Lending Club Case Study

Group Details :

Karthick Chetti – Group Facilitator

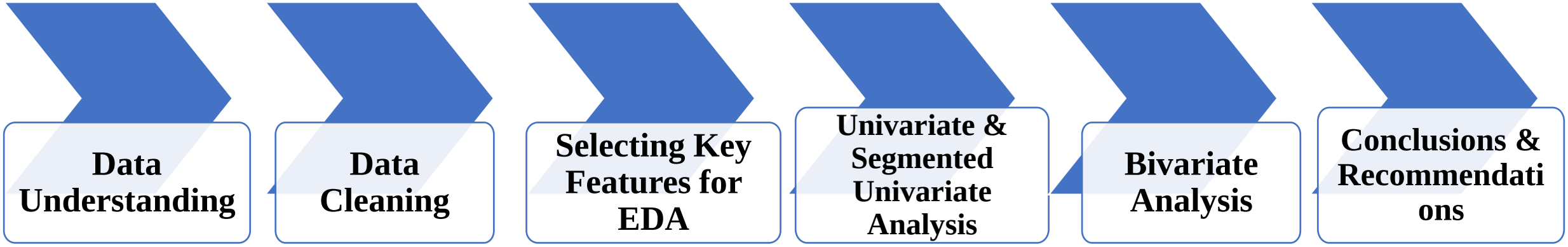
Anirudh KVC - Collaborator

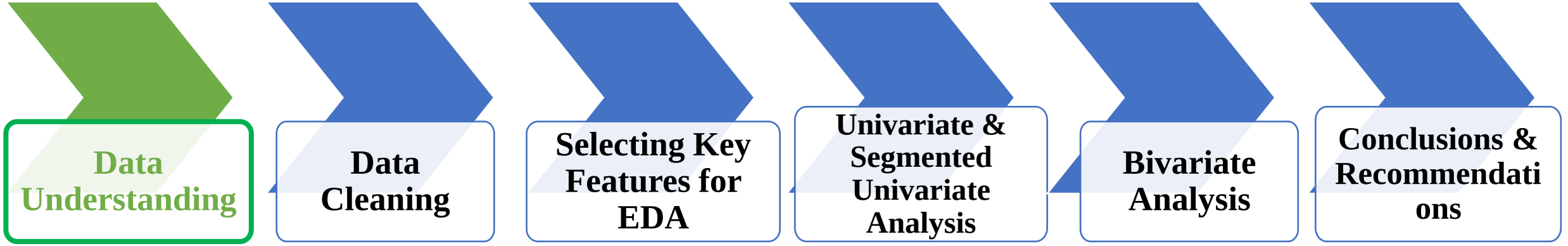
Business Objective & Approach

Objectives:

- Understanding the driving factors behind loan default thereby aiming to reduce credit loss
- Data Driven inferences to prevent charging off for a potential loan applicant

Approach:





- Understanding the meaning of variables with the help of data dictionary
- Trying to group the variables and understanding their relevance:
  - Demographic variables (details about the person *ex. state*)
  - Loan variables(details about the current loan *ex. int_rate*)
  - Applicant credit profile(details about the past credit lines and credibility *ex. delinq_2_years*)
- Target variable is **loan status**

Assumption : For our objective, the customers with loan status as ***Current*** are not important for the study, since we want to analyze the driving factors for the completed loans. The analysis is carried out only for ***Fully Paid*** and ***Charged Off*** Customers

Data Understanding | **Data Cleaning** | Selecting Key Features for EDA | Univariate & Segmented Univariate Analysis | Bivariate Analysis | Conclusions & Recommendations

1. Removing Redundant Columns

Total Number of Columns initially in the loan dataset = 111

- Having more than 50% missing values in them *(57 columns*)
- Removing Columns with same values *(8 columns*)
- Removing Columns with either 0s or missing values in them *(3 columns*)
- Removing descriptive columns *(6 columns)*

Output - 37 statistically meaningful columns highlighted

2. Identifying most relevant data features

- Based on business judgement and data understanding, these columns can help us identify the trends for customer charge-off

3. Data manipulation

String manipulation is used to remove % from **interest rate** and convert them from string to float values, similarly in earliest credit line, we are interested in years, so we extracted years from dates

4. Checking missing values & imputing data

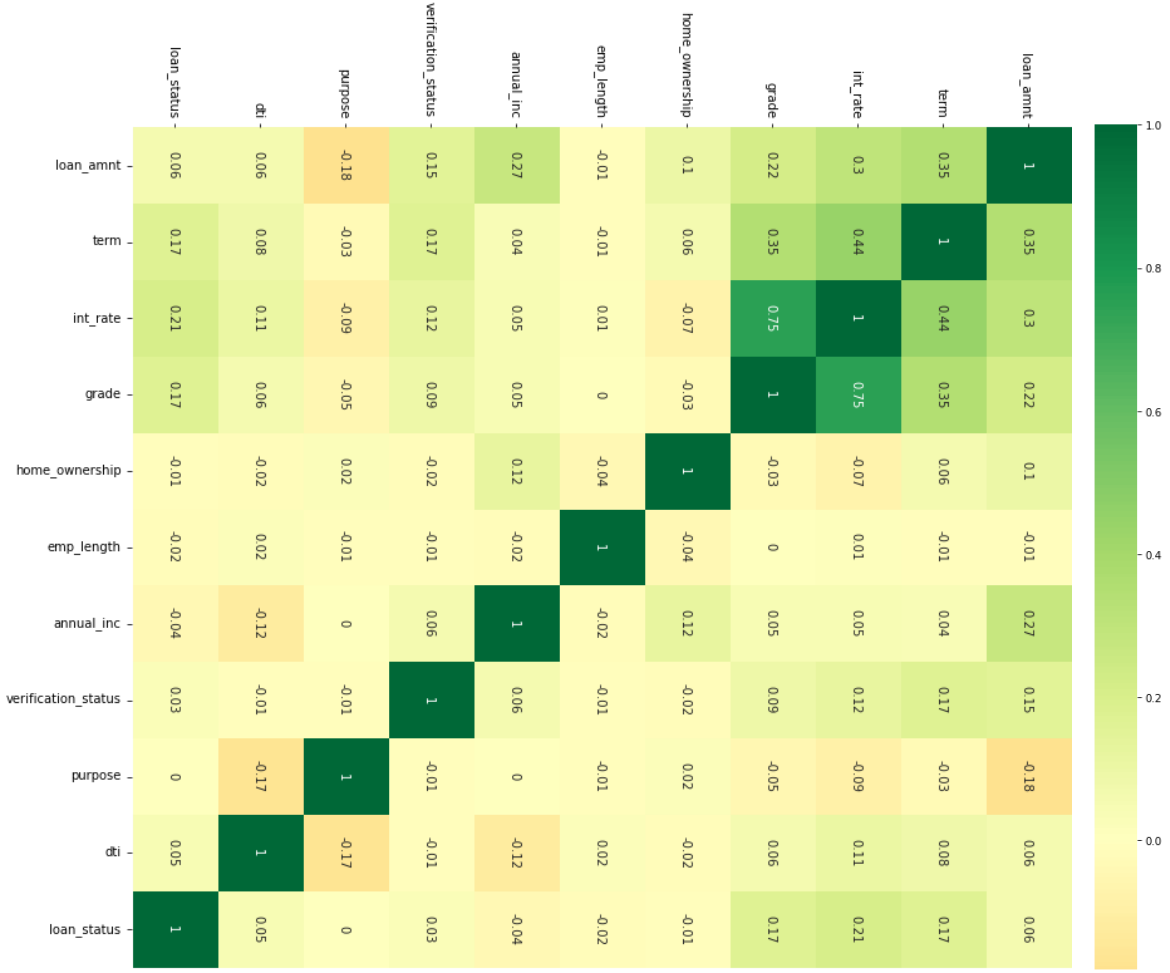
- It is observed that within the selected columns for analysis, *emp_length* column has about 2.7% missing values from total values.
- Missing data is within 5%, we chose to impute rather than deleting those rows
- Since it's a categorical variable, it is imputed by the mode of the column which was 10+ years

Output : Selected dataset has no missing values

5. Checking outliers and removing them

- For the numeric data columns, box plots were plotted to understand the distribution and outliers.
- Other than *dti*, rest of the columns showed outliers
- It was removed using standard outlier limits but Q1 & Q3 were taken at 10 & 90 percentile respectively to minimize the deletion of data points

Data Understanding → Data Cleaning → **Selecting Key Features for EDA** → Univariate & Segmented Univariate Analysis → Bivariate Analysis → Conclusions & Recommendations

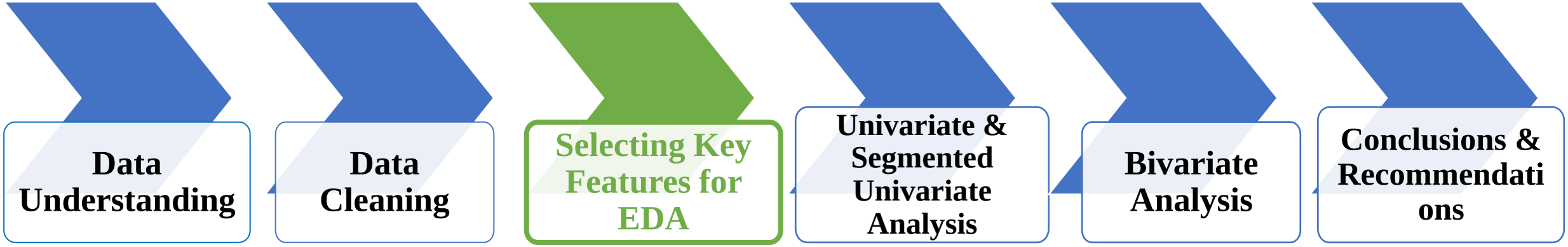


Correlation coefficient depicted by R

- The categorical variables are encoded with numbers for each category to check correlation
- As the input features are not highly correlated among themselves (i.e. R < 0.9) the selected input features can be taken for the EDA analysis
- Amongst the variables, loan_amnt, term, int_rate & grade are show some positive correlation amongst them, with grade and int_rate being the most correlated (R=0.75)

Key Features Selected for EDA →

1. **Numeric Data** :
   loan_amt,int_rate,annual_inc,dti
2. **Ordered Categorical Data** :
   term,grade,emp_length
3. **Unordered Categorical Data** :
   home_ownership,purpose,verification_status,loan_status



Data Driven Metrics

Business Driven Metric :

- Lending club would not be interested in exact number of years of experience of the employee, so employee length is categorized as <1 year(New employees), 1-3, 3-6, 6-9 years(Different experience groups), >10 years(Most experienced employees)

Data Driven Metric:

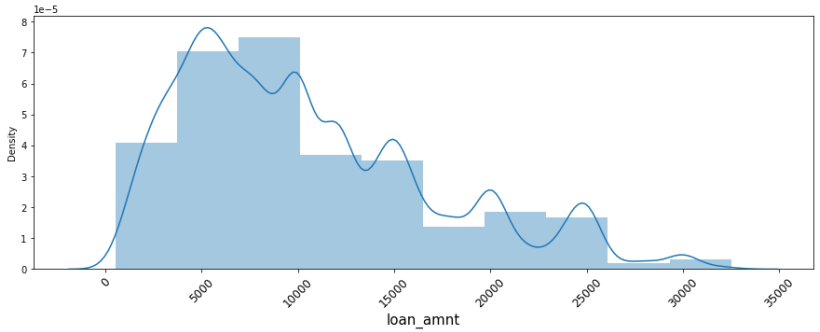
- The numeric variables are binned into certain categories to aid better analysis of numeric variables such as loan_amnt, int_rate, annual income and dti. (Ex. int_rate is binned into categories like 5-10%(low), 10-15%(medium),15-20%(high) etc. to derive insights on the derived groups of interest rates

Type Driven Metric:

In the earliest credit line column, we are interested in years, so we extracted years from dates and then binned them in groups of 10 years to analyze their impact on loan status

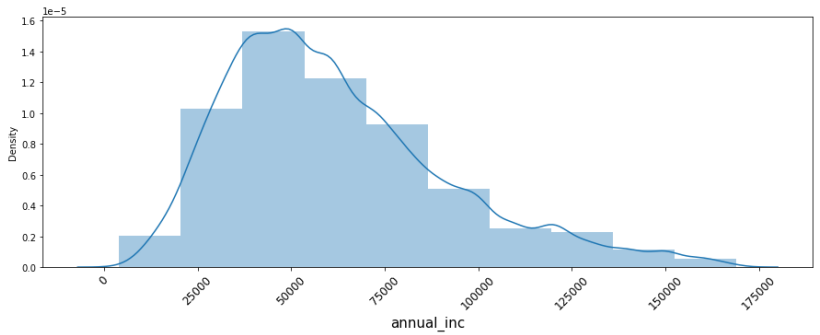
Univariate Analysis – Numeric Variables

Loan Amount



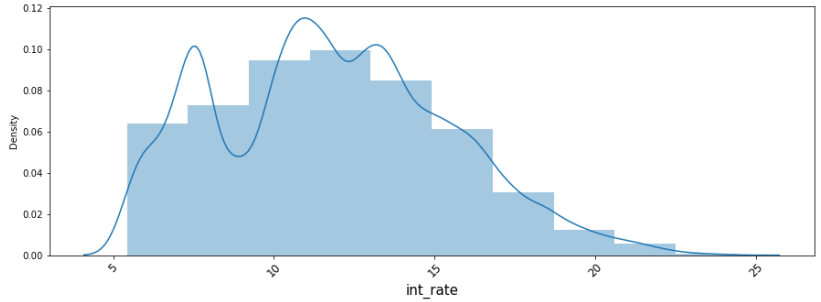
Most of the people prefer loan amount ranging from 4000 to 10000. The median of loan amount is 9200

Annual Income



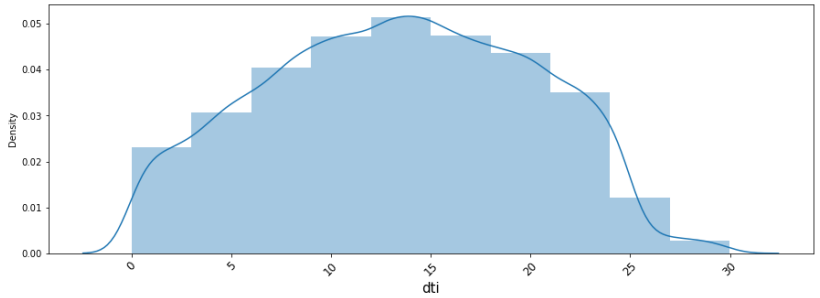
The median of annual income is 56100. It follows a near normal distribution which is right-skewed

Interest Rate



The average interest rate is 11.8 %. As the interest rate increases beyond 15%, number of customers taking loans reduces significantly

Debt to income ratio



The average dti ratio is 13.38. The dti ratio is almost following normal distribution with mean near to 14 dti

Data Understanding

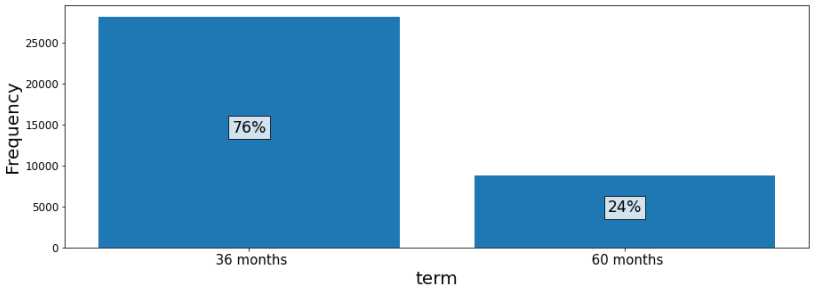
Data Cleaning

Selecting Key Features for EDA

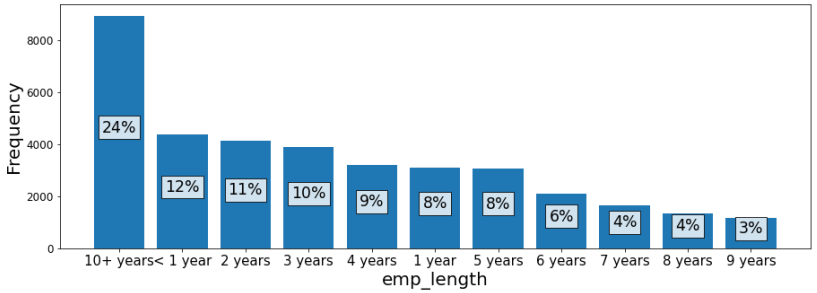
Univariate & Segmented Univariate Analysis

Bivariate Analysis

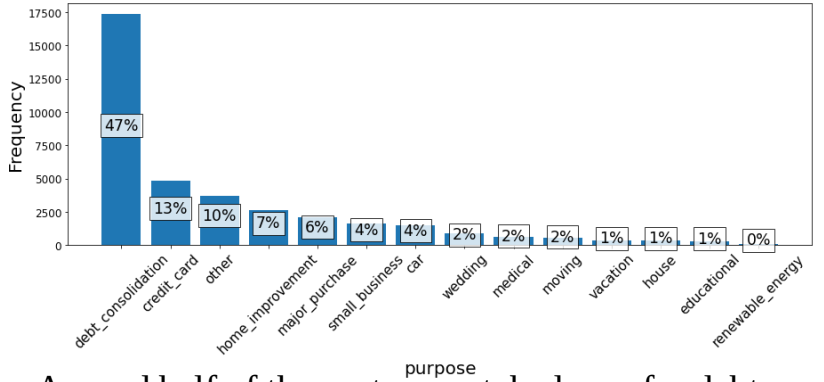
Conclusions & Recommendations



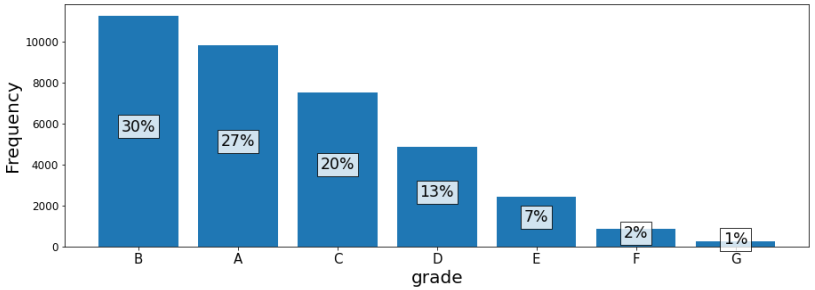
About three quarters of the customers prefer 36 months of loan duration



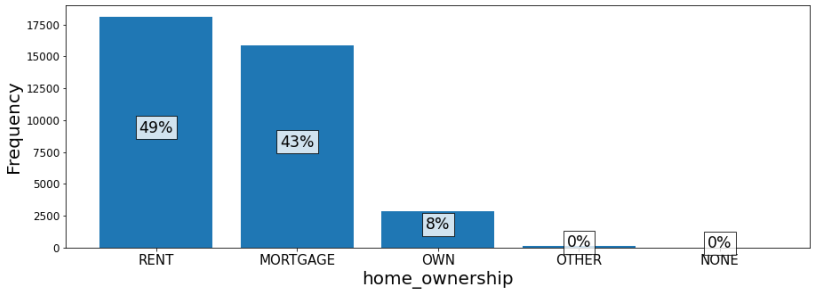
About a quarter of loans are taken by customers having 10+ years of experience



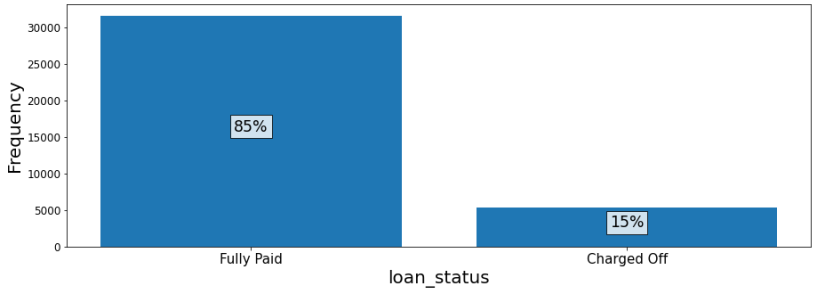
Around half of the customers take loans for debt consolidation



About 80% of the customers take loans in grades A,B & C



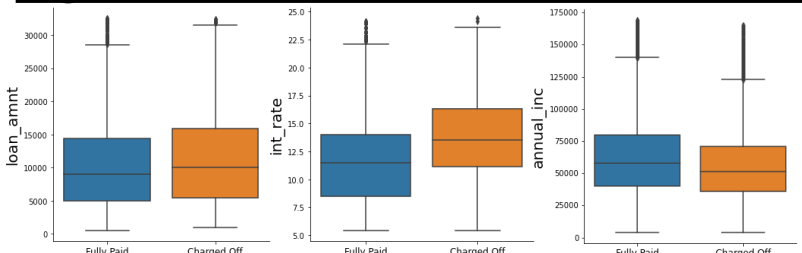
About 90% of the people who have taken the loan are on rent & mortgage



About 85% of the loans are fully paid, charge off percentage is 15%, this leads to credit loss

Data Understanding | Data Cleaning | Selecting Key Features for EDA | **Univariate & Segmented Univariate Analysis** | Bivariate Analysis | Conclusions & Recommendations

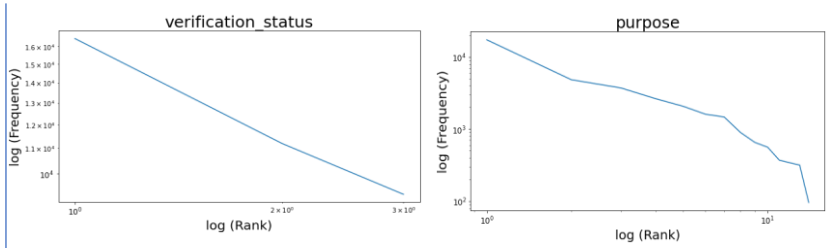
Segmented Univariate – Numeric Variables



Median of →	Loan Amount	Interest Rate	Annual income	dti
Fully Paid	9000	11.48	57700	13.34
Charged Off	10000	13.49	51200	14.35

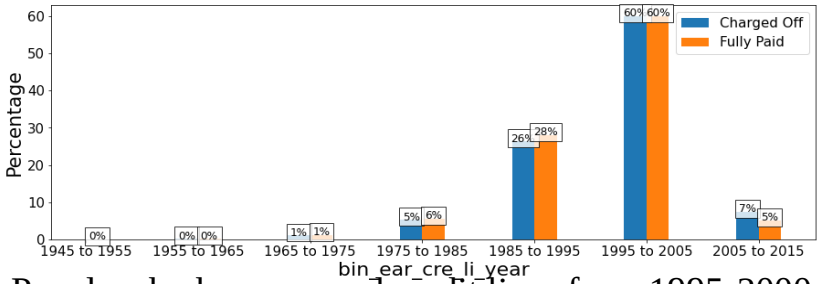
Charged off customers take more loan amount, have more interest rate on loan, have more debt to income ratio and have less annual income than fully paid customers

Univariate Analysis – Unordered Categorical



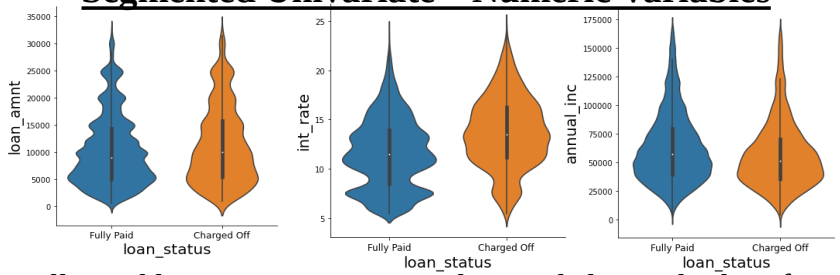
From the rank vs frequency plots it can be inferred that **verification status** follow power law distribution & **purpose** follows near power law dist.

Segmented Univariate –Driven Metrics



People who have opened credit lines from 1995-2000 have taken about 60% loans, however **new customers(earliest credit line 2005-2015) have charged off more than fully paid the loan, they might be risky**

Segmented Univariate – Numeric Variables



Fully paid loan amounts are majorly rounded in multiples of 5000(ex. 15k,20k)

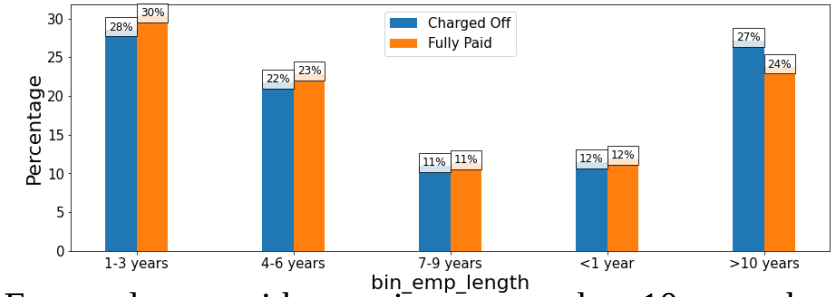
Fully paid customers take loans at interest rates of 6-7% and 10-15% while charged off loan interest rates are majorly more than 10%

Creating plots for categorical variables

The plots have been made considering:

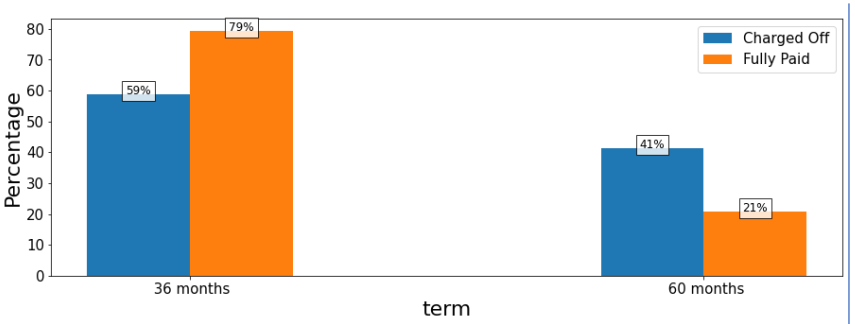
- The percentages of total fully-paid or charged off instead of frequency
- This is done to scale down the plot and understand the difference between them
- We are not using frequency as the y axis scale because data is biased towards fully paid customers (which is expected for a lending company)

Segmented Univariate –Driven Metrics

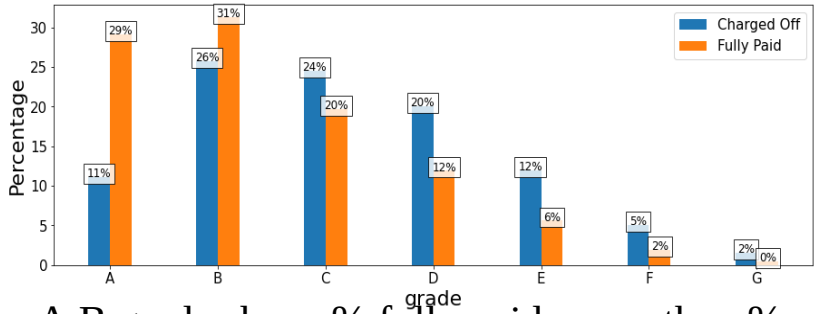


For employees with experience more than 10 years the percentage of charged off customers is higher than fully paid, **so giving loans to such employees might be risky**

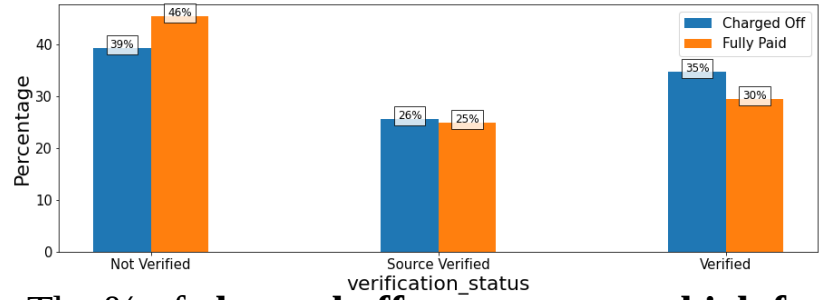
Data Understanding | Data Cleaning | Selecting Key Features for EDA | **Univariate & Segmented Univariate Analysis** | Bivariate Analysis | Conclusions & Recommendations



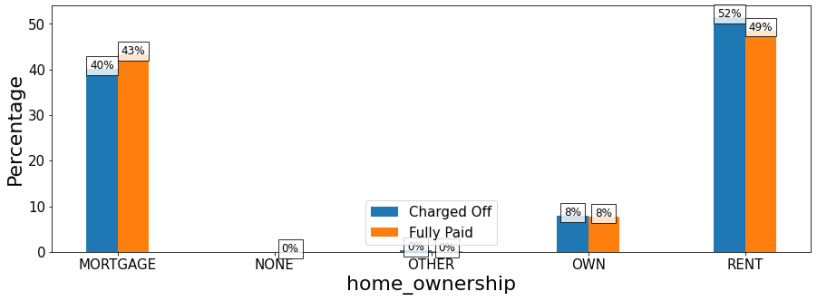
Charged off customers are significantly higher than fully paid for 60 months term, **making 5 years loan tenure a risky option**



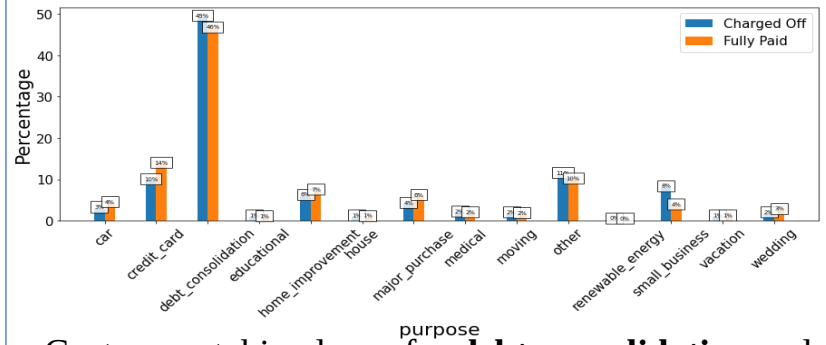
A,B grades have % fully paid more than % charged off where as the trend is opposite from grade C to G, making it risky loan grades



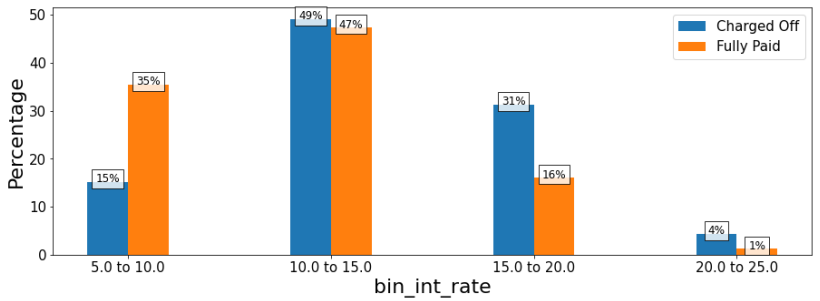
The % of **charged off customers are high for verified applications** compared to fully paid, logically this should have lower charged off



Customers who have a **rented house** are risky because they data shows that this category has more charged off than fully paid customers

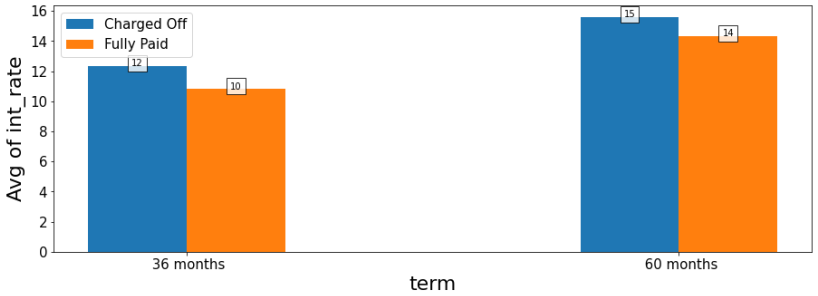


Customers taking loans for **debt consolidation** and **small business** have high percentage of charged off customers compared to fully paid

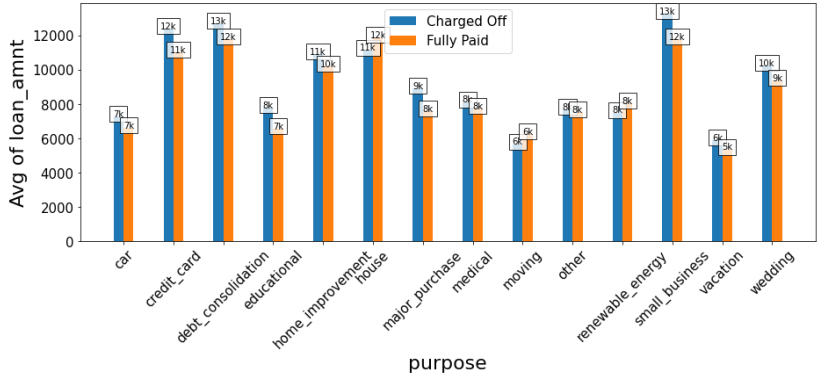


For interest rates greater than 10%, the percentage of charged off customers is more than that of fully paid with **15-25% being most risky interest rates**

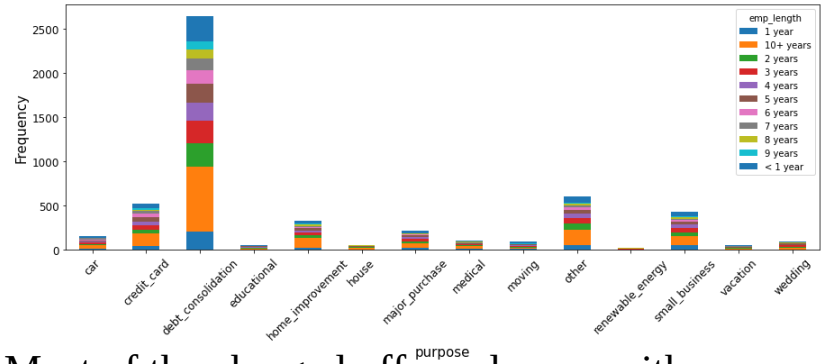
Data Understanding → Data Cleaning → Selecting Key Features for EDA → Univariate & Segmented Univariate Analysis → **Bivariate Analysis** → Conclusions & Recommendations



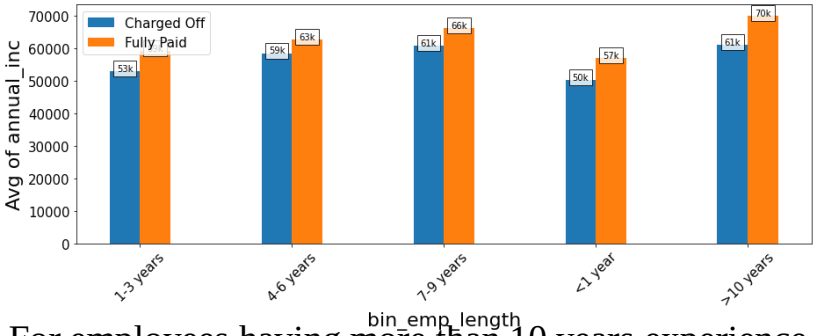
The average **interest rate** for **60 months is more than that of 36 months** for all customers and it is the **highest(15%) for charged off customers taking loans for 60 months**



The average loan amount is higher for **small business, debt consolidation and credit card** - LC should be more cautious in offering loans for these purposes

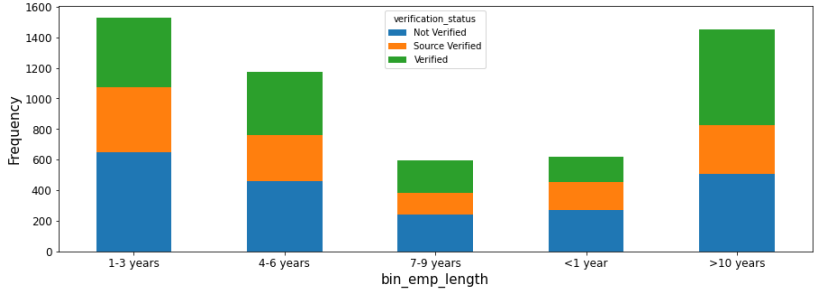


Most of the charged off employees with more than 10 years of experience take loans for debt consolidation



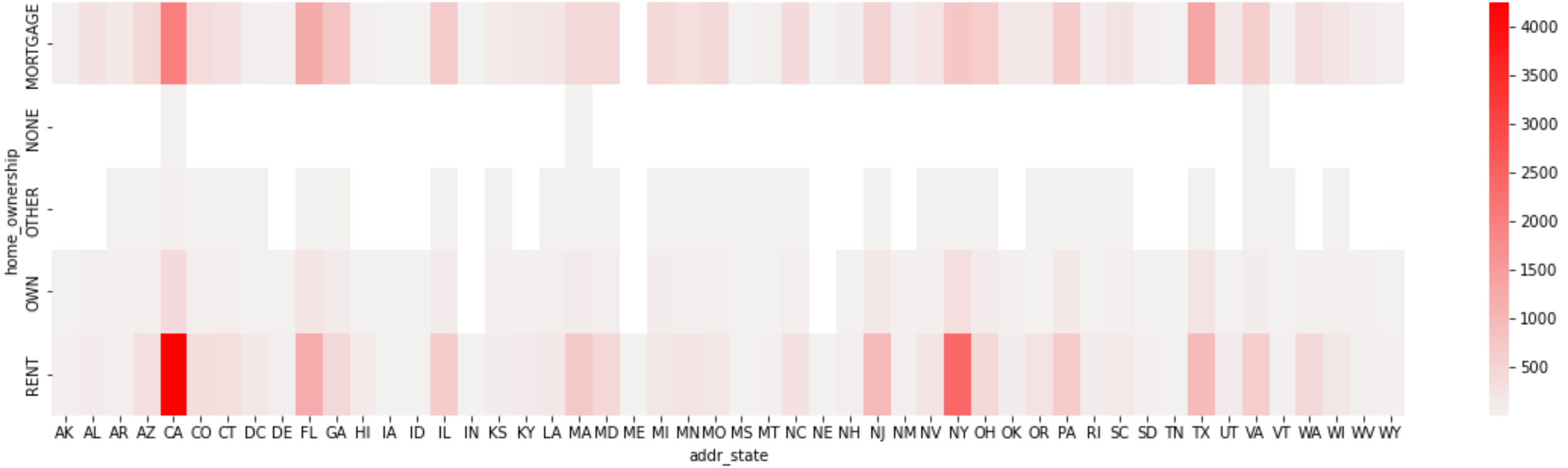
For employees having more than 10 years experience, the average annual income of the fully paid customers is **70,000** while for charged off customers it is **61,000**

To further analyze the major reasons for charge off with the help of bivariate analysis**, we focus our analysis on only charged off customers**



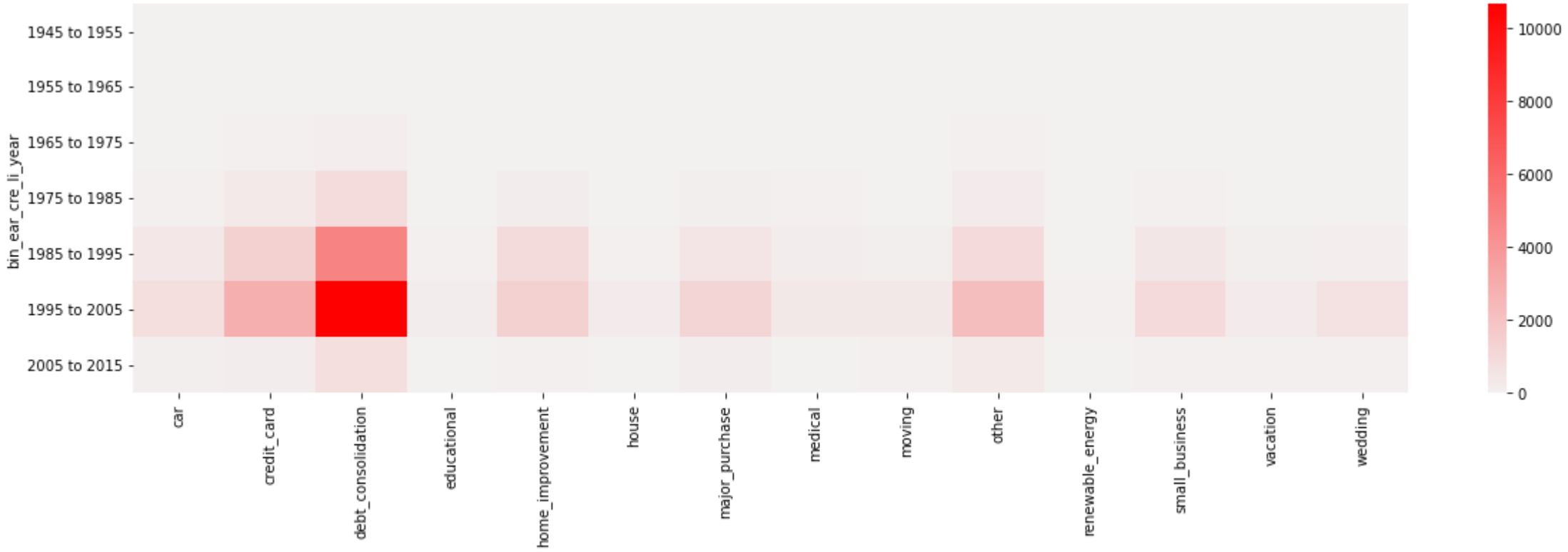
For charged off customers with experience less than equal to 1 year, **most of them have not verified documents** while those **with more than 10 years experience have verified documents**

Data Understanding → Data Cleaning → Selecting Key Features for EDA → Univariate & Segmented Univariate Analysis → **Bivariate Analysis** → Conclusions & Recommendations

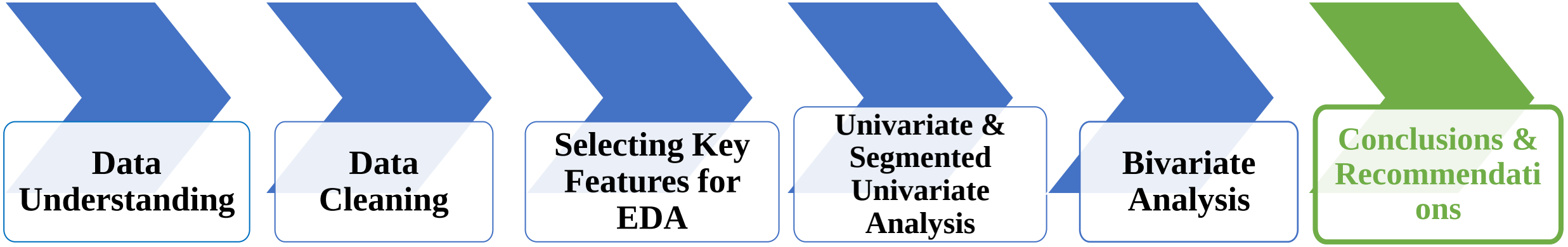


- It is observed the customers living on rent in California are charging off more.
- This might be probably due to high rental amounts in California

Data Understanding | Data Cleaning | Selecting Key Features for EDA | Univariate & Segmented Univariate Analysis | **Bivariate Analysis** | Conclusions & Recommendations



- The customers who have earliest credit lines starting from the year 1985 to 2005 and taking loans for debt consolidation are charging off more
- This might be probably to repay the earlier loans taken



The major factors for charging off loans :

1) ***Loan amount*** – **For amounts greater than 10000**
2) ***Annual income*** **- For income less than 50000**
3) ***Interest rate*** **- For interest rates greater than 10%**
4) ***Debt to Income Ratio*** – **For dti greater than 15**
5) ***Term*** – **For loan tenure of 60 months**
6) ***Employee Length*** – **For employees with more than 10 years of experience**
7) ***Purpose*** – **Debt consolidation and small business**
8) ***Earliest Credit Line*** – **For customers starting credit lines from 1985 to 2005**
9) ***Address State*** – **Customers from California and Florida**

"It is recommended for the lending club to be more cautious in providing loans to customers who fall under any of the above criteria"