Student Loan Repayment Rate Analysis

Kabiru Murtala, July 2017

1. Executive Summary

This paper is an analysis of data containing repayment rate for student loans among students of higher institution in the United States of America. The data is made up of 15096 observations (8705 training observations, and 6391 test observations), and 443 predictor variables, all publicly available and published in the United States Department of Education website.

The data cleaning, exploration, initial visualization and summary statistics were done using R language for statistical computing. Using the same software, correlation analysis was done to identify potential relationship between the predictor variables and student loan repayment rate. A boosted decision tree regression model was used to predict student loan repayment rate from the selected features.

Based on the analysis conducted, some important correlates of student loan repayment rate include the following:

- **Midpoint of the ACT cumulative score**: Schools that admit students with a lower average median ACT cumulative score have a lower repayment rate, and schools that admit students with a higher median ACT cumulative score have a higher repayment rate.
- **Average SAT equivalent score of students admitted**: Schools that admit students with a lower average overall SAT equivalent score have a lower repayment rate, and schools that admit students with a higher average overall SAT equivalent score have a higher repayment rate.
- **Level of institution**: Schools that with 2 year duration of study have higher repayment rates than schools with less than 2 year duration of study. Also, schools with 4 year duration of study have higher repayment rate than school with 2 year duration of study.

- **School Ownership**: Privately owned not-for-profit schools have higher median repayment rate than public schools. Also, public schools have higher median repayment rate than privately owned for-profit schools.
- **Region:** Service schools and schools in the New England Region have the highest median repayment rates. Schools in the Southeast and Southwest regions have the lowest repayment rates.
- **Median family income in real 2015 dollars:** Schools where students have a high median family income tend to have a high repayment rate, and schools where students have a low median family income tend to have a low repayment rate.
- **Share of first-generation students**: Schools with lower share of first generation student have high repayment rate and those with higher share of first generation students have lower repayment rate.
- **Share of dependent students**: Schools with higher share of dependent student have high repayment rates and those with lower share of dependent students have low repayment rates.
- **Highest Degrees Awarded**: Students at non degree granting schools had the lowest repayment rate and those at graduate schools had the highest repayment rates.
- **Predominant Degrees Awarded**: Students of schools where the predominant degree awarded is Bachelors had the highest repayment rate. The lowest repayment rates were in schools where the predominant degree awarded is unclassified or certificate degree.

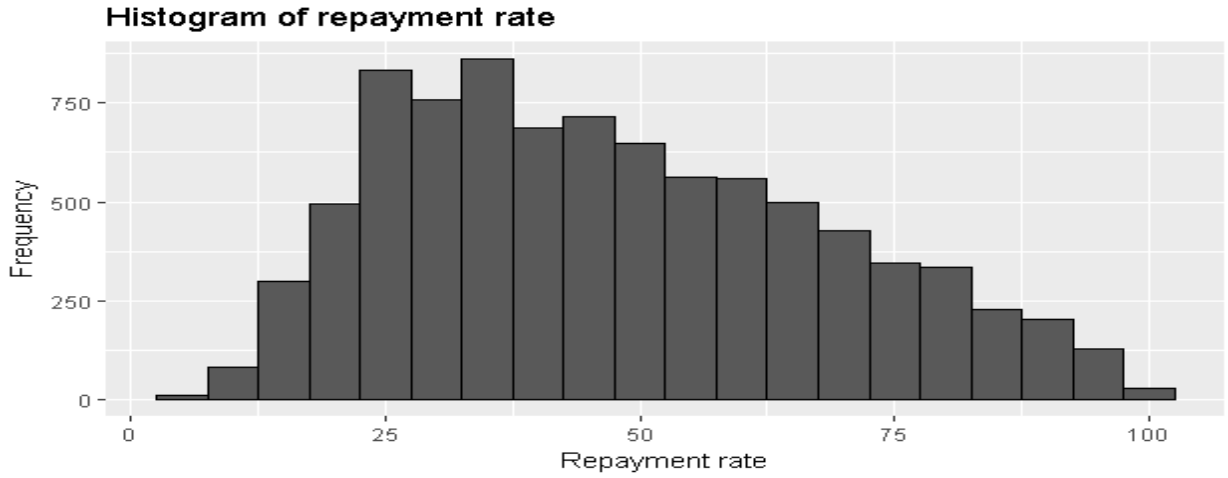
2. Initial Data Exploration

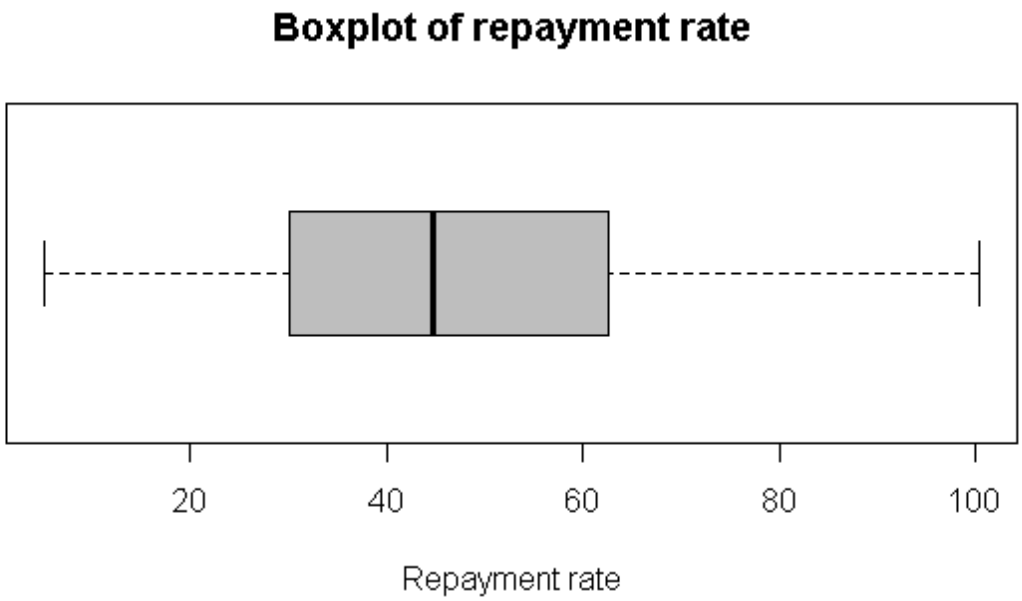
The first step in the analysis if the data involved summarizing the main characteristics of the data set. For numeric data, the minimum, maximum values as well as the mean, median, and standard deviation were calculated. The number of missing values were also noted. For categorical dataset, a table was constructed to view the distribution of variables.

Individual summary statistic of important numeric features

Feature	Min	Max	Mean	Median	Std Dev	NAs
Midpoint of the ACT cumulative score	14.0	34.0	24.5	26.0	3.01	9283
Average SAT equivalent score of students admitted	714	1505	1110	1171	116.7	9080
Median family income in real 2015 dollars	0	122446	35555	23582	27810.77	13
Share of first-generation students (%)	0.06	0.92	0.48	0.50	0.16	629
Share of dependent students (%)	0.01	0.99	0.47	0.42	0.25	282
Loan Repayment Rate (%)	5.16	100.47	47.37	44.86	20.99	0

The mean value of the loan repayment rate is higher than the median value. This indicates perhaps that the distribution is skewed to the right. There are no outliers.





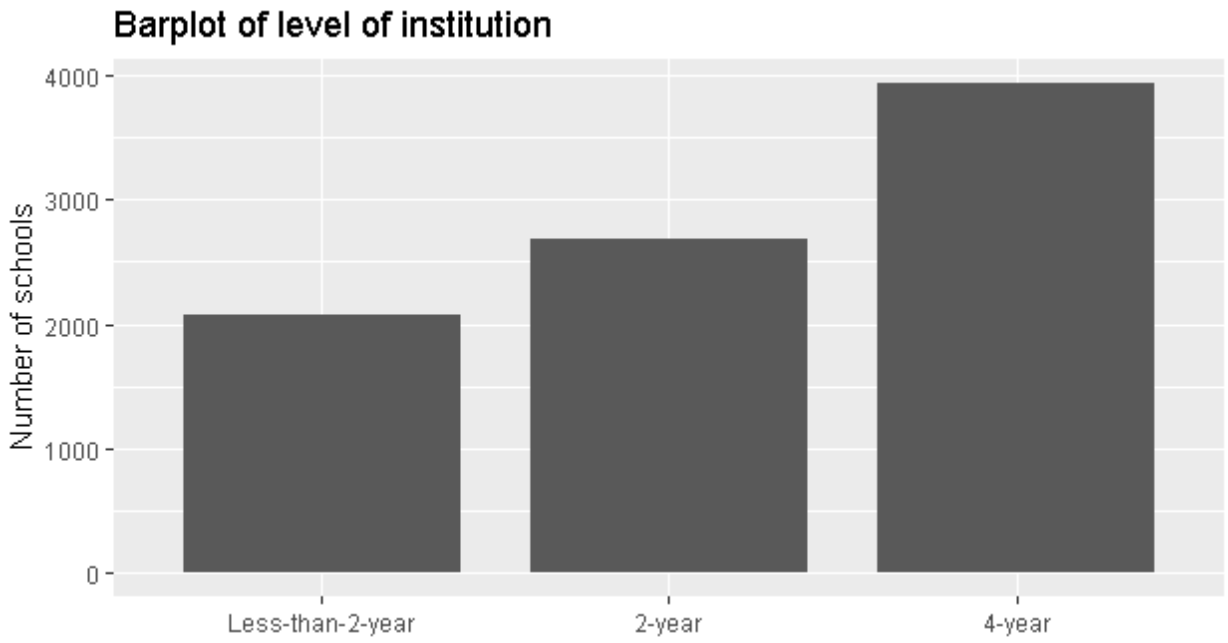
Descriptive statistic of important categorical features

The significant categorical features include:

- **Level of institution**: 4-year, 2-year, Less-than-2-year
- **School Ownership**: Private nonprofit, Private for-profit, Public
- **Region**: Rocky Mountains, Plains, Southeast, Southwest, Mid East, New England, Far West, Great Lakes, Outlying Areas, and U.S. Service Schools
- **Highest Degrees Awarded**: Non-degree-granting, Certificate degree, Associate degree, Bachelor's degree, Graduate degree
- **Predominant Degrees Awarded**: Not classified, Predominantly associate's-degree granting, Predominantly certificate-degree granting, Predominantly bachelor's-degree granting, Entirely graduate-degree granting

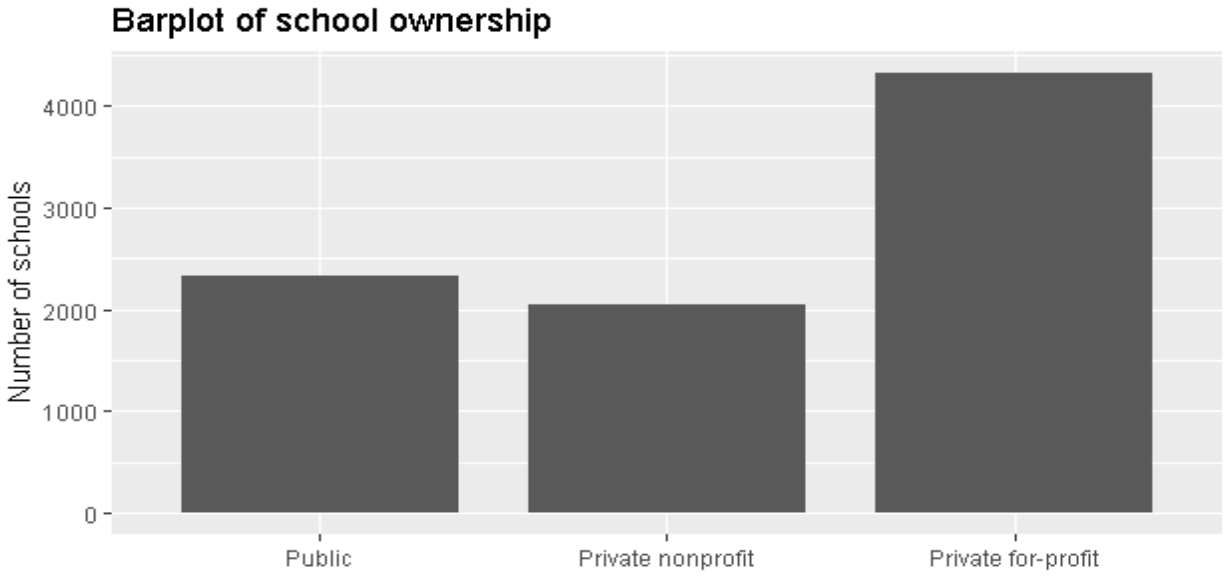
i. Level of institution

As shown in the bar chart below, 4-year colleges are the most common type of institution, followed by 2-year colleges and less than 2 year colleges



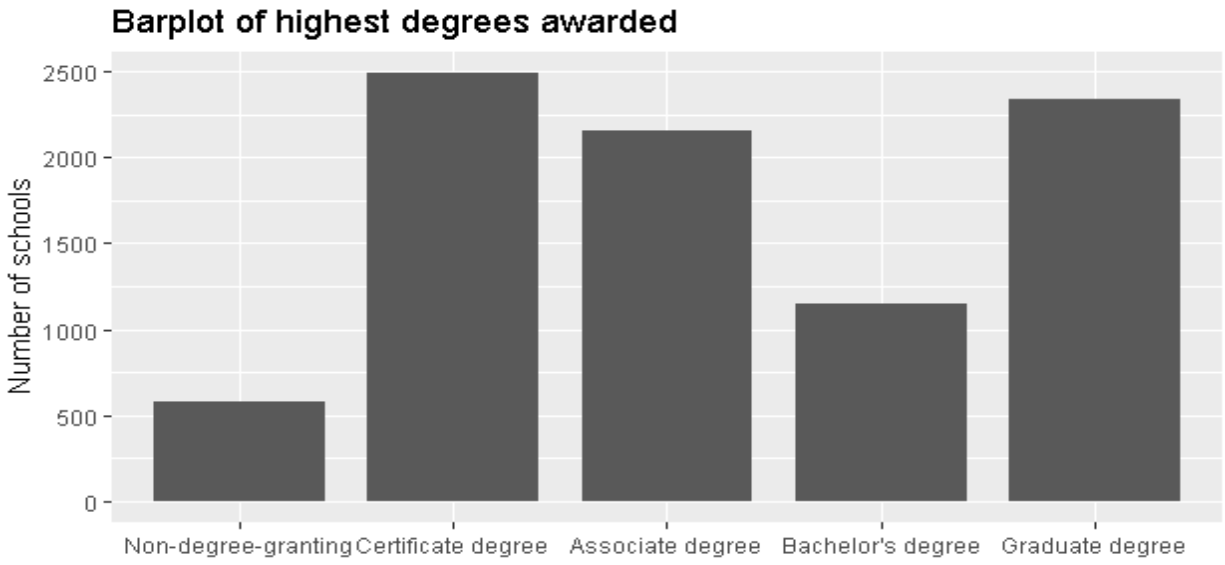
ii. School ownership

As shown in the bar chart below, Private for-profit schools are the most common type of institution, followed by Public schools and Private nonprofit schools.



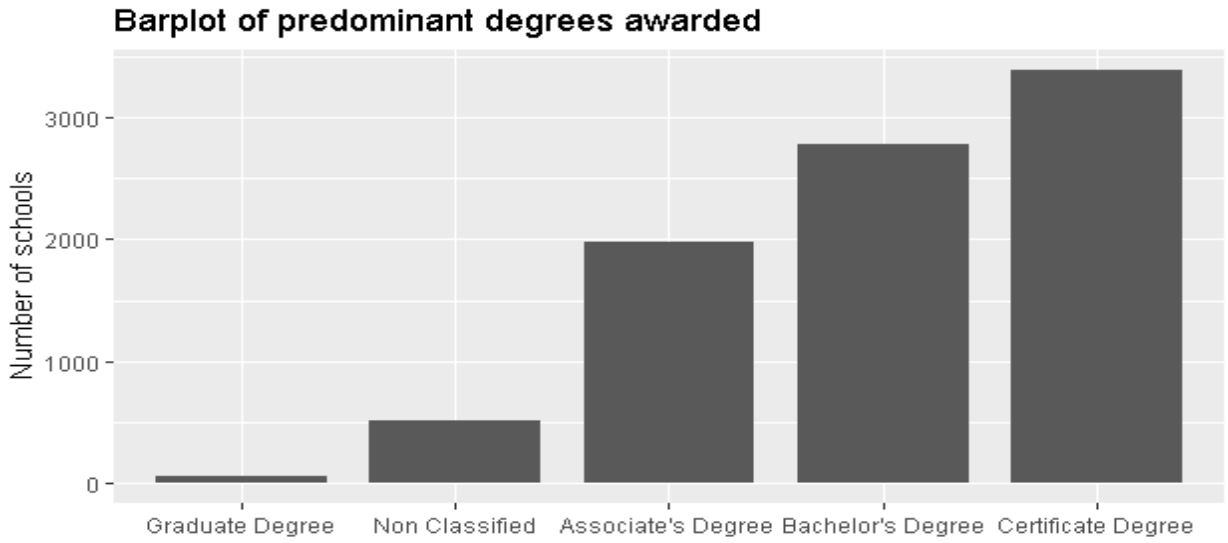
iii. Highest Degrees Awarded

As shown below, based on highest degrees awarded, certificate degrees are the highest, followed by graduate degrees, associate degree, bachelor's degree, and non-degree granting institution.



iv. Predominant Degrees Awarded

As shown in the barplot below, certificate degree were the most predominant degree awarded, followed by Bachelor's degree, Associate's degree, Non Classified, and Graduate degree being the least.



3. Key Findings

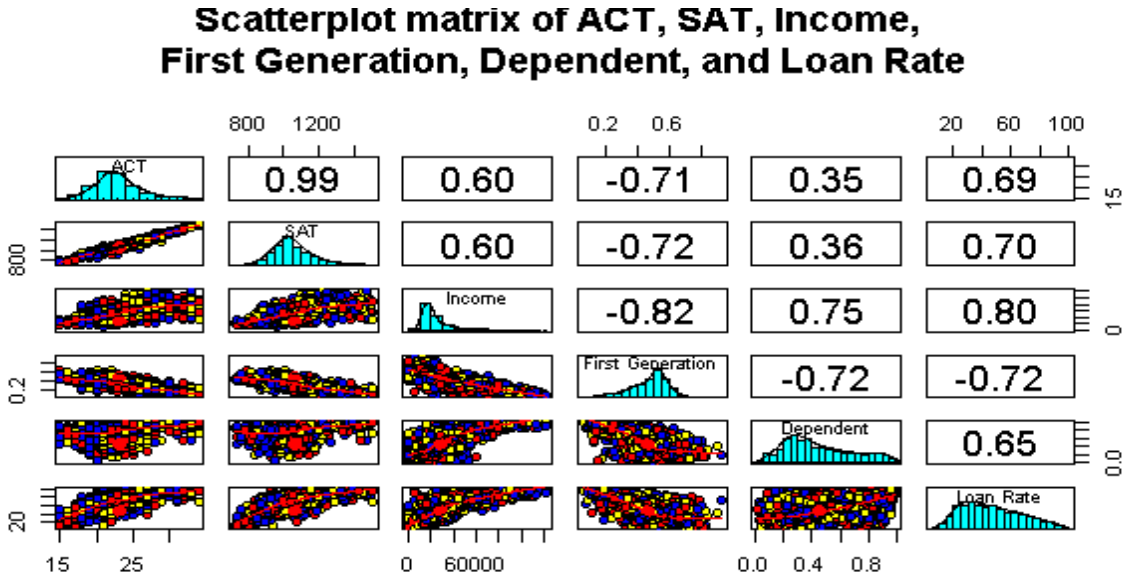
a. Numeric relationships

Following the initial data exploration, a scatterplot matrix was used to compare correlations between various numeric data as well as the correlations between loan repayment rate and each of the numeric data

The numeric data in this case are: midpoint of the ACT cumulative score (ACT), average SAT equivalent score of students admitted (SAT), median family income in real 2015 dollars (Income), Share of first-generation students (First-Generation), Share of dependent students (Dependent), and Loan Repayment Rate (Loan Rate).

As shown in the scatterplot matrix above, there is an almost perfect positive correlation (r = 0.99) between the midpoint of the ACT cumulative score (ACT), average SAT equivalent score of students admitted (SAT).

There was also observed strong positive relationship (r =0.75) between median family income in real 2015 dollars (Income) and Share of dependent students (Dependent). Also noted was the strong negative relationship (r = -0.82) between median family income in real 2015 dollars (Income) and Share of first-generation students (First-Generation).



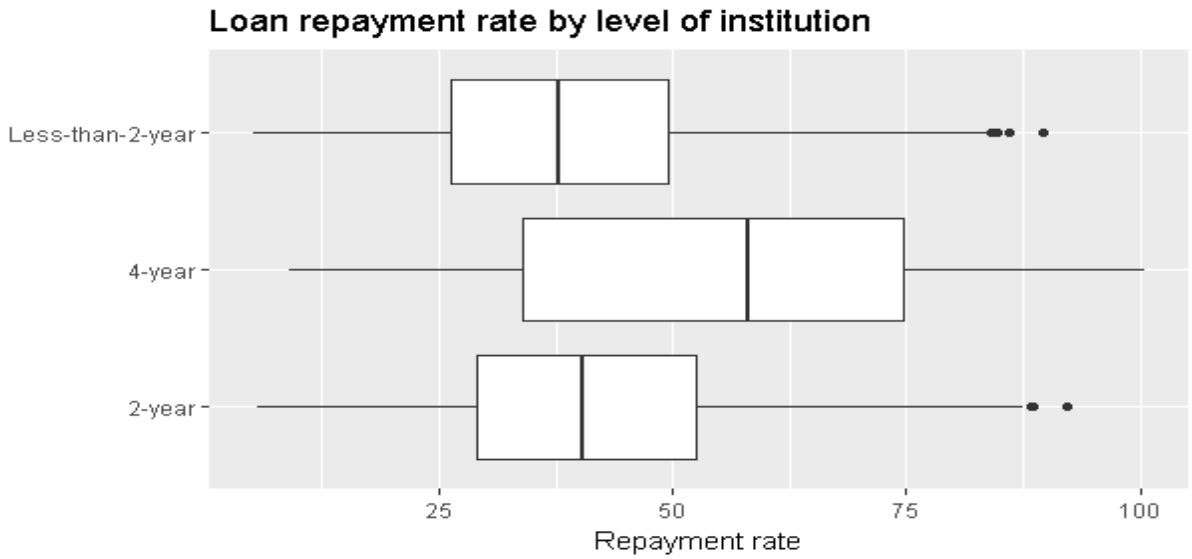
It can also be seen from the scatterplot matrix that there is a strong positive relationship between loan repayment rate and the following numeric features: median family income in real 2015 dollars (r = 0.80), average SAT equivalent score of students admitted (r=0.70), midpoint of the ACT cumulative score (r=0.69), and share of dependent students (r=0.65). Specifically, as median family income in real 2015 dollars, average sat scores of students admitted, midpoint of ACT scores, and share of dependent students increase, the loan repayment rate increases.

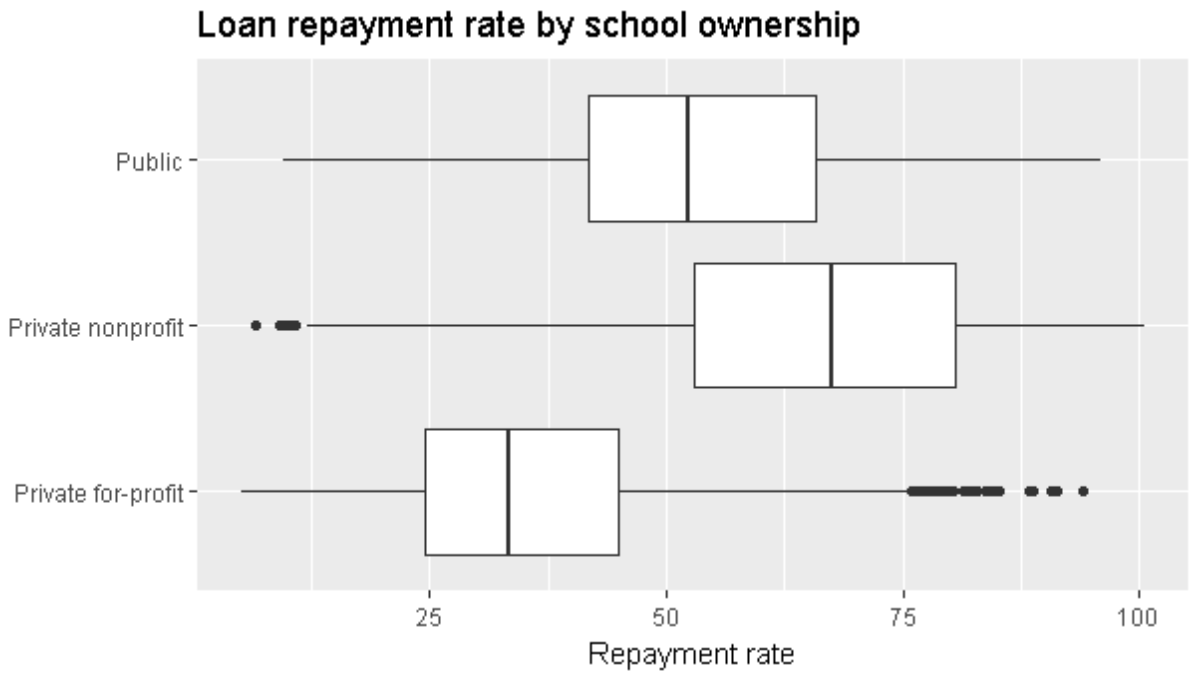
Also noted is the strong negative relationship between loan repayment rate and share of first-generation students (r=-0.72). As the share of first-generation students increases, the loan repayment rate decreases.

b. Categorical relationships

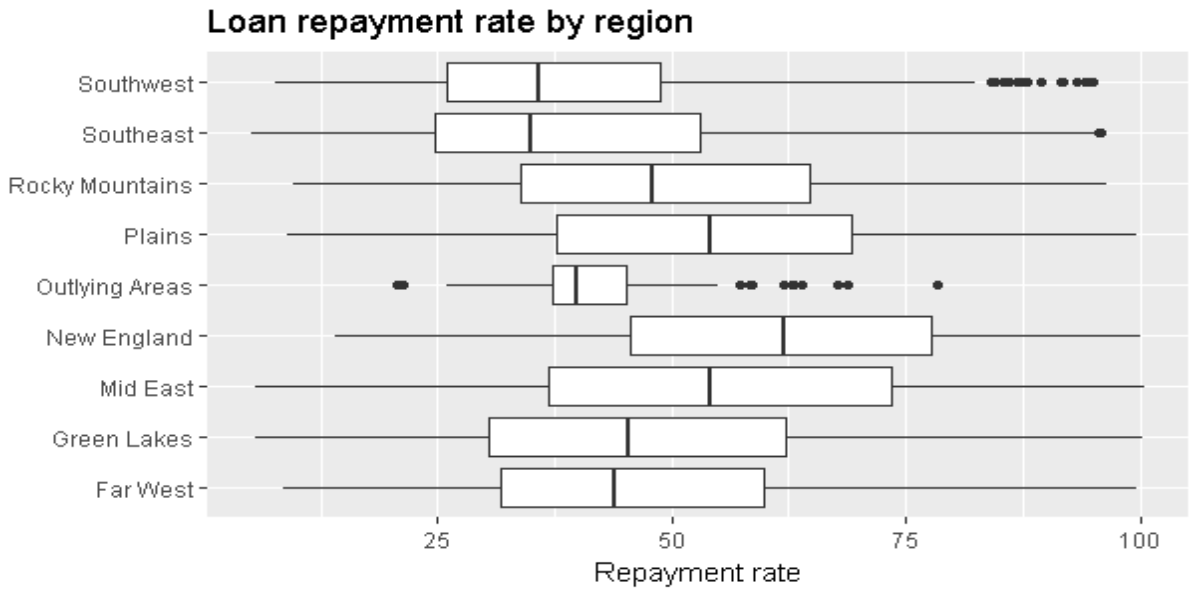
As seen in the boxplot below, students at 4 year colleges have significantly higher median loan repayment rates than students at 2 year colleges, who in turn have higher median loan repayment rates than students at less than 2 year colleges.

Also, students at private for profit schools have the lowest median loan repayment rates, followed by students at public schools. The students at private nonprofit schools had the highest median loan repayment rates.

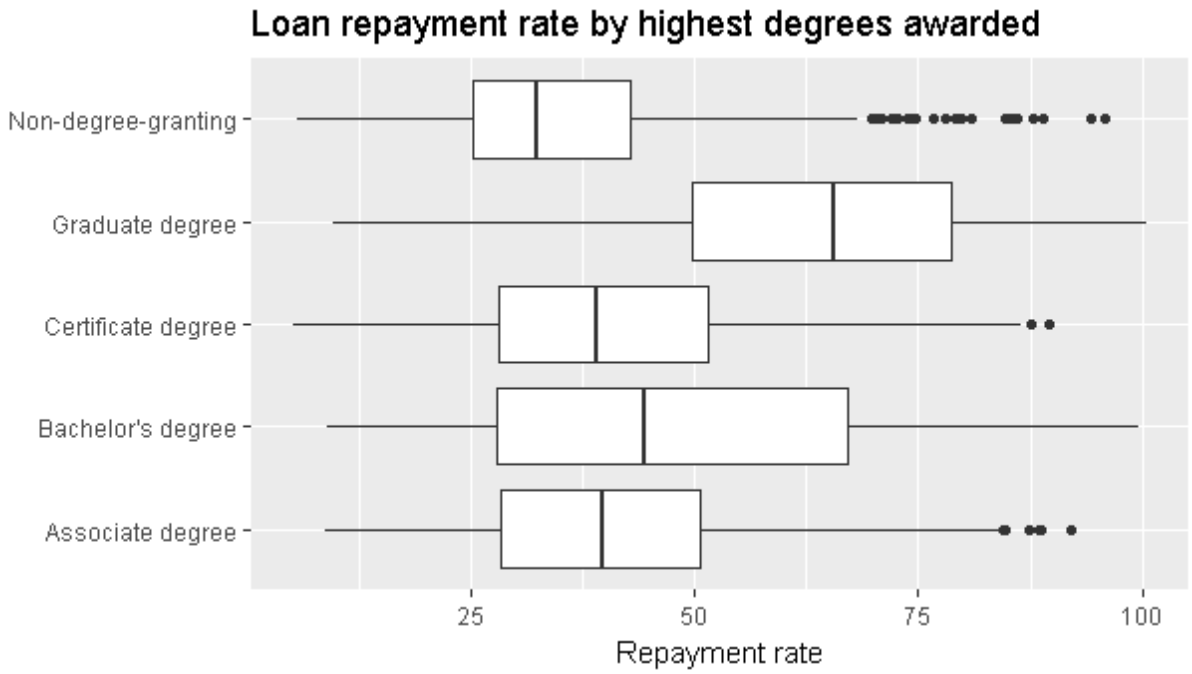




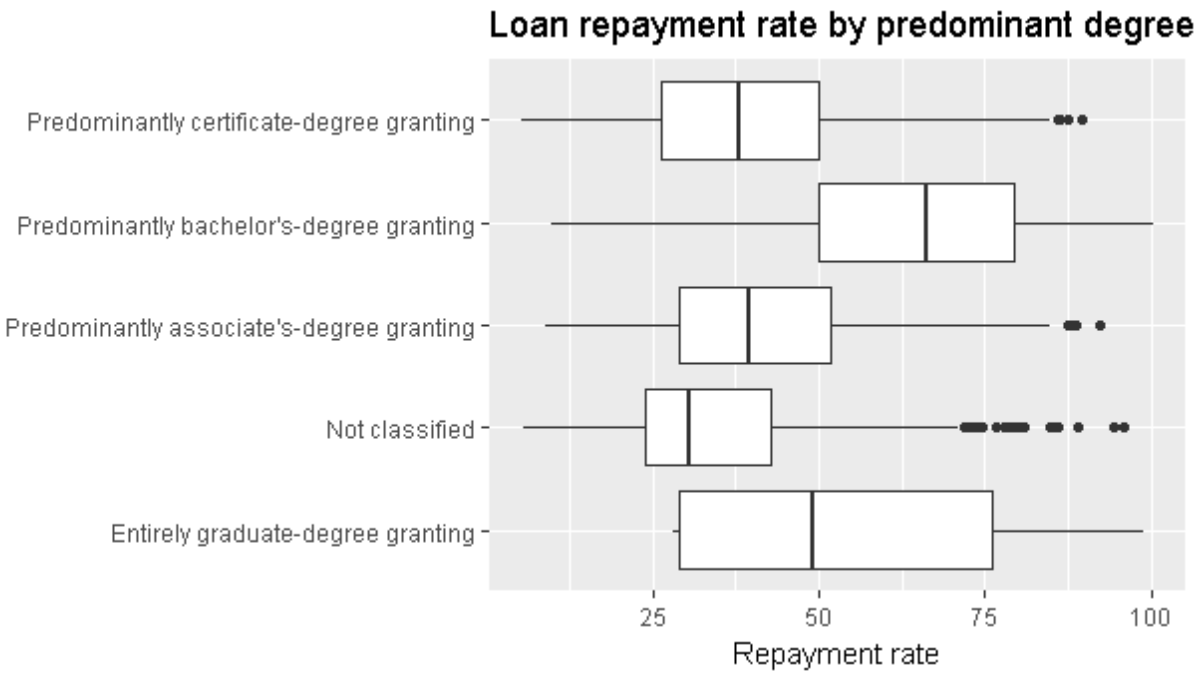
Also, analysis of median loan repayment rate by region shows that schools in the Southeast, Southwest, and Outlying Areas have the lowest loan repayment rates. Schools in New England, Plains, and Mid East have the highest repayment rates, with schools in the Far West, Great Lakes, and Rocky Mountains in between.



Analysis of median loan repayment rate by highest degree awarded shows that students of schools that award graduate degrees have the highest loan repayment rates. Students that attend non-degree granting schools have the lowest repayment rates.



Finally, analysis of median loan repayment rate by predominant degree awarded showed that students in schools that have no classification have the lowest median repayment rates, followed by predominantly certificate-degree granting, predominantly associate's-degree granting, entirely graduate-degree granting, and predominantly bachelor's-degree granting, in increasing order.



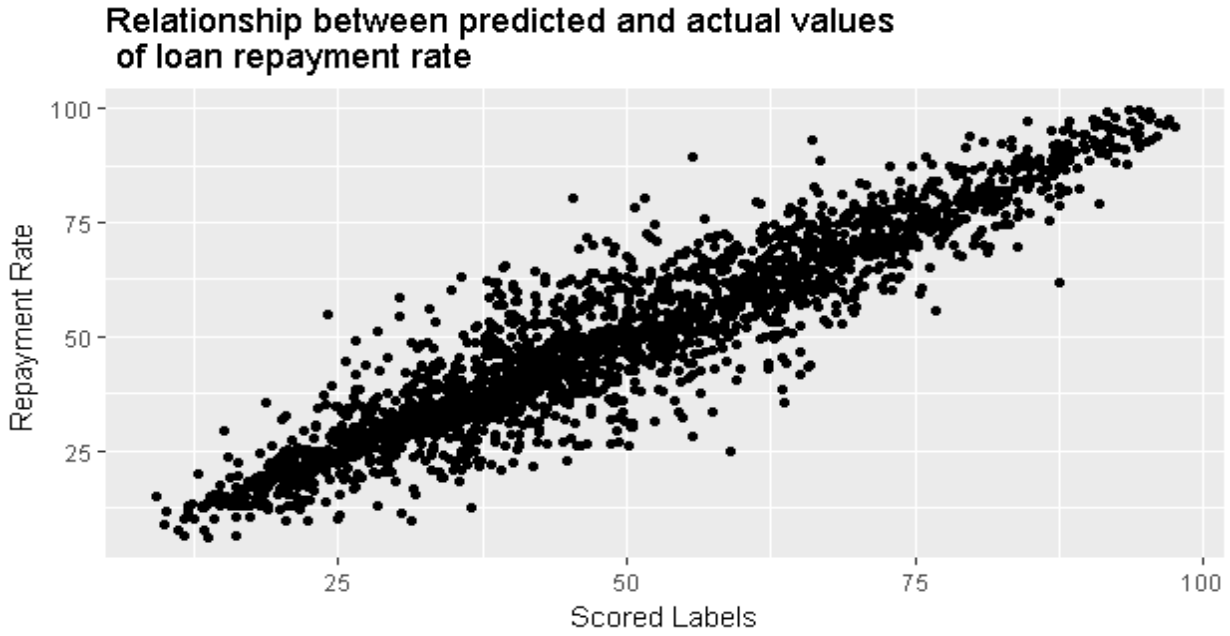
c. Regression

Before regression analysis, an initial data cleaning was done using R language for statistical computing. In the training data frame, features with more than 10% missing data were excluded. After this, 303 features were left out of an initial 444 features. The missing data in the remaining features were replaced by imputation using the R mice package. The repayment rate column from the training labels data frame was then appended to the training data frame.

Subsequent analysis was done using Microsoft Azure Machine learning platform. The training and test dataset were imported into Azure Machine Learning and categorical features were transformed using the "Convert to Indicator Values" module.

In the first step, a boosted decision tree regression model was used to predict the loan repayment rates. The "Tune Model Hyperparameters Module" was used to select optimal parameters for the regression model. Since the test dataset had no labels, the model as trained with 70% of the training dataset and tested with 30% of the training dataset.

A scatter plot showing predicted repayment rates versus actual repayment rates is shown below



The plot shows a strong linear relationship between the predicted and actual values of the test dataset. The Root Mean Square Error (RMSE) for the test result is 7.076148 and the Coefficient of Determination – Adjusted R Square is 0.88536.

In the second and last step, all the 8705 observation of the training data set was used to train a model to predict the repayment rate of the 6391 observations of the test data set using the boosted decision tree regression model as in the first step enumerated above.

The summary result of the predicted values is as follows:

Min: 6.784

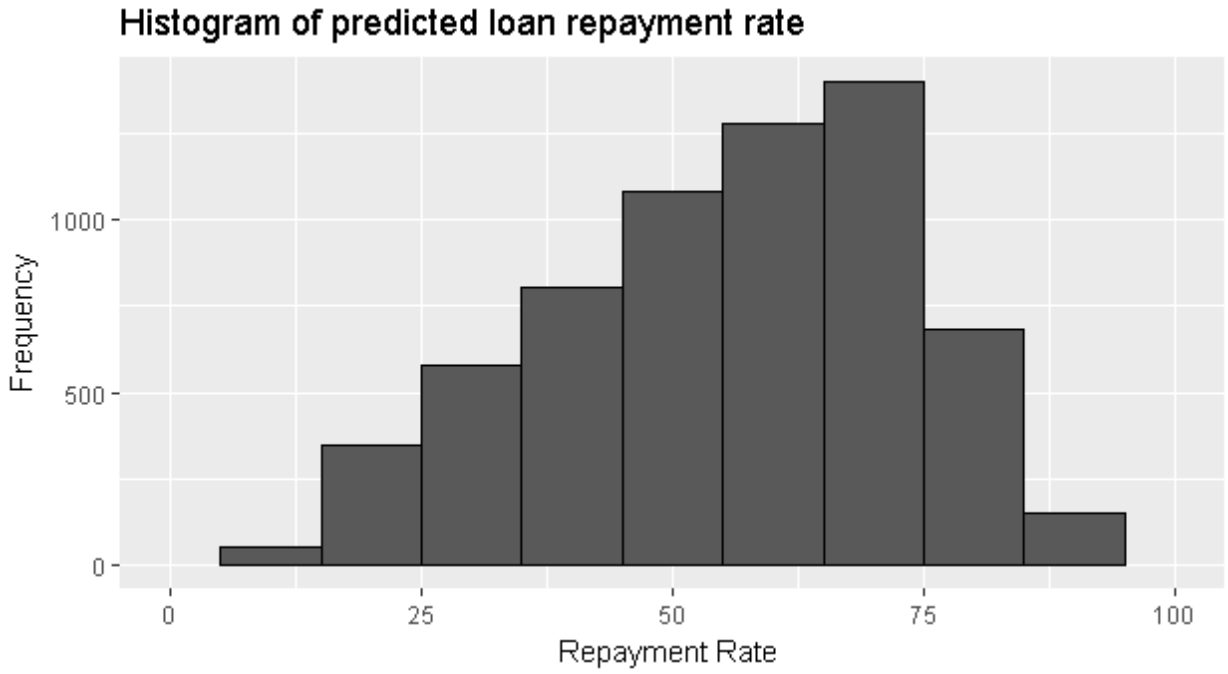
Max: 98.131

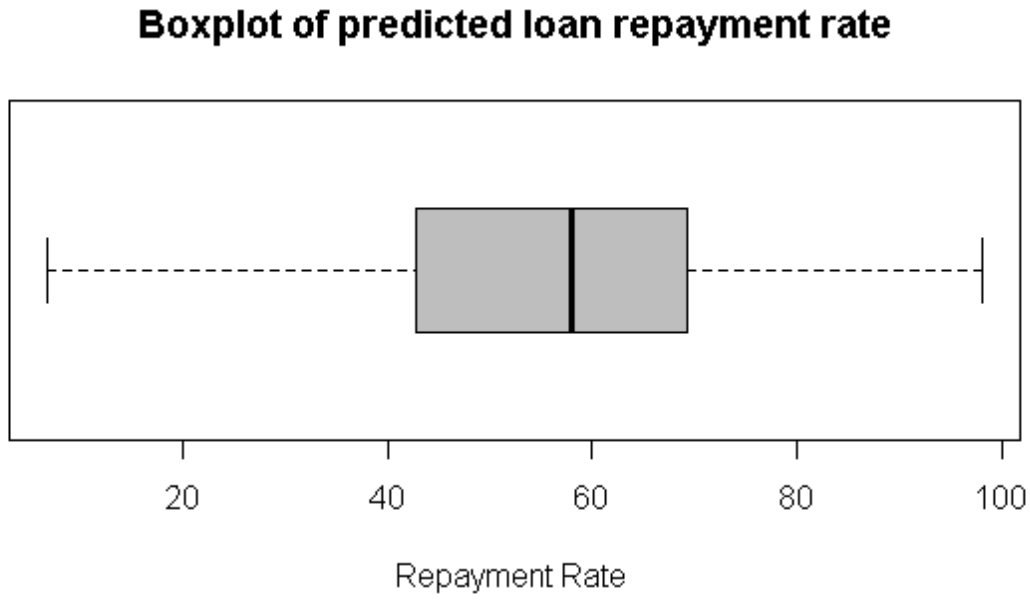
Mean: 55.636

Median: 57.909

Standard Deviation: 17.74026

The histogram and boxplot are also shown below:





4. Conclusion

This data analysis has shown that the repayment rate for the student loans given to students at United States institutions of higher education can be predicted to a very high degree from features / variables in publicly available data. Such features include but are not limited to: midpoint of the ACT cumulative score, average SAT equivalent score of students admitted, level of institution, school Ownership, region,, median family income in real 2015 dollars, share of first-generation students, share of dependent students, highest degrees awarded, and predominant degrees awarded.

5. Recommendations

The author recommends that this model can be used by students and their parents to determine how whether the loan they are about to take will be worthwhile investment. It can also be used by school financial advisors, student loan servicers, as well as public and private student loan lenders to better understand the ability of students to pay back their school loans. This will go a long way in mitigating the student loan debt crisis that has engulfed the United States of America over the past couple of decades.