EDA for loan defaults

Understand driving factors behind loan defaults for a corporate finance company using exponential data analysis

Steps performed in EDA

- Data Loading
- Data Cleaning
- Univariate Analysis
  - Data Description
  - Segmented Analysis
- Bivariate Analysis
  - Categorical Attributes vs Loan Default Ratio
  - Continuous Attributes vs Loan Default Ratio
- Multivariate Analysis
  - Correlation Matrix

Tools and Technologies Used

- Python 3
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Excel

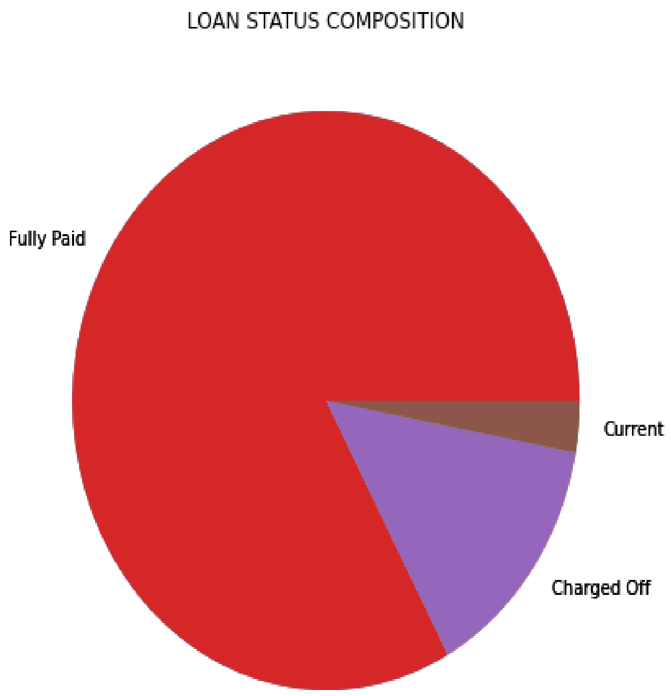
Dataset

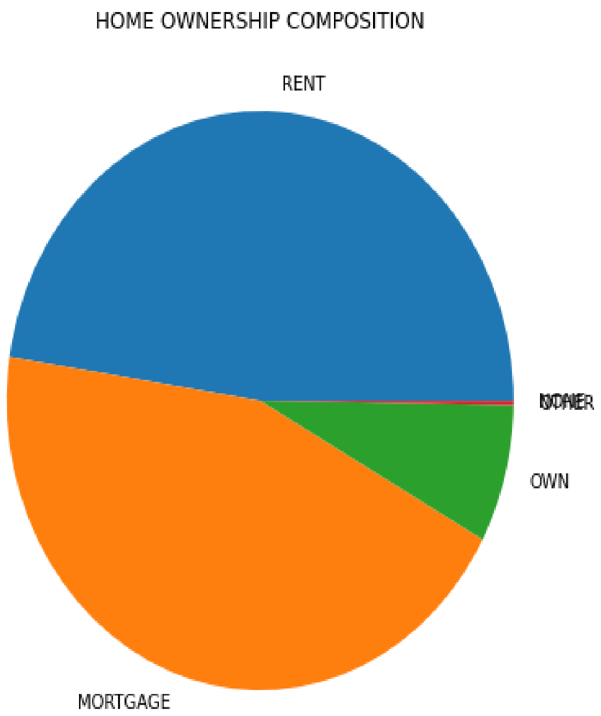
- Data set is record of all loans issued by a corporate finance company from year 2007 to 2011

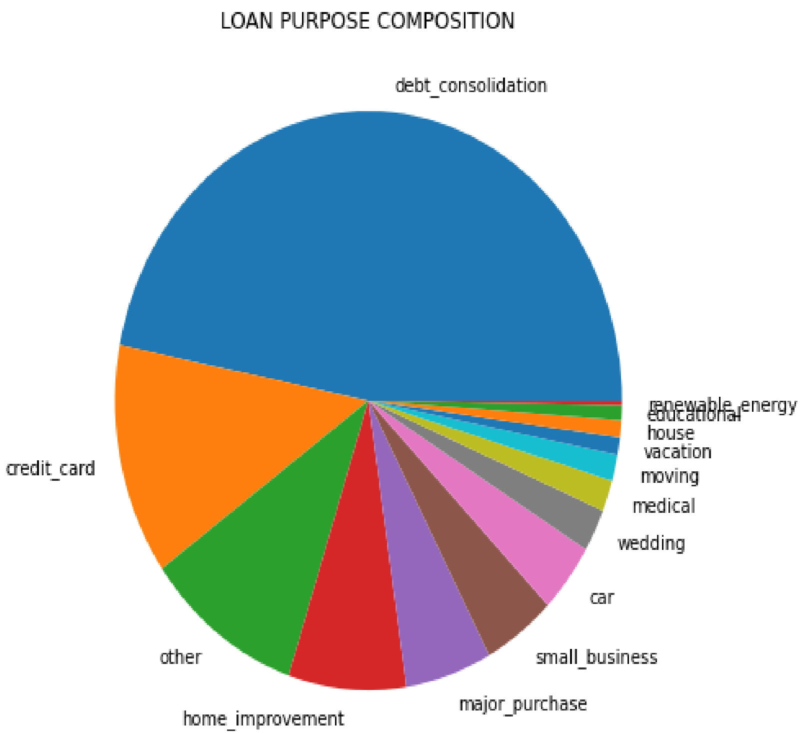
Data Cleaning

- Cleaning empty rows and columns
- Standardising Data
- Removing attributes not relevant to analysis

Univariate Analysis



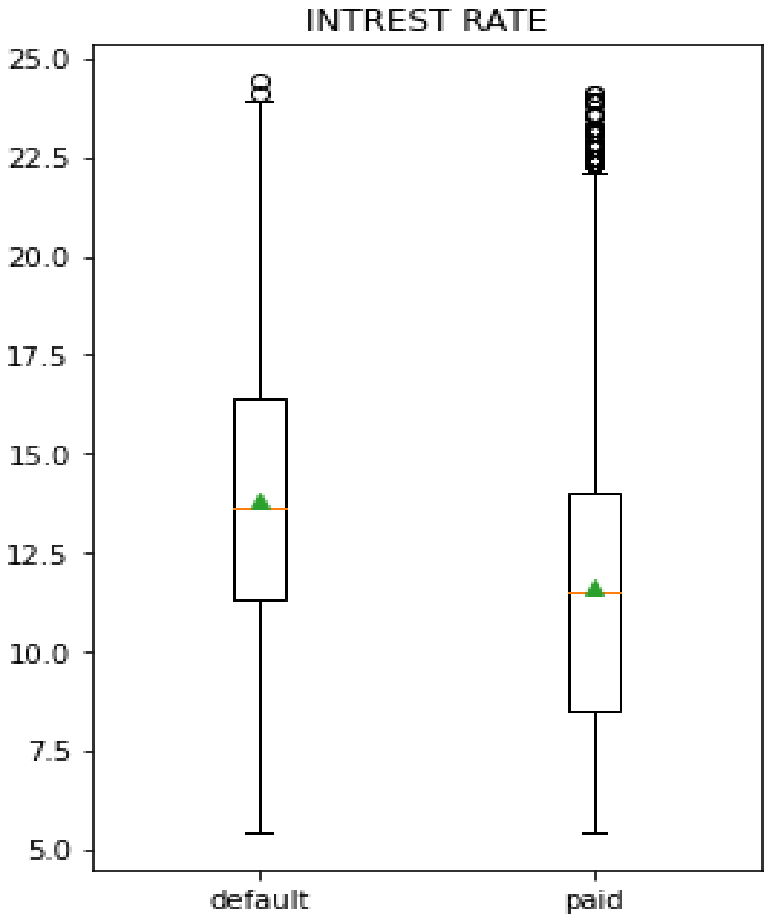




Univariate analysis of categorical variables reaveal

- Majority of loan records are fully paid
- Most loan applicants are living on rent or have mortgaged house
- Debt Consolidation is most common purpose for taking loans

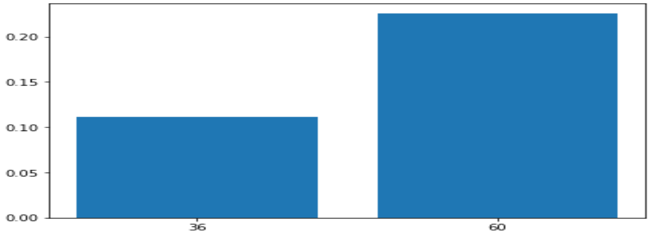
Univariate Analysis - Segmented



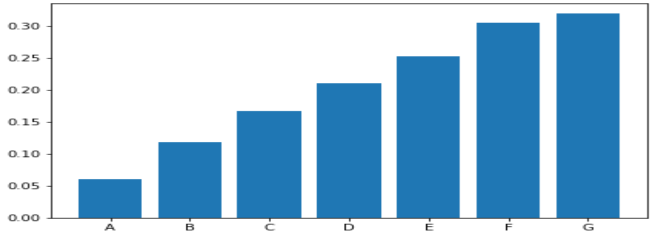
Box plot of interest rate for paid vs defaulted loan reveals that defaulted loan on average have high interest rate

Bivariate Analysis - Categorical

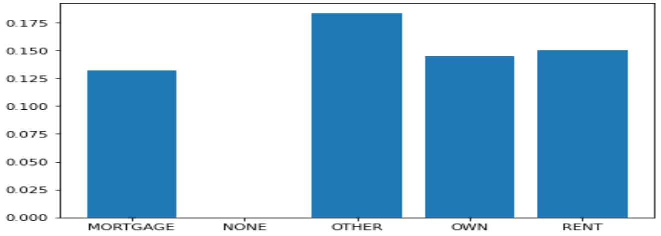
TERM vs LOAN DEFAULT RATIO



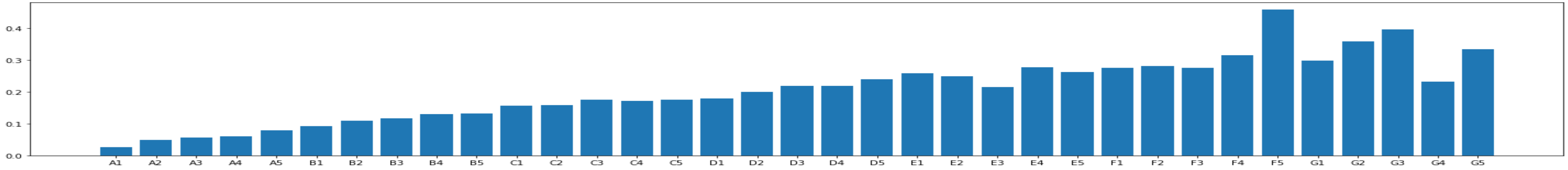
GRADE vs LOAN DEFAULT RATIO



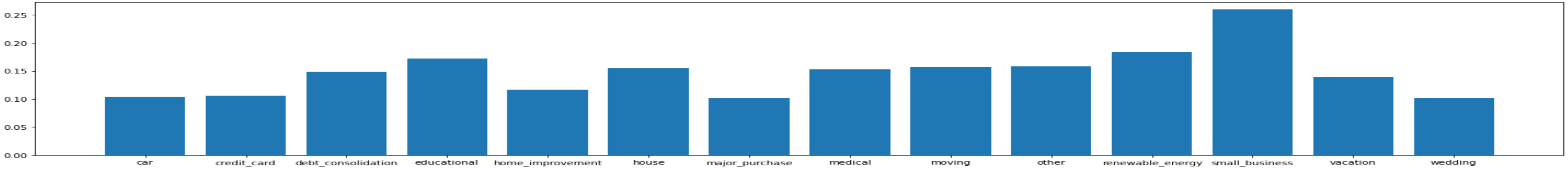
HOME OWNERSHIP vs LOAN DEFAULT RATIO



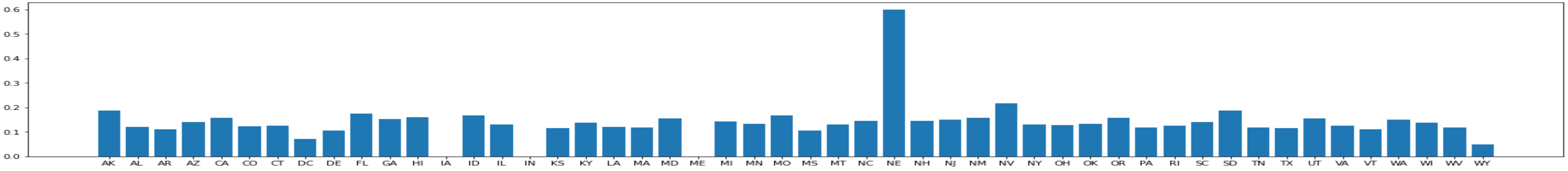
SUB GRADE vs LOAN DEFAULT RATIO



LOAN PURPOSE vs LOAN DEFAULT RATIO



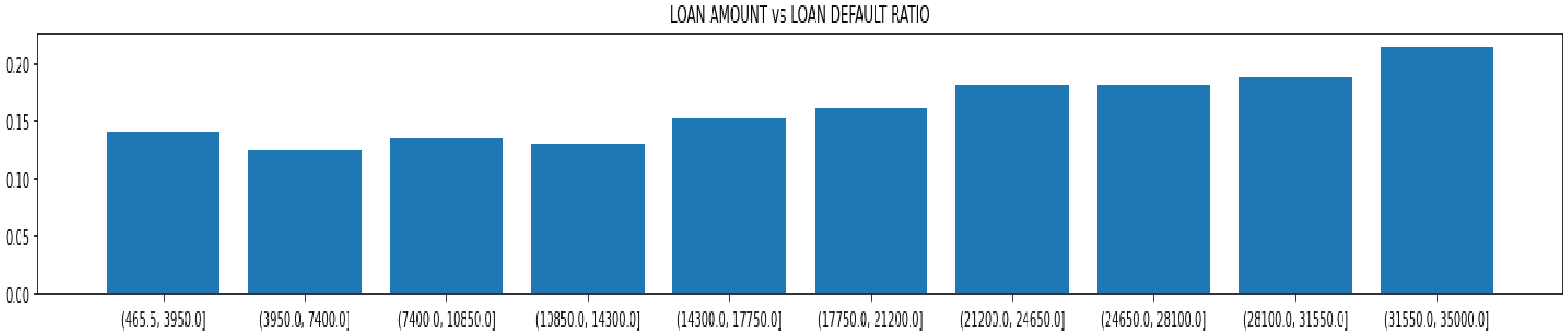
STATE vs LOAN DEFAULT RATIO

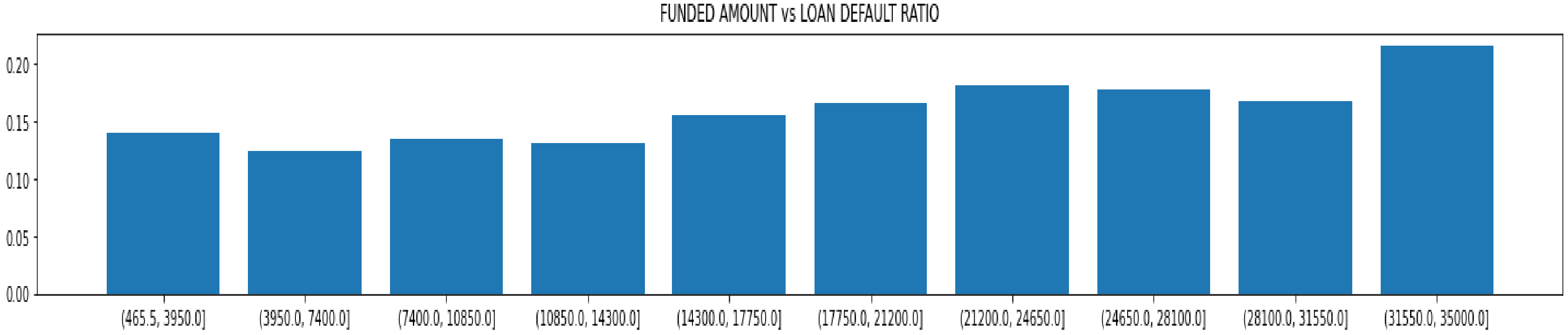


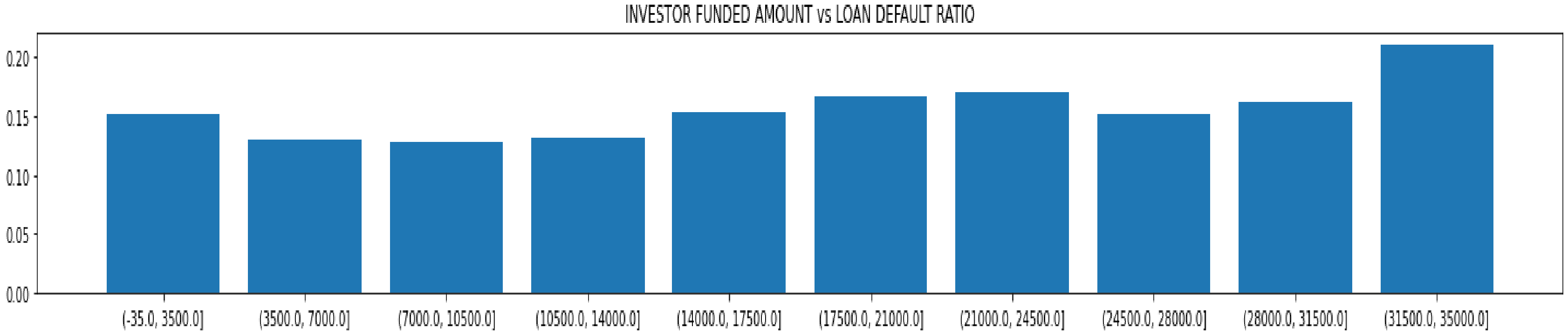
Bivariate Analysis – Categorical(contd.)

- Long term loans in general have higher default rates. This can be attributed to chance of financial distress being higher in long duration.
- Grade and Subgrade both show positive correlation with loan default ratio
- Home Ownership doesn’t have much influence on Loan Default.
- Loan taken for small businesses have significantly higher default rate. This can be attributed to unpredictability with successful running of small business. However rest of the loan purpose are close to average.

Bivariate Analysis - Continuous

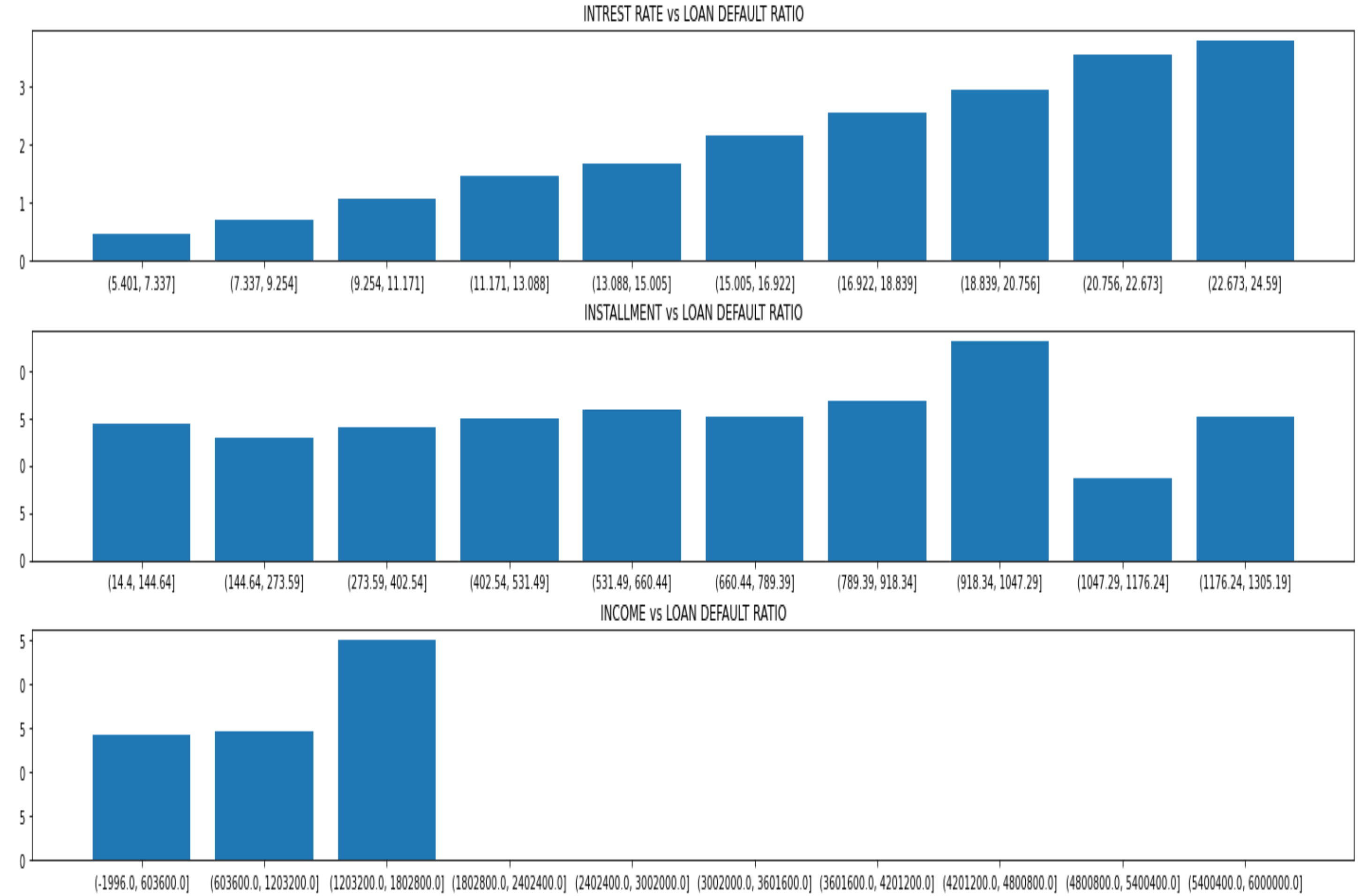






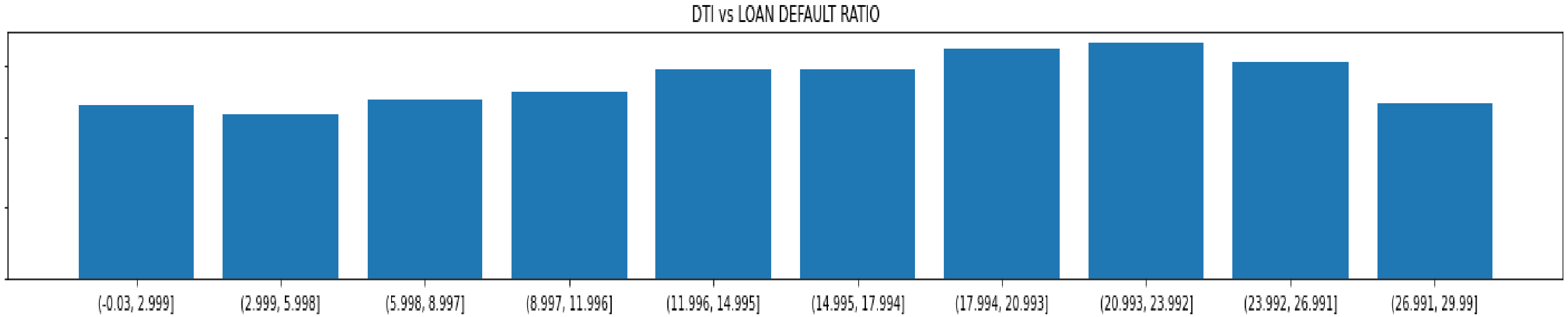
- Higher Loan Amount, Funded Amount, Investor Funded Amount result in higher default.

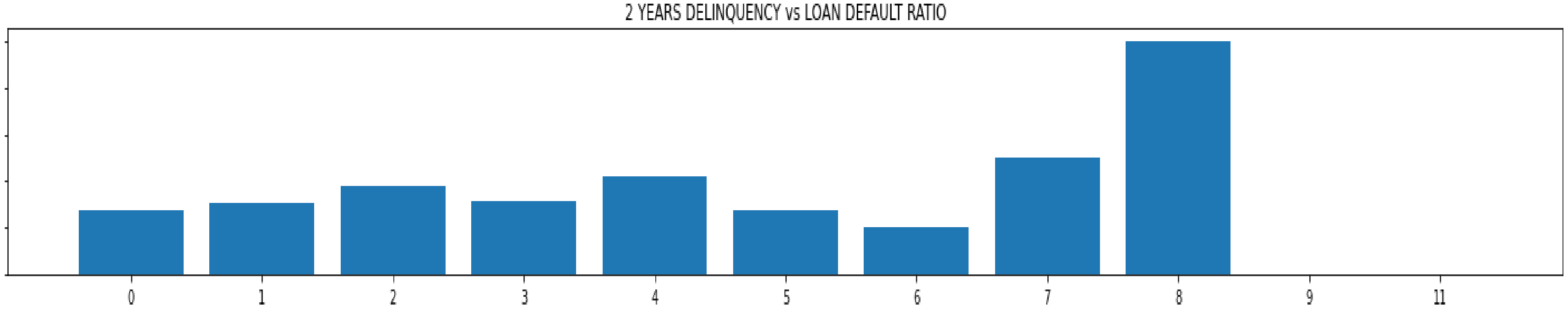
Bivariate Analysis – Continuous(contd.)

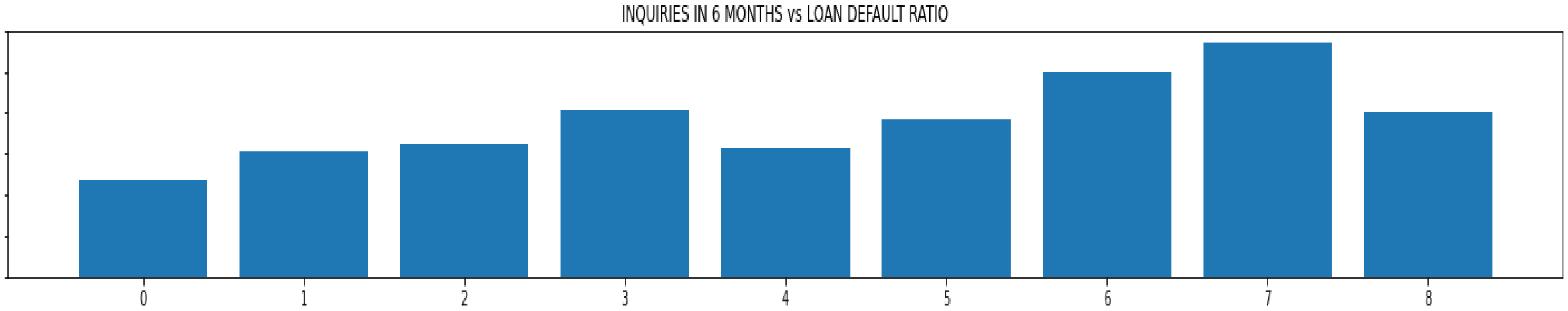


- Loan Default Rate increases with interest rate
- Installment and loan default ratio show neutral relation
- Higher Income Groups have negligible loan defaults

Bivariate Analysis – Continuous(contd.)

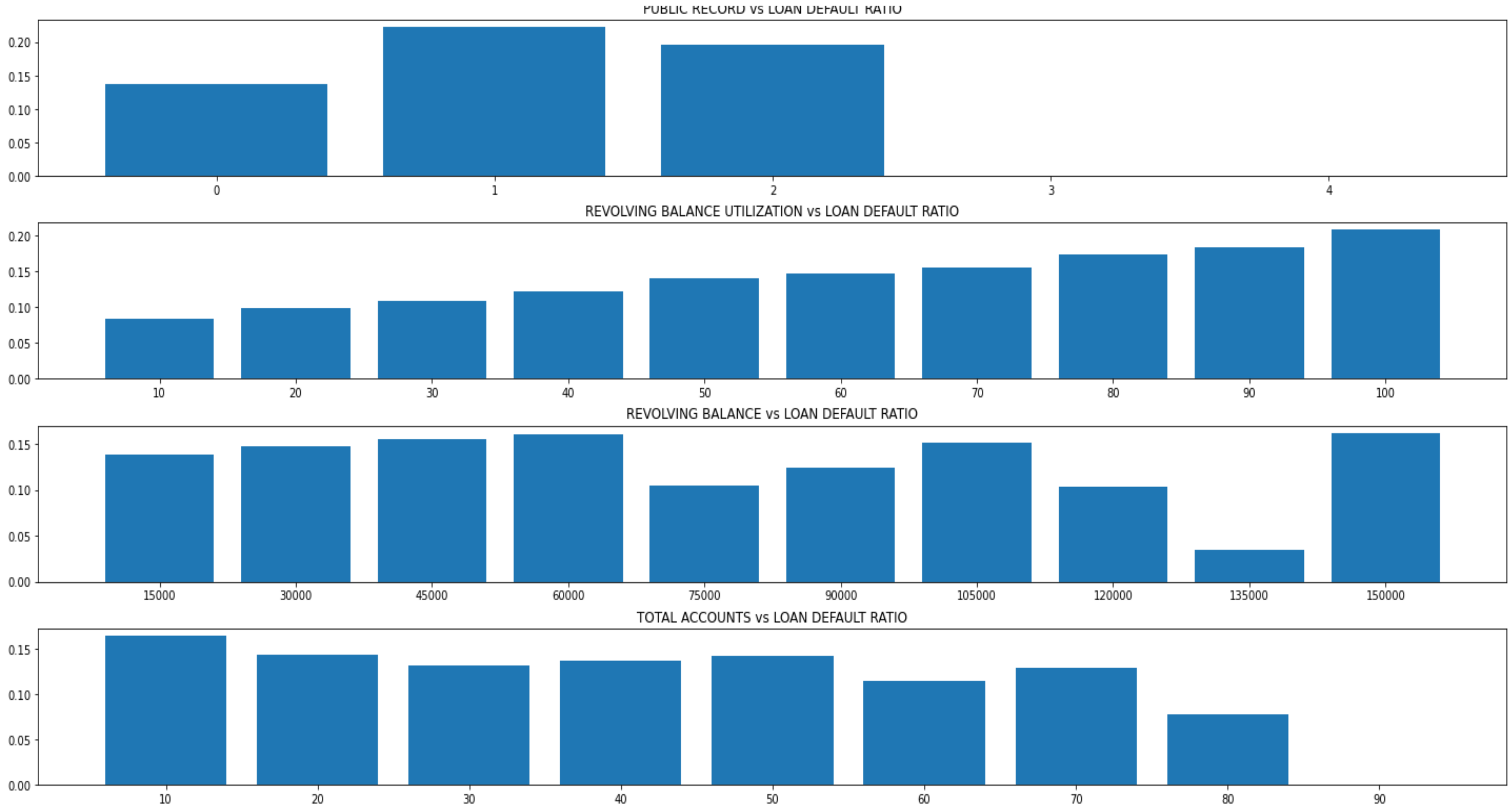






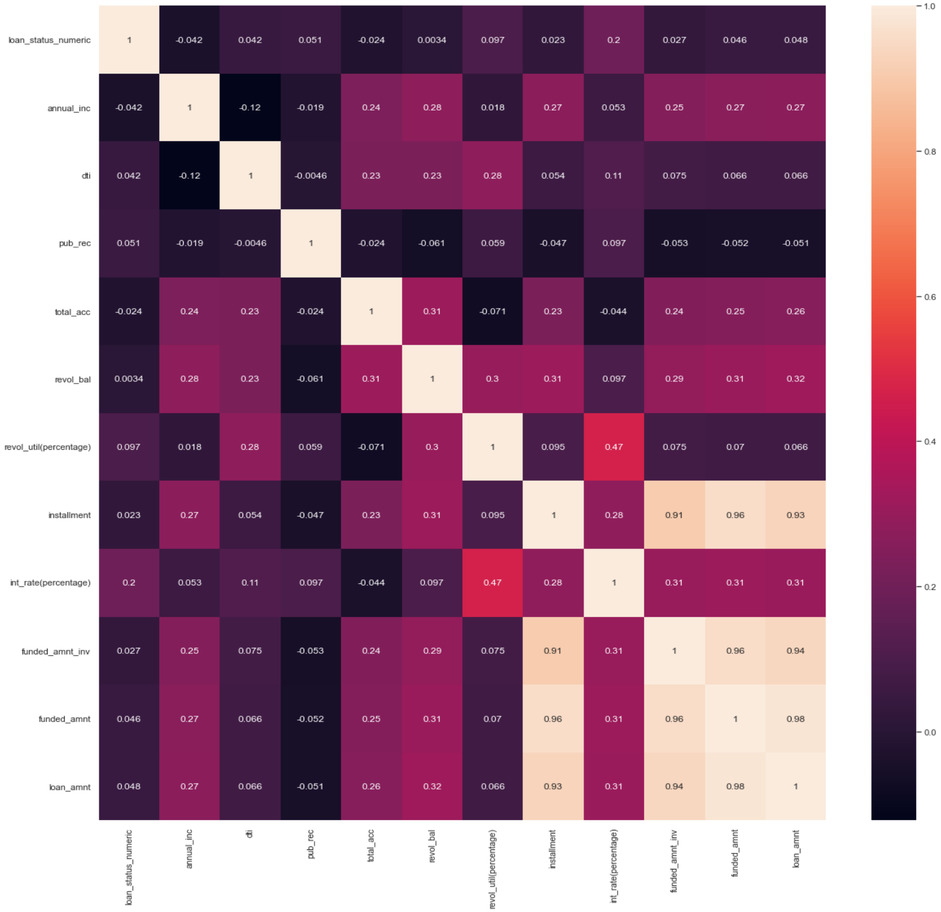
- DTI, 2 years delinquency and inquiries in 6 month doesn’t show any relation with loan default ratio

Bivariate Analysis – Continuous(contd.)



- Loan Default Rate increases with Revolving Balance Utilization.
- Public Record, Revolving Balance, Total accounts show no relation.

Multivariate Analysis



	loan_status_numeric	annual_inc	dti	pub_rec	total_acc	revol_bal	revol_util(percentage)	installment	int_rate(percentage)	funded_amnt_inv	funded_amnt	loan_amnt
loan_status_numeric	1	-0.042	0.042	0.051	-0.024	0.0034	0.097	0.023	0.2	0.027	0.046	0.048
annual_inc	-0.042	1	-0.12	-0.019	0.24	0.28	0.018	0.27	0.053	0.25	0.27	0.27
dti	0.042	-0.12	1	-0.0046	0.23	0.23	0.28	0.054	0.11	0.075	0.066	0.066
pub_rec	0.051	-0.019	-0.0046	1	-0.024	-0.061	0.059	-0.047	0.097	-0.053	-0.052	-0.051
total_acc	-0.024	0.24	0.23	-0.024	1	0.31	-0.071	0.23	-0.044	0.24	0.25	0.26
revol_bal	0.0034	0.28	0.23	-0.061	0.31	1	0.3	0.31	0.097	0.29	0.31	0.32
revol_util(percentage)	0.097	0.018	0.28	0.059	-0.071	0.3	1	0.095	0.47	0.075	0.07	0.066
installment	0.023	0.27	0.054	-0.047	0.23	0.31	0.095	1	0.28	0.91	0.96	0.93
int_rate(percentage)	0.2	0.053	0.11	0.097	-0.044	0.097	0.47	0.28	1	0.31	0.31	0.31
funded_amnt_inv	0.027	0.25	0.075	-0.053	0.24	0.29	0.075	0.91	0.31	1	0.96	0.94
funded_amnt	0.046	0.27	0.066	-0.052	0.25	0.31	0.07	0.96	0.31	0.96	1	0.98
loan_amnt	0.048	0.27	0.066	-0.051	0.26	0.32	0.066	0.93	0.31	0.94	0.98	1

Correlation matrix show higher positive correlation with interest rate, grade, subgrade and revolving balance utilization

Conclusion

Most important deciding factors according to analysis are

- Loan Amount, Funded Amount, Investor Funded Amount
- Interest Rate
- Term
- Grade and Subgrade
- Revolving Balance Utilization