Lending Club Case Study

Predicting defaults at the time of loan approval

Missing Values

- The last 57 columns are completely null, and were removed, except for past bankruptcies
- Employment length was assumed to be zero if null, meaning retirement or unemployed.
- Dropped rows with no revolving utilization balance

Redundant or did not fit objective

- Member_id'
- Desc
- Acc_now_delinq'
- Chargeoff_within_12_mths'
- Delinq_amnt'
- Tax_liens
- emp_title
- Ccollections_12_mths_ex_med'
- Title
- Policy_code'
- Initial_list_status'
- Url
- Delinq_2yrs'
- Id
- next_pymnt_d

- Earliest_cr_line
- Pymnt_plan
- Inq_last_6mths
- Pub_rec
- 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
- application_type

Filled, edited, or omitted Values

- Removed all rows which were ‘current’ on loan status, and hadn’t finished their term yet
- Filled null values of employment length with 0
- interest rate column and revolving utilization changed from object/string % to float
- Dropped columns months since last delinquency/public record because they were mostly null and this data is recent and not known at the time of approval
- Created new column from Loan Status with boolean values 0 for fully paid, 1 for default
- Filled all NA values in revolving utilization with 0 after confirming all but 1 had 0 balance
- Removed 1 row with non-zero balance but NA utilization
- Removed all rows which were funded less than $100 by investors, because it was assumed these loans did not affect investors

Potentially useful columns:

```
Index(['loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate',
       'installment', 'grade', 'sub_grade', 'emp_length', 'home_ownership',
       'annual_inc', 'verification_status', 'issue_d', 'loan_status',
       'purpose', 'zip_code', 'addr_state', 'dti', 'open_acc', 'revol_bal',
       'revol_util', 'total_acc', 'pub_rec_bankruptcies', 'LS_bool'],
      dtype='object')
```

Outliers

- Removed rows which Home Ownership was ‘Other’ or ‘None’ because there were less than 1% of the data and too ambiguous to fill with other data which was split evenly between renting and owning.

Outliers & Skewness of numerical data

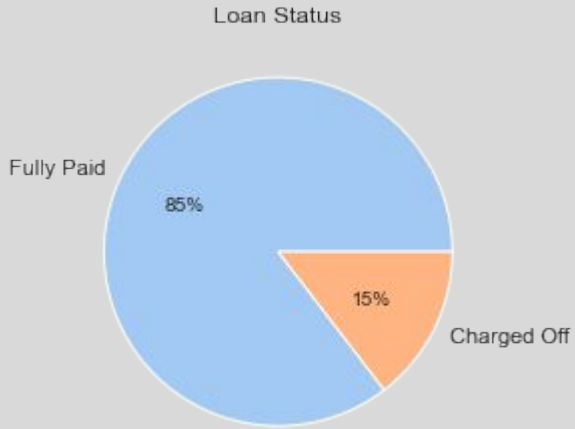
- Numerical data was split into its own dataframe and the interquartile range was determined and all values below Q1 - 1.5 * IQR and above Q3 + 1.5 * IQR were removed
- Histograms were made for all numerical data and for skewed variables np.log was applied to unskew the data

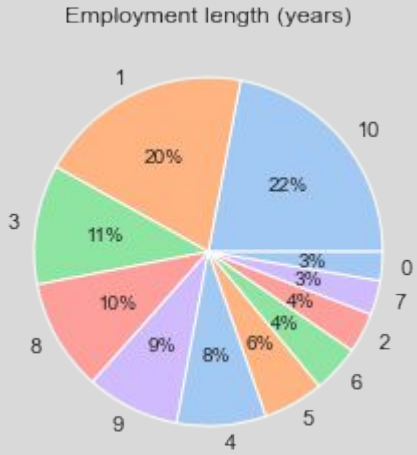
Univariate Analysis

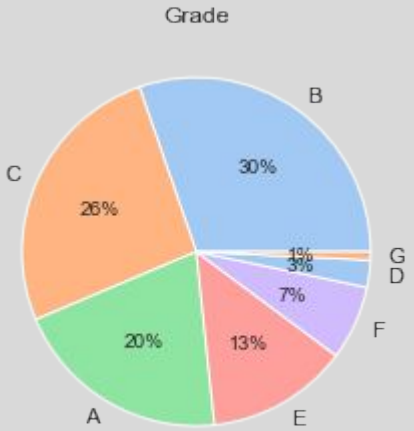
- In total investors on LendingClub lent $61,134,661 to people who would eventually be charged-off or default

- Lending to borrowers who own a home means recuperation of funds is more likely, less for mortgage owners, and even lesser for renters.

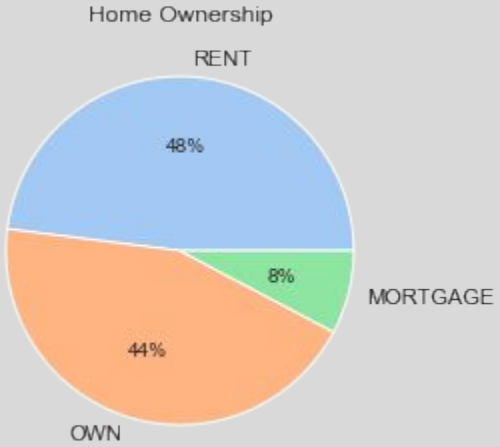
Univariate Analysis

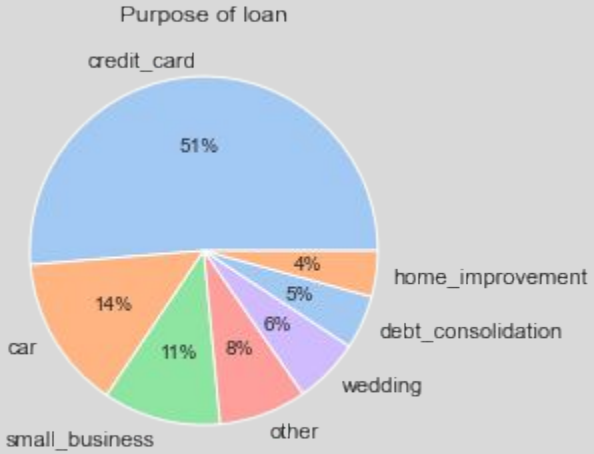


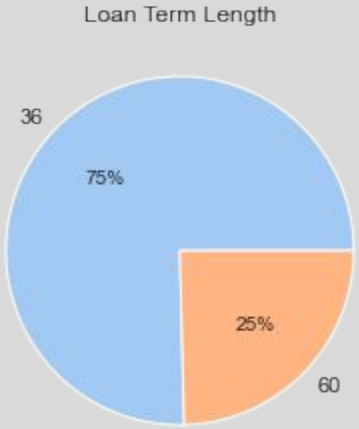




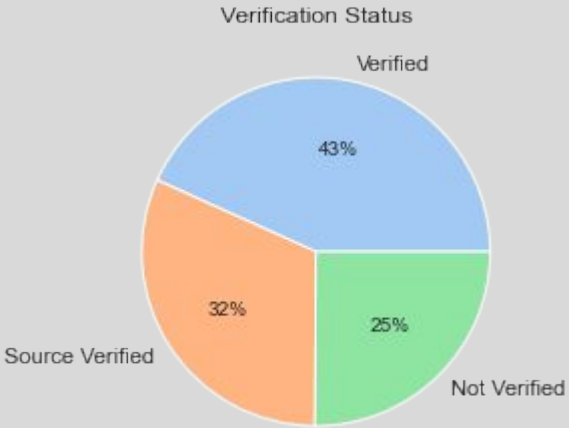
Univariate Analysis



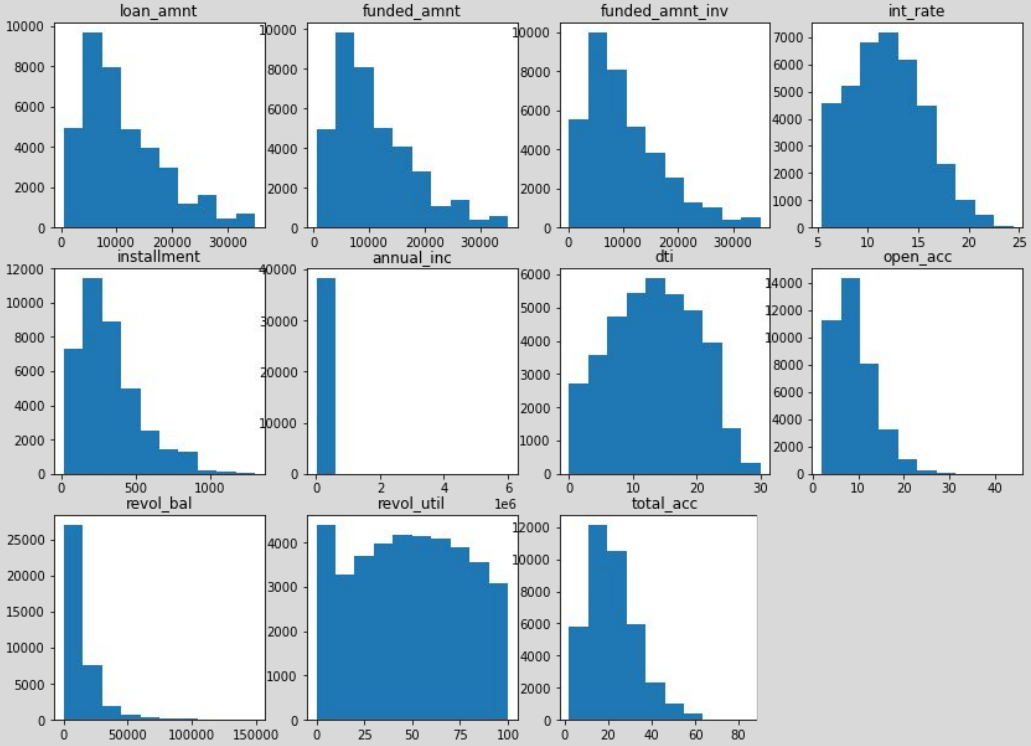




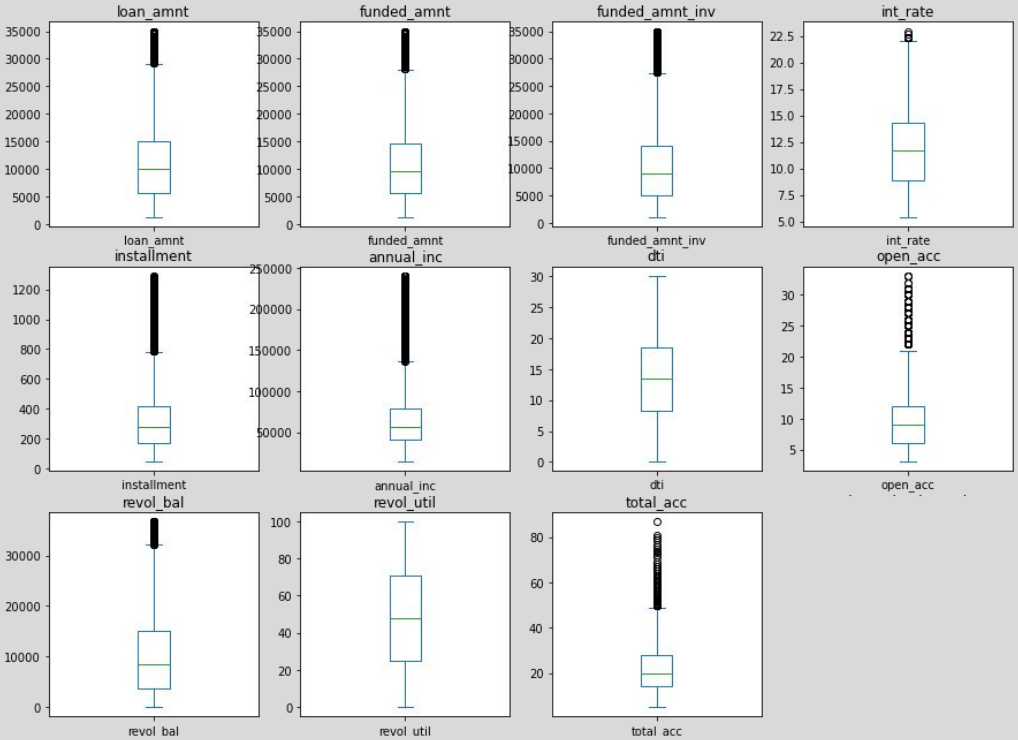
Univariate Analysis



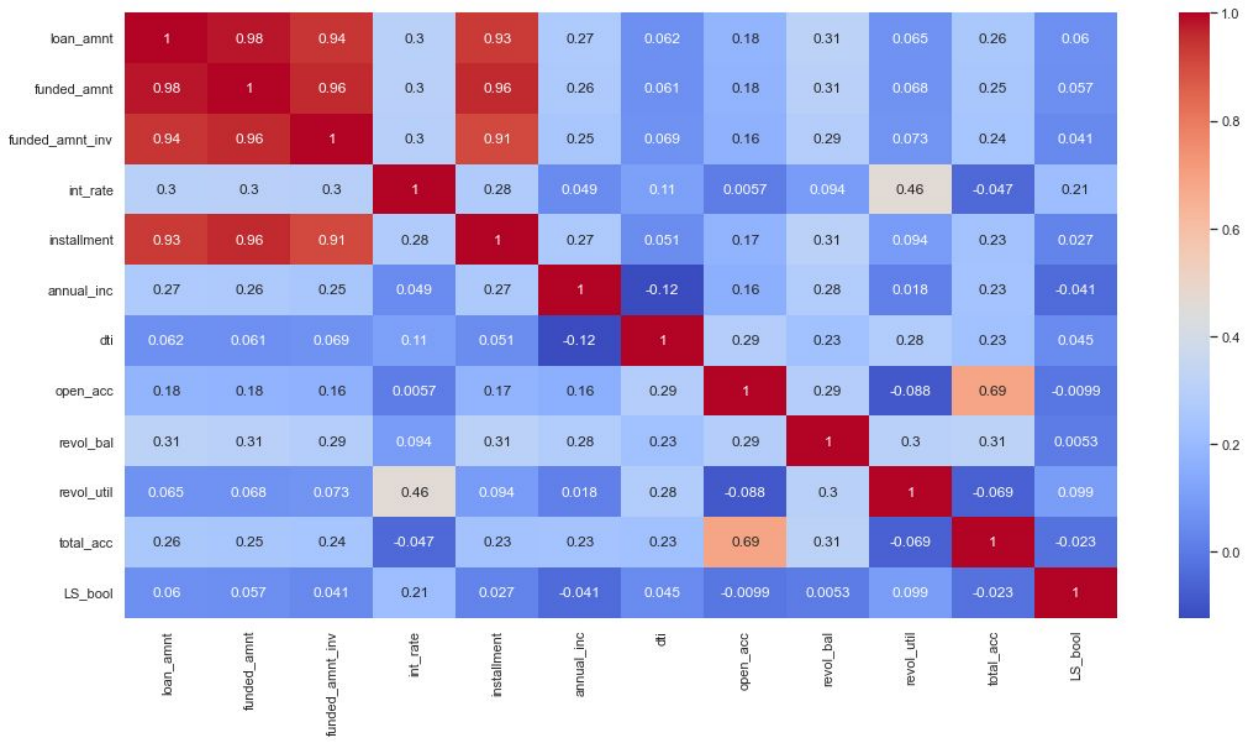
Univariate Analysis



Univariate Analysis



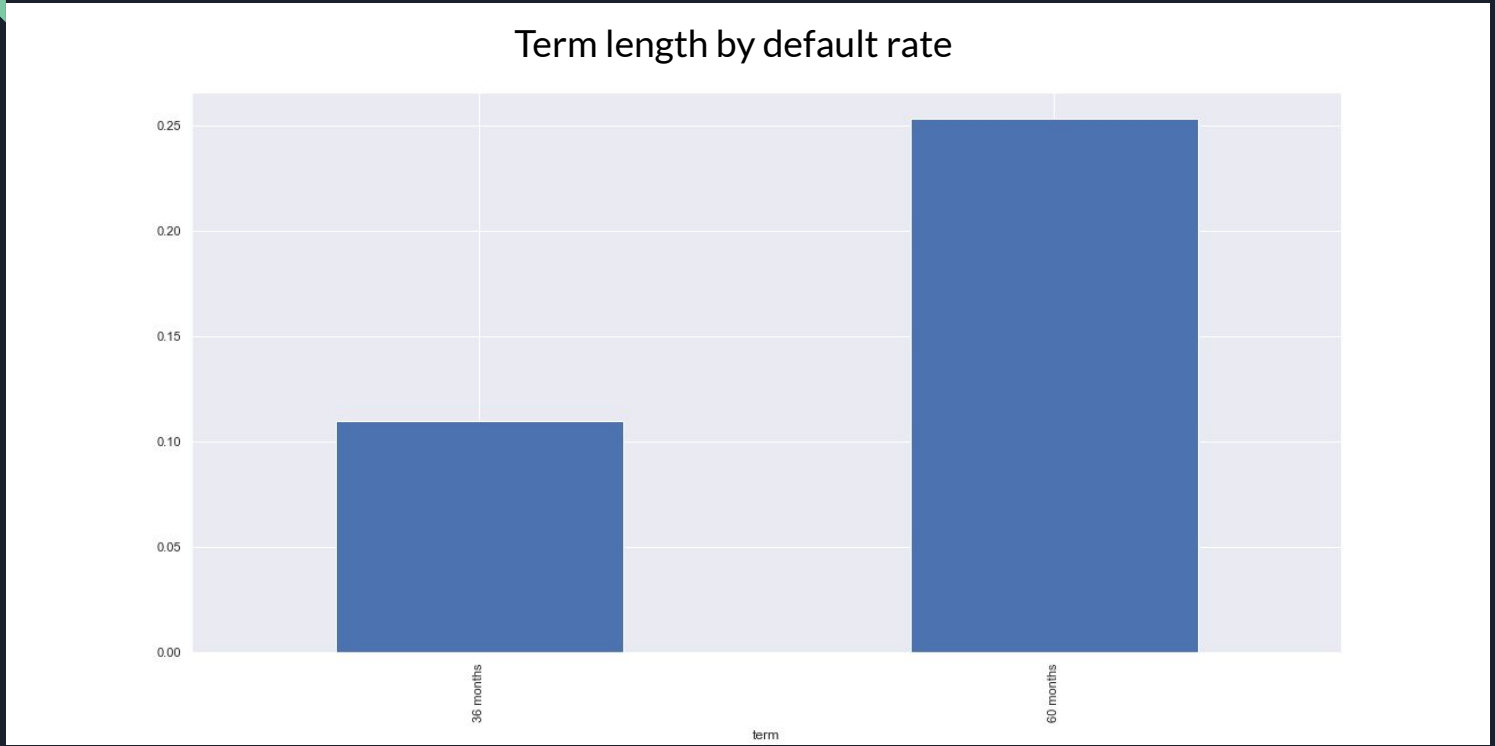
Bivariate Analysis



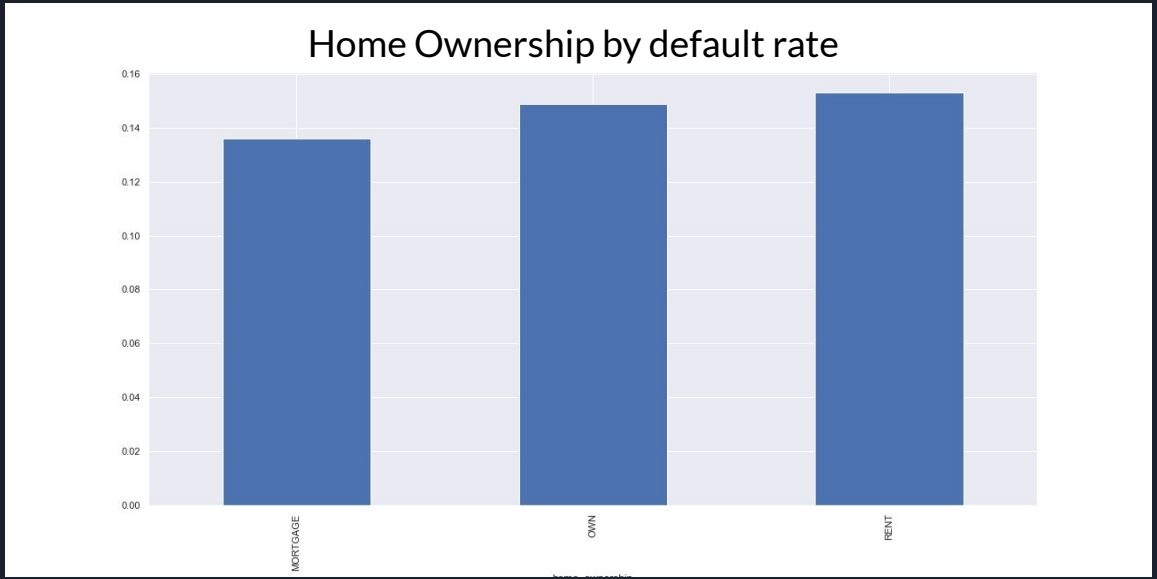
Bivariate Analysis

- Found that public record of bankruptcies actually had no correlation with defaults, in addition these people probably defaulted after loan approval so it was not available at the time of approval. Dropped column
- Found that zip code had too many unique variables to be useful in the dataframe, however there were a few zip_codes which stood out as the highest offenders for default. Namely: 935xx, 891xx, 890xx, 906xx, 641xx, 907xx, 302xx, 333xx
- The general trend of defaults versus time was that 2007 had the least defaults and 2011 had the most, and each year in-between increased gradually.
- Highest risk purpose for a loan was small-business, lowest risk for purpose was large purchase
- Lending to people who were unemployed was highest risk length of employment
- Lending to people who had lowest grade credit was by far the riskiest of all categories

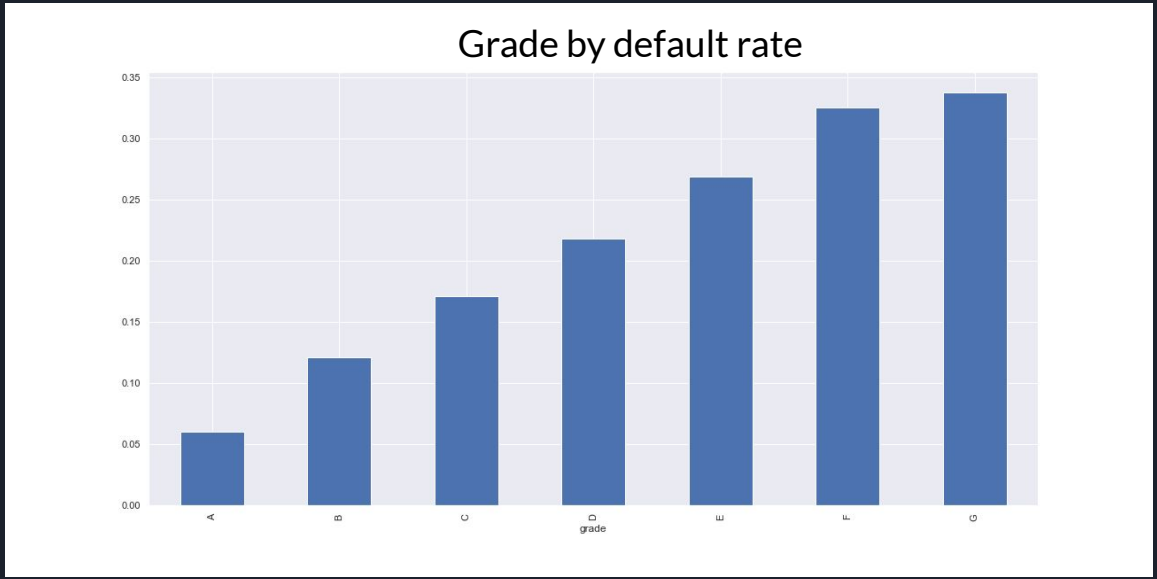
Bivariate Analysis



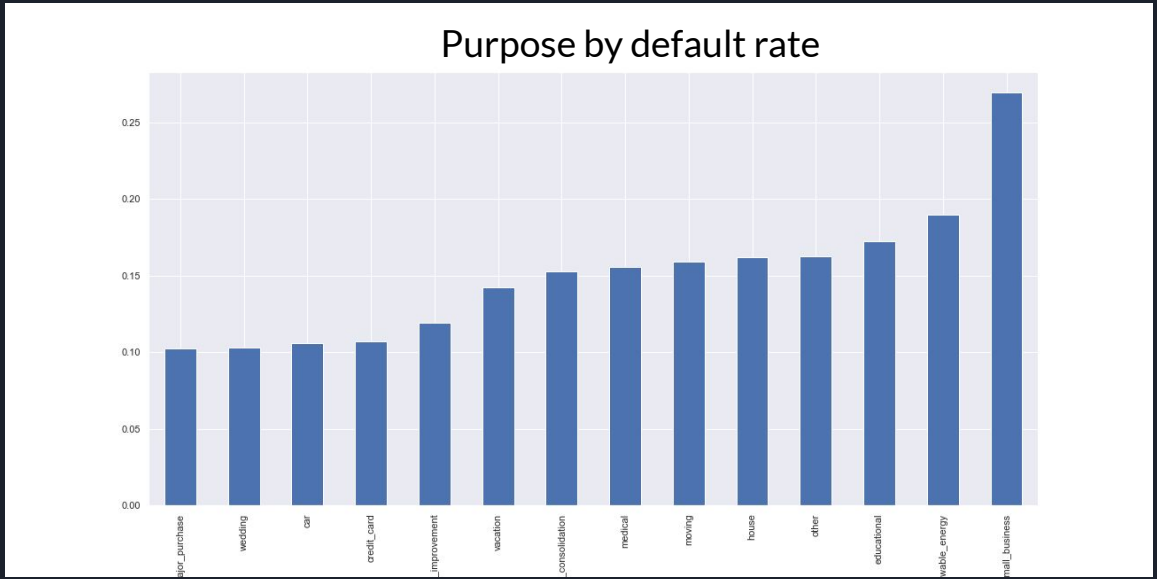
Bivariate Analysis



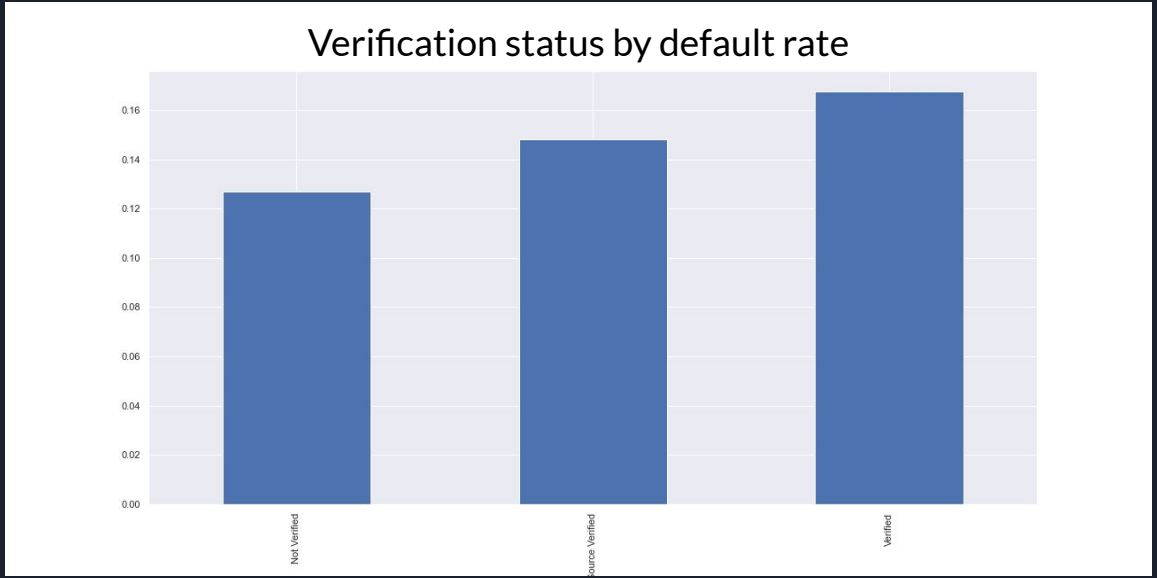
Bivariate Analysis



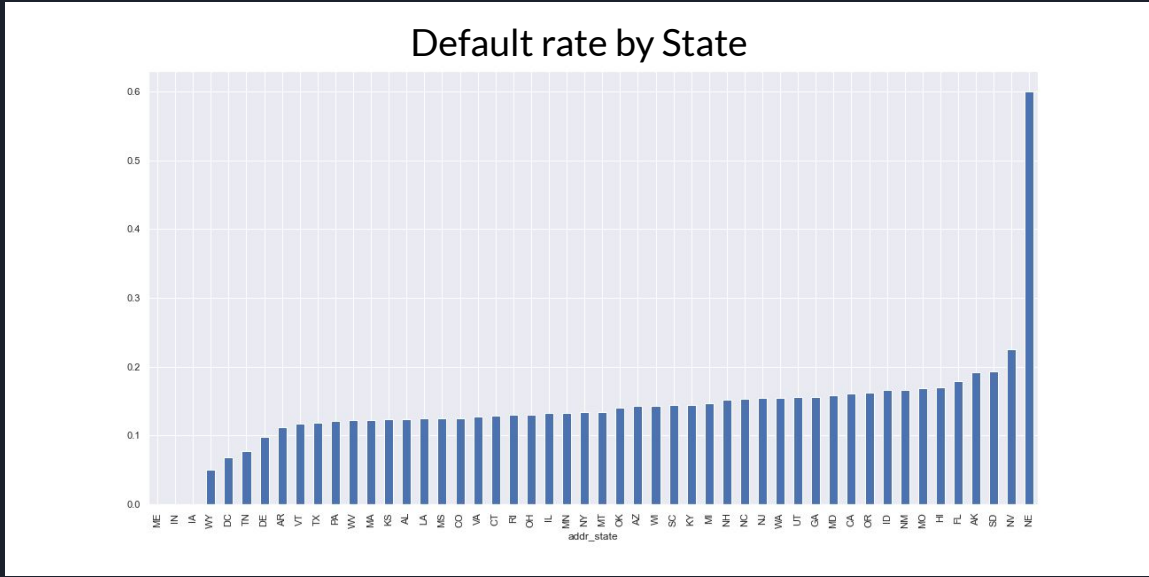
Bivariate Analysis



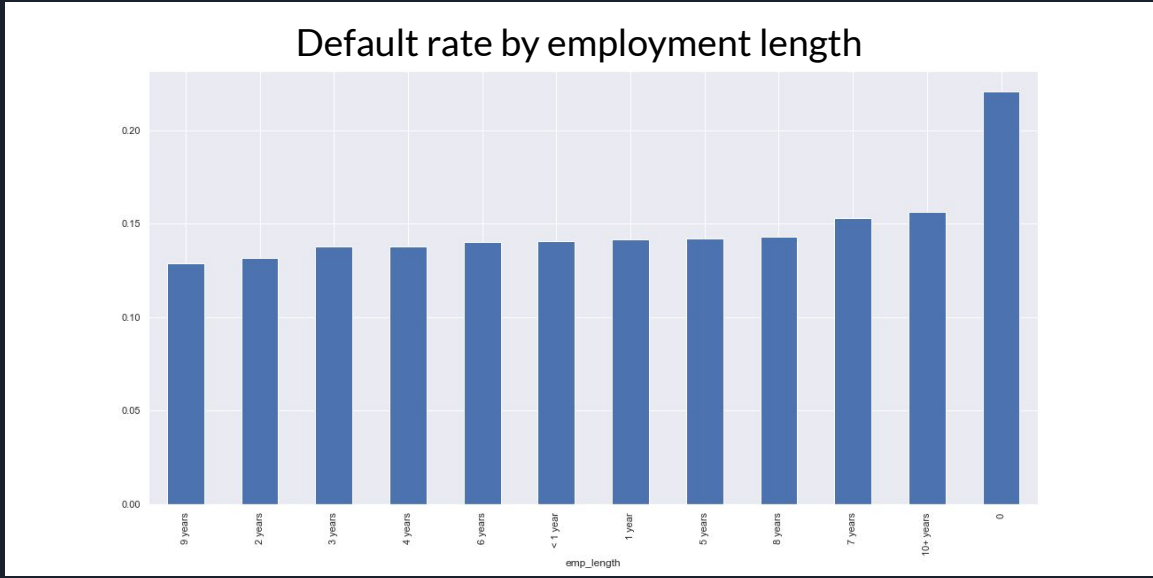
Bivariate Analysis



Bivariate Analysis



Bivariate Analysis



Multivariate Analysis

- New column was created for Credit Score
- Random weights were assigned to 11 significant variables 10,000 times to generate a credit score column which had 25.7% correlation with loan status

Thank you!