# Lending Club Case Study

Suresh Kumar Yatirajula

Sundar Shankar

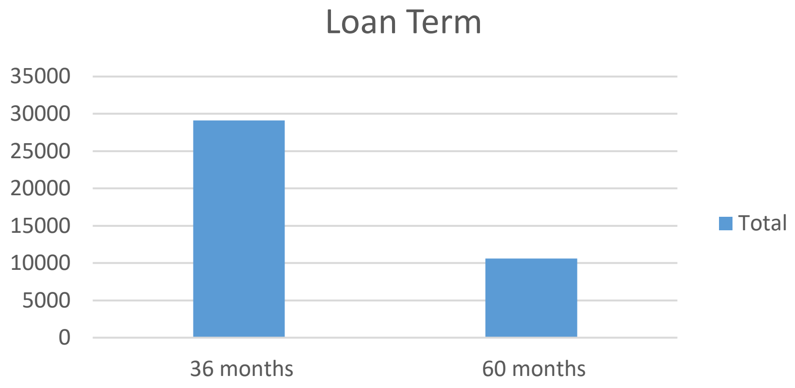
Srinivasa Marappa

Prateek Pande

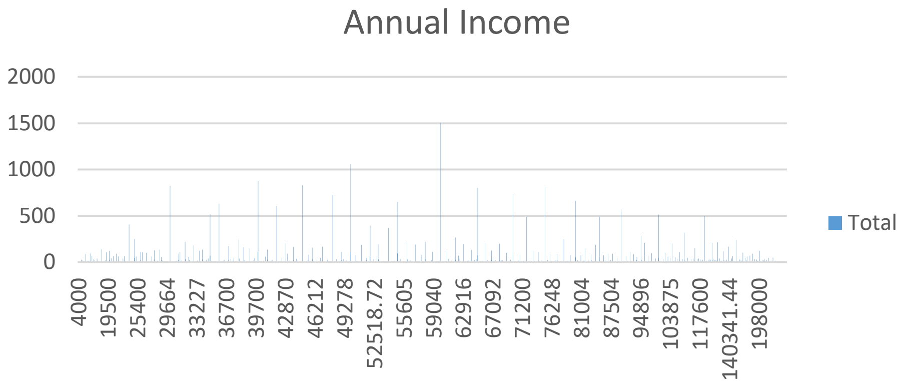
# Data Understanding

# Some Interesting Observations from Data

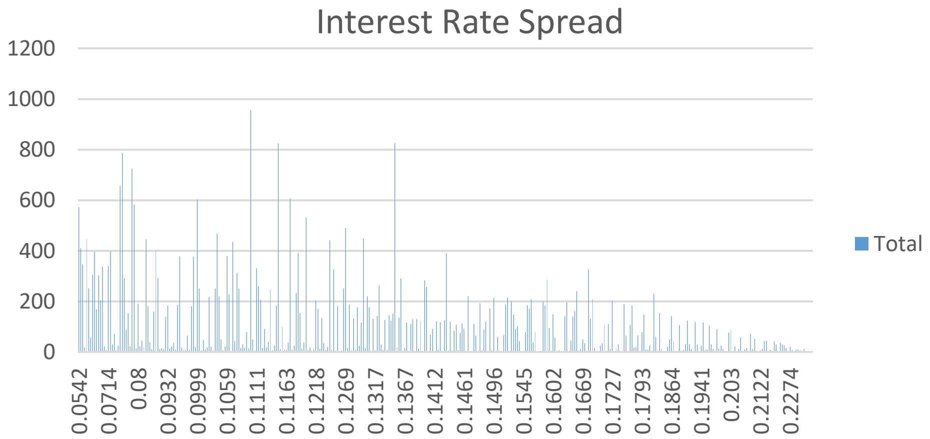
- “Loan Term” is either 36 months or 60 months



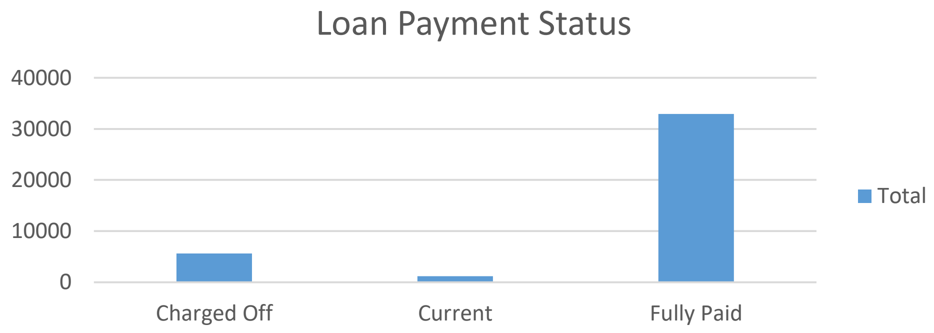
- Income of borrowers range from 4K to 6 million, with a mean salary around 68K



- Minimum interest is 5.42% and max interest rate is 24.59%

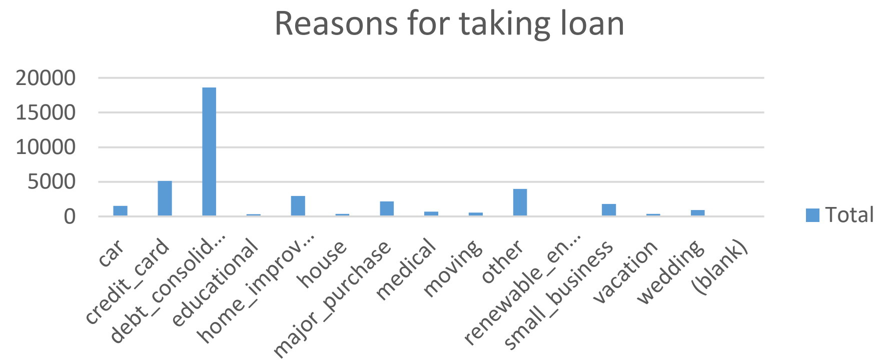


- Loan Delinquent Borrowers Percentage - ~ 14%

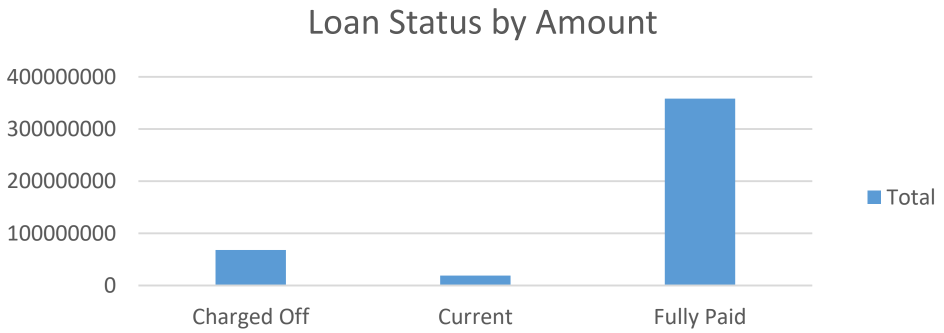


# Some Interesting Observations from data

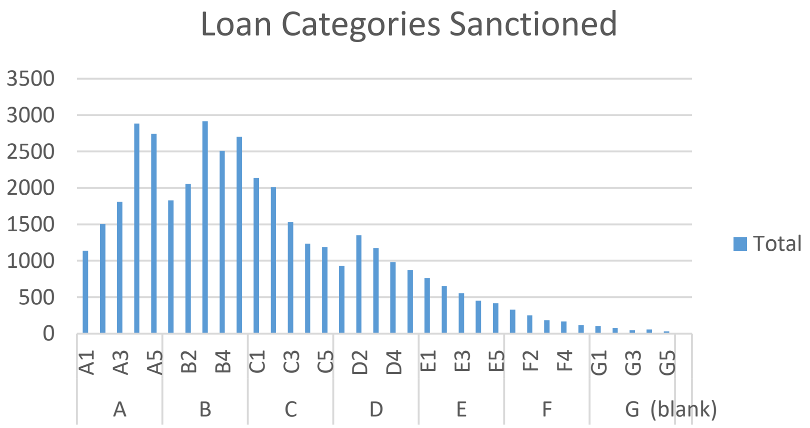
- The most common reason for taking loan is: “Debt Consolidation”



- Charged Off Loans in Dollars is around ~15%



- The loans are categorized from A to Z and then further sub-graded from 1 to 5



# Data Understanding

- The input loan data file has 39,718 records and 111 columns
- Columns with all NULL values:
  - There are around 54 columns where all values are NULL/NaN

```
'mths_since_last_major_derog', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'tot_coll_amt', 'tot_cur_bal',
'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m',
'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths',
'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op','mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl',
'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq',
'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl',
'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m',
'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit'
```

- Columns with Majority NULL values:
  - Out of remaining 57 columns, there are 3 columns, where more than 90% values are NULL

```
'mths_since_last_delinq','mths_since_last_record','next_pymnt_d'
```

- Columns with only 1 Distinct Value:
  - There are 6 columns where there is only a one distinct value

```
['pymnt_plan', 'initial_list_status', 'policy_code', 'application_type', 'acc_now_delinq', 'delinq_amnt']
```

# Data Understanding

- Columns with NULL values in more than 90% values
  - Few columns have more than 90% of data as NULL. These columns are not very useful in the analysis

```
Index(['pymnt_plan', 'initial_list_status', 'policy_code', 'application_type', 'acc_now_delinq',
'delinq_amnt']
```

- Columns with either 0 or NaN values:
  - Some columns have only either 0 or NaN as values. These columns are not very useful in the analysis.

```
'chargeoff_within_12_mths', 'pub_rec_bankruptcies', 'tax_liens'
```

- Columns with Missing Values:
  - There are few columns with missing values. These columns needs to be filled for further analysis

- Categorical data which can be converted to Numerical Data:
  - There are some categorical columns like “Interest Rate percentage”, “loan term”, “employment length” etc… which can be converted to numerical fields for better analysis.

# Data Cleaning and Manipulation

# Data Cleaning and Manipulation

- The following cleaning/Manipulation steps are done on the input data:

  - The Columns following properties are dropped :
    - All values as Null or Empty
    - only 1 Distinct Value
    - Values either NaN or 0
    - Less than 10% of valid values

  - The following columns that have redundant or unwanted data are dropped:
    - URL: The information available in url are already part of other columns and hence redundant
    - loan_amt: Similar information is available in column “funded_amt” and doesn’t seem to add any value
    - funded_amt_inv, out_prncp_inv , total_pymnt_inv : These columns are strongly correlated to “funded_amt” and doesn’t seem to add any value

  - The following categorical columns are converted to Numerical columns for better analysis
    - Term: remove the string “months” from “36 months” & “60 months”. So, the values of the column now become numerical 36 & 60.
    - int_rate: Remove the symbol “%” from the values.
    - revol_util: Remove the symbol “%” from the values.
    - emp_length: Remove the string “Years” and special characters like “<” & “+”.

# Data Cleaning and Manipulation

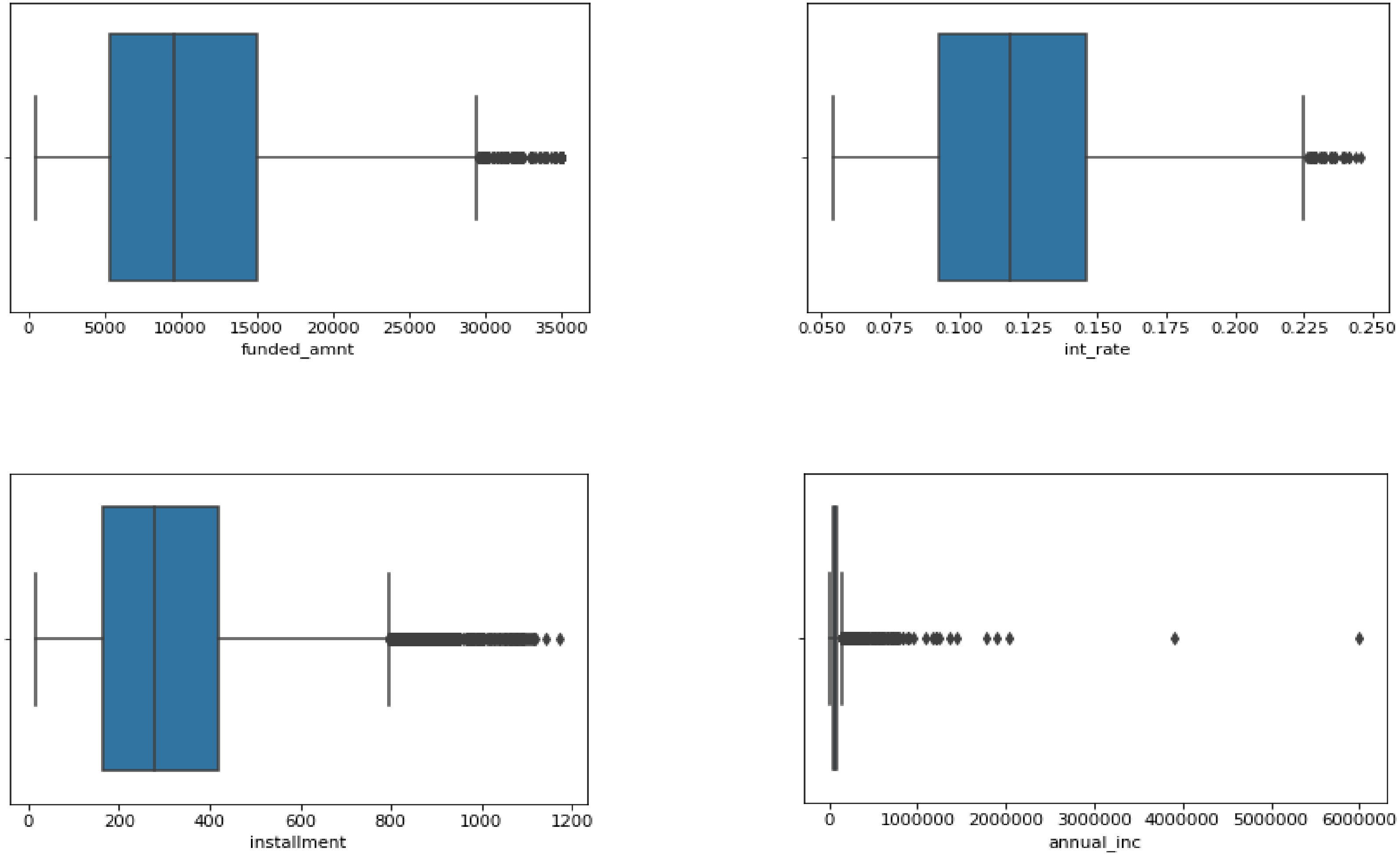
- Format and Convert Data Type of Date Fields:
  - The columns issue_d, last_pymnt_d, last_credit_pull_d & earliest_cr_line are formatted and converted to Date data type

- Fill missing values:
  - last_credit_pull_d: whenever values of this date field are missing, then the values from issue_d are taken
  - pub_rec_bankruptcies: Missing values of this field are filled with 0
  - emp_title: This is a string data type column and hence missing values are filled with “NA”

- Remove Outliers:
  - Outlier are identified and removed for the following numerical columns:
    - funded_amnt
    - int_rate
    - Installment
    - int_rate
    - annual_inc
    - delinq_2yrs
    - inq_last_6mths
    - open_acc
    - pub_rec
    - revol_bal

# Data Cleaning and Manipulation

- Example Outliers Box Plots for “funded_amt”,”int_rate”,”installment”,”annual_inc”:



# Data Analysis

# Data Analysis

After cleaning and manipulating data, the number of columns have comedown from 111 to 37. The data analysis is done on this cleaned data and some important steps and results are presented below.

- Finding Numerical and Categorical Features:
  - There are 23 numerical and 14 categorical features in the data:

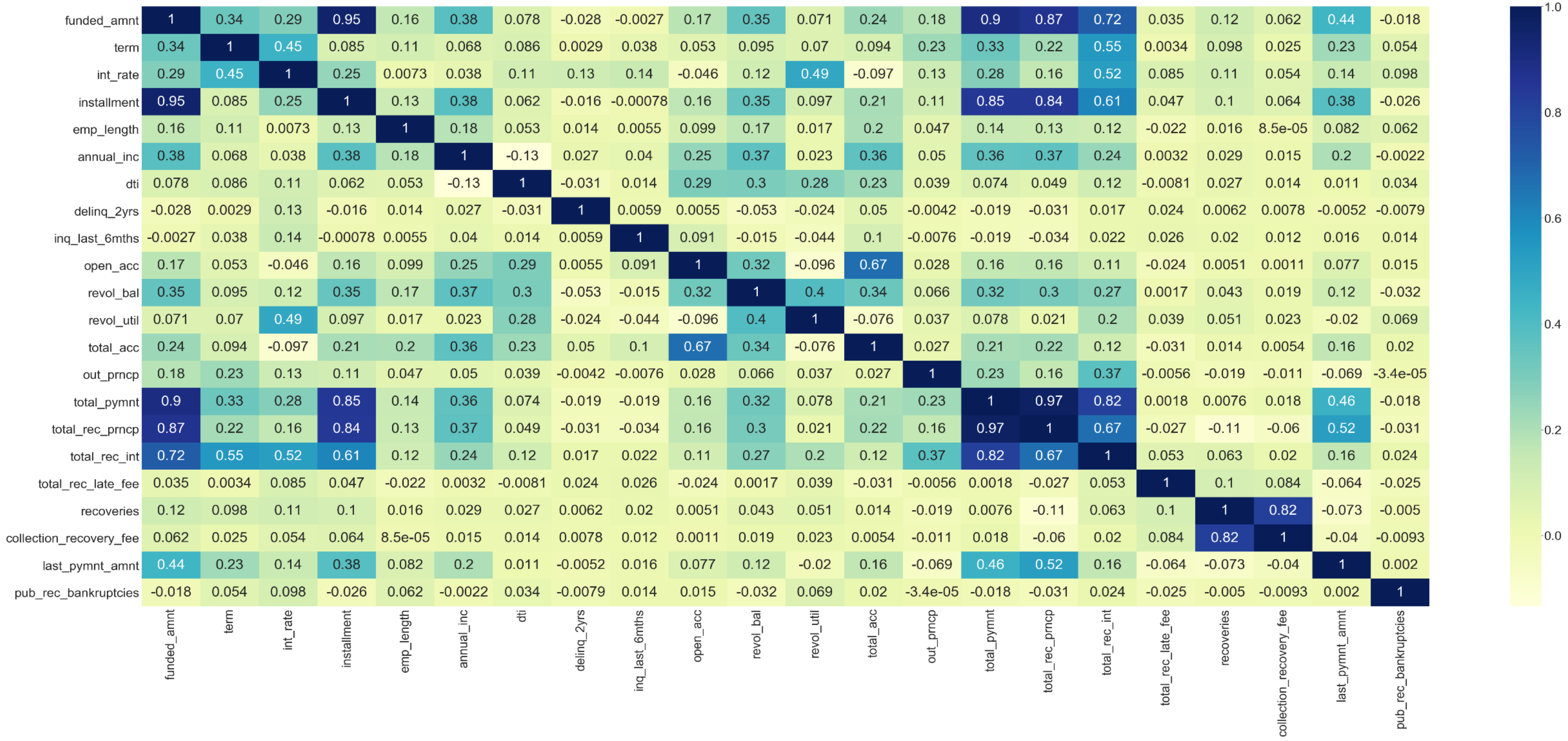
```
Numerical Fields:
Index(['funded_amnt', 'term', 'int_rate', 'installment', 'emp_length', 'annual_inc', 'dti', 'delinq_2yrs',
'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'out_prncp', 'total_pymnt',
'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_amnt',
'pub_rec_bankruptcies'], dtype='object')

Categorical Fields:
Index(['id', 'grade', 'sub_grade', 'emp_title', 'home_ownership', 'verification_status', 'issue_d', 'loan_status',
'purpose', 'title', 'addr_state', 'earliest_cr_line', 'last_pymnt_d', 'last_credit_pull_d'], dtype='object')
```

This data helps in understanding the type of variables and possible analysis like uni-variant/bi-variant analysis on them. This data also helps in determining the columns for correlation analysis.

- Analysis: Further analysis is done on various column combinations to find and arrive at the correlation between different variables and results are presented in following slides:

# Data Analysis

## Correlation plot for numerical data

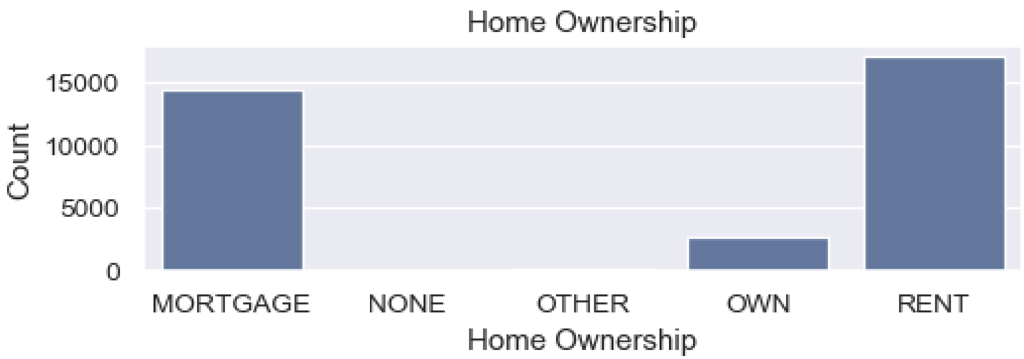


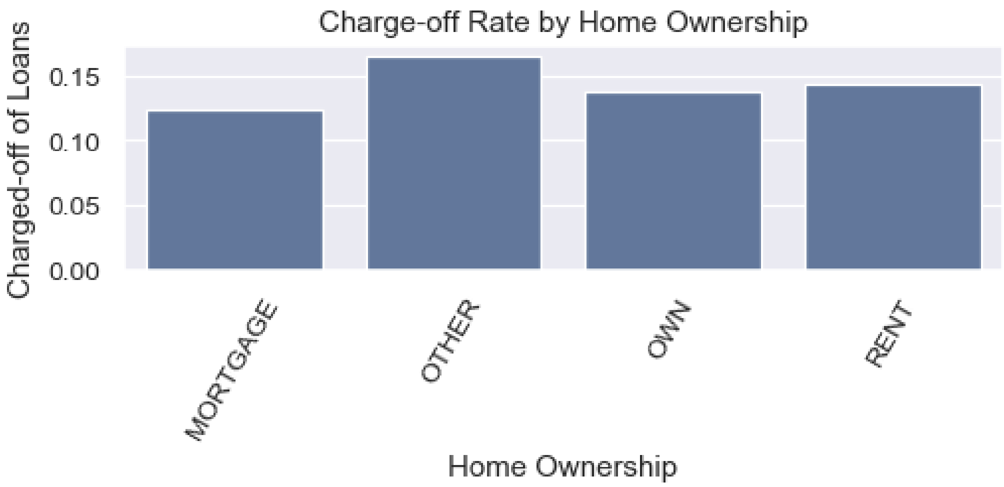
| | funded_amnt | term | int_rate | installment | emp_length | annual_inc | dti | delinq_2yrs | inq_last_6mths | open_acc | revol_bal | revol_util | total_acc | out_prncp | total_pymnt | total_rec_prncp | total_rec_int | total_rec_late_fee | recoveries | collection_recovery_fee | last_pymnt_amnt | pub_rec_bankruptcies |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| funded_amnt | 1 | 0.34 | 0.29 | 0.95 | 0.16 | 0.38 | 0.078 | -0.028 | -0.0027 | 0.17 | 0.35 | 0.071 | 0.24 | 0.18 | 0.9 | 0.87 | 0.72 | 0.035 | 0.12 | 0.062 | 0.44 | -0.018 |
| term | 0.34 | 1 | 0.45 | 0.085 | 0.11 | 0.068 | 0.086 | 0.0029 | 0.038 | 0.053 | 0.095 | 0.07 | 0.094 | 0.23 | 0.33 | 0.22 | 0.55 | 0.0034 | 0.098 | 0.025 | 0.23 | 0.054 |
| int_rate | 0.29 | 0.45 | 1 | 0.25 | 0.0073 | 0.038 | 0.11 | 0.13 | 0.14 | -0.046 | 0.12 | 0.49 | -0.097 | 0.13 | 0.28 | 0.16 | 0.52 | 0.085 | 0.11 | 0.054 | 0.14 | 0.098 |
| installment | 0.95 | 0.085 | 0.25 | 1 | 0.13 | 0.38 | 0.062 | -0.016 | -0.00078 | 0.16 | 0.35 | 0.097 | 0.21 | 0.11 | 0.85 | 0.84 | 0.61 | 0.047 | 0.1 | 0.064 | 0.38 | -0.026 |
| emp_length | 0.16 | 0.11 | 0.0073 | 0.13 | 1 | 0.18 | 0.053 | 0.014 | 0.0055 | 0.099 | 0.17 | 0.017 | 0.2 | 0.047 | 0.14 | 0.13 | 0.12 | -0.022 | 0.016 | 8.5e-05 | 0.082 | 0.062 |
| annual_inc | 0.38 | 0.068 | 0.038 | 0.38 | 0.18 | 1 | -0.13 | 0.027 | 0.04 | 0.25 | 0.37 | 0.023 | 0.36 | 0.05 | 0.36 | 0.37 | 0.24 | 0.0032 | 0.029 | 0.015 | 0.2 | -0.0022 |
| dti | 0.078 | 0.086 | 0.11 | 0.062 | 0.053 | -0.13 | 1 | -0.031 | 0.014 | 0.29 | 0.3 | 0.28 | 0.23 | 0.039 | 0.074 | 0.049 | 0.12 | -0.0081 | 0.027 | 0.014 | 0.011 | 0.034 |
| delinq_2yrs | -0.028 | 0.0029 | 0.13 | -0.016 | 0.014 | 0.027 | -0.031 | 1 | 0.0059 | 0.0055 | -0.053 | -0.024 | 0.05 | -0.0042 | -0.019 | -0.031 | 0.017 | 0.024 | 0.0062 | 0.0078 | -0.0052 | -0.0079 |
| inq_last_6mths | -0.0027 | 0.038 | 0.14 | -0.00078 | 0.0055 | 0.04 | 0.014 | 0.0059 | 1 | 0.091 | -0.015 | -0.044 | 0.1 | -0.0076 | -0.019 | -0.034 | 0.022 | 0.026 | 0.02 | 0.012 | 0.016 | 0.014 |
| open_acc | 0.17 | 0.053 | -0.046 | 0.16 | 0.099 | 0.25 | 0.29 | 0.0055 | 0.091 | 1 | 0.32 | -0.096 | 0.67 | 0.028 | 0.16 | 0.16 | 0.11 | -0.024 | 0.0051 | 0.0011 | 0.077 | 0.015 |
| revol_bal | 0.35 | 0.095 | 0.12 | 0.35 | 0.17 | 0.37 | 0.3 | -0.053 | -0.015 | 0.32 | 1 | 0.4 | 0.34 | 0.066 | 0.32 | 0.3 | 0.27 | 0.0017 | 0.043 | 0.019 | 0.12 | -0.032 |
| revol_util | 0.071 | 0.07 | 0.49 | 0.097 | 0.017 | 0.023 | 0.28 | -0.024 | -0.044 | -0.096 | 0.4 | 1 | -0.076 | 0.037 | 0.078 | 0.021 | 0.2 | 0.039 | 0.051 | 0.023 | -0.02 | 0.069 |
| total_acc | 0.24 | 0.094 | -0.097 | 0.21 | 0.2 | 0.36 | 0.23 | 0.05 | 0.1 | 0.67 | 0.34 | -0.076 | 1 | 0.027 | 0.21 | 0.22 | 0.12 | -0.031 | 0.014 | 0.0054 | 0.16 | 0.02 |
| out_prncp | 0.18 | 0.23 | 0.13 | 0.11 | 0.047 | 0.05 | 0.039 | -0.0042 | -0.0076 | 0.028 | 0.066 | 0.037 | 0.027 | 1 | 0.23 | 0.16 | 0.37 | -0.0056 | -0.019 | -0.011 | -0.069 | -3.4e-05 |
| total_pymnt | 0.9 | 0.33 | 0.28 | 0.85 | 0.14 | 0.36 | 0.074 | -0.019 | -0.019 | 0.16 | 0.32 | 0.078 | 0.21 | 0.23 | 1 | 0.97 | 0.82 | 0.0018 | 0.0076 | 0.018 | 0.46 | -0.018 |
| total_rec_prncp | 0.87 | 0.22 | 0.16 | 0.84 | 0.13 | 0.37 | 0.049 | -0.031 | -0.034 | 0.16 | 0.3 | 0.021 | 0.22 | 0.16 | 0.97 | 1 | 0.67 | -0.027 | -0.11 | -0.06 | 0.52 | -0.031 |
| total_rec_int | 0.72 | 0.55 | 0.52 | 0.61 | 0.12 | 0.24 | 0.12 | 0.017 | 0.022 | 0.11 | 0.27 | 0.2 | 0.12 | 0.37 | 0.82 | 0.67 | 1 | 0.053 | 0.063 | 0.02 | 0.16 | 0.024 |
| total_rec_late_fee | 0.035 | 0.0034 | 0.085 | 0.047 | -0.022 | 0.0032 | -0.0081 | 0.024 | 0.026 | -0.024 | 0.0017 | 0.039 | -0.031 | -0.0056 | 0.0018 | -0.027 | 0.053 | 1 | 0.1 | 0.084 | -0.064 | -0.025 |
| recoveries | 0.12 | 0.098 | 0.11 | 0.1 | 0.016 | 0.029 | 0.027 | 0.0062 | 0.02 | 0.0051 | 0.043 | 0.051 | 0.014 | -0.019 | 0.0076 | -0.11 | 0.063 | 0.1 | 1 | 0.82 | -0.073 | -0.005 |
| collection_recovery_fee | 0.062 | 0.025 | 0.054 | 0.064 | 8.5e-05 | 0.015 | 0.014 | 0.0078 | 0.012 | 0.0011 | 0.019 | 0.023 | 0.0054 | -0.011 | 0.018 | -0.06 | 0.02 | 0.084 | 0.82 | 1 | -0.04 | -0.0093 |
| last_pymnt_amnt | 0.44 | 0.23 | 0.14 | 0.38 | 0.082 | 0.2 | 0.011 | -0.0052 | 0.016 | 0.077 | 0.12 | -0.02 | 0.16 | -0.069 | 0.46 | 0.52 | 0.16 | -0.064 | -0.073 | -0.04 | 1 | 0.002 |
| pub_rec_bankruptcies | -0.018 | 0.054 | 0.098 | -0.026 | 0.062 | -0.0022 | 0.034 | -0.0079 | 0.014 | 0.015 | -0.032 | 0.069 | 0.02 | -3.4e-05 | -0.018 | -0.031 | 0.024 | -0.025 | -0.005 | -0.0093 | 0.002 | 1 |

# Data Analysis

- Uni-Variant and Bi-Variant analysis is done on different columns to analyze the data and their relationship
- Some important plots of various columns in relationship with loan default/charge-off

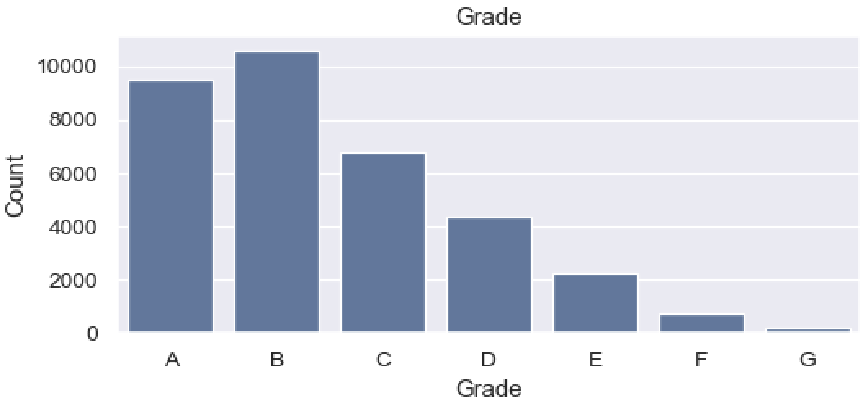
- <u>Plot: Charge-off rate by Home Ownership</u>

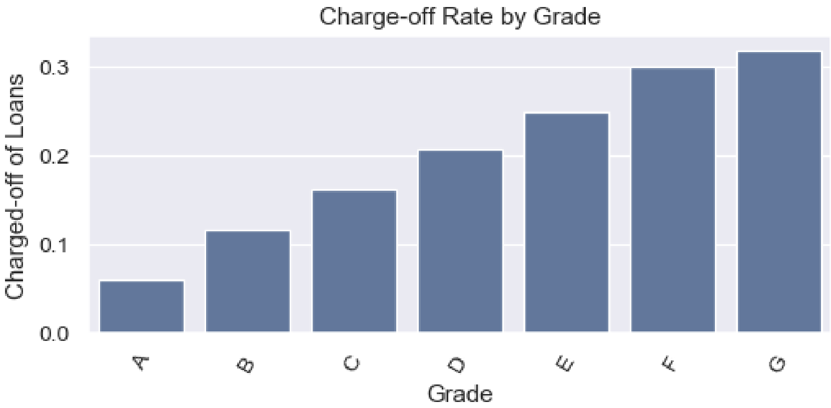


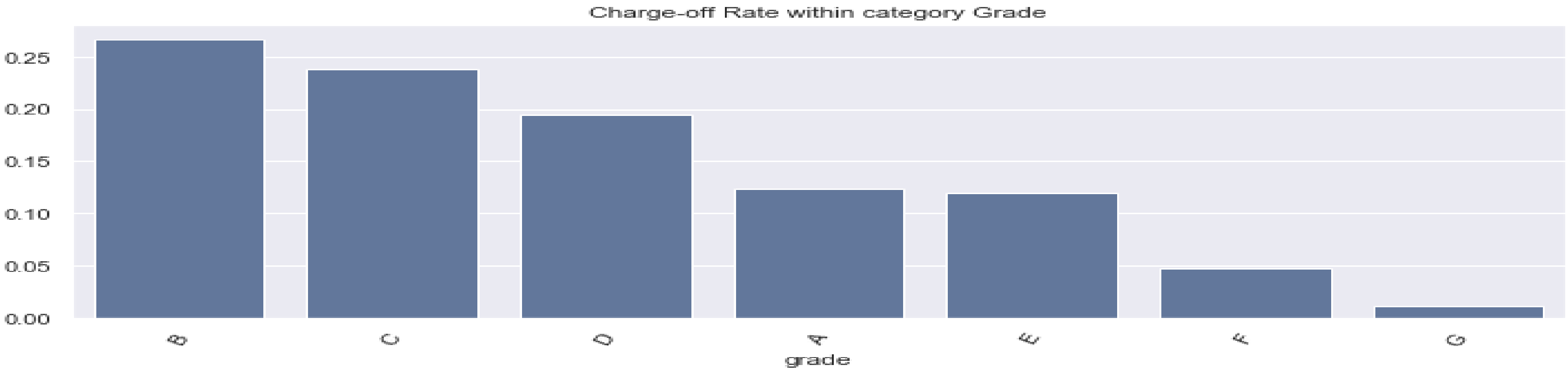


# Data Analysis

- Plot: Charge-off rate by "Loan Grade"

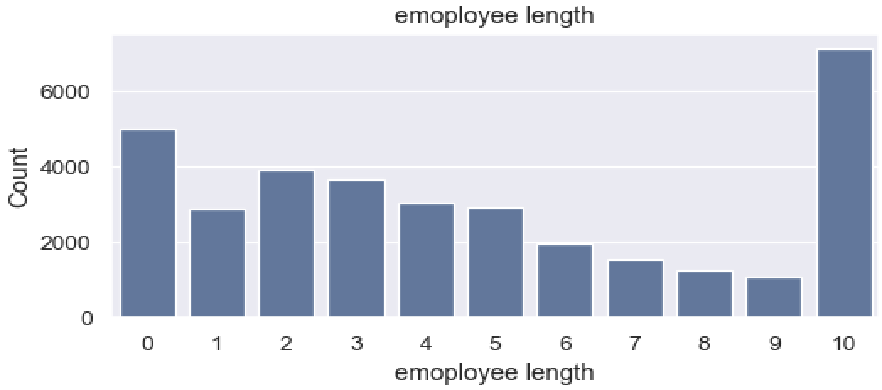


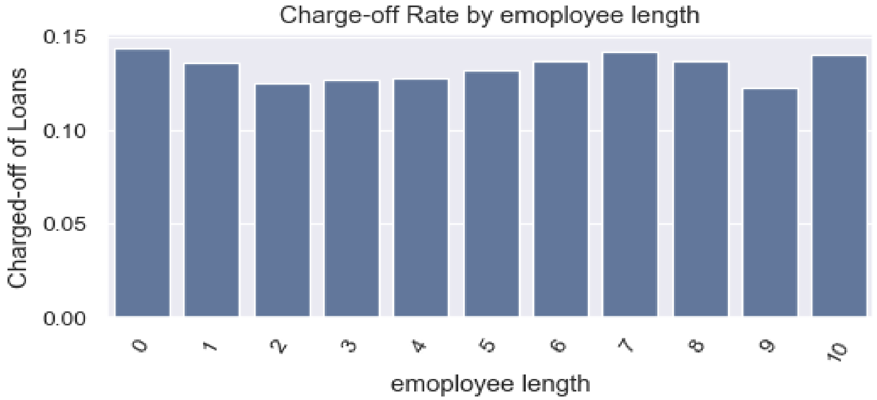


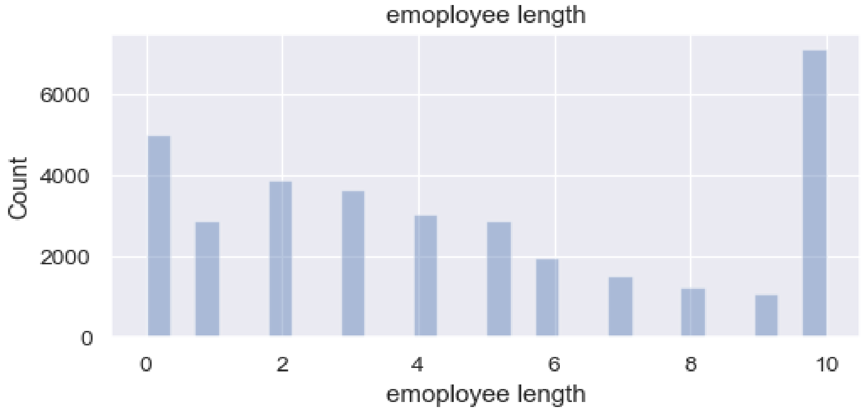


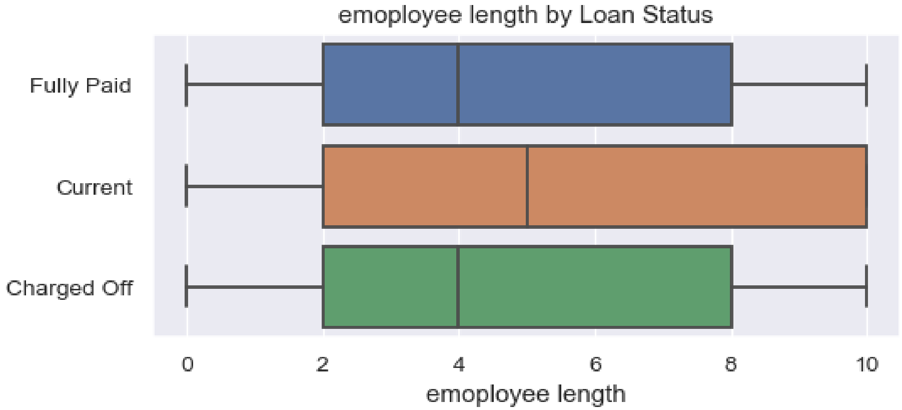
# Data Analysis

- Plot: Charge-off rate by “Employment Length”



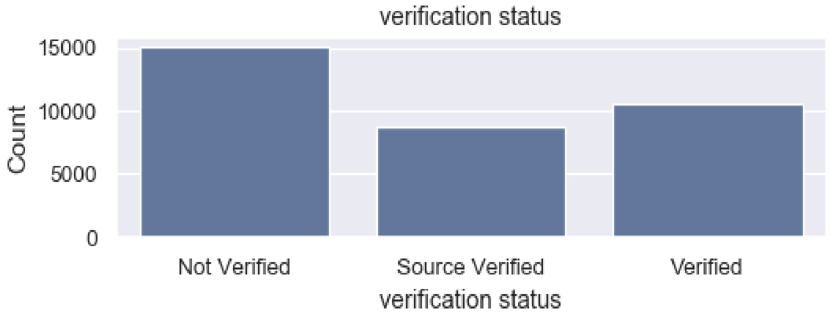


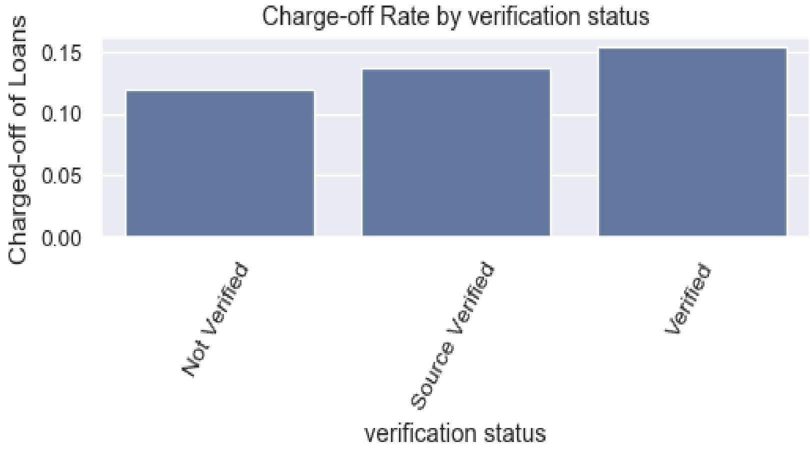


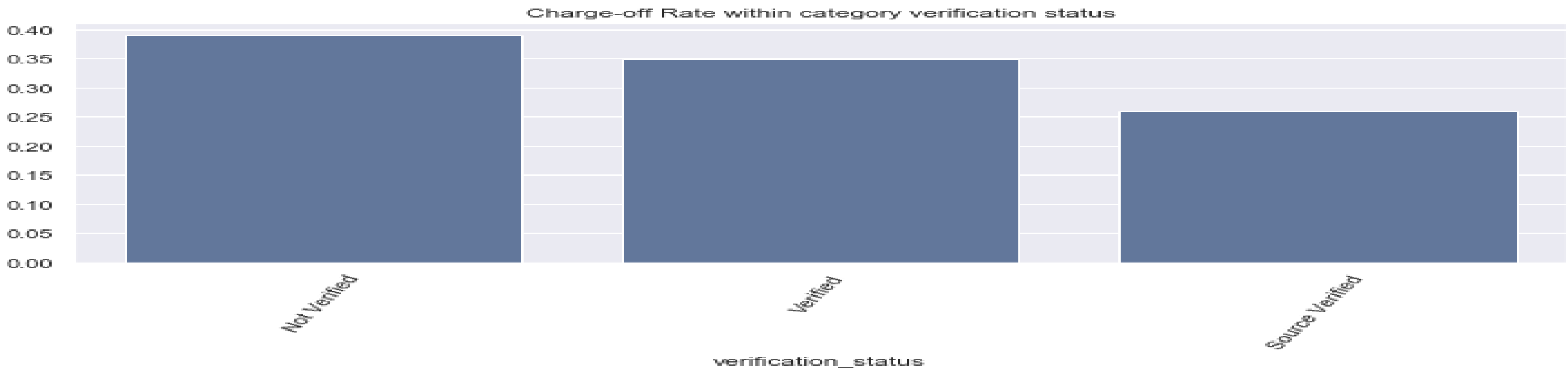


# Data Analysis

- Plot: Charge-off rate by “Verification Status”

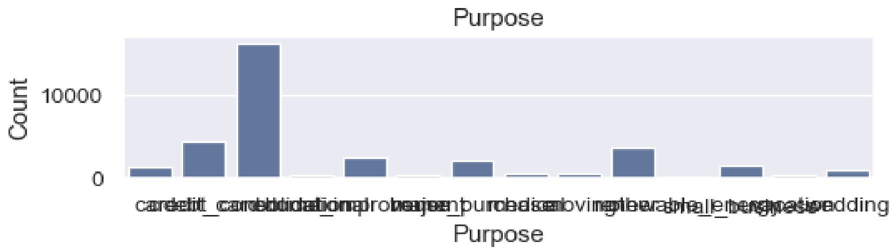


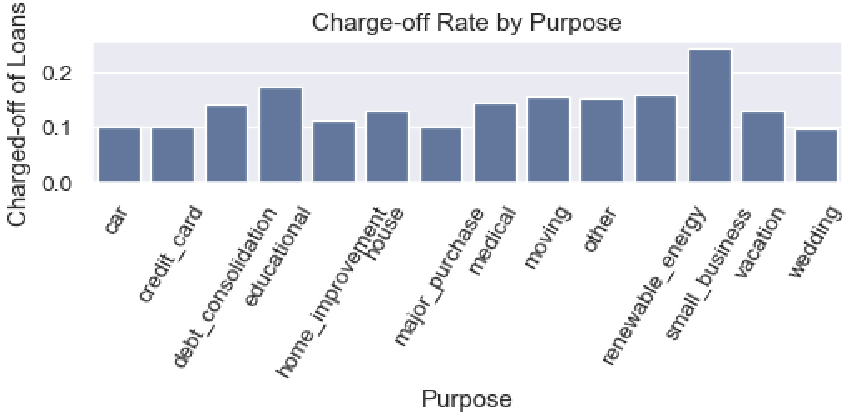


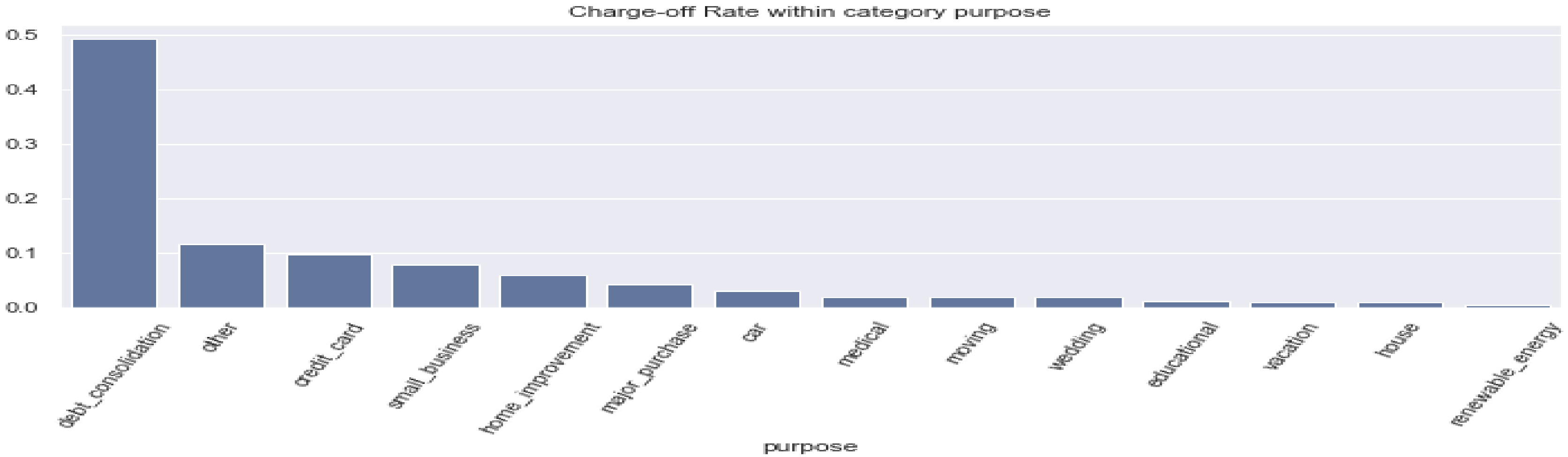


# Data Analysis

- <u>Plot: Charge-off rate by “Purpose of Loan”</u>

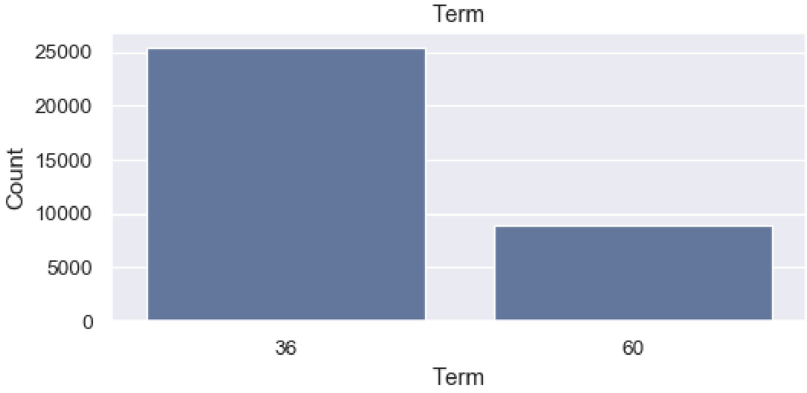


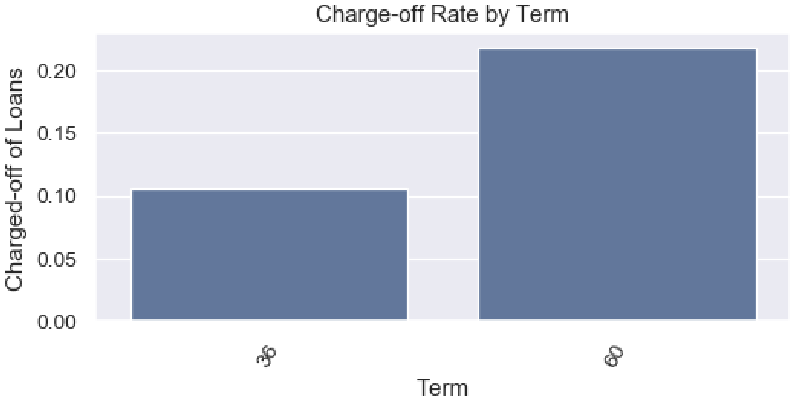




# Data Analysis

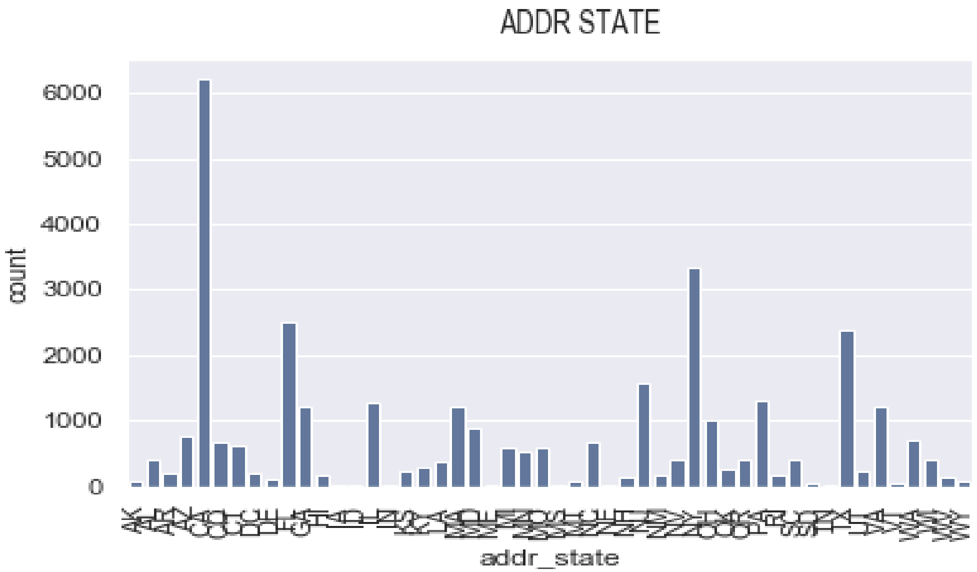
- Plot: Charge-off rate by “Loan Term”

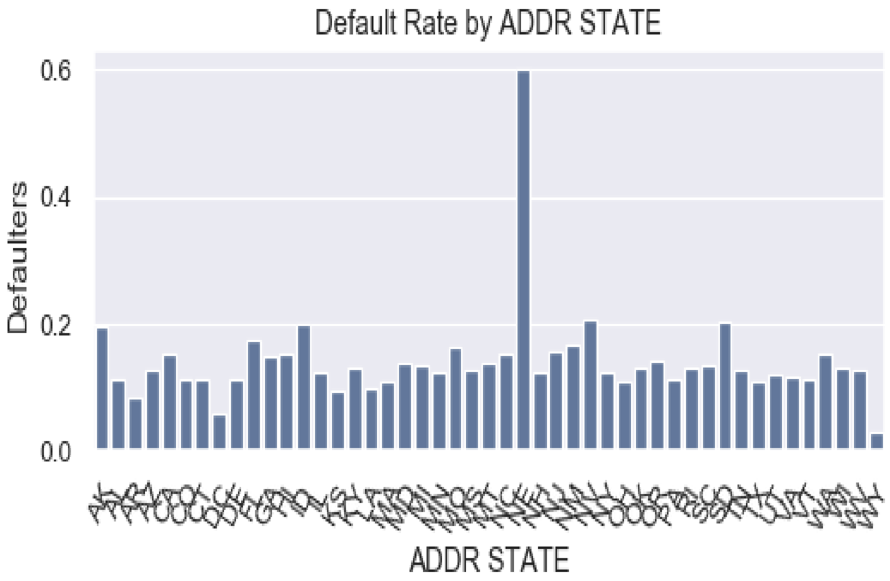


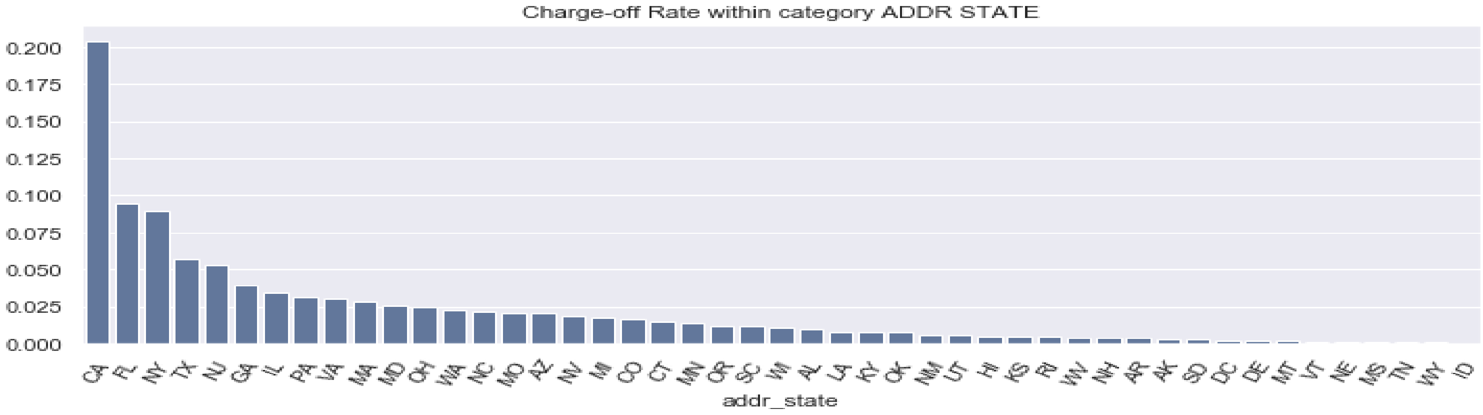


# Data Analysis

- Plot: Charge-off rate by “State”

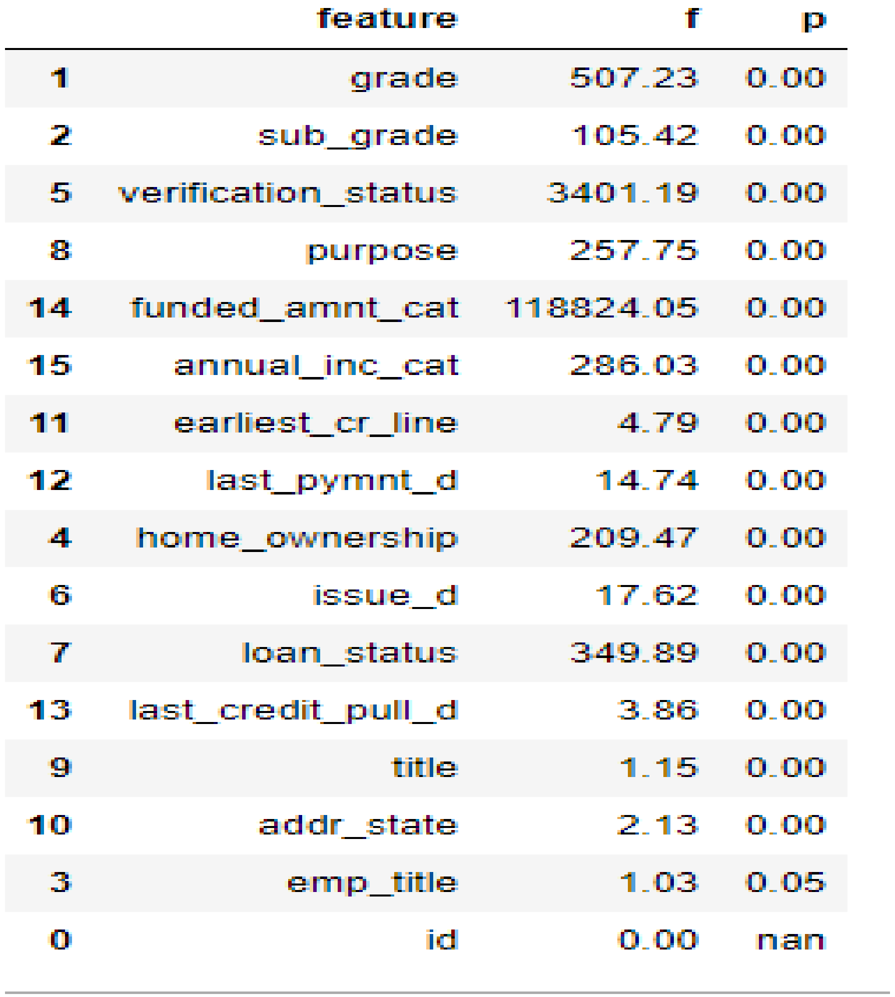

ADDR STATE


Default Rate by ADDR STATE


Charge-off Rate within category ADDR STATE

# Data Analysis

- Annova Analysis on Loan



| | feature | f | p |
|---|---|---|---|
| 1 | grade | 507.23 | 0.00 |
| 2 | sub_grade | 105.42 | 0.00 |
| 5 | verification_status | 3401.19 | 0.00 |
| 8 | purpose | 257.75 | 0.00 |
| 14 | funded_amnt_cat | 118824.05 | 0.00 |
| 15 | annual_inc_cat | 286.03 | 0.00 |
| 11 | earliest_cr_line | 4.79 | 0.00 |
| 12 | last_pymnt_d | 14.74 | 0.00 |
| 4 | home_ownership | 209.47 | 0.00 |
| 6 | issue_d | 17.62 | 0.00 |
| 7 | loan_status | 349.89 | 0.00 |
| 13 | last_credit_pull_d | 3.86 | 0.00 |
| 9 | title | 1.15 | 0.00 |
| 10 | addr_state | 2.13 | 0.00 |
| 3 | emp_title | 1.03 | 0.05 |
| 0 | id | 0.00 | nan |

# Results & Recommendations

# Inferences, Results & Recommendations

Based on the above analysis:

Top inferences and driving factors behind loan defaults are marked in blue color:

Loan Provider Factors:

- Median of funded amount is higher for current loan than fully paid loan or charged off loan.
- Loans are not given for DTI more than 30.
- Also very less loans are provided for DTI higher than 25.

Loan Consumer Factors:

- People living in rent or mortgage tend to take more loans.
- **Percentage of people who are living in rent have highest defaults. After that people with mortgage and least are people with own house or others.**
- **Number of default is more in small business.**
- Maximum number of loans is taken from California.
- **Also highest default rate is from CA.**
- **As funding amount increases above 20K risk of default increases.**

Loan Consumer Behavior Factors -

- **Frequency of people getting loans is higher for higher grade. Also people with better grades have lower defaults.**
- People with 0 or 10+ years of experience take more number of loans
- **Within defaulters people with 0 or 10+ years experience tends to default more.**
- **Default rate is highest in loan for the purpose of debt-consolidation.**
- Maximum loans are availed by people with income below 100,000.
- As the income increase default rates are coming down.
- More number of people have taken loan with less term of 36 month.
- **Higher defaulters are found in long term loan of 60 months.**