**Data Science**

**Capstone Project-**

**Predict Student Loan Repayment**

**July 2017**

Executive Summary:

The goal is to predict repayment rate given the details of the students at United States institutions of higher education. This observation is based on large number of observations and various features. This specific dataset covers two particular years.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between characteristics feature and repayment rate were identified. After analyzes and exploring data, linear regression to predict repayment rate from its feature was created.

Important features found in the dataset to predict repayment rate are

* **Ownership of the Institute** - Control of institution

((Possible Values: Private nonprofit, Private for-profit, Public)

* **Institutional Characteristics level**- Level of institution
* **Region** - Region (IPEDS)
* **Loan Principal**-The median original amount of the loan principal upon entering repayment
* **Family Income**-Number of students in the family income cohort
* **SAT Scores**-Average SAT equivalent score of students admitted
* **Student Average Income**- Average family income of dependent and independent students in real 2015 dollars.
* **Marital Status**-Share of married students
* **Retention Rate**- Student retention rate

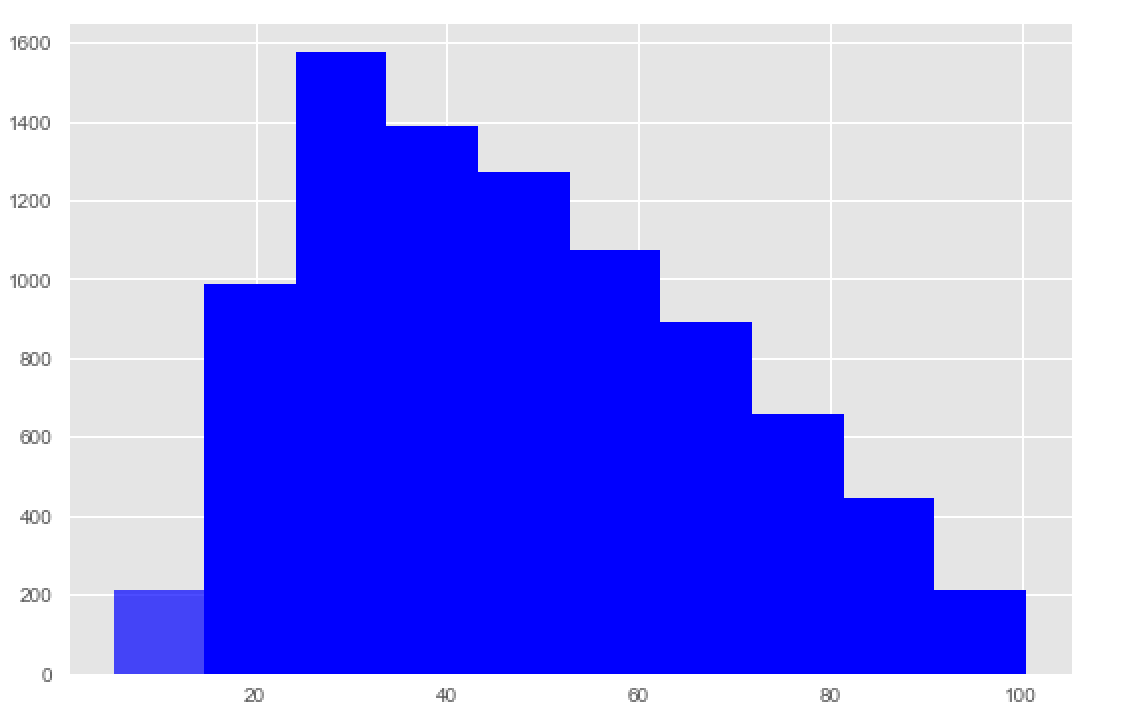
Exploring and Analyzing

Initial Data Exploration

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Family Income | Retention rate | Sat Scores | Dep. Income | Repayment rate |
| count | 8705 | 8705 | 8705 | 8705 | 8705 |
| mean | 38529.887008 | 0.595932 | 1117.806807 | 51158.216 | 47.37 |
| std | 22786.362412 | 0.135296 | 68.485513 | 22260.37 | 20.98 |
| min | 4420.343869 | 0 | 720.000000 | 7854.26 | 5.16 |
| 25% | 23245.915410 | 0.541978 | 1134.295372 | 35054.67 | 30.22 |
| 50% | 30387.174410 | 0.541978 | 1134.295372 | 45550.7 | 44.85 |
| 75% | 48358.342100 | 0.6362 | 1134.295372 | 65790.07 | 62.62 |
| max | 142461.5397 | 1 | 1505 | 151735.9541 | 100.47 |

Histogram of Repayment rate

Repayment rate vs Frequency  

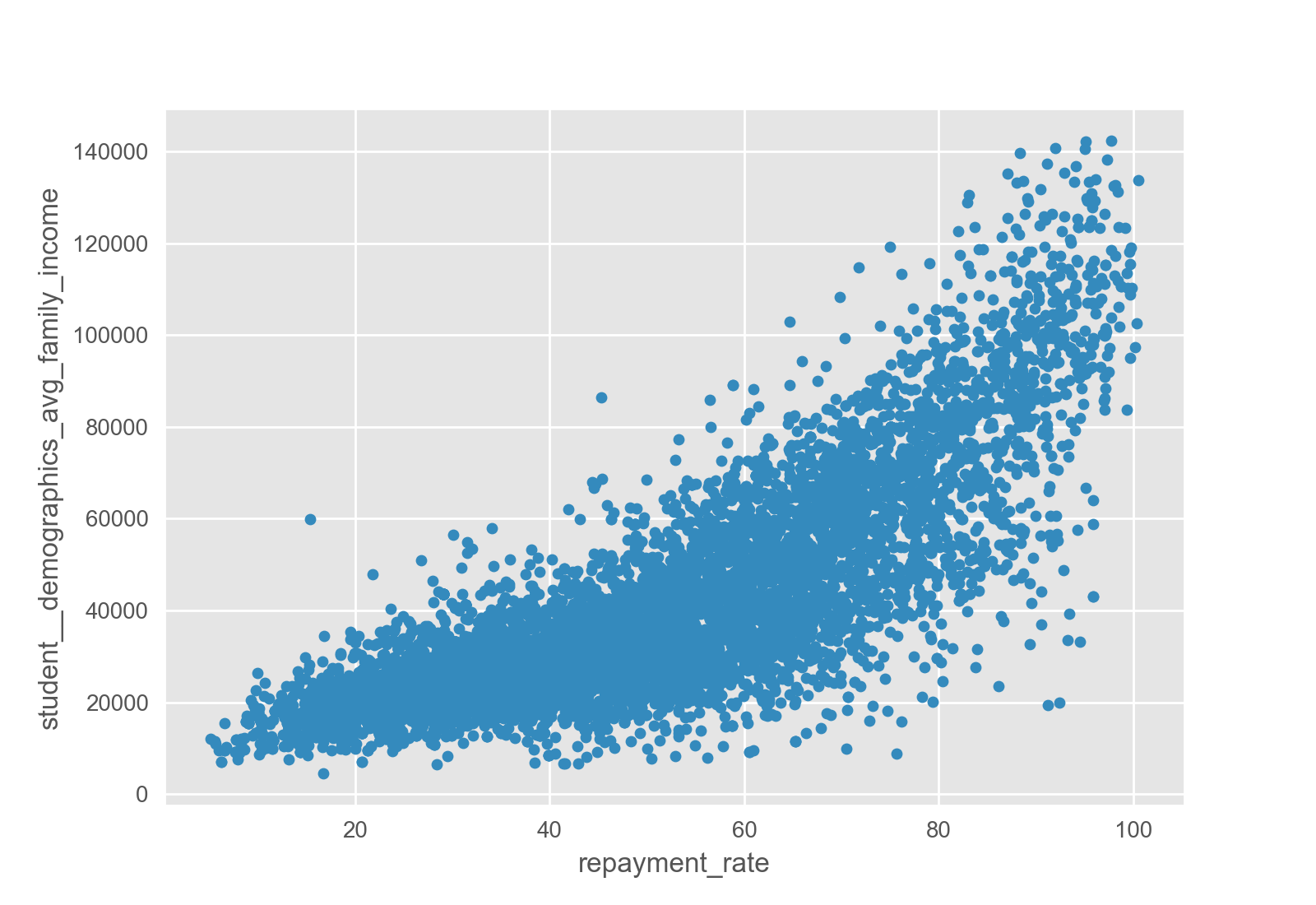
Based on the distribution of repayment rate, we can determine the frequency of the students who repay more than 50% of the loan keeps decreasing and the students who repay loan completely is very less.

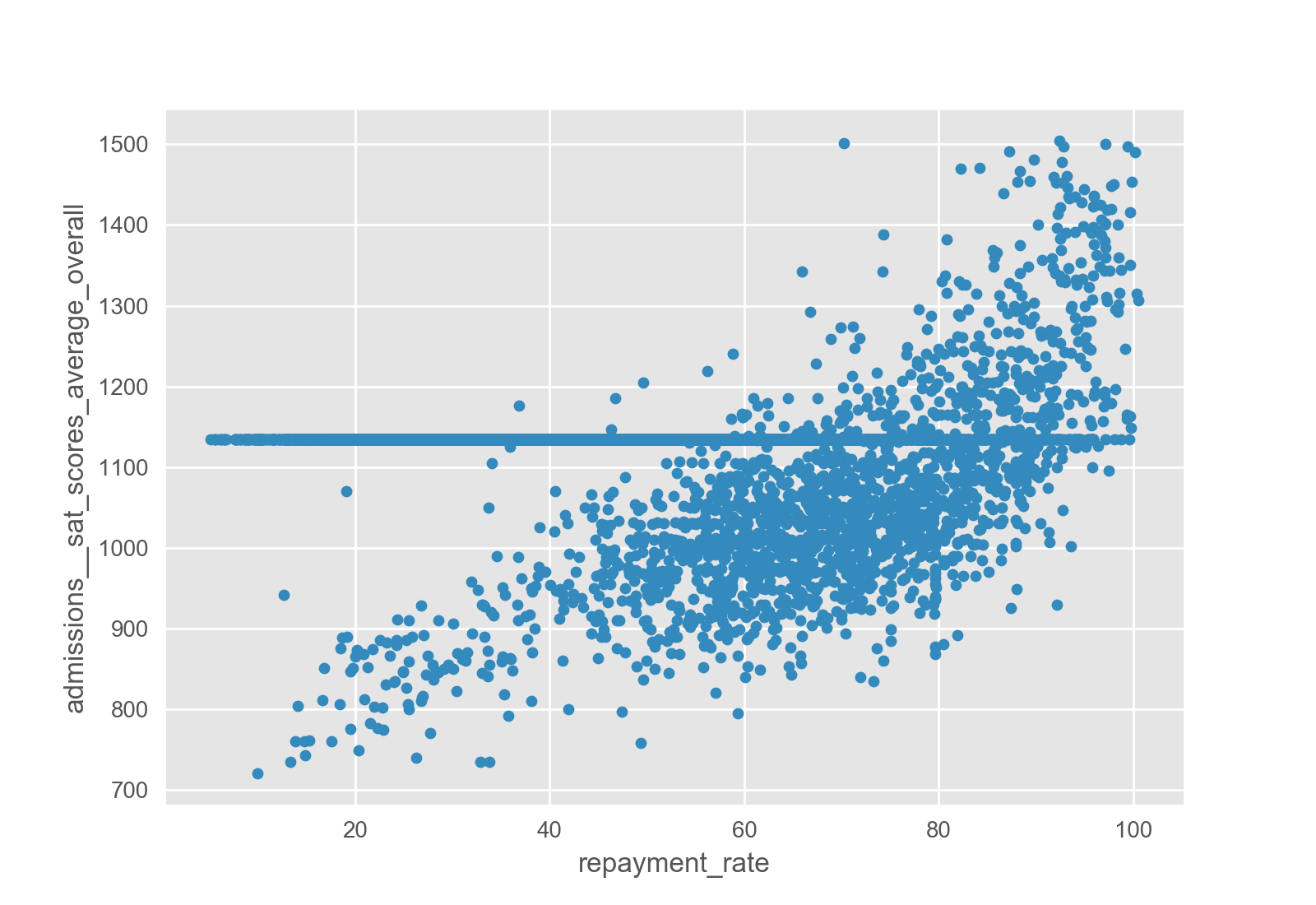
Correlation and Apparent Relationships

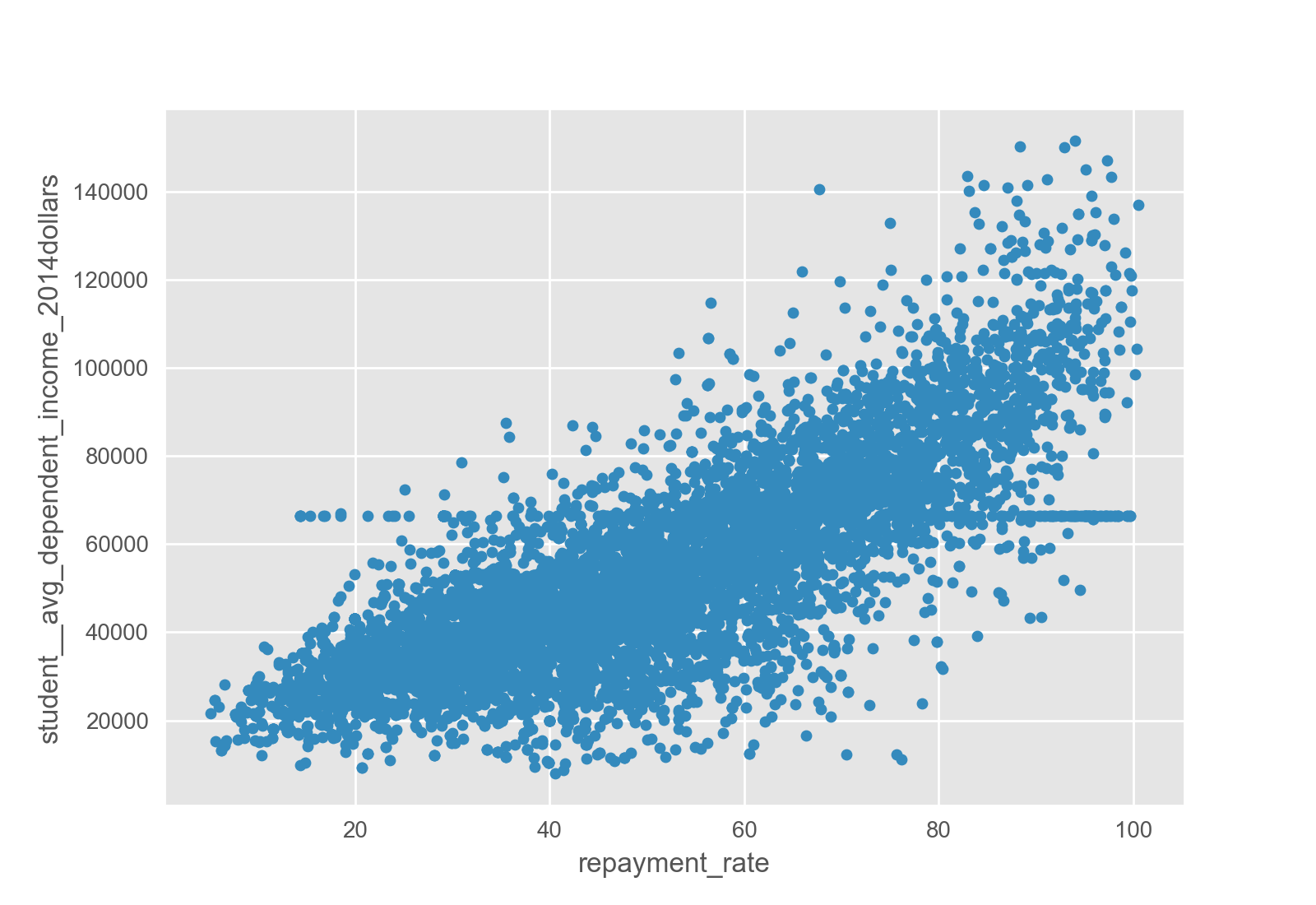
After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between **repayment rate** and the other features.

### Numeric Relationships

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







The correlation between the numeric columns was then calculated with the following results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | family\_income | retention\_rate | avg\_dependent\_income | repayment\_rate |
| family\_income | 1 | 0.653333 | 0.907799 | 0.822844 |
| retention\_rate | 0.653333 | 1 | 0.550179 | 0.54478 |
| avg\_dependent\_income | 0.907799 | 0.550179 | 1.000000 | 0.803872 |
| repayment\_rate | 0.822844 | 0.54478 | 0.803872 | 1 |

Categorical Relationships

In addition to the numeric values, the observations include categorical features, including:

* **Ownership of the Institute** - Control of institution

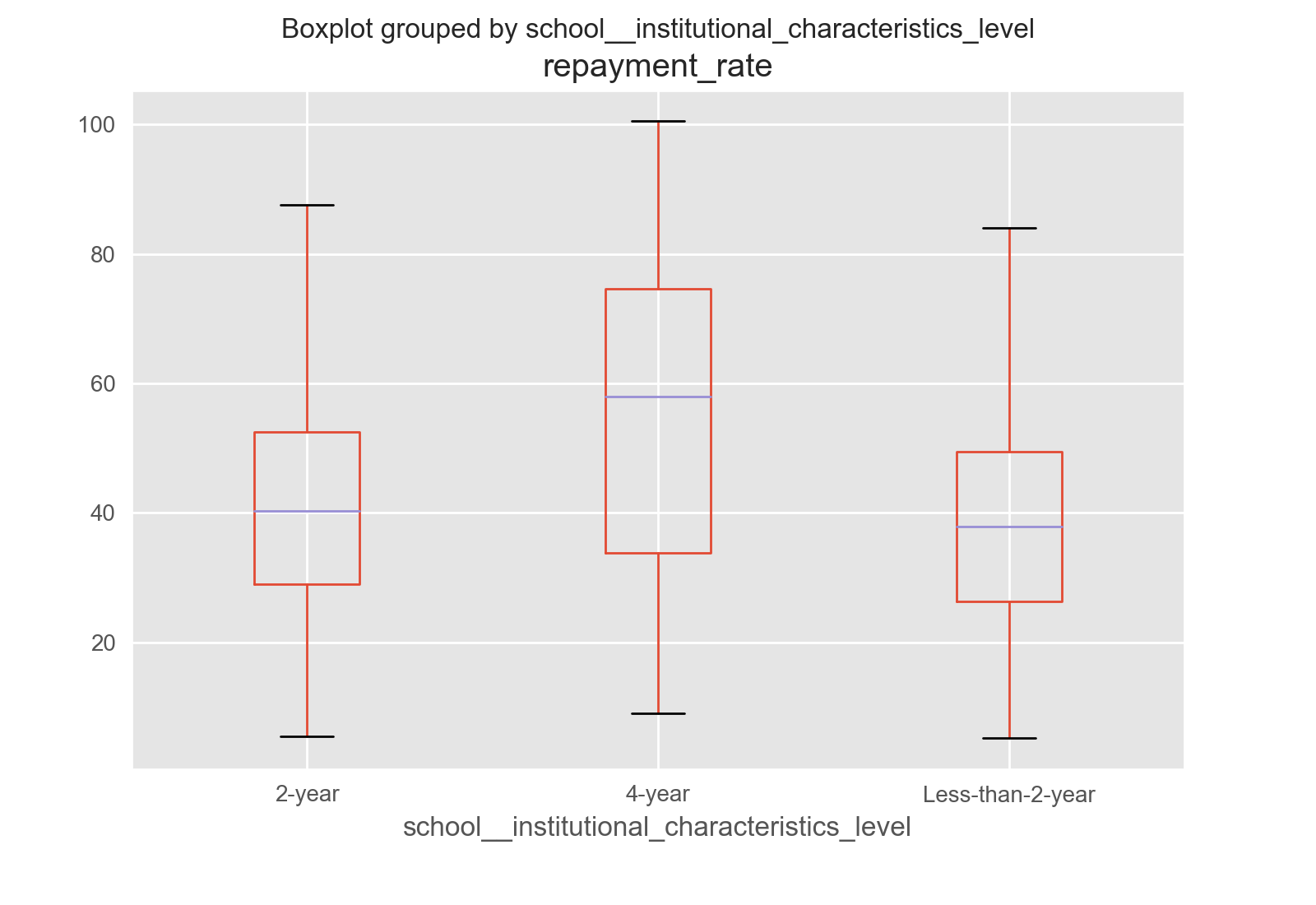
((Possible Values: Private nonprofit, Private for-profit, Public)

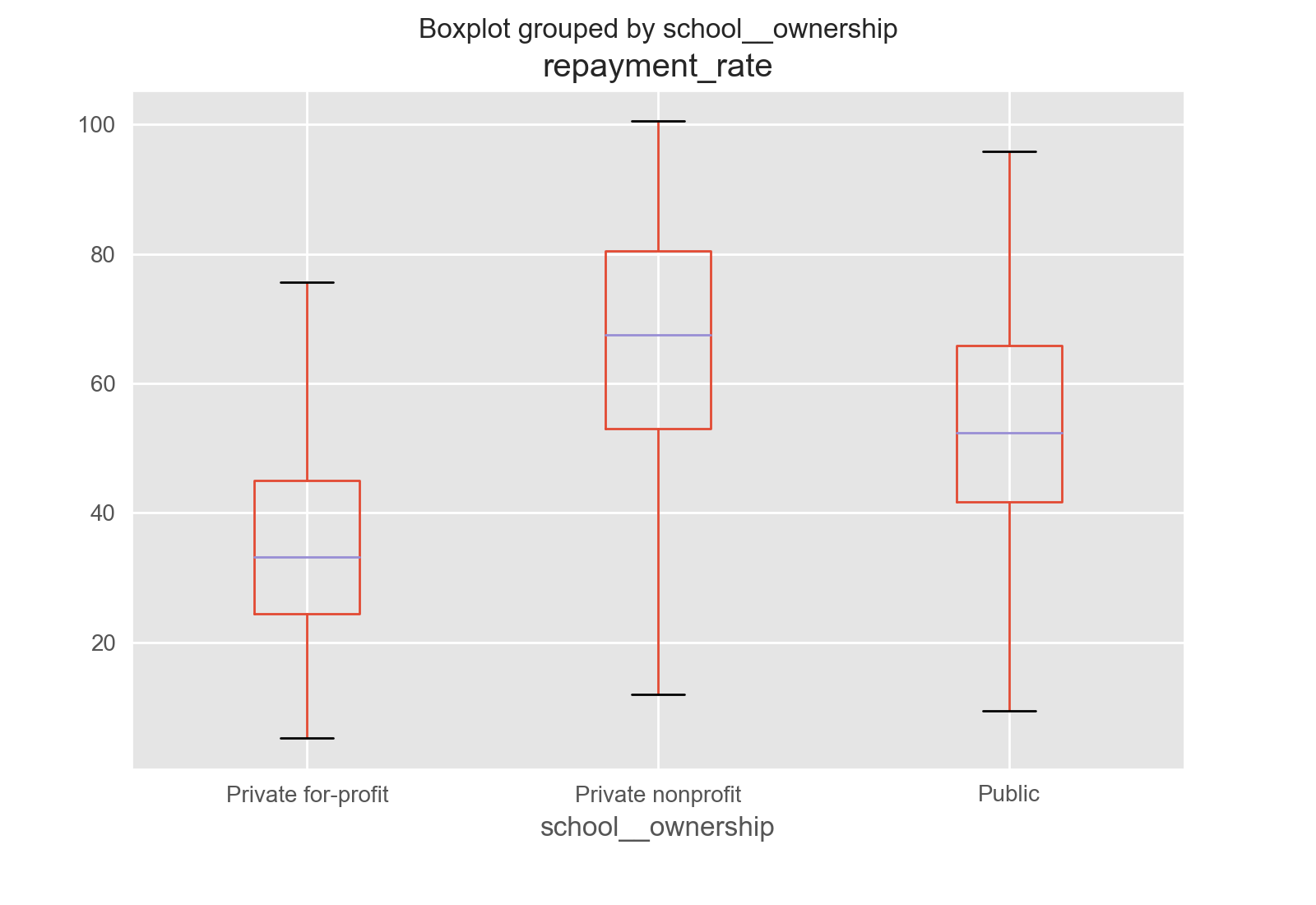
* **Institutional Characteristics level**- Level of institution
* **Region** - Region (IPEDS)

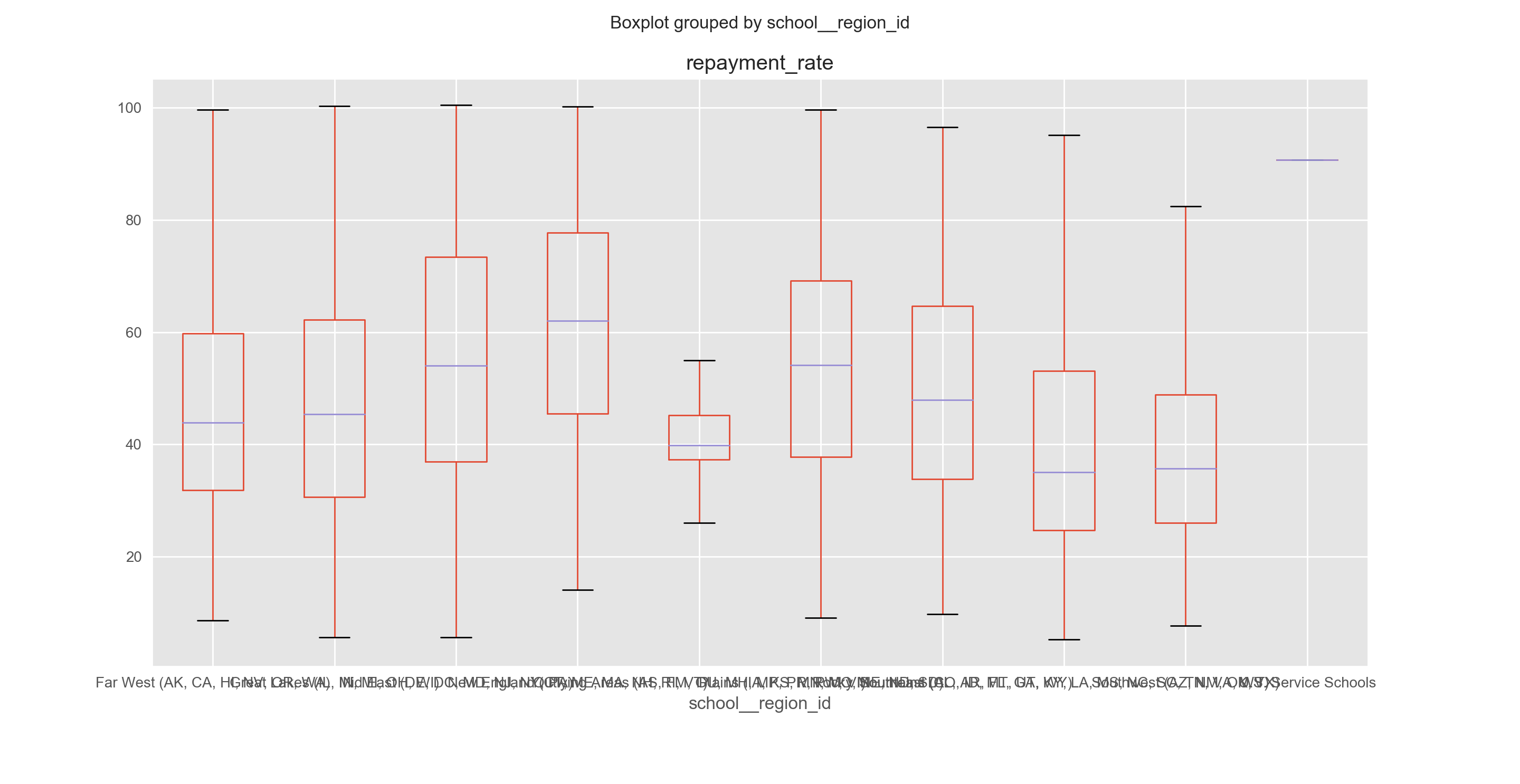
Having explored the relationship between repayment rate and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and repayment rate. The following box- plots show the categorical columns that seem to exhibit a relationship with the repayment rate:

Box plots were created to show frequency of these features, and indicate the following:

* Student repayment rate depends on the region in which the institute are located
* Student who attend the Private non-profit tend to have low repayment rate
* Student who complete full 4 years tend to have high repayment rate







Prediction Algorithms:

## Linear Regression

After exploring and analyzing classification and numeric features of the given observations, a regression model to predict the actual repayment rate was created. Based on the apparent relationships identified when analyzing the data, a linear regression model was created to predict the repayment rate. The model was trained with 70% of the data, and tested with the remaining 30%.

For the actual prediction of repayment rate using linear regression many features were used, only the key significant features are described above. Other features which were used to determine repayment rate are

|  |
| --- |
| aid\_\_students\_with\_any\_loan |
| cost\_\_avg\_net\_price\_private |
| cost\_\_avg\_net\_price\_public |
| aid\_\_loan\_principal |
| school\_\_faculty\_salary |
| student\_\_demographics\_avg\_family\_income |
| student\_\_demographics\_avg\_family\_income\_independents |
| student\_\_demographics\_married |
| student\_\_retention\_rate\_four\_year\_full\_time |
| admissions\_\_sat\_scores\_average\_overall |
| student\_\_avg\_dependent\_income\_2014dollars |
| student\_\_avg\_independent\_income\_2014dollars |

The Root Mean Square Error (RMSE) for the test results is 8.5749.

Conclusion:

This analysis has shown that the repayment rate of a student can be confidently predicted from its characteristics. In particular, the ownership of institute, institute characteristics level, region, family income, sat scores, loan principal.