Lending Club Case Study

Team Members

Rahul Kumar

Ranganath P V V

Business Objective

- Informative decision to avoid ‘credit-loss’ to the company for loan applications
  - Analyze the driving factors behind loan default based on loan data available.
  - Analyze the ‘consumer attributes’ and ‘loan attributes’
  - Identify the category of the loan-applicant
  - Informed decision for granting loan based on above analysis

Process

Data Sourcing	Data is available in this case via CSV – normal process: source from company Loading CSV
Data Understanding	Understand data dictionary Shape of data Mean, Quantiles
Data Cleaning	Fix Rows and Columns, Remove Outliers Standardize values Filter data
Data Analysis - EDA	Univariate Analysis Bivariate Analysis
Observations/Conclusion	Identify category of defaulters Identify traits of defaulters

Data understanding

- Glance through the data to understand what is present in the data
- Understand the meaning of each of the columns to know which columns are of interest and need analysis on
- Understand if there are null values, duplicate values in any of the columns

```
##Check the Data Set
loanData.shape
```

```
(39717, 111)
```

```
#to understand the variation of data
loanData.describe()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	...
count	3.971700e+04	3.971700e+04	39717.000000	39717.000000	39717.000000	39717.000000	3.971700e+04	39717.000000	39717.000000	39717.000000	...
mean	6.831319e+05	8.504636e+05	11219.443815	10947.713196	10397.448868	324.561922	6.896893e+04	13.315130	0.146512	0.869200	...
std	2.106941e+05	2.656783e+05	7456.670694	7187.238670	7128.450439	208.874874	6.379377e+04	6.678594	0.491812	1.070219	...
min	5.473400e+04	7.069900e+04	500.000000	500.000000	0.000000	15.690000	4.000000e+03	0.000000	0.000000	0.000000	...
25%	5.162210e+05	6.667800e+05	5500.000000	5400.000000	5000.000000	167.020000	4.040400e+04	8.170000	0.000000	0.000000	...
50%	6.656650e+05	8.508120e+05	10000.000000	9600.000000	8975.000000	280.220000	5.900000e+04	13.400000	0.000000	1.000000	...
75%	8.377550e+05	1.047339e+06	15000.000000	15000.000000	14400.000000	430.780000	8.230000e+04	18.600000	0.000000	1.000000	...
max	1.077501e+06	1.314167e+06	35000.000000	35000.000000	35000.000000	1305.190000	6.000000e+06	29.990000	11.000000	8.000000	...

8 rows × 87 columns

Data Cleaning

- Remove/drop columns with missing values, 0 or Nan
- Remove empty rows
- Remove outliers
- Remove rows/columns not useful for analysis
- Standardize values: convert to correct date-time format

```
### Count of Missing Values in Each Column

missing_values = loanData.isnull().sum()
print(missing_values)
```

```
id                              0
member_id                       0
loan_amnt                       0
funded_amnt                     0
funded_amnt_inv                 0
                              ...
tax_liens                      39
tot_hi_cred_lim             39717
total_bal_ex_mort           39717
total_bc_limit              39717
total_il_high_credit_limit  39717
Length: 111, dtype: int64
```

```
##Filter the columns with unique values less than 5
categorical_columns = unique_values[unique_values < 5]
print(categorical_columns)
```

```
term                          2
verification_status           3
loan_status                   3
pymnt_plan                    1
initial_list_status           1
collections_12_mths_ex_med    1
policy_code                   1
application_type              1
acc_now_delinq                1
chargeoff_within_12_mths      1
delinq_amnt                   1
pub_rec_bankruptcies          3
tax_liens                     1
dtype: int64
```

```
# Drop the columns from the dataframe
loanData_cleaned = loanData.drop(columns=columns_to_drop)

# Display the cleaned dataframe
loanData_cleaned.shape
```

```
(39717, 54)
```

```
#will going to remove the rows with pub_rec_bankruptcies is null, it will help us in analysis
loanData_cleaned = loanData_cleaned[~loanData_cleaned.pub_rec_bankruptcies.isnull()]
loanData_cleaned.shape
```

```
(39020, 38)
```

```
#Will remove the columns with 0,nan values as they are not useful for analysis
loanData_cleaned = loanData_cleaned.drop(columns=['pymnt_plan', 'initial_list_status', 'collections_12_mths_ex_med',
                                                  'policy_code', 'application_type', 'acc_now_delinq', 'chargeoff_within_12_mths'

unique_values = loanData_cleaned.nunique()
print(unique_values)
```

```
#will going to remove the columns which are not useful for analysis
loanData_cleaned = loanData_cleaned.drop(columns=['id', 'member_id', 'url', 'zip_code', 'funded_amnt', 'funded_amnt_inv'])
loanData_cleaned.shape
```

```
(39717, 38)
```

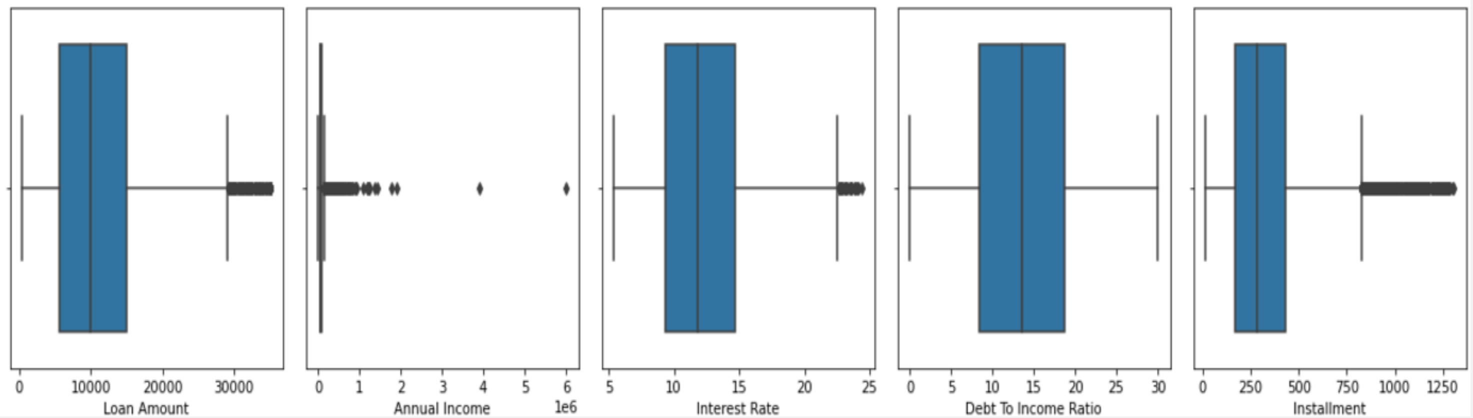
Data Analysis – Understand outliers and remove

- Create box plot for various columns and remove the outliers

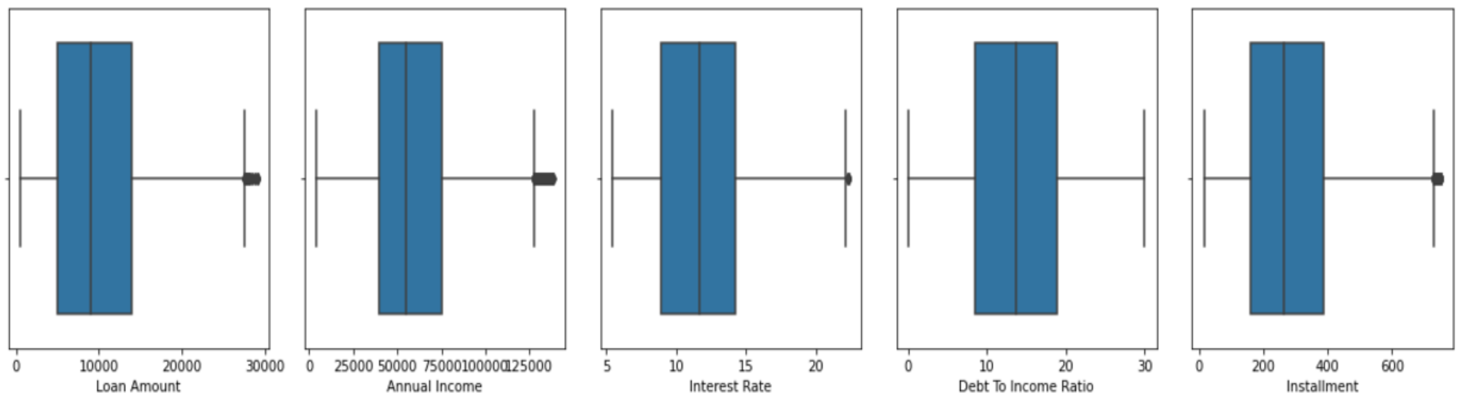
Identifying outliers

After removing outliers

```
#using createBoxPlot function to create the boxplot for multiple columns
createBoxPlot(final_loan, ['loan_amnt', 'annual_inc', 'int_rate', 'dti', 'installment'])
```



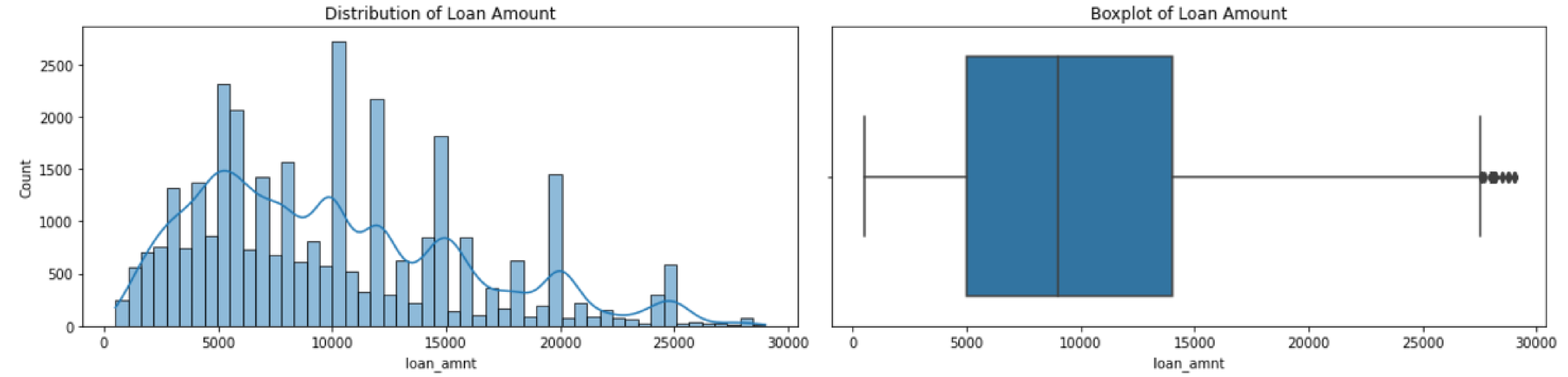

```
#using createBoxPlot function to create the boxplot for multiple columns
createBoxPlot(final_loan, ['loan_amnt', 'annual_inc', 'int_rate', 'dti', 'installment'])
```




```
# Remove outliers from the columns
# Define the columns to remove outliers
columns = ['loan_amnt', 'annual_inc', 'int_rate', 'dti', 'installment']
final_loan = removeOutliersBasedOnIqr(final_loan, columns, 1.5)
```

Data Analysis – Univariate analysis on loan amount

- Create distribution and box plot of loan amount

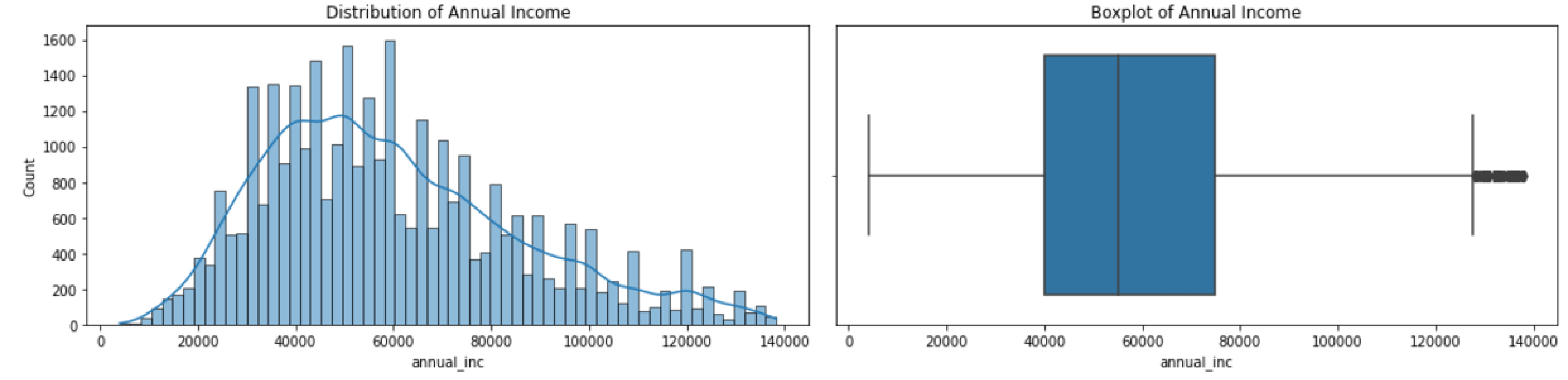


Observations/Analysis:

- Most of the loans are between 5000 to 15000
- Few loans which are above 35000 and less than 5000

Data Analysis – Univariate analysis on annual income

- Create distribution and box plot of Annual income

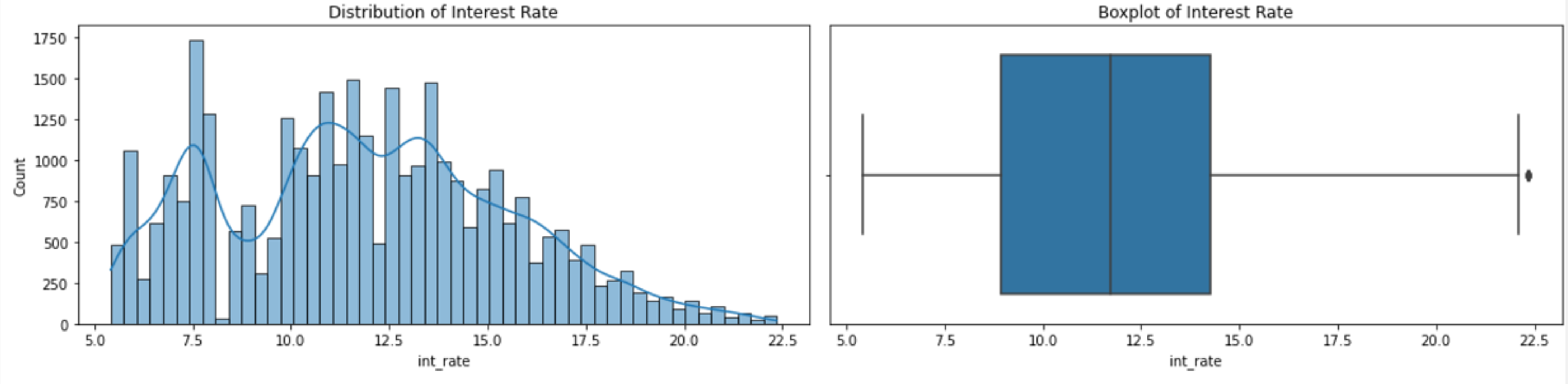


Observations/Analysis:

- Most of the people applying for loans have an annual income between 40,000 and 80,000
- Average annual income is around 60,000

Data Analysis – Univariate analysis on interest rate

- Create distribution and box plot of interest rate



Observations/Analysis:

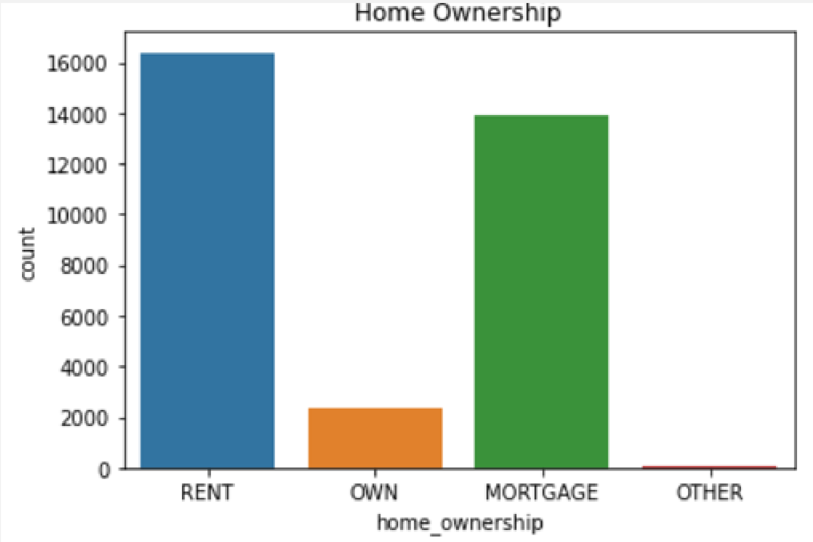
- Most of the loans have interest rate between 8 to 13%
- Few have interest rates above 20%
- Average interest rates is around 12%

Categorical Analysis – Home ownership

- Plot bar plot for home ownership

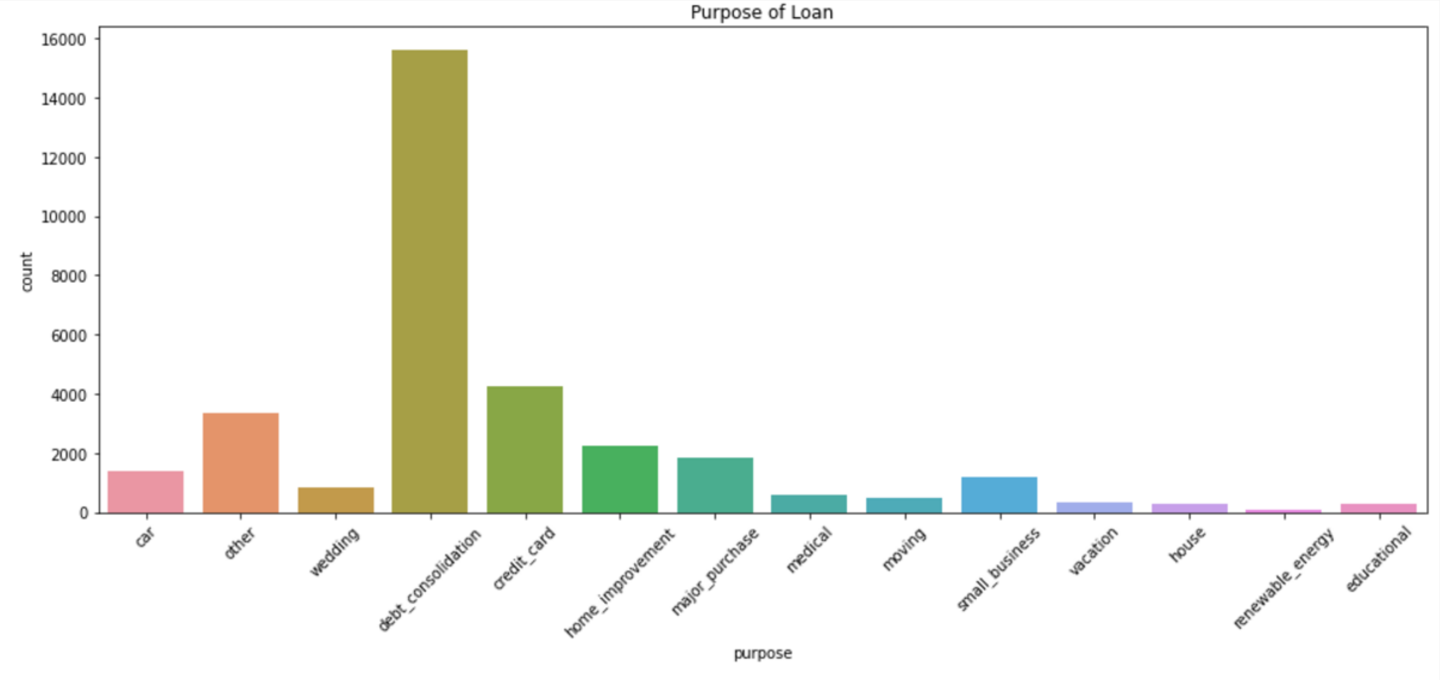
Observations/Analysis:

- Most of the loans are being taken by people who have rented/mortgaged homes



Categorical Analysis – Purpose of loan

- Plot bar plot for purpose of loan

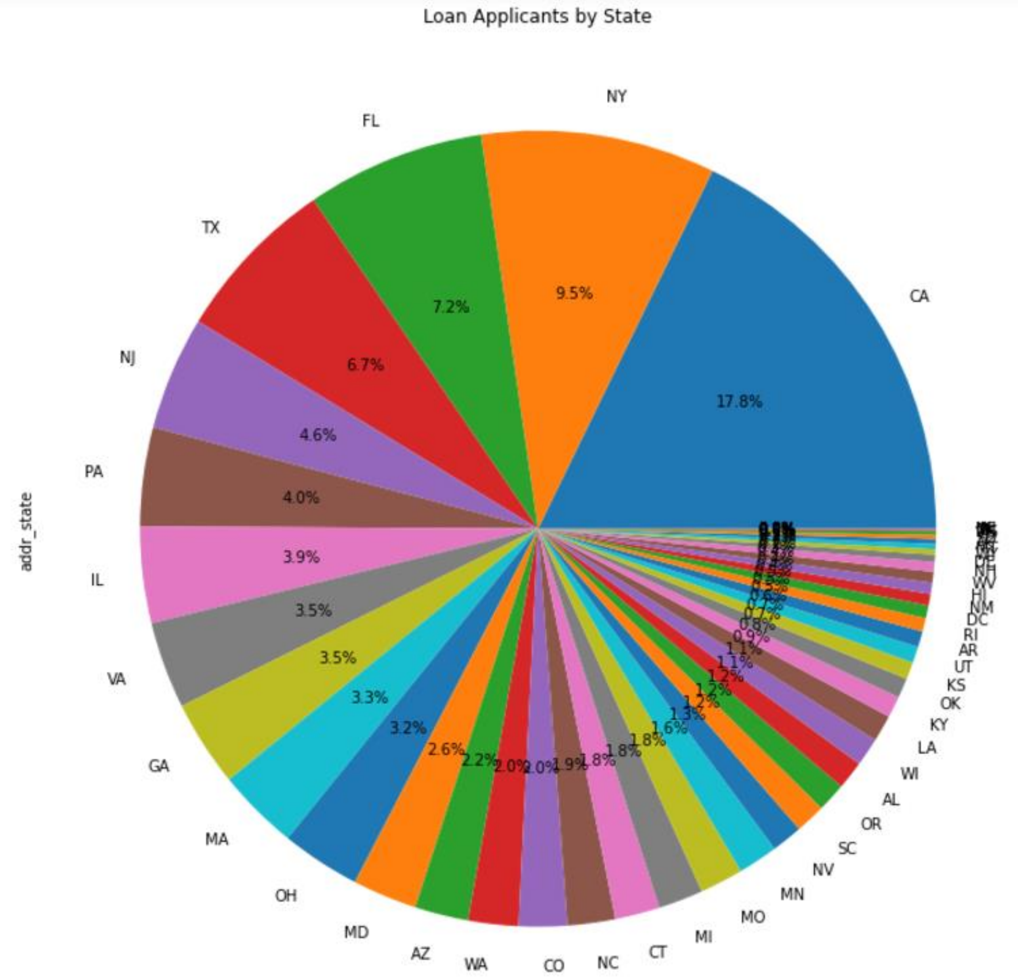


Observations/Analysis:

- Most of the loans are being taken for ‘debt consolidation’, followed by ‘credit card’ payment and ‘others’
- Vast difference between the purpose of ‘debt consolidation’ and ‘credit card’
- ‘Wedding’ is not one of the higher purpose for taking loan

Categorical Analysis – Loan applicants by state

- Pie chart of loan applicants by state



Observations/Analysis:

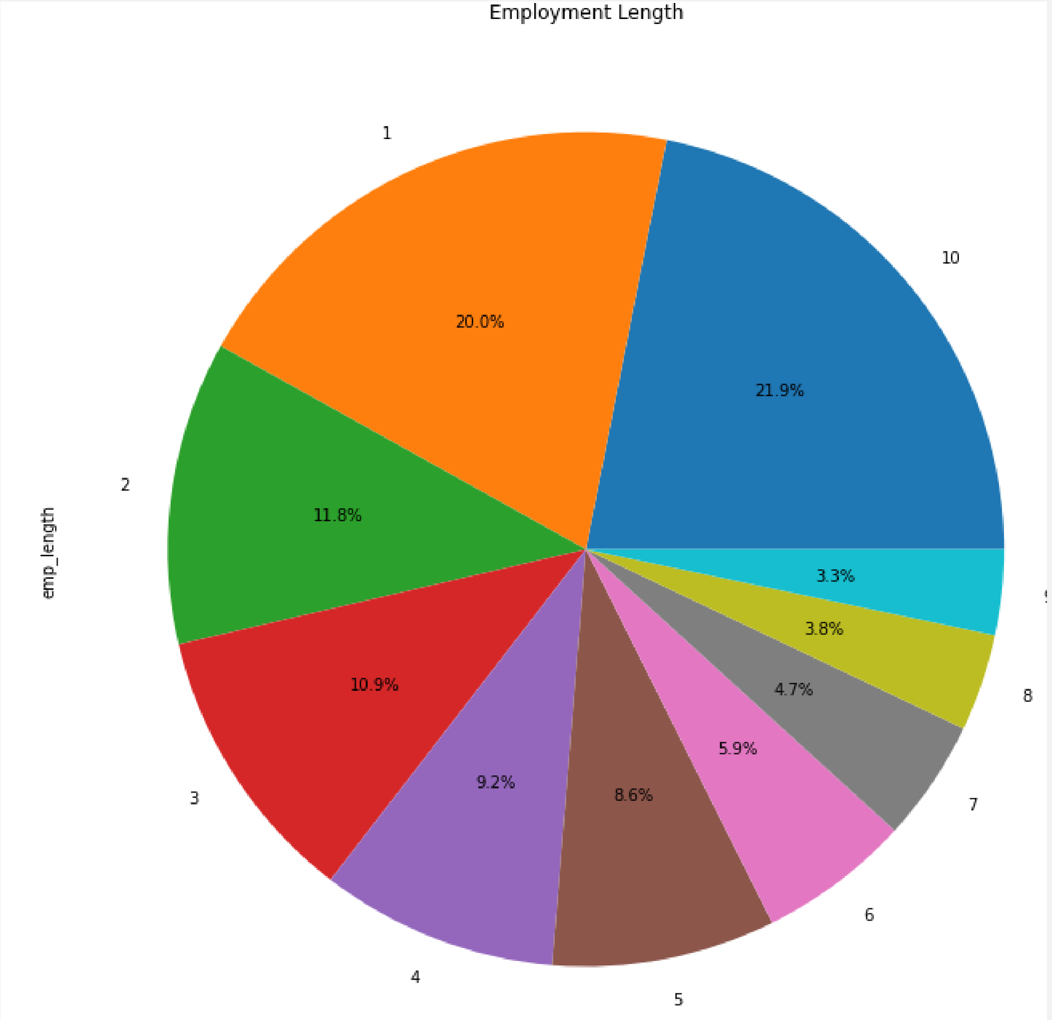
- Most of the loans are being taken from states of CA, NY, FL, TX

Categorical Analysis – Employment length

- Pie chart by employment length

Observations/Analysis:

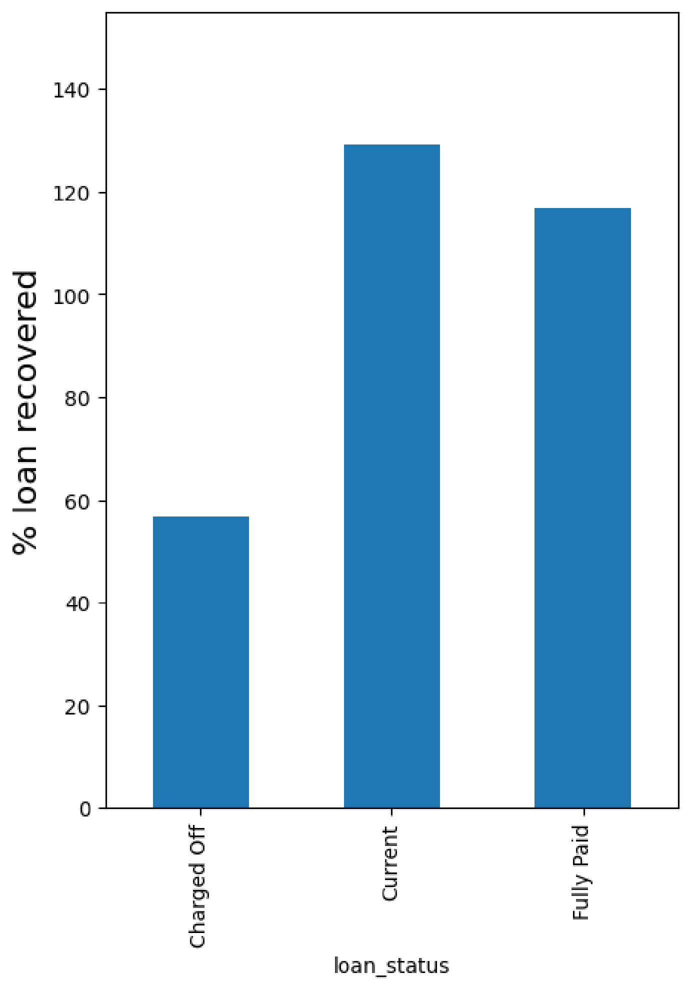
- Employees with about 1 yr of experience and10yrs of experience are more likely to take loan



Bivariate Analysis – loan status vs %loan recovered

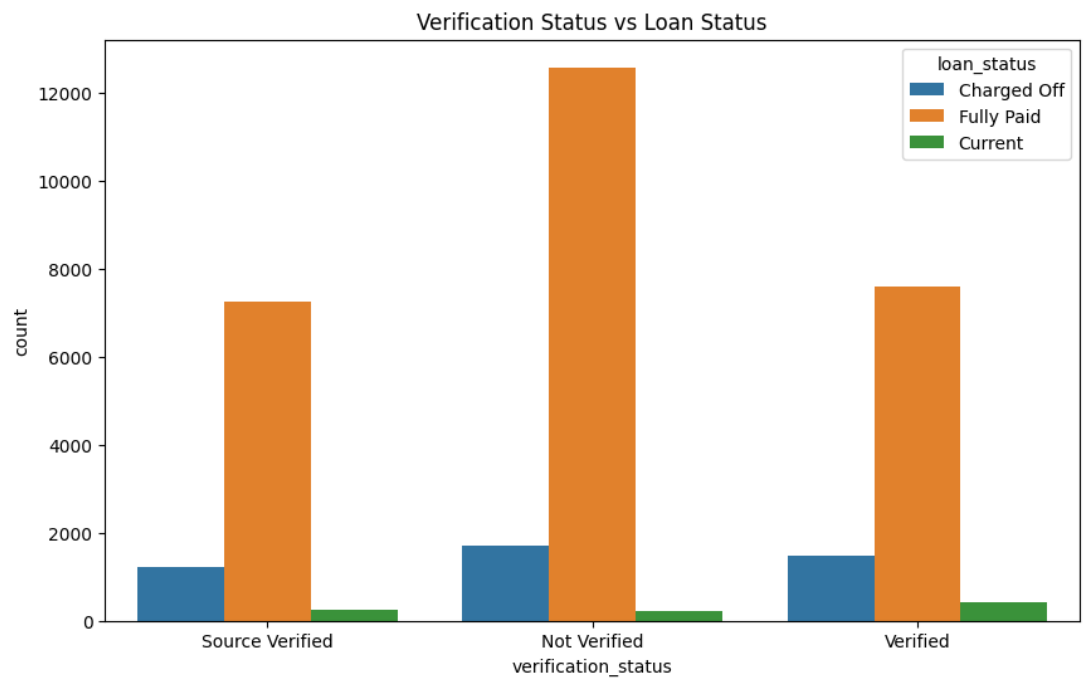
- Bar plot of loan status vs %loan recovered

Observations/Analysis:
- There is profit in case the loan is fully paid, but there is loss in case of charged off and default



Bivariate Analysis – Verification Status and Loan Status

- Bar plot of verification status vs loan status

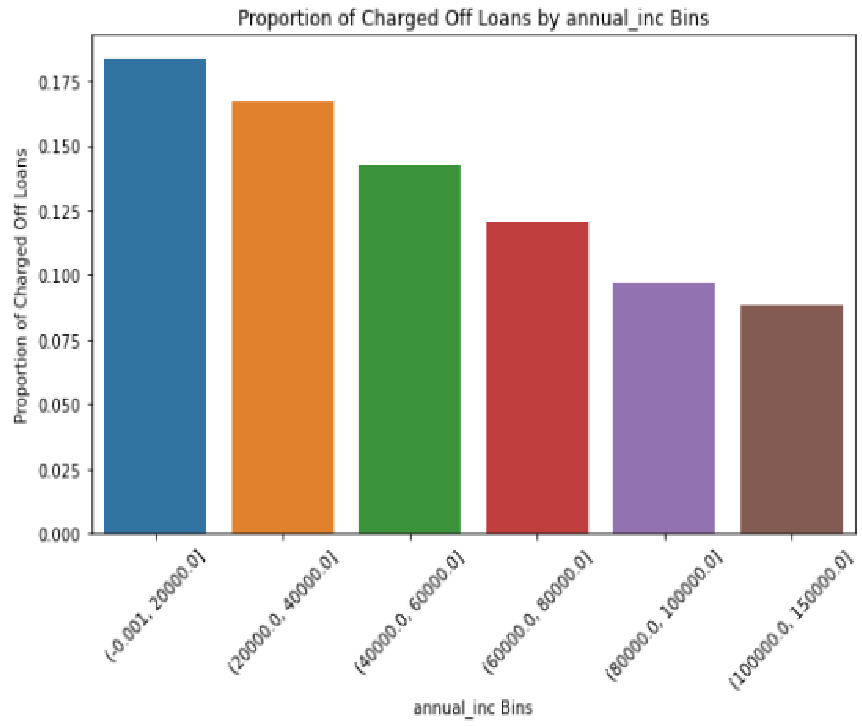


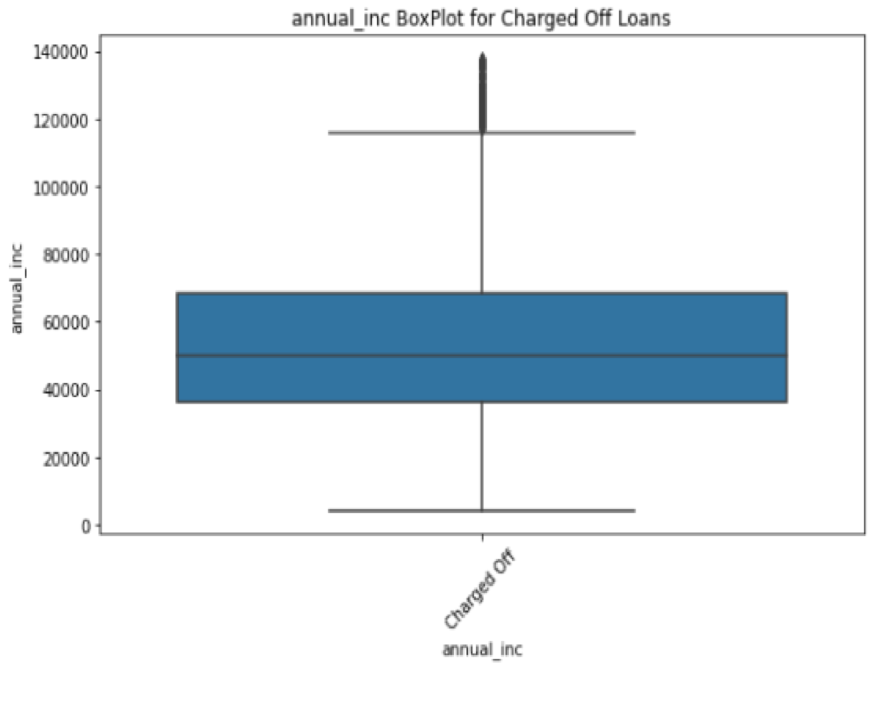
Observations/Analysis:

- There are loans which are verified and fully paid, but there are loans which are verified and charged off
- Not verified loans are the major contributor to charged-off and default

Bivariate Analysis – Annual income vs Charged-off

- Bar plot of annual income vs charged-off



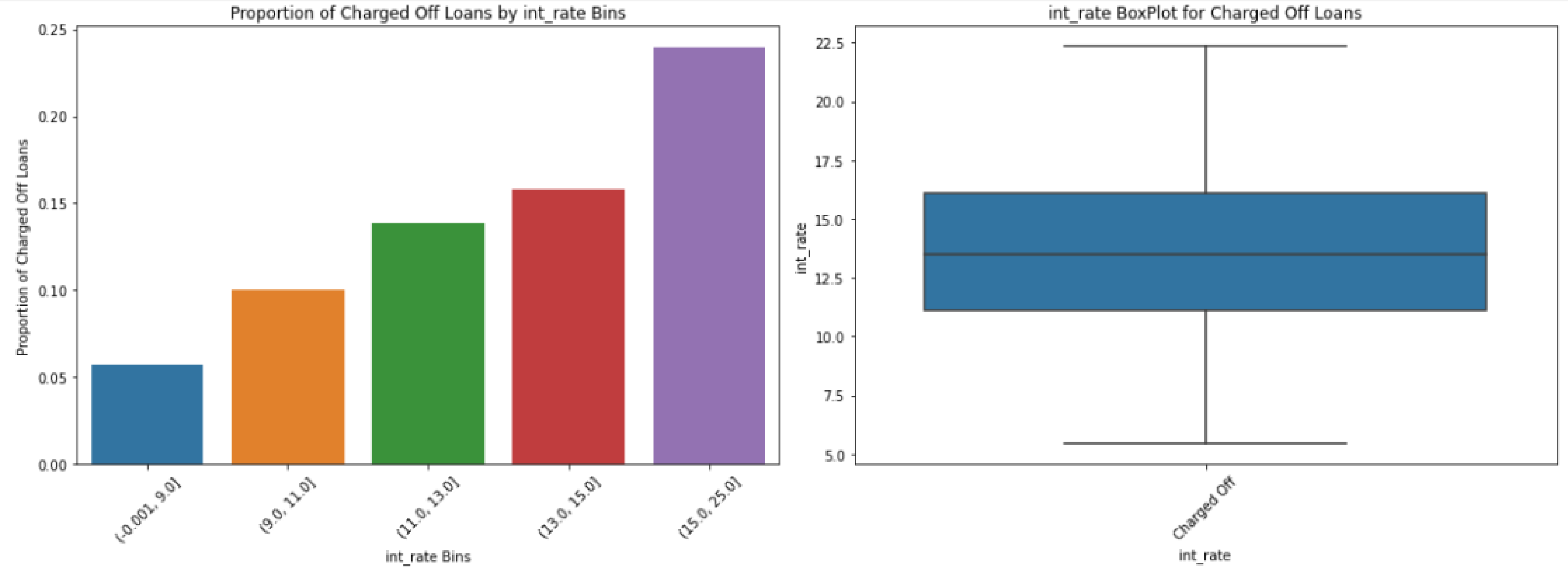


Observations/Analysis:

- People with lesser annual income (0-20,000) are more likely to default
- People with higher annual income (1,00,000-1,50,000) are less likely to default

Bivariate Analysis – Interest rate vs Charged-off

- Bar plot of Interest rate vs charged-off

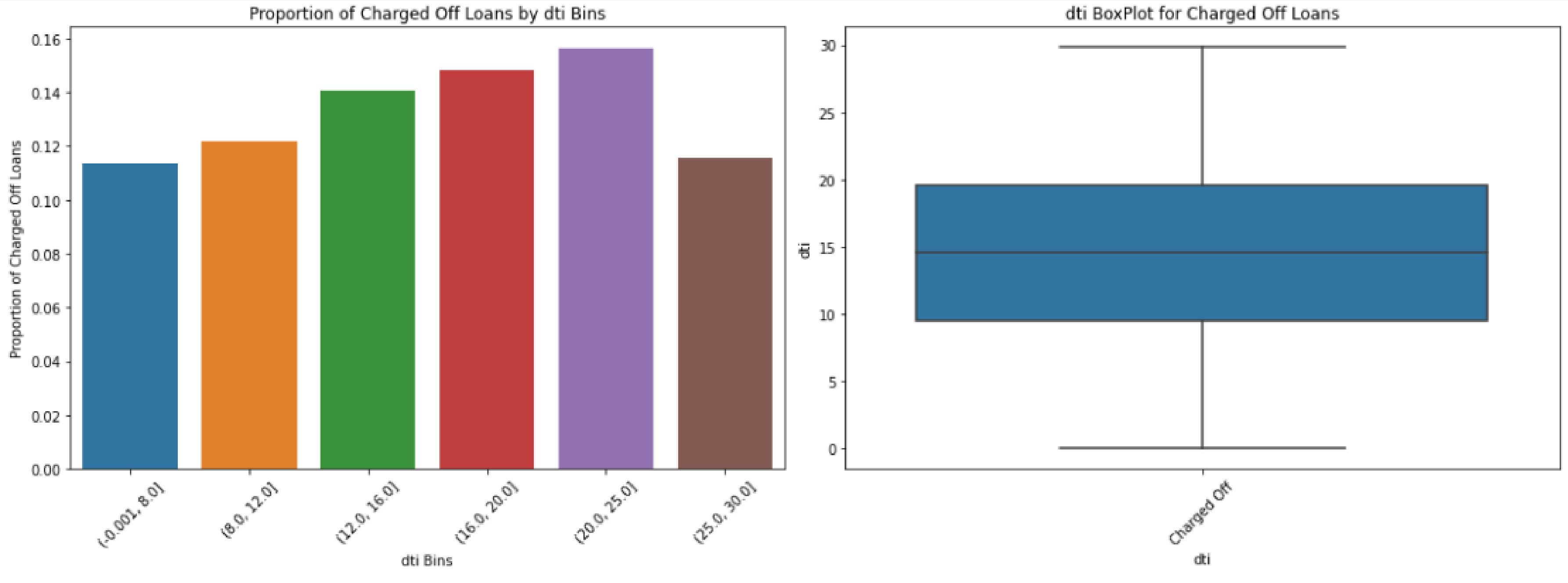


Observations/Analysis:

- People with interest rate between 15 to 25 are more likely to default, as the interest rate is high in this range
- People with interest rate between 0 to 9 are less likely to default, as the interest rate is low in this range

Bivariate Analysis – DTI vs Charged-off

- Bar plot of DTI vs charged-off

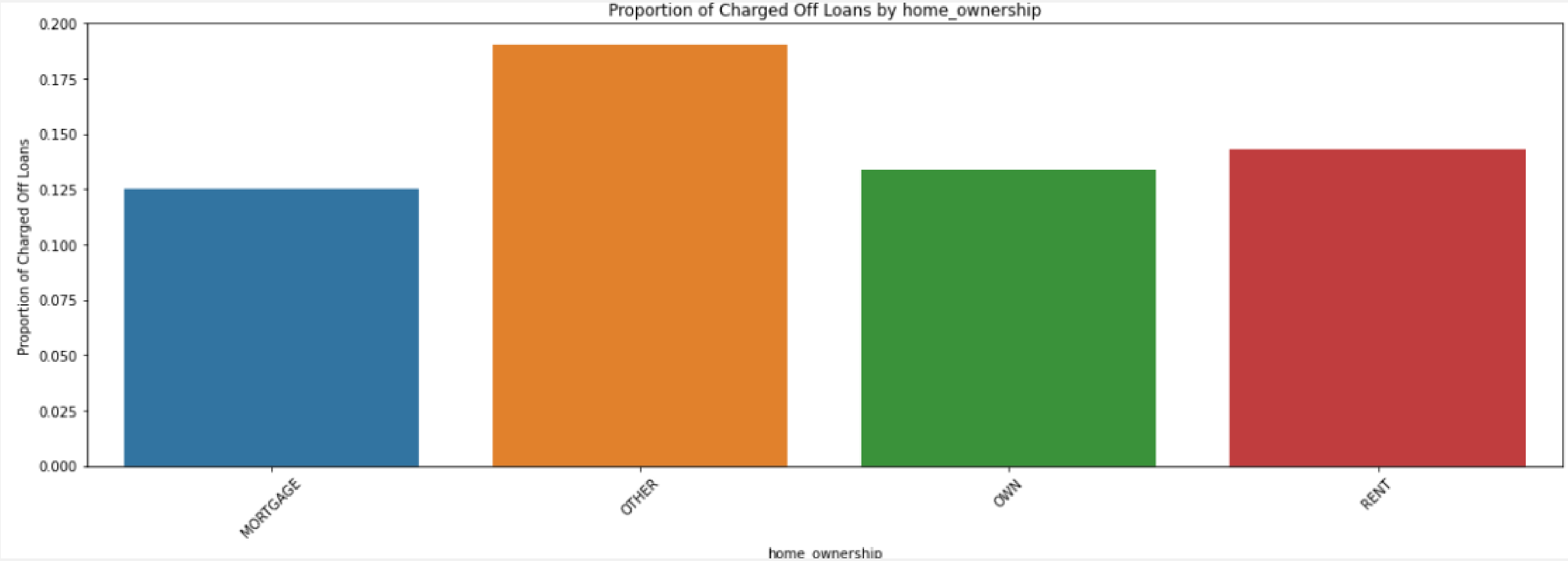


Observations/Analysis:

- People with DTI between 20 to 25 are more likely to default
- People with DTI between 0 to 8 are less likely to default, as the DTI is low in this range

Bivariate Analysis – Home ownership vs Charged-off

- Bar plot of home ownership vs charged-off

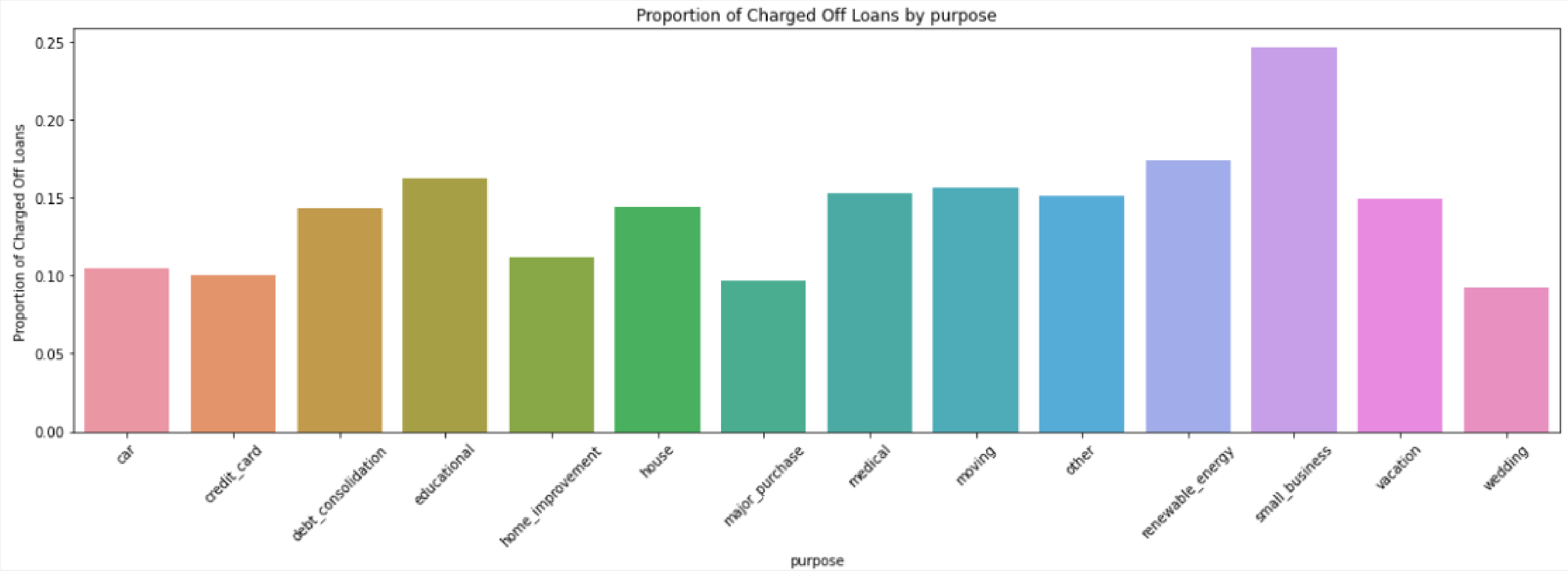


Observations/Analysis:

- People with home ownership as ‘other’ are more likely to default
- Not a major difference in defaulters in terms of categories. All the categories have significant defaulters.
- People with ‘Rent’ are more likely to default compared to ‘Mortgage’ and ‘Own’ house

Bivariate Analysis – Purpose vs Charged-off

- Bar plot of purpose of loan vs charged-off

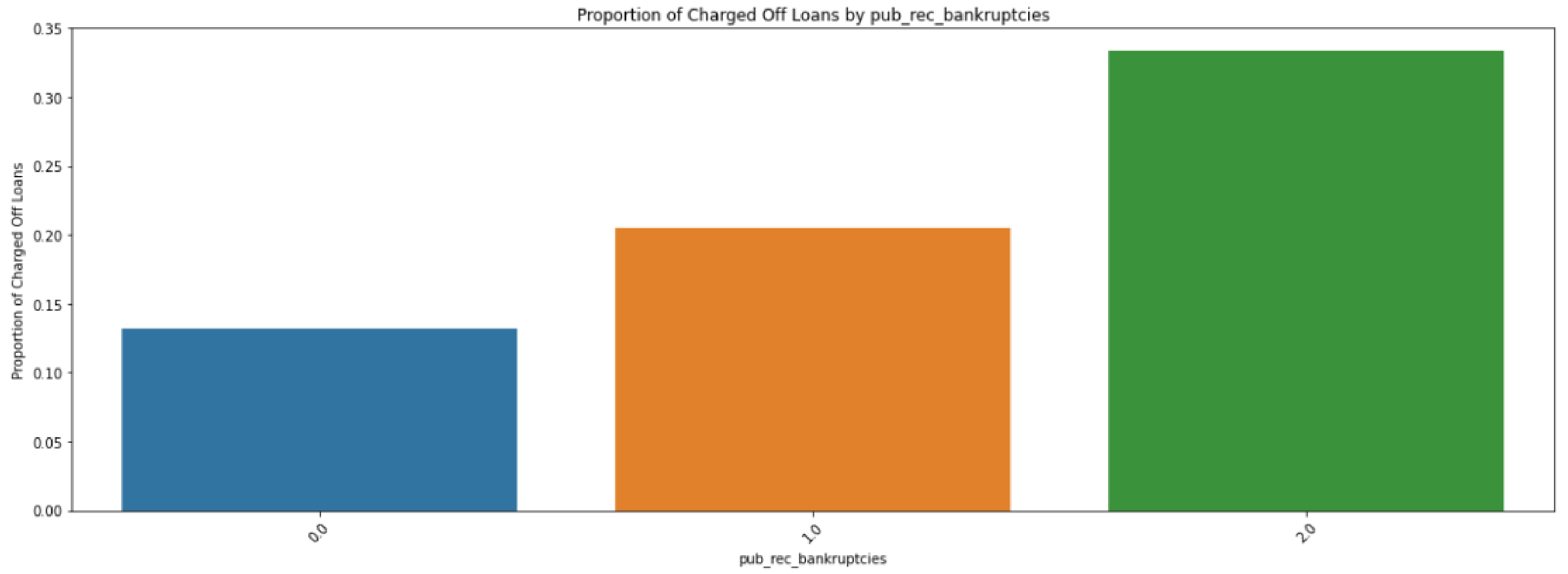


Observations/Analysis:

- People with small business are more likely to default
- People with wedding are less likely to default, and are more likely to pay the loan

Bivariate Analysis – Public record bankruptcies vs Charged-off

- Bar plot of 'Public record bankruptcies' vs charged-off

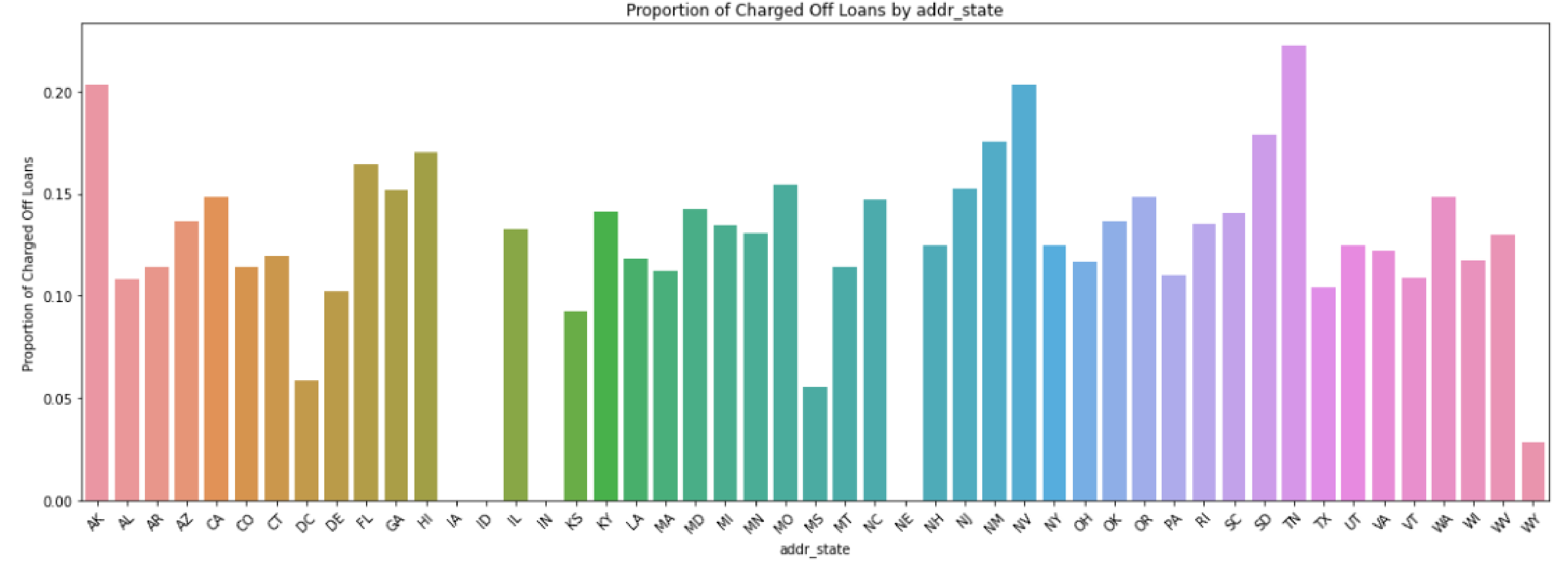


Observations/Analysis:

- People with 2 public record bankruptcies are more likely to default
- Lower the public record bankruptcies, less likely to default

Bivariate Analysis – Public record bankruptcies vs Charged-off

- Bar plot of loans by state vs charged-off

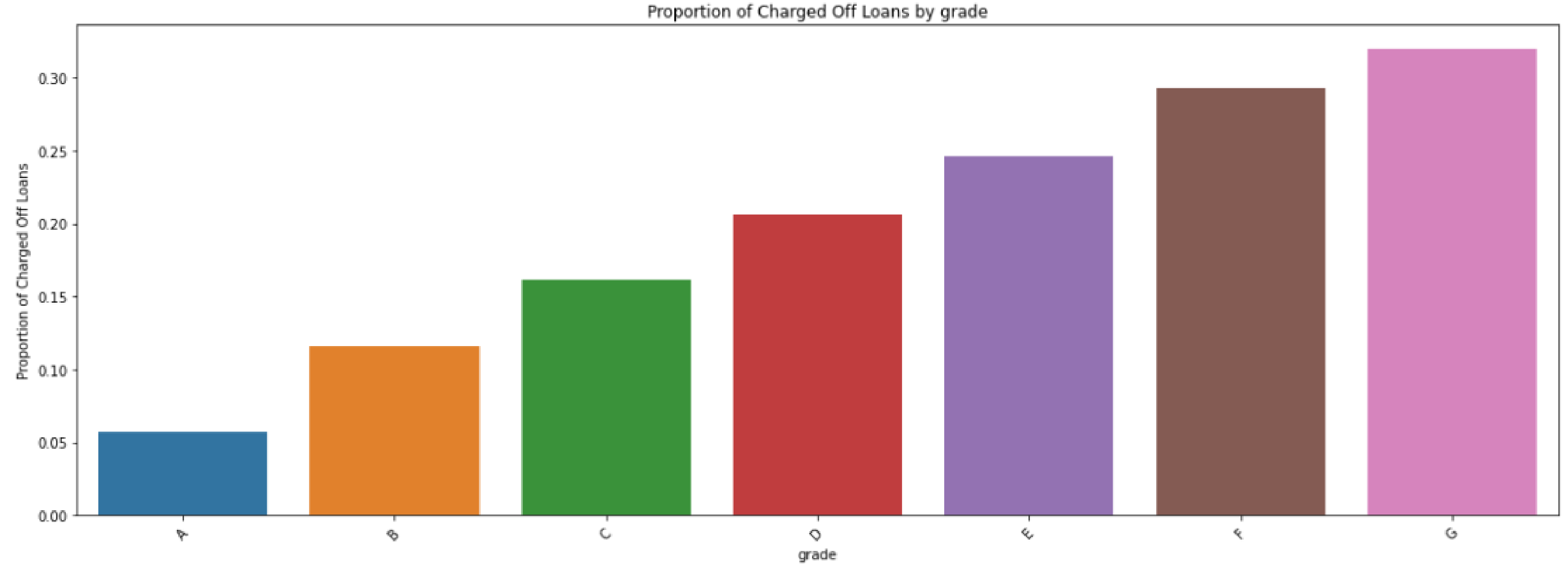


Observations/Analysis:

- People from TN, NV, AK are more likely to default
- People from WY are less likely to default

Bivariate Analysis – Grade vs Charged-off

- Bar plot of Grade vs charged-off

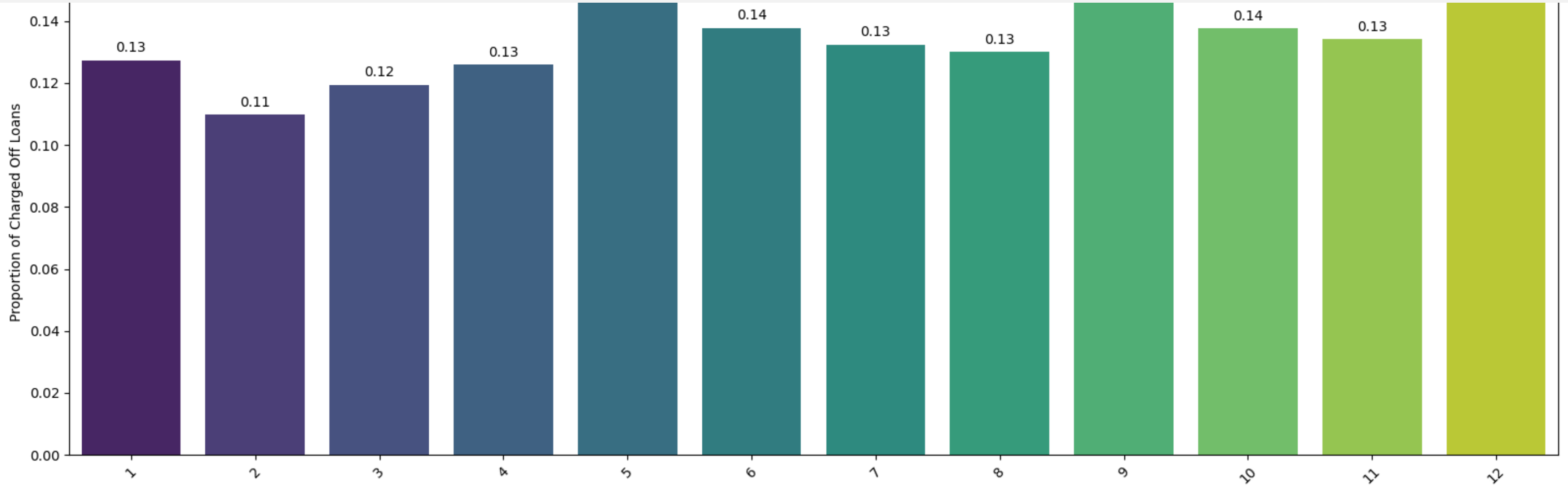


Observations/Analysis:

- People from Grade G are more likely to default
- People from Grade A are less likely to default

Bivariate Analysis – Issue Month Vs Charged Off Loan

- Bar plot of Issue Months vs Charged-off loan

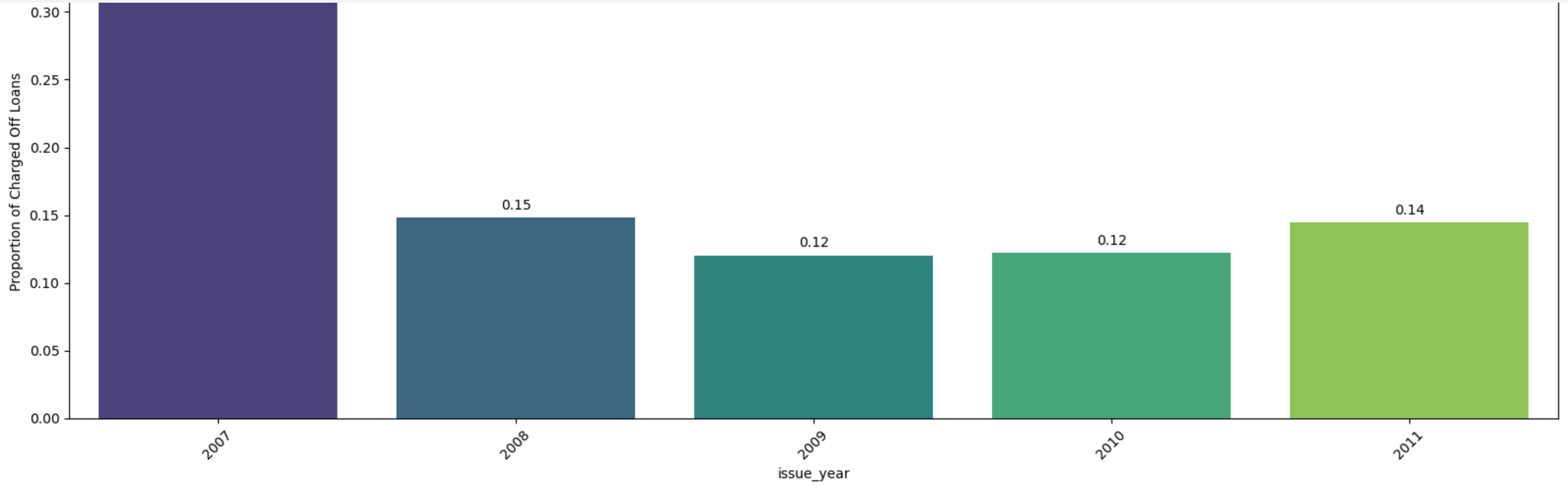


Observations/Analysis:

- People who have taken loan in month 12 are more likely to default

Bivariate Analysis – Issue Year Vs Charged Off Loan

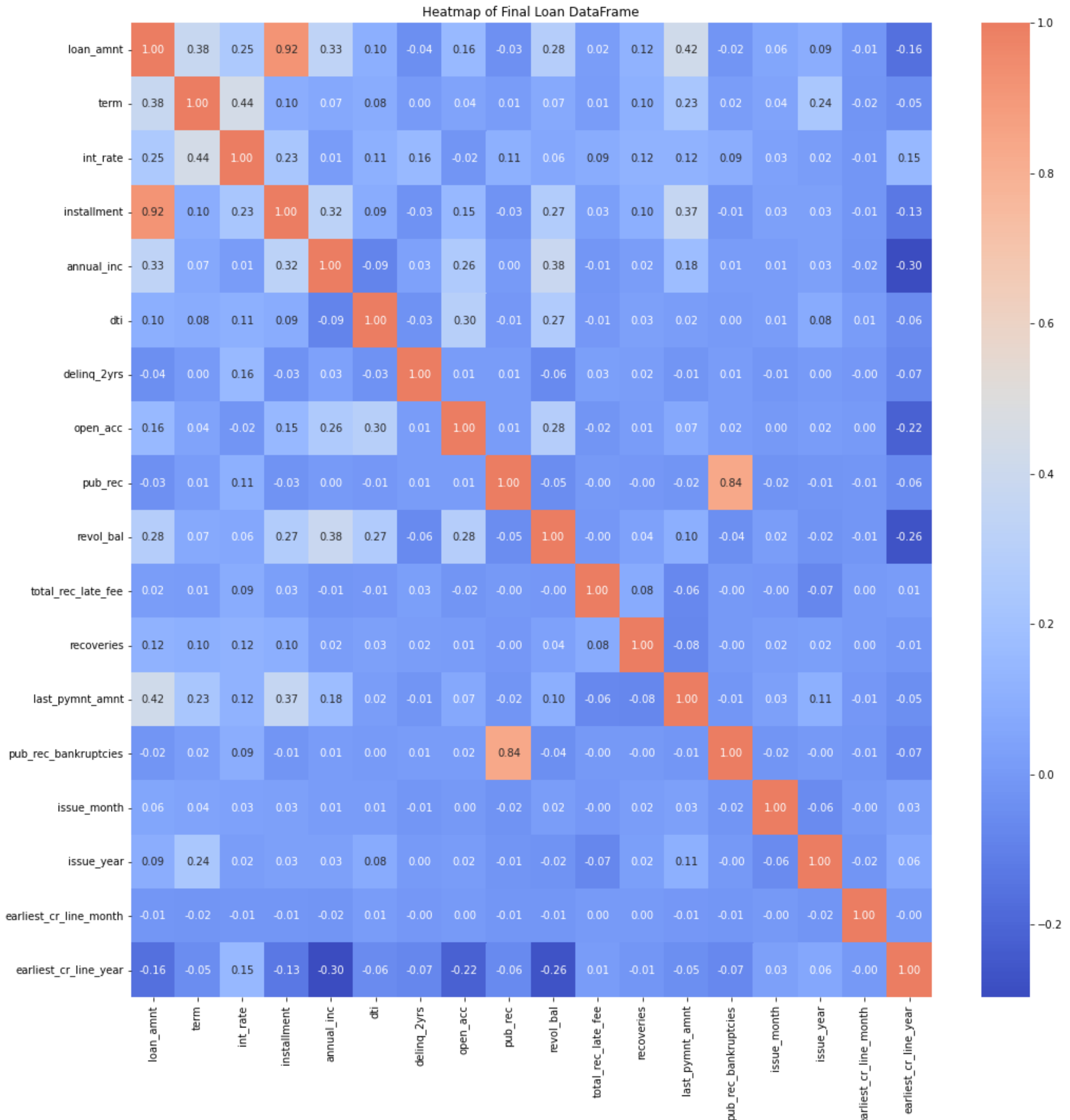
- Bar plot of Issue Year vs Charged-off loan



Observations/Analysis:

- Year 2007 has highest proportion of charged off loans

Heat map


Heatmap of Final Loan DataFrame
loan_amnt 1.00 0.38 0.25 0.92 0.33 0.10 -0.04 0.16 -0.03 0.28 0.02 0.12 0.42 -0.02 0.06 0.09 -0.01 -0.16
term 0.38 1.00 0.44 0.10 0.07 0.08 0.00 0.04 0.01 0.07 0.01 0.10 0.23 0.02 0.04 0.24 -0.02 -0.05
int_rate 0.25 0.44 1.00 0.23 0.01 0.11 0.16 -0.02 0.11 0.06 0.09 0.12 0.12 0.09 0.03 0.02 -0.01 0.15
installment 0.92 0.10 0.23 1.00 0.32 0.09 -0.03 0.15 -0.03 0.27 0.03 0.10 0.37 -0.01 0.03 0.03 -0.01 -0.13
annual_inc 0.33 0.07 0.01 0.32 1.00 -0.09 0.03 0.26 0.00 0.38 -0.01 0.02 0.18 0.01 0.01 0.03 -0.02 -0.30
dti 0.10 0.08 0.11 0.09 -0.09 1.00 -0.03 0.30 -0.01 0.27 -0.01 0.03 0.02 0.00 0.01 0.08 0.01 -0.06
delinq_2yrs -0.04 0.00 0.16 -0.03 0.03 -0.03 1.00 0.01 0.01 -0.06 0.03 0.02 -0.01 0.01 -0.01 0.00 -0.00 -0.07
open_acc 0.16 0.04 -0.02 0.15 0.26 0.30 0.01 1.00 0.01 0.28 -0.02 0.01 0.07 0.02 0.00 0.02 0.00 -0.22
pub_rec -0.03 0.01 0.11 -0.03 0.00 -0.01 0.01 0.01 1.00 -0.05 -0.00 -0.00 -0.02 0.84 -0.02 -0.01 -0.01 -0.06
revol_bal 0.28 0.07 0.06 0.27 0.38 0.27 -0.06 0.28 -0.05 1.00 -0.00 0.04 0.10 -0.04 0.02 -0.02 -0.01 -0.26
total_rec_late_fee 0.02 0.01 0.09 0.03 -0.01 -0.01 0.03 -0.02 -0.00 -0.00 1.00 0.08 -0.06 -0.00 -0.00 -0.07 0.00 0.01
recoveries 0.12 0.10 0.12 0.10 0.02 0.03 0.02 0.01 -0.00 0.04 0.08 1.00 -0.08 -0.00 0.02 0.02 0.00 -0.01
last_pymnt_amnt 0.42 0.23 0.12 0.37 0.18 0.02 -0.01 0.07 -0.02 0.10 -0.06 -0.08 1.00 -0.01 0.03 0.11 -0.01 -0.05
pub_rec_bankruptcies -0.02 0.02 0.09 -0.01 0.01 0.00 0.01 0.02 0.84 -0.04 -0.00 -0.00 -0.01 1.00 -0.02 -0.00 -0.01 -0.07
issue_month 0.06 0.04 0.03 0.03 0.01 0.01 -0.01 0.00 -0.02 0.02 -0.00 0.02 0.03 -0.02 1.00 -0.06 -0.00 0.03
issue_year 0.09 0.24 0.02 0.03 0.03 0.08 0.00 0.02 -0.01 -0.02 -0.07 0.02 0.11 -0.00 -0.06 1.00 -0.02 0.06
earliest_cr_line_month -0.01 -0.02 -0.01 -0.01 -0.02 0.01 -0.00 0.00 -0.01 -0.01 0.00 0.00 -0.01 -0.01 -0.00 -0.02 1.00 -0.00
earliest_cr_line_year -0.16 -0.05 0.15 -0.13 -0.30 -0.06 -0.07 -0.22 -0.06 -0.26 0.01 -0.01 -0.05 -0.07 0.03 0.06 -0.00 1.00
loan_amnt
term
int_rate
installment
annual_inc
dti
delinq_2yrs
open_acc
pub_rec
revol_bal
total_rec_late_fee
recoveries
last_pymnt_amnt
pub_rec_bankruptcies
issue_month
issue_year
earliest_cr_line_month
earliest_cr_line_year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Conclusion

- Loans having higher interest rates are more likely to default
- Loans provided to lower income group are more likely to default
- Loans provided for debt consolidation are more likely to default
- Loans provided to applications from TN are more likely to default
- Informative decision should be taken based on the income vs purpose of loan, interest rate and state of residency for granting of the loan
- Background check on the nature of employment and other past debts/loans taken by the person and repayment history may also serve as an additional input if the data is available