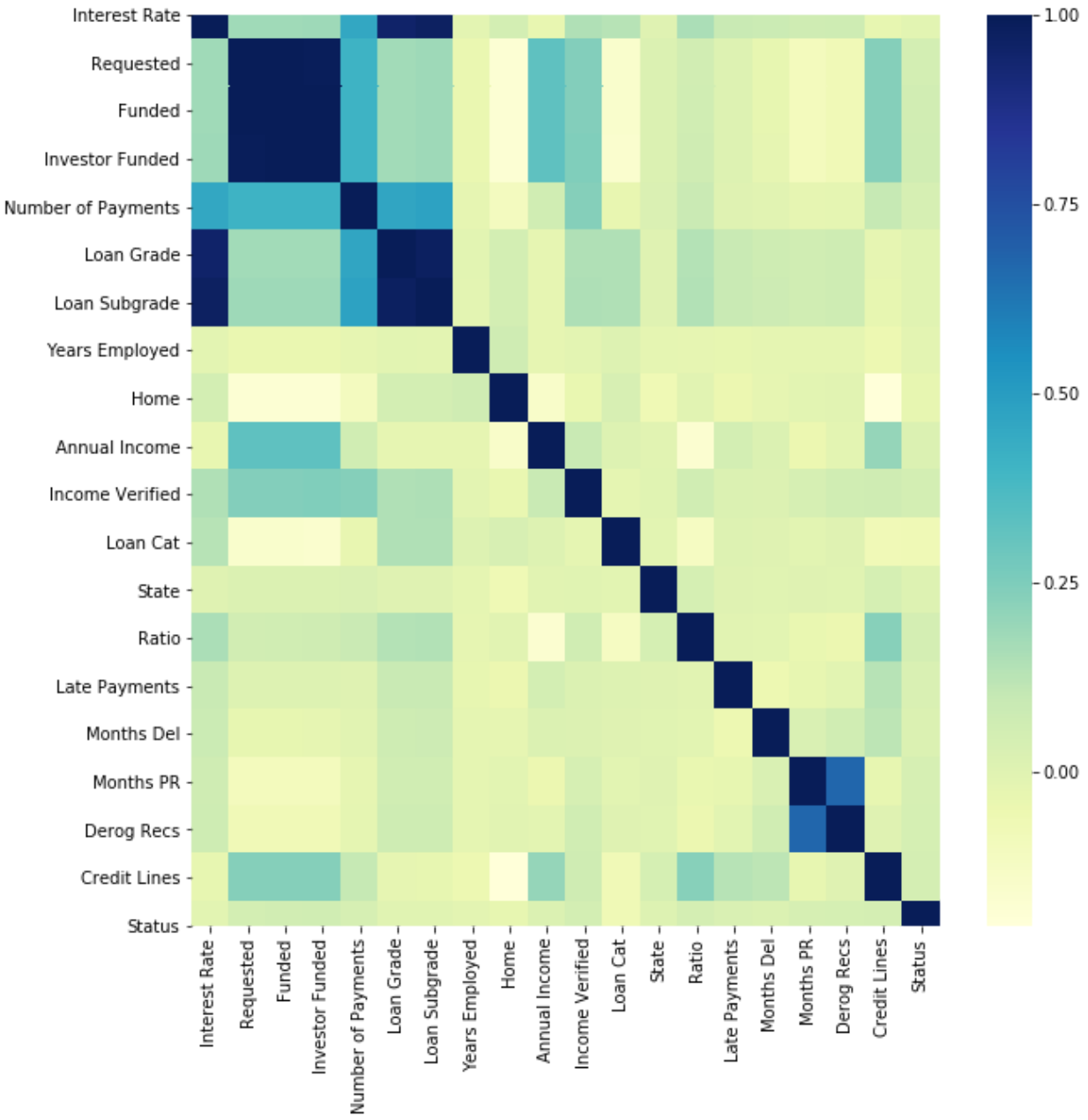
Summary of Loan Interest Rate Analysis

Introduction

Given information about loans and borrowers, the business would like a model to accurately estimate interest rates for loans.

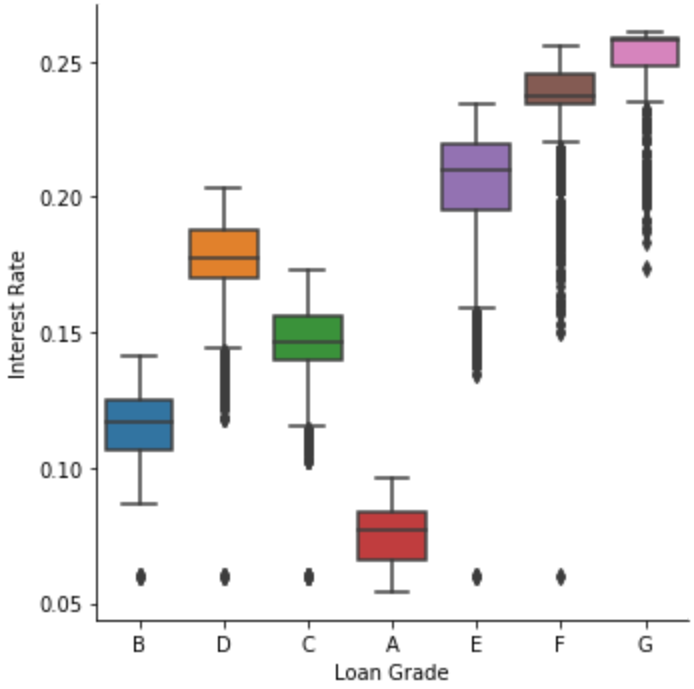
Data

- Data received from business consisted of a csv file with 400,000 records with 26 feature columns and one target column, which was the interest rate.
- Columns consisted of the following (metadata provided by business):
    - Interest Rate (target)
    - Loan Id
    - Borrower Id
    - Requested
    - Funded
    - Investor Funded
    - Number of Payments (36 or 60 months)
    - Loan Grade
    - Loan Subgrade
    - Job (provided by borrower)
    - Years Employed
    - Home (Rent, Own, Mortgage, Other)
    - Annual Income
    - Income Verified
    - Loan Date
    - Reason (provided by borrower)
    - Loan Category (provided by borrower)
    - Loan Title
    - State
    - Ratio
    - Late Payments
    - Credit Line Date
    - Months Del (months since borrower's last delinquency)
    - Months PR (months since borrower's last public record)
    - Derog Recs (number of derogatory public records)
    - Credit Lines (total number in borrower's credit file)
    - Status (initial listing status of loan, W or F)



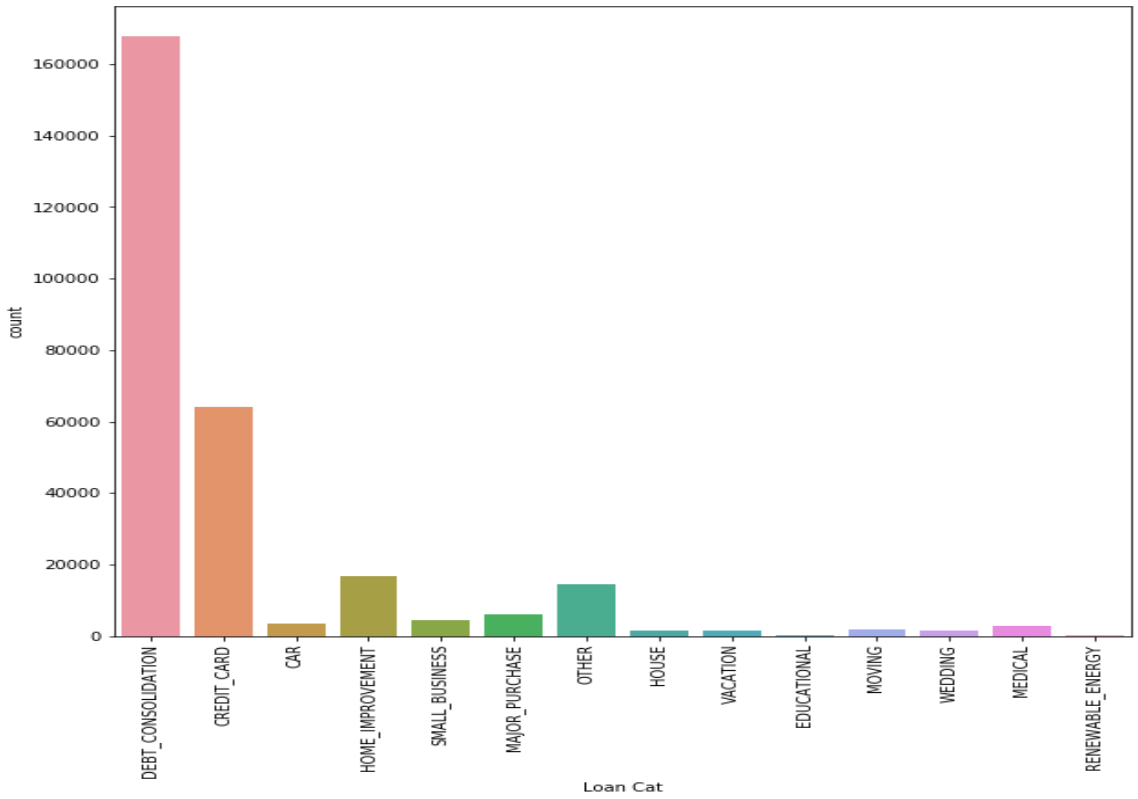
- There does not seem to be a strong correlation between the interest rate and any of the numeric values.

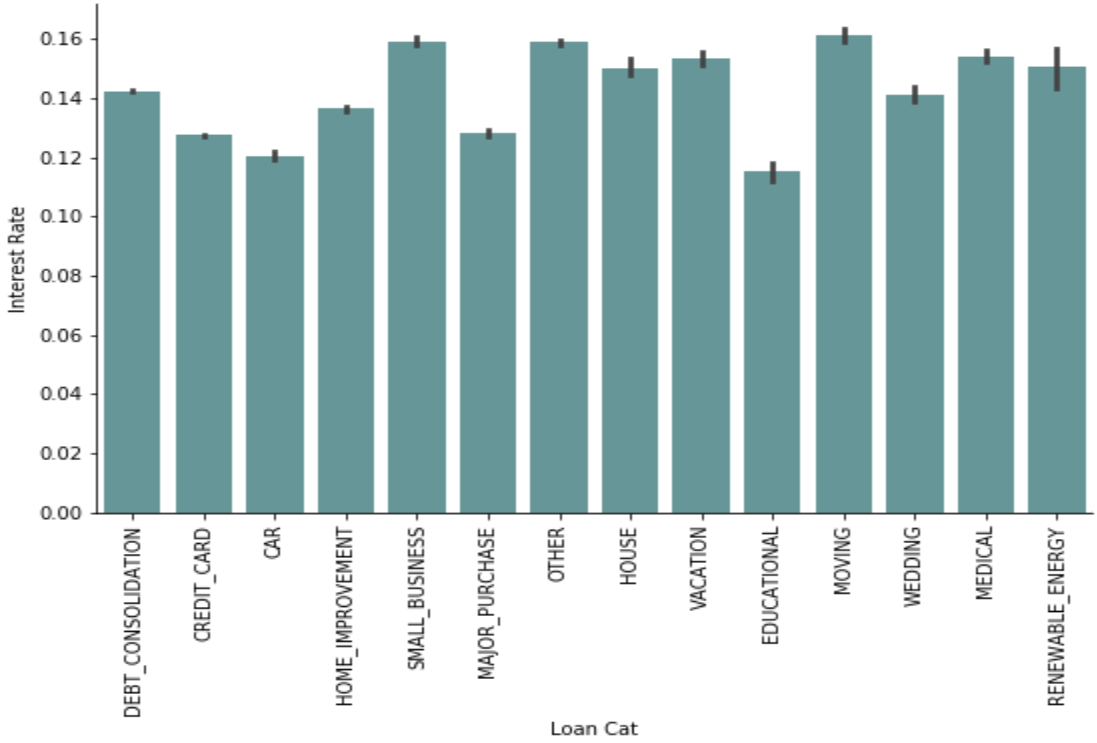
- Loan Grade and interest rate appear to be positively correlated.



- The data for states ID, IA, NE, and ME is in the single digits. At first glance, it appears that these states have significantly higher or lower interest rates, but if you consider the lack of data, this probably isn't the case.

- Debt Consolidation and Credit Card are by far the largest categories, but don't have the highest interest rates.





Data Preparation

- There are a number of columns that are probably not useful to analysis due to too many values, such as Reason, Borrower Id, Loan Id, Job, and Loan Title. These columns are removed from analysis.
- Some data needs to be cleaned - dates that are backwards and unclear, dollar and percent signs removed, percent values converted, removal of null values, etc.
- From the heatmap, Loan Grade and Loan Subgrade are highly correlated and Requested, Funded, and Investor Funded are highly correlated. With this information, Loan Subgrade, Funded, and Investor Funded will be dropped from analysis.
- If the interest rate is null, the row will not contribute to the analysis. These rows are removed for analysis.
- For home values, any and none categories were less than 31 records. These values were changed to other.
- The most common value of Years Employed, 10+ Years, was used to fill null values in Years Employed.
- Average income was used to fill null values in Annual Income.
- Forward fill was used to fill null values in Home.

Methodology

- Models tested include Linear Regression, Ridge Regression, Lasso Regression and Elastic Net.
- Evaluation metric used was the root mean squared error (RMSE) because it is sensitive to outliers and the same metric as interest rate.
- Linear and Ridge had similar and the best value of RMSE of approximately .012. Lasso and Elastic Net had RMSE of approximately .04.

Results

- From the RMSE and the residual graph, both Linear and Ridge Regression are appropriate models.
- Using coefficients, it is determined that the features that are the most influential to the interest rate with the other features included are the following:
    - Loan Grade
    - Loan Category
    - Income Verified

Opportunities

- Adding other factors, such as economic measures and credit scores, may help predict borrower interest rate.
- Adding functionality to the point of record entry to restrict fields, such as job, to certain values.
- Requiring borrowers to give other demographic data at point of record entry.

- If a field such as Reason and Job are not restricted at the point of entry, natural language processing is a technique to consider in these fields.
- Date was removed from analysis, but a technique to consider is to break down the date into year, month and day for analysis and/or visualizations.