Upgrad PG Diploma in Machine Learning and AI- October 2020

Lending Club Case Study

Team : Uhapriya K S , Rajeev Sawant

Agenda

Objective

Approach

Data understanding

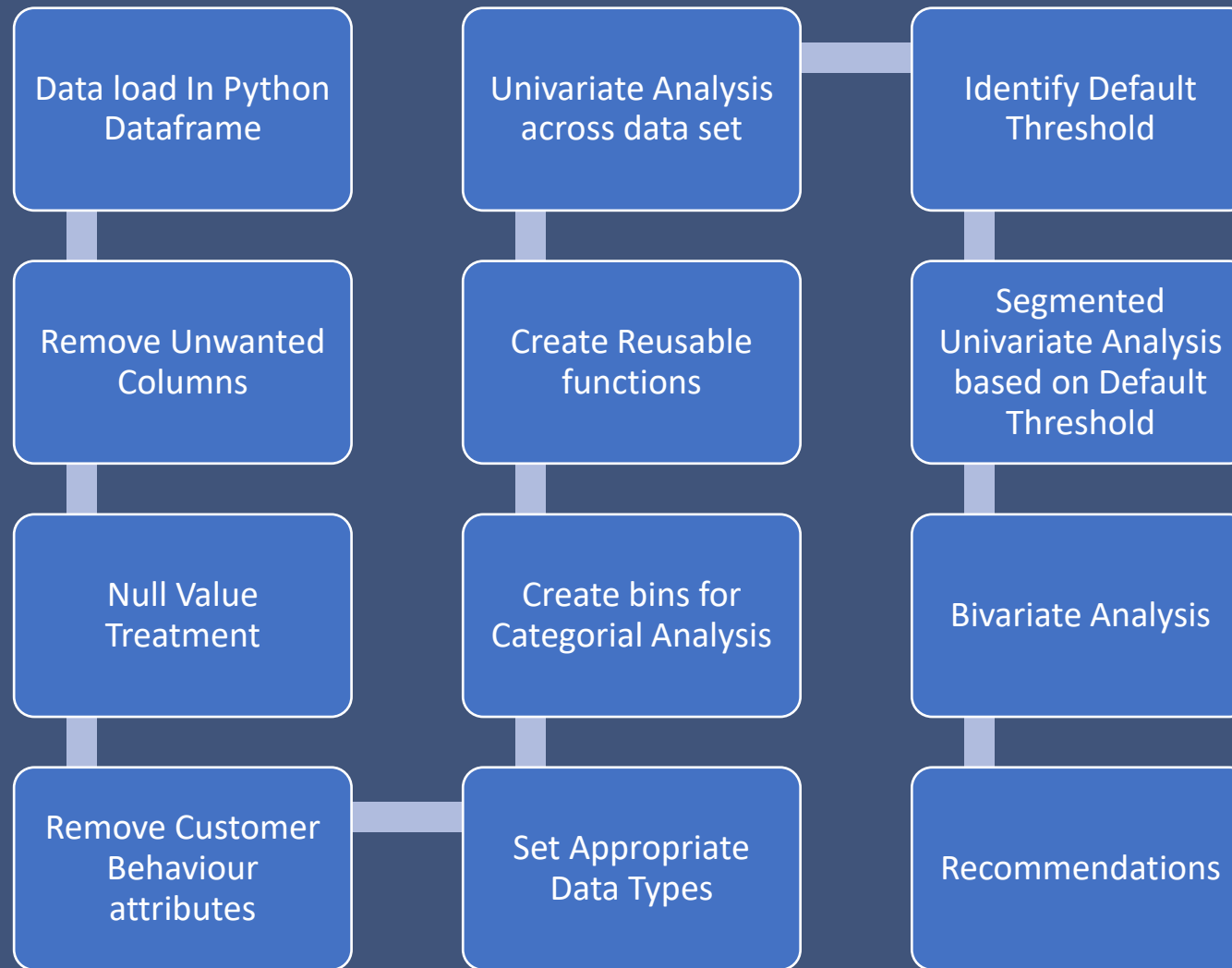Data Cleaning & Manipulation

Data analysis

Recommendations

Objective

AS a Consumer finance company use exploratory data analysis to understand how **consumer attributes** and **loan attributes** influence the tendency of default

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Approach used for Analysis

- Data load In Python Dataframe
- Remove Unwanted Columns
- Null Value Treatment
- Remove Customer Behaviour attributes

- Univariate Analysis across data set
- Create Reusable functions
- Create bins for Categorial Analysis
- Set Appropriate Data Types

- Identify Default Threshold
- Segmented Univariate Analysis based on Default Threshold
- Bivariate Analysis
- Recommendations

Data Understanding

Key Observations

- Data file : loan.csv
- Shape of the dataset : (39717, 111) :
  - Number of Columns : 111
  - Number of Rows : 39717

- There are many data attributes with more than 90 % null value
- There are many attributes with one value only
- There are few attributes like id/ member_id which are not relevant for the analysis
- There are few attributes which are related to customer behavior and not suitable for analysis
- There are many values with status as current which is not suitable for analysis

Data Cleaning & Manipulation

Key Actions

- Remove columns with more than 90 % null values
- Drop customer behaviour attributes and keep only Loan attributes & customer attributes
  ["delinq_2yrs", "earliest_cr_line", "inq_last_6mths", "open_acc", "pub_rec", "revol_bal", "revol_util", total_acc", "out_prncp", "out_prncp_inv", "total_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "total_rec_late_fee", "recoveries", "collection_recovery_fee", "last_pymnt_d", "last_pymnt_amnt", "last_credit_pull_d"]
- Remove columns with missing values
- Remove irrelevant columns
- Remove rows which has loan_status as current
- Convert Strings to numeric : emp_length, term, int_rate
- Rounded Floats : funded_amnt_inv
- Date conversion to Months & Years : issue_d -> issue_month & issue_year
- Derived columns : funded_to_income (ratio of funded amount to Annual income
- Loan status conversion : Fully Paid = 0 , Charged off = 1
- Create Data bins : installment_bins, funded_to_income_bins, int_rate_bins, dti_bins, funded_amnt_inv_bins, dti_5_bins, annual_inc_bins, funded_amnt_bins

Univariate Analysis
Key Observations

- The % of charged off loans increases substantially as we go up the loan amount buckets. Most loans are below 20000 amount. The higher loans, though lesser in number, carry a substantially higher risk of default.

- There are 2 terms - 36 months and 60 months in the entire dataset.
- ~75% of loans have 36 months term and ~25% have 60 months term.

- There are 3 categories - Verified, not verified and source verified
- ~43% are not verified.

- Debt_consolidation and credit card comprises more than 50% of loan.

- Employee with less than or equal to one year and more than 9-year experience are likely to apply for loans.

- People living in rental and having mortgages are more likely to apply for loan.

- Grade with A, B and C comprises more than 70% of the loan.

- Steady increase in month-on-month loan request as a year progresses.

- Substantial increase in loan as the year on year.
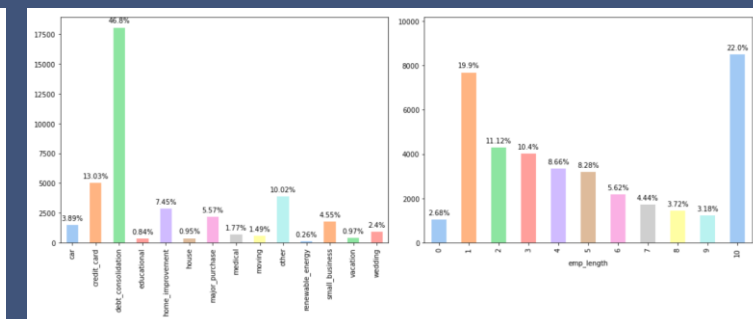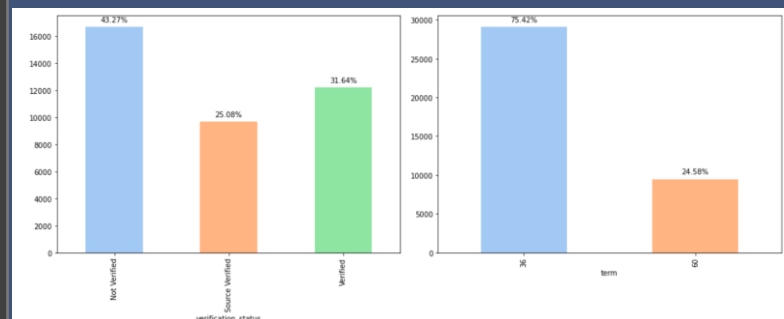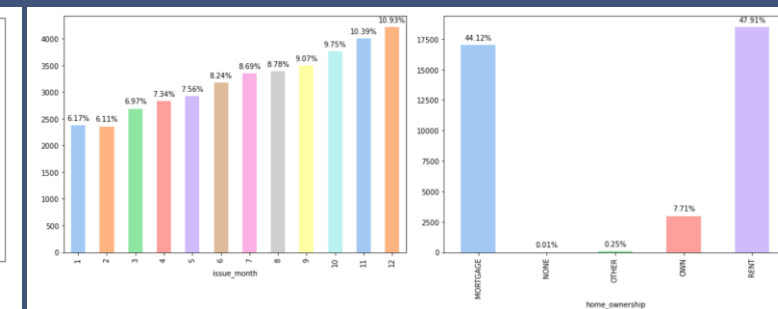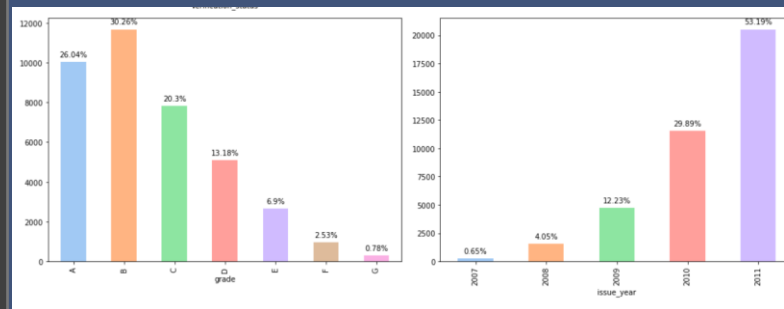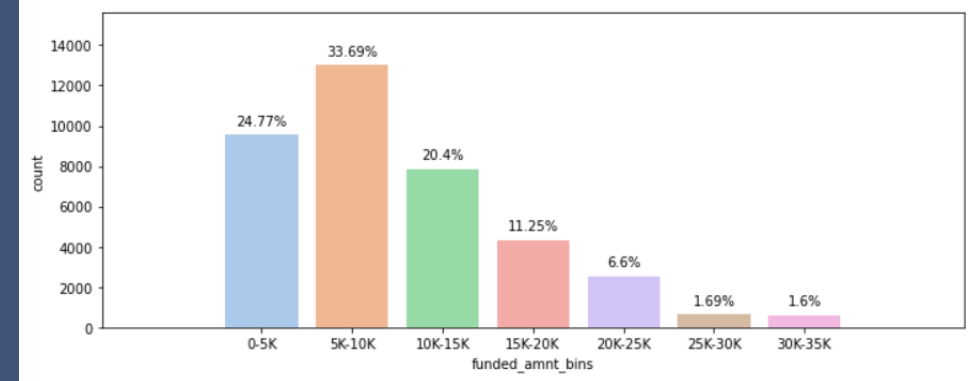
Univariate Analysis
Key Observations Continued

- People with annual income less than 100k are more likely to apply for loan.
- More than 70% funded amount is less than 15k.
- More than 70% of loan funded amount will be 30% of the annual income of requestor.

- 70% of the requestors likely to choose 5% to 15% interest rate.
- Almost 90% of the loans have installment less than 600.
- More than 70% funded amount by investors is less than 15k.
- Almost 94% of the requestors are not having public records of bankruptcies.

Segmented Univariate Analysis

- Higher the installment rate, higher the default rate.
- As grade increases, risk factor is increasing at almost 5% per grade from A to F. There is high probability of default from grade C to grade G.
- There is a linear increase. F5 have almost 50% defaulters and G3 have 42% defaulters and clearly standing out.
- There are no right correlation between emp_length and risk factor. 0 is imputed for null values.
- Loans with 60-month term are likely to be defaulted.
- debt_consolidation,educational,house,medical,moving,other,renewable_energy and small_business are the purposes which are likely to be defaulted. People availing loan for small business have 30% probability of default.
- Other and rented are above the threshold.
- There are high chances of default with existing bankruptcies record. People with 2 bankruptcies record are having 40% probability of default.
- As the funded_to_income ratio increases, the risk of default increases. The ratio above 20% is above the threshold for defaulters.
- Funded amount invested above 15k are likely to be defaulted.
- The dti values within 15 to 27 are above the threshold.
- Higher the interest rate, higher the chance of default. Default percent increases exponentially as the interest rate increases. Interest rate above 15% are above threshold and likely to be defaulted.
- People with annual income less than 50k and between 500k-1000k are above threshold.
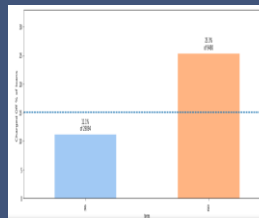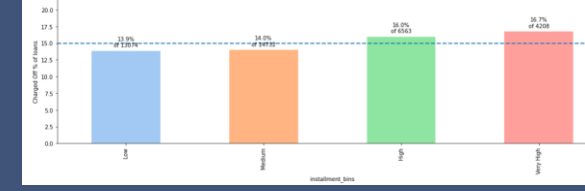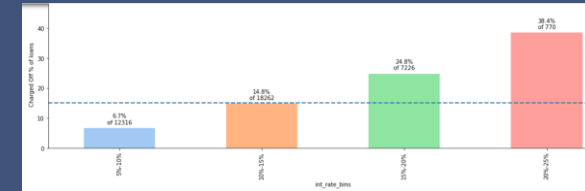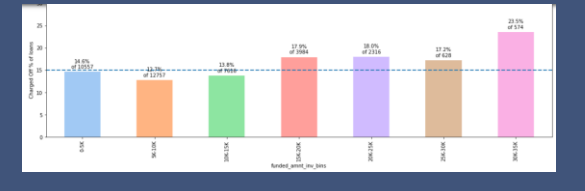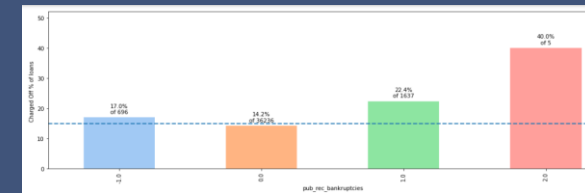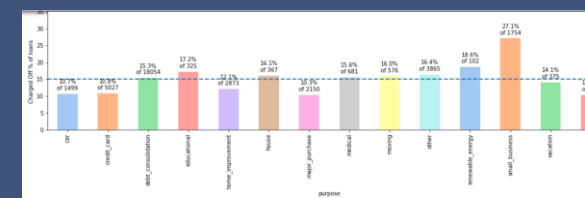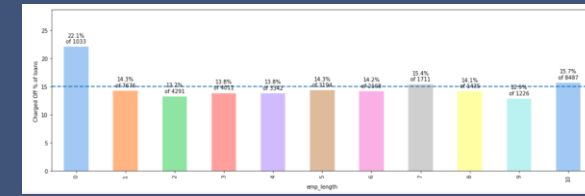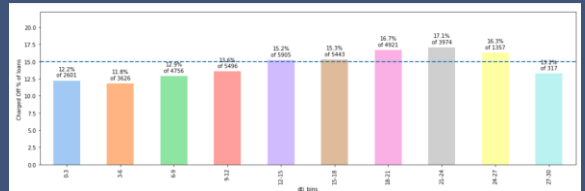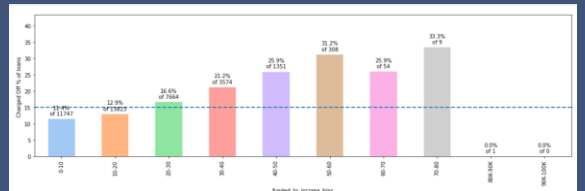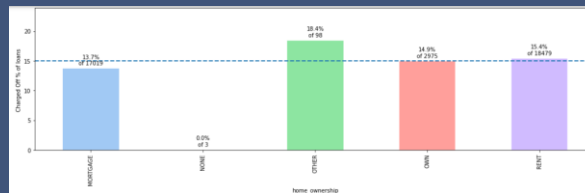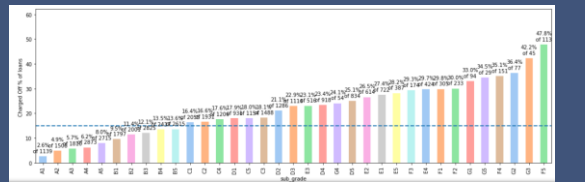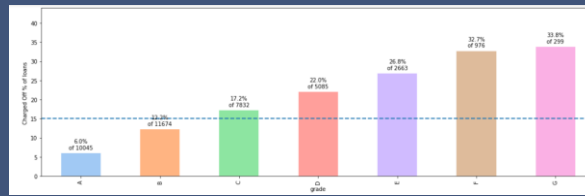
# Data Analysis : Segmented Univariate Analysis

States above Threshold

State	Percentage Defaulters
New Hampshire	15%
North Carolina	15%
Washington	16%
New Jersey	16%
Georgia	16%
Maryland	16%
Utah	16%
California	16%
Oregon	16%
New Mexico	16%
Idaho	17%
Hawaii	17%
Missouri	17%
Florida	18%
Alaska	19%
South Dakota	19%
Nevada	23%
Nebraska	60%
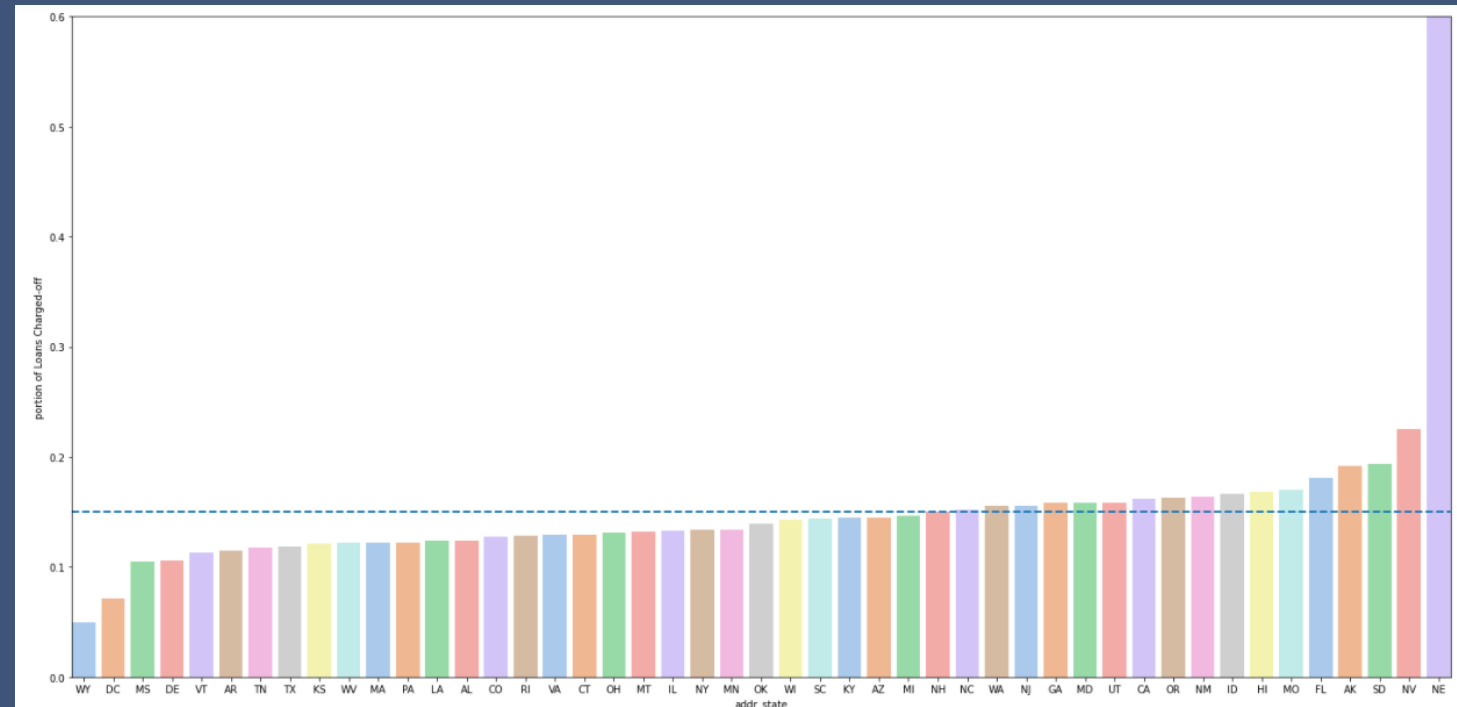
Key Observations

- Verified loans are funded high.
- There is positive correlation between grade and funded_amnt_inv.
- Funded_amnt_inv is high for 60 months term.
- Mortgage and own are getting higher funded_amnt_inv
- Even though there are public record bankruptcies, investors are lending money to requestors.
- Investors are lending money for low and medium dti.
- As funded_amnt_inv increases, interest rate increases.
- As annual income increases, funded amount by investors increases.
- Even though verification done for loans. we could see that there are defaulters mapping to the high invested amount.
- There are people with average income lower than 50000 taking loans of 25000 or higher. These would-be risky loans.
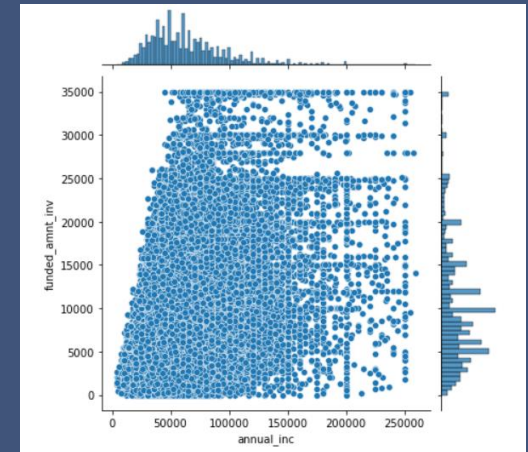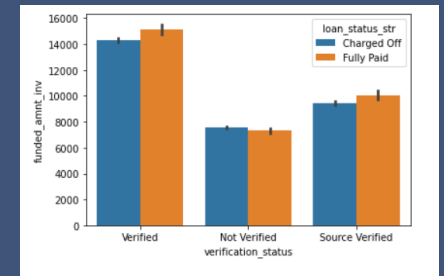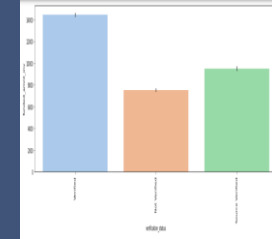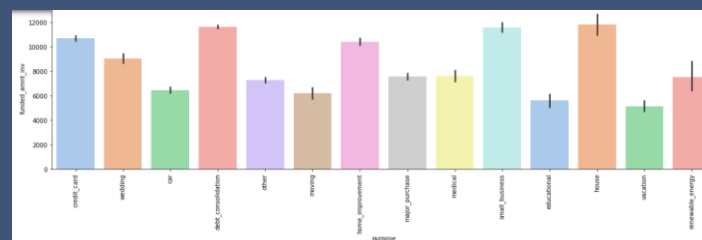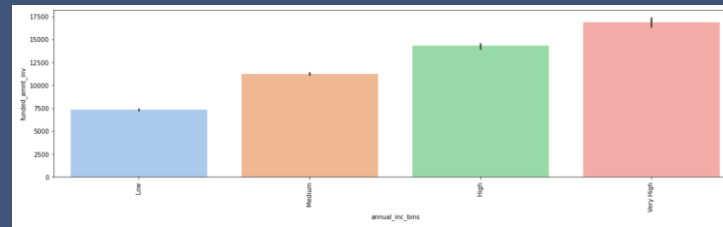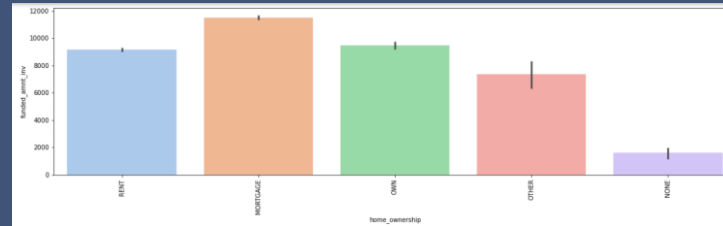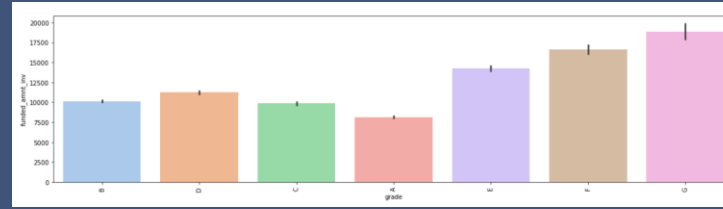
Heat map created for Charged off data set

- High correlation between
  - Loan Amount & Interest Rate
  - Loan Amount & Installment
  - Loan Amount & Annual Income
  - Interest Rate & Installments
  - Installment & Annual Income

- Negative correlation between
  - Loan Amount & Employment Length
  - Loan Amount & Public Bankruptcies
  - Interest Rate & Employment Length
  - Installment & Employment Length
  - Annual Income & Employment Length
  - Annual Income & DTI
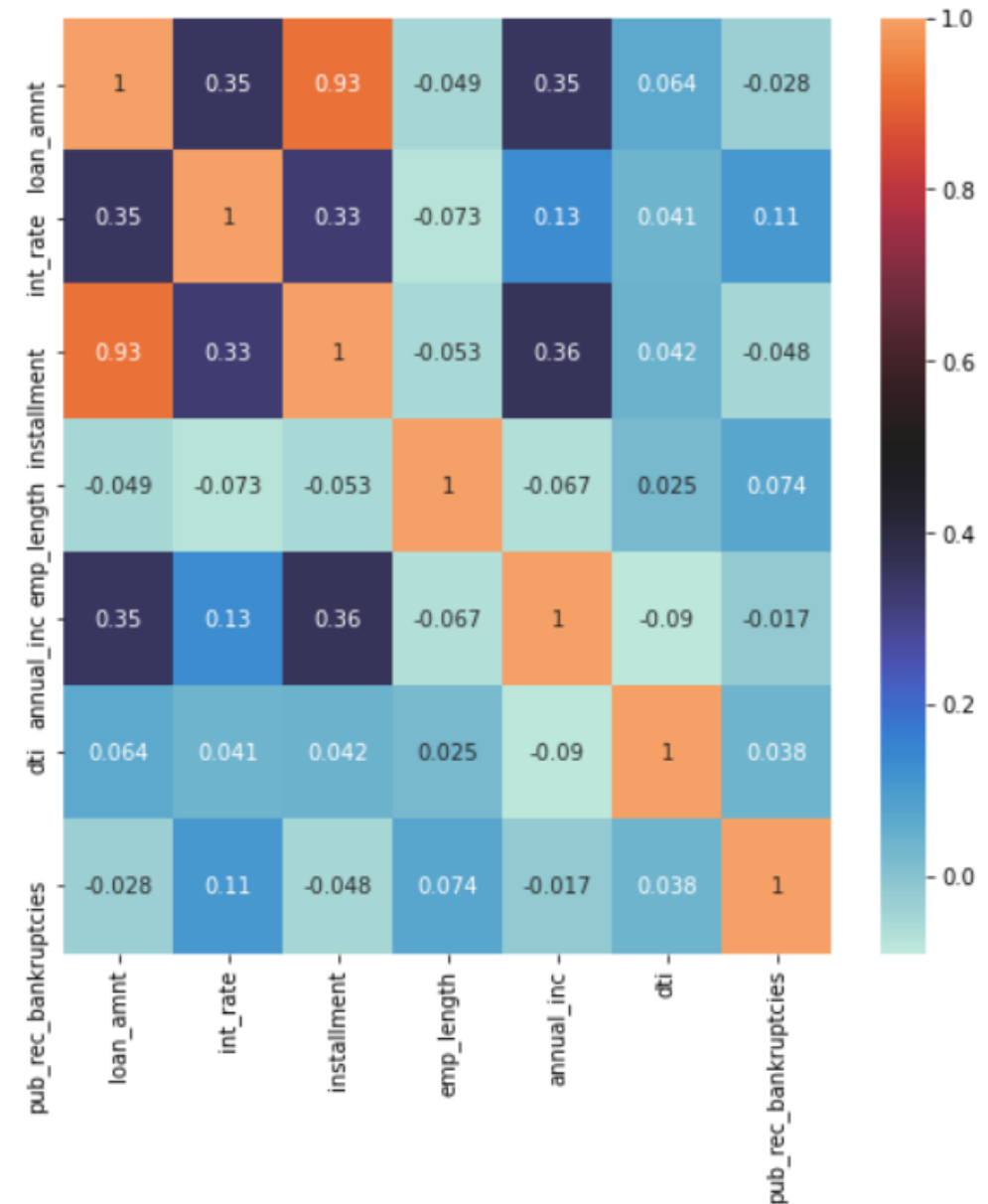
Recommendations

Below are High Risk Metrics
Due diligence / reduction in funded loan amount is recommended for the same

Attributes	High Risk Items	Recommendation
Grade	C,D,E,F,G	Avoid high grade loans
Interest Rate	Above 15 %	Keep interest rate below 15 %
Home Ownership	Rented	Check other parameters in the list for rented loan applicant
Purpose	Small business & renewable energy	Avoid loans to small business set up as it has more than 25 % defaulters
Term	60 Months	Avoid 60 Months term , Go for 36 months term
Public Record Bankruptcies	> 0	Avoid loans to applicants with public records bankruptcies
State	Alaska , South Dakota, Nevada, Nebraska	Do thorough checking before giving loans to applicants from these states
Funded Amount	> 15000	Avoid giving loans with high amount specially more than 25000
DTI	15 – 27	Do thorough check if DTI values are in this range
Annual Income	< 50000 500000 - 1000000	There is a high risk in giving loans to applicant with annual income less than 500000 also reduce funded amount if loan range is between 500000-1000000
Funded Amount to Annual income	> 30	Avoid loan imbursement of high amount compared to less annual income. Ideally fund amount to annual income should be less than 25 %

Thank You