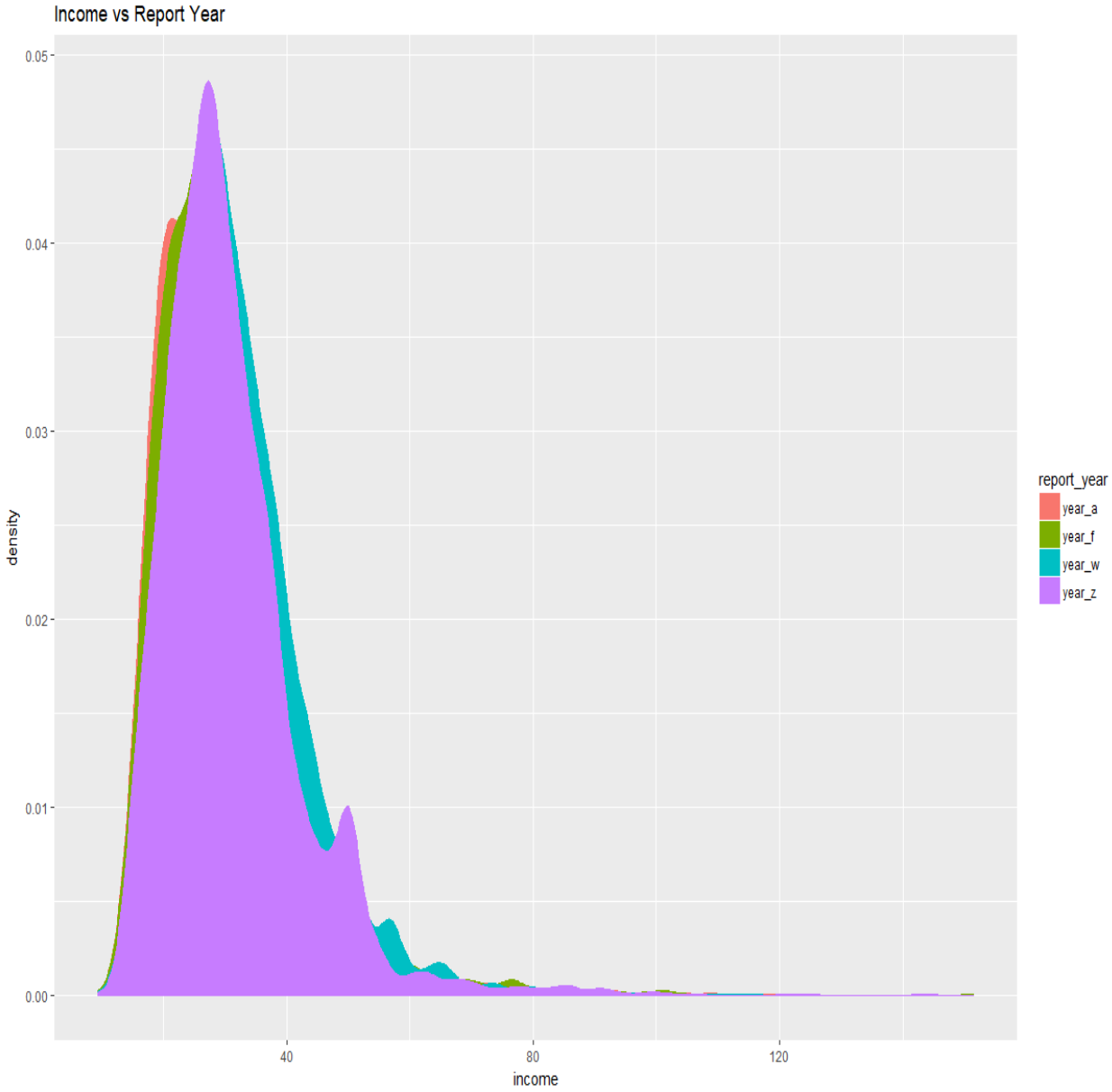
4.2.1 Income Vs. School ownership

The chart shows that there is the difference in income based on the school ownership. The income of students who studied in **Private Non-profit school** is higher than other institutions.



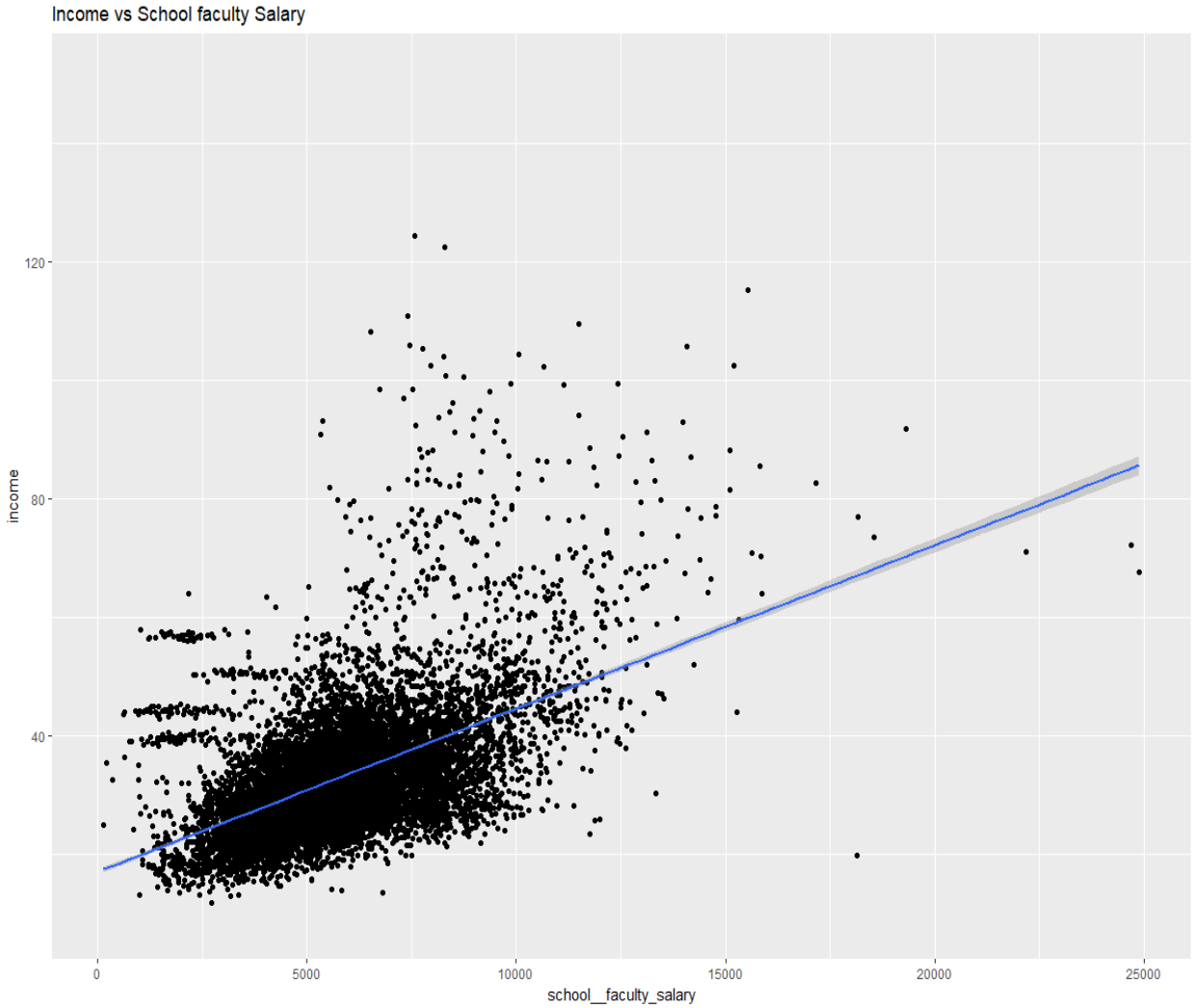
4.2.2 *Income Vs. Report Year*

Student studies during Report year "**year_w**" seems to have higher median income than other report years.



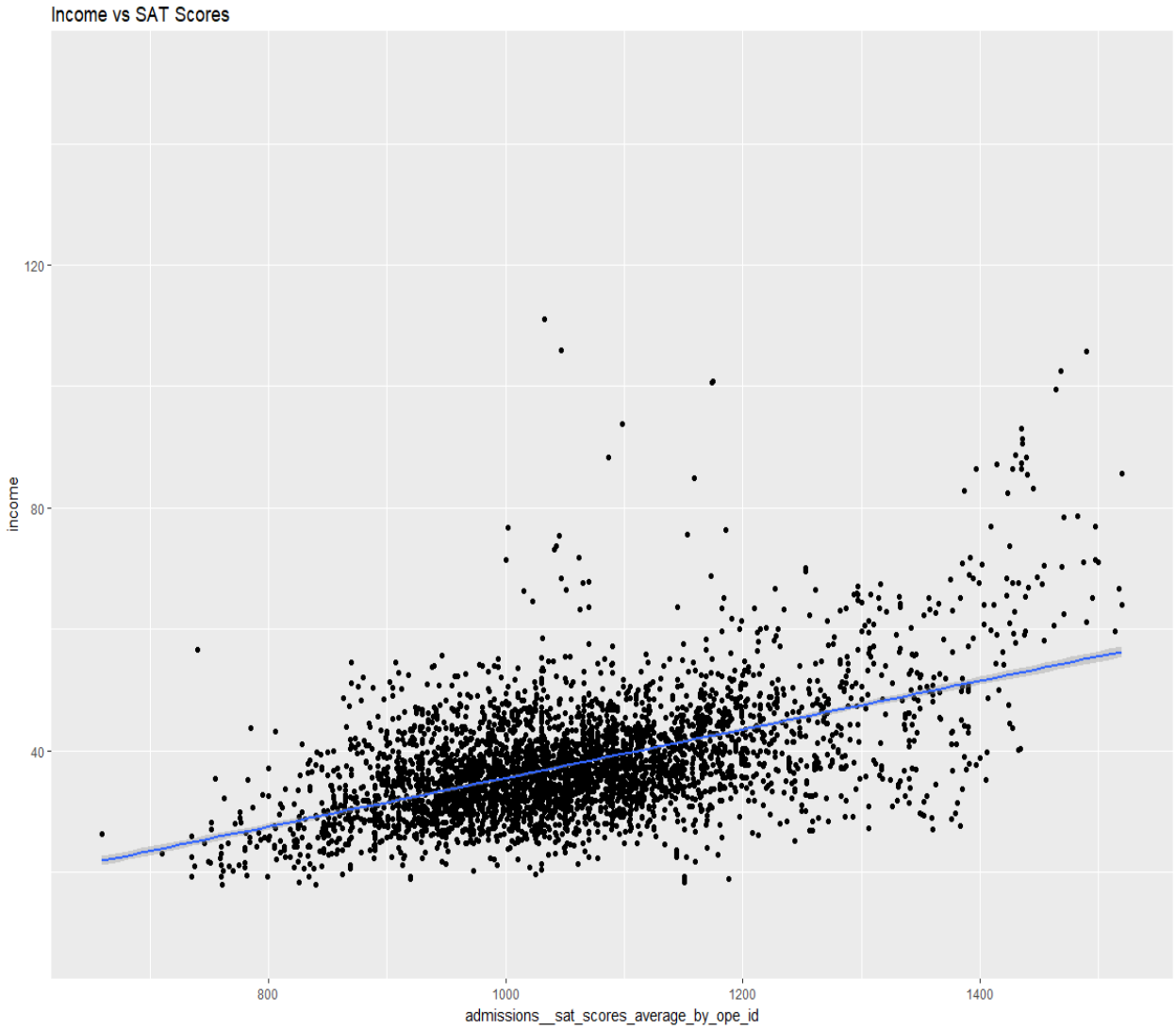
4.2.3 *Income Vs. School faculty Salary*

Faculty Salary has a **positive** correlation with the student's income. The following chart shows a positive linear relationship.



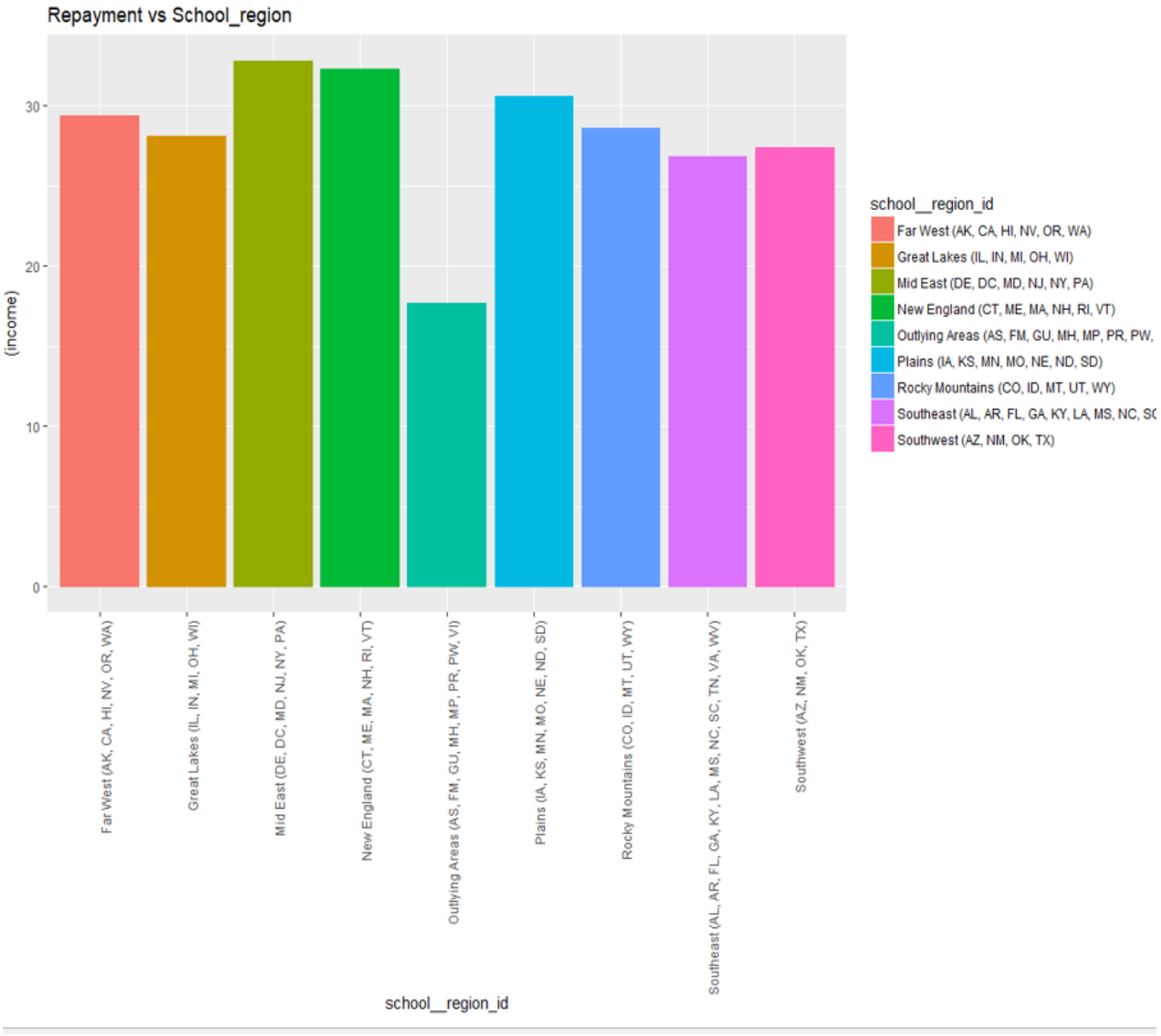
4.2.4 *Income Vs. Admissions SAT Scores Average by OPE ID*

The following chart shows a **positive** linear relationship between income and SAT scores average by OPE ID.



4.2.5 *Income Vs. School Region*

Students are graduating from schools in Mid east region have a higher income than other regions.

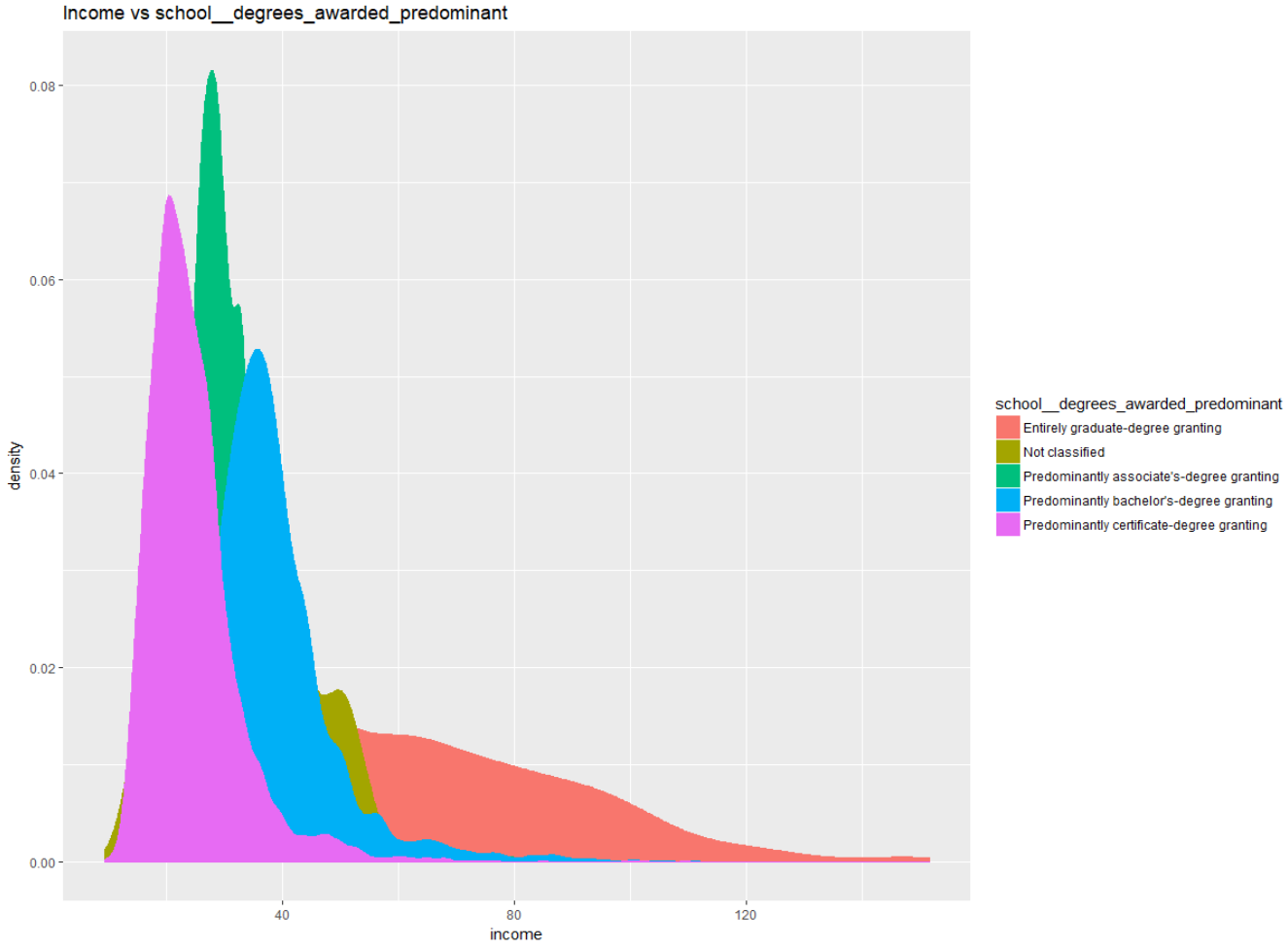


4.2.6 *Income Vs. School degrees awarded predominant*

The median income of students from schools which receives entirely graduate degree granting has higher incomes compared to other schools.

	school__degrees_awarded_predominant	income
1	Entirely graduate-degree granting	60.3
2	Not classified	30.4
3	Predominantly associate's-degree granting	28.4
4	Predominantly bachelor's-degree granting	36.2
5	Predominantly certificate-degree granting	23.1

The chart shows the variation in income by degree awarded predominant



4.2.7 *Income Vs. School degrees awarded predominant recoded*

The median income of students from schools with school degree awarded predominant recoded (3) has higher incomes compared to other schools

	school__degrees_awarded_predominant_recoded	income
1	1	22.9
2	2	27.9
3	3	36.45

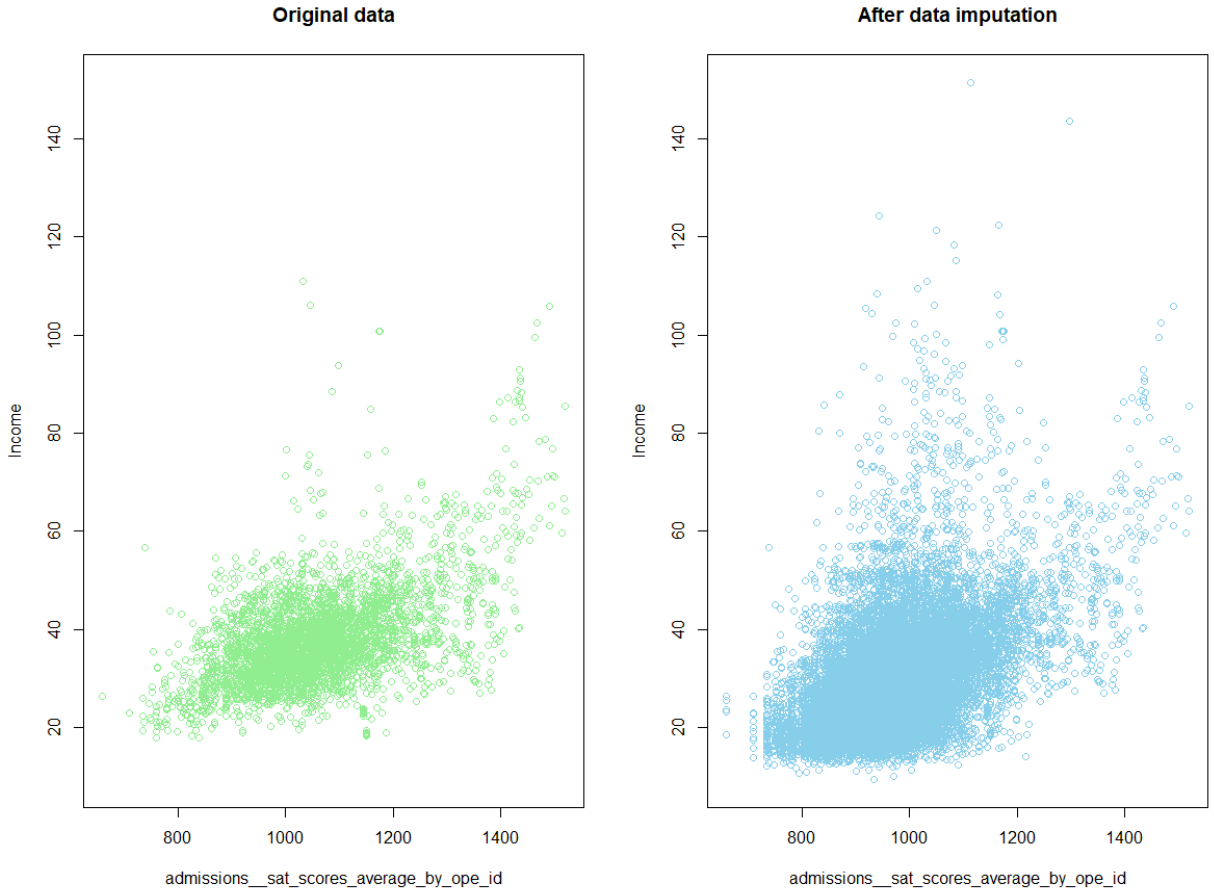
5. Data Imputation

There were close to **35% of train data that had NA a**nd had no complete cases. Most of the feature sets were missing data. Inorder to understand the correlation and proper modeling, Data imputation was used. Data were imputed using the **Predictive mean matching ("pmm")** method to calculate the missing values. After imputation only five feature sets has NA and the overall percentage of NA was **reduced from 35% to 7%.** As an example check the following statistics for the feature **"admissions__sat_scores_average_by_ope_id"**

Before Imputation: admissions__sat_scores_average_by_ope_id	
NA Values	Data
12768 /17107 = ~75%	4339/17107= ~25%

After Imputation: admissions__sat_scores_average_by_ope_id	
NA Values	Data
0	17107 = **100%**

The following charts show the effect of data imputation for the feature set admissions__sat_scores_average_by_ope_id. The Blue charts have more denser points showing data imputation.

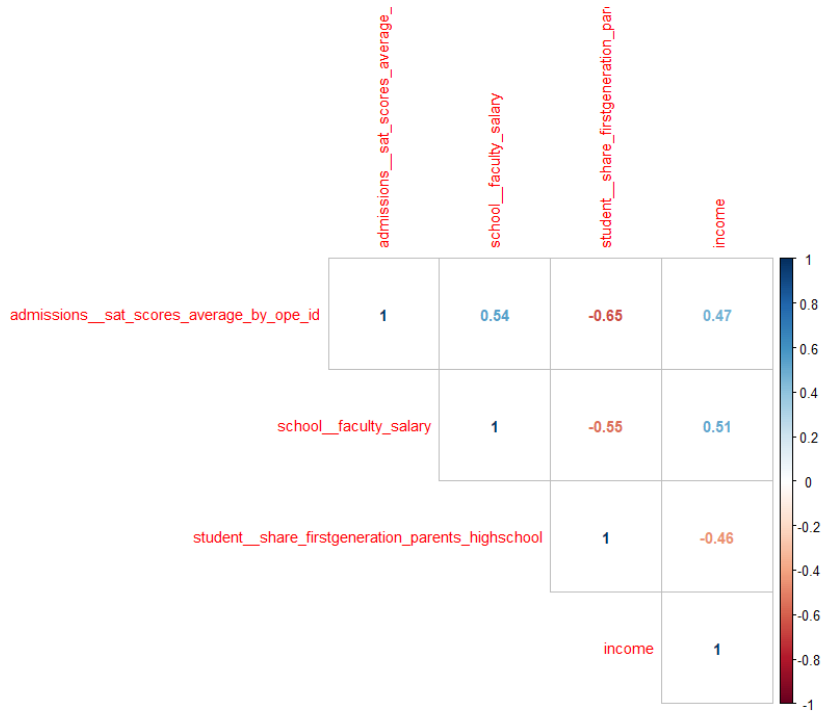


6. Correlation: Numerical features

After exploring the data, an attempt was made to understand the correlation between income and other numeric features. Following plot shows the relationship between income and other features.

The darker the square, the more the correlation between the features. Color Blue represents positive correlation, and RED represents negative correlation.

The intercorrelation among features is evident from this plot. The model is updated removing inter-correlated features.



Key numerical features that are considered for final modeling are

1. admissions__sat_scores_average_by_ope_id
2. school__faculty_salary
3. student__share_firstgeneration_parents_highschool

7. Categorical variable: Significance

Chi-squared Test was performed to understand the significance levels of "categorical" features. Features that had a resultant p-value **<0.05 were selected.**

Following were the 8 "**categorical**" features that had a **p-value < 0.05** which were significant.

#	P-Value	Feature
1	0	school__institutional_characteristics_level
2	0	school__ownership
3	0	school__degrees_awarded_highest
4	0	school__degrees_awarded_predominant
5	0	school__degrees_awarded_predominant_recoded
6	1.40E-112	school__region_id
7	4.68E-24	school__main_campus
8	4.04E-14	report_year

8. Regression Models

Based on the apparent relationships identified when analyzing the data, three models were created to predict the income. A linear regression model, a Random Forest model and several Randome Forest Model with fine tuned mtry, Tunelength parameter models.

As a final step the Test data was used to predict the income, and the final RMSE for each model was captured. The model with minimum RMSE was selected as the best model which in this case is "Random forest model with 500 decision tree, Mtry= 20 and tune length = 15".

Model	RMSE	Comments
Random Forest	8.1656	Mtry=3
Linear Regression	21.7193	0.6746627
Random	6.7462	Mtry=20, tunelength=20

9. Conclusion

This analysis shows that the students income can be confidently predicted using the feature sets. The income is **positively** correlated to

1. school__faculty_salary
2. admissions__sat_scores_average_by_ope_id

Negatively correlated to

1. student__share_firstgeneration_parents_highschool

And impacted by these significant **categorical** variables

2. school__institutional_characteristics_level
3. school__ownership
4. school__degrees_awarded_highest
5. school__degrees_awarded_predominant
6. school__degrees_awarded_predominant_recoded
7. school__region_id
8. school__main_campus
9. report_year