LENDING CLUB CASE STUDY

SUBMITTED BY:

TANISHA SINHA

SUMANTH AN

WHAT IS LENDING CLUB?

LendingClub is a financial services company headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. At its height, LendingClub was the world's largest peer-to-peer lending platform.

Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

LCCS: PROBLEM STATEMENT

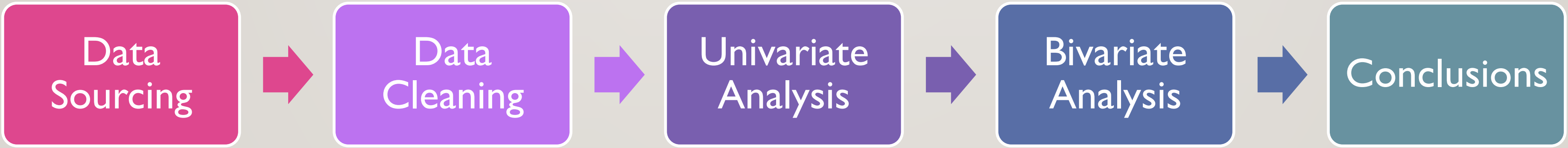
- You work for a **consumer finance company** which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:
- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

LCCS: BUSINESS OBJECTIVES

- The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- 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 utilise this knowledge for its portfolio and risk assessment.

LCCS: EXPLORATORY DATA ANALYSIS

Exploratory data analysis is the first and foremost step to analyze any kind of data. Rather than a specific set of procedures, EDA is an approach, or a philosophy, which seeks to explore the most important and often hidden patterns in a data set. In EDA, we explore the data and try to come up with a hypothesis about it which we can later test using hypothesis testing. Statisticians use it to take a bird’s eye view of the data and try to make some sense of it.



Data Sourcing → Data Cleaning → Univariate Analysis → Bivariate Analysis → Conclusions

LCCS: DATA SOURCING

Data Sourcing is the process of data collection from the different sources for a specific target and to achieve a specific goal.

For this case study, we are referring to a private source of data(csv file) which contains application/applicant details.

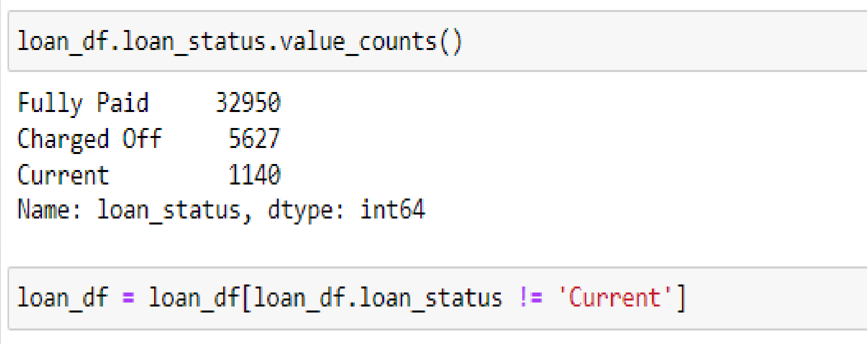
DATA CLEANING

Data cleaning in EDA refers to the process of identifying and correcting or removing inaccuracies, inconsistencies, and errors in a dataset. It involves activities such as handling missing values, handling outliers, identifying and removing duplicate data, correcting data types, resolving inconsistencies, and formatting data. The purpose of data cleaning is to improve the quality of the data and ensure that the analysis is based on accurate and reliable data.

LCCS: DATA CLEANING
DATA STANDARDIZATION

- We have 39717 rows and 111 columns in the dataset.
- Considering the loan_status column which indiicates the current status of loan has below 3 values:
  - Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
  - Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
  - Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

Removing the applicants whose loan status is current as we cannot absolutely determine if they are going to default or not and this might cause uncertainty in the result set.



```
loan_df.loan_status.value_counts()
```

```
Fully Paid     32950
Charged Off     5627
Current         1140
Name: loan_status, dtype: int64
```

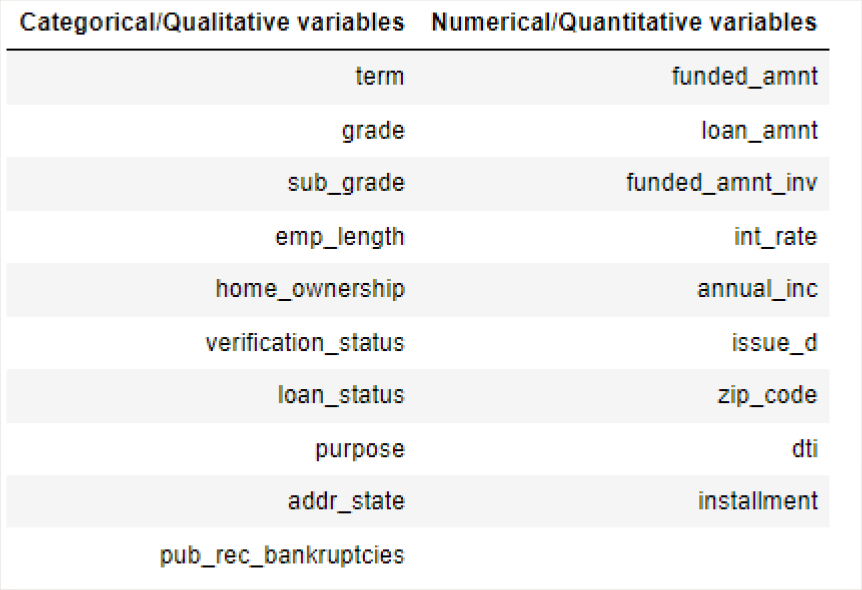
```
loan_df = loan_df[loan_df.loan_status != 'Current']
```

LCCS: DATA CLEANING
DATA STANDARDIZATION

- There are 21 customer behavior variables which are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.
- Removing such columns as they would not make an impact on the analysis.
- Dropping columns which have more than 30% of null values as imputing such data would generate a biased result. There are 58 such columns.
- Dropping columns which has only 1 unique value as it cannot be used as an indicator. There are 8 such columns.
- Dropping 3 columns with all unique value such as identification columns as it does not impact the analysis.
- Dropping rows with more than 50% null values. There are no such rows
- We are left with 38577 rows and 21 columns that will be used.
- Converted int_rate column to float by removing ‘%’ so that it can be used as numerical variable.

LCCS: VARIABLE TYPES

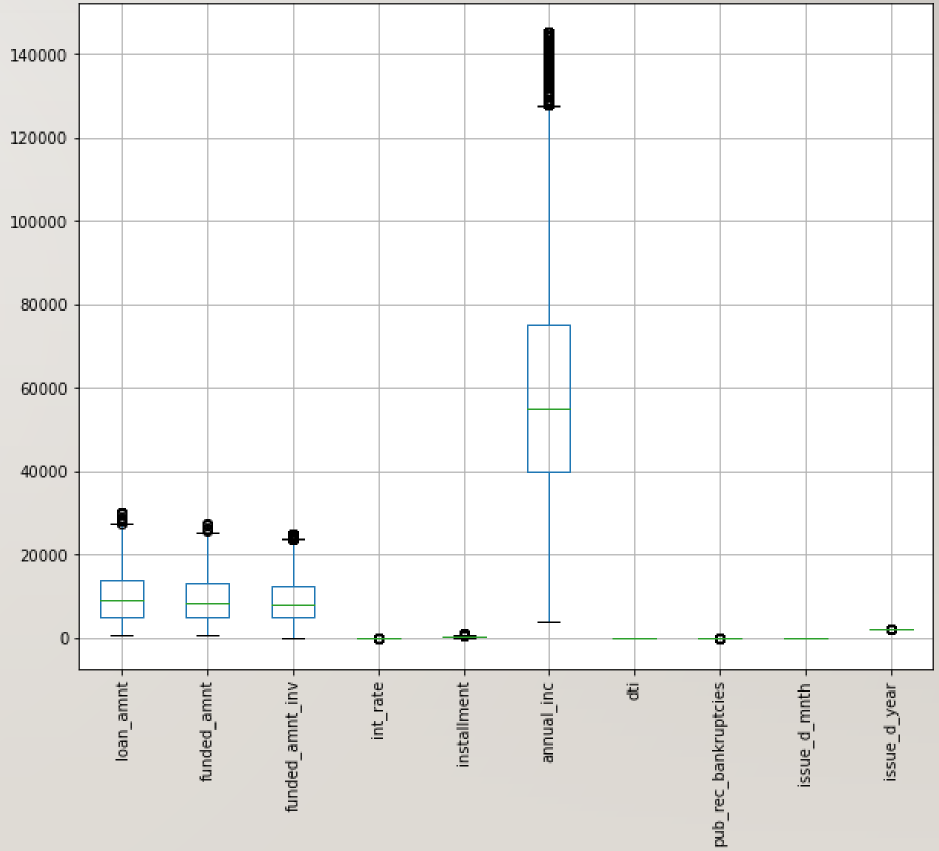
- Broadly we have 2 type of variables namely, ’Categorical’ and ’Numerical’.
- The dataset consists of 10 categoriacal and 9 numerical variables.



Categorical/Qualitative variables	Numerical/Quantitative variables
term	funded_amnt
grade	loan_amnt
sub_grade	funded_amnt_inv
emp_length	int_rate
home_ownership	annual_inc
verification_status	issue_d
loan_status	zip_code
purpose	dti
addr_state	installment
pub_rec_bankruptcies	

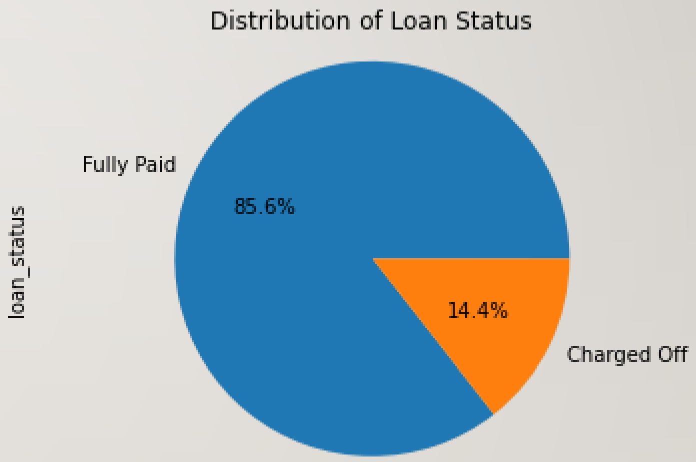
LCCS: UNIVARIATE ANALYSIS
REMOVAL OF OUTLIERS

- Using boxplot for data visualization of numerical data outliers.
- Outliers skew the data distribution and affects the accuracy and reliability of predictive models as they may lead to overfitting or underfitting of data.
- We can see outliers for annual_inc, loan_amnt, funded_amnt, funded_amnt_inv.
- Using the upper and lower inner fence to remove the outliers.



LCCS: UNIVARIATE ANALYSIS DISTRIBUTION OF LOAN STATUS

- The distribution chart shows us that 14.4% of loan applications are charged off as compared to 85.6% of full paid loan applications.



LCCS: UNIVARIATE ANALYSIS
CATEGORICAL VARIABLE SUBPLOTS

The analysis from above subplots is as follows:

1. Most applicants are likely to apply for a loan term of 36 months.

2. Applicants of Grade B and A are more likely to apply for loan followed by C, D, E and F. Grade G applicants are least likely to apply for a loan.

3. Applicants with more than 10 years of experience and most likely to apply for the loan.

4. debt_consolidation is the most frequently used purpose for applying for loans.

5. Of all the applications most of the applications are not verified which indicates that most applications source of income was not verified by LC.

6. Most applicants with rented or mortgage property apply for loan rather than the one's owning the property.

7. Applicants with 0 public record bankruptcies have the maximum number of applications.

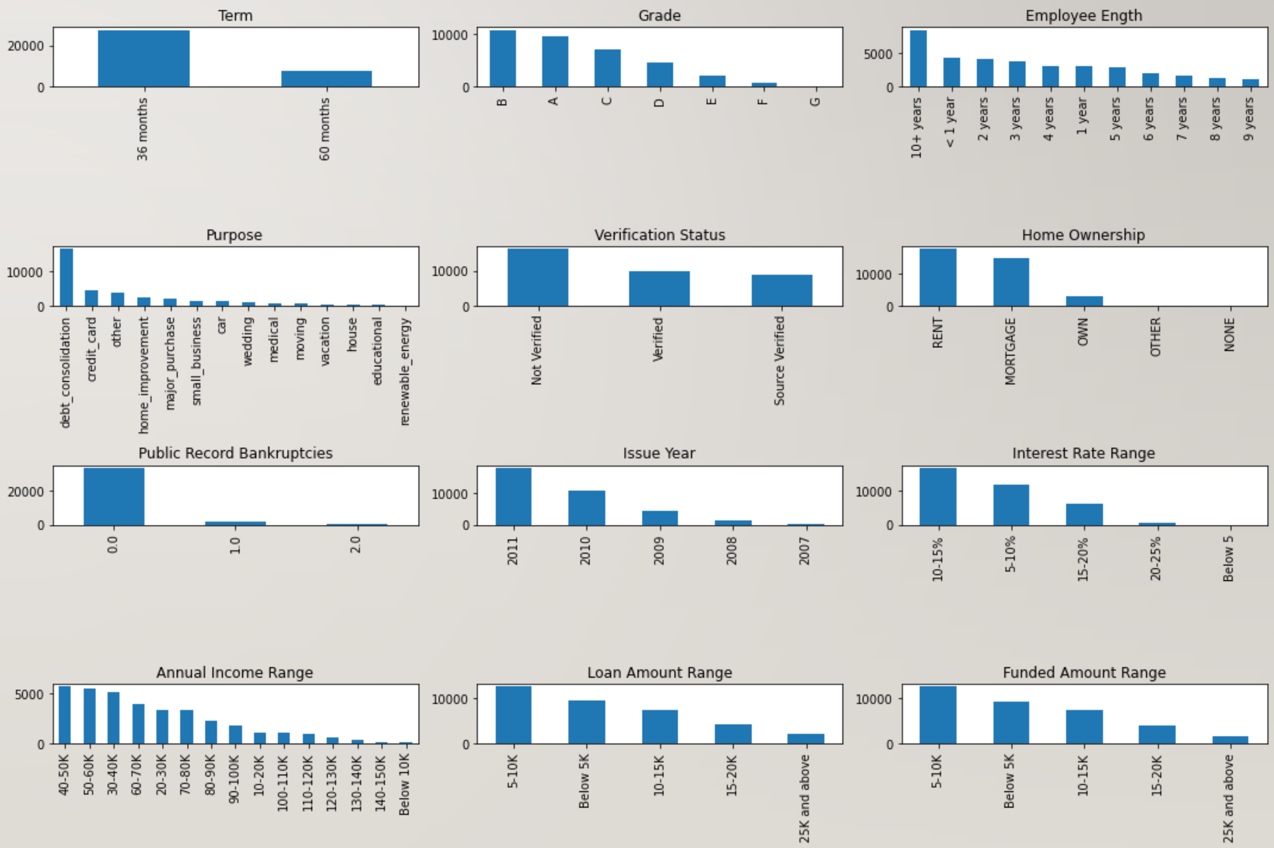
8. Maximum applications are from the year 2011.

9. Maximum applications have been applied with interest rate range 10-15%.

10. The applicants with annual income range 30K to 60 K mostly apply for loans.

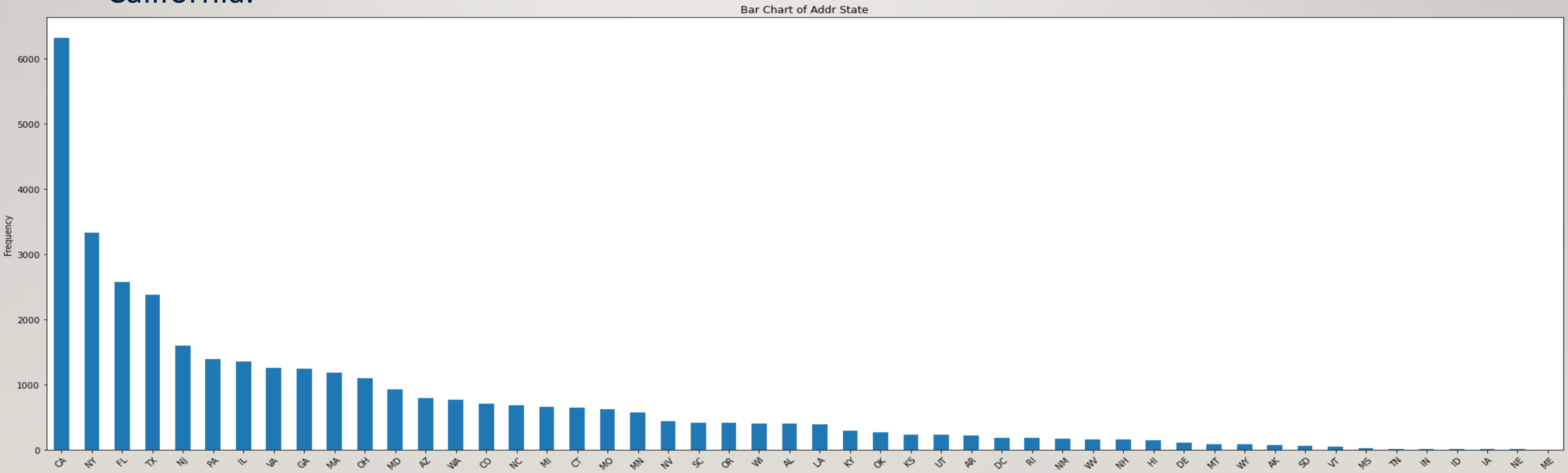
11. The frequency of application of loan amount 5-10K is the highest.

12. The frequency of application of funded amount 5-10K is the highest.



LCCS: UNIVARIATE ANALYSIS FREQUENCY OF APPLICATIONS BY ADDR_STATE

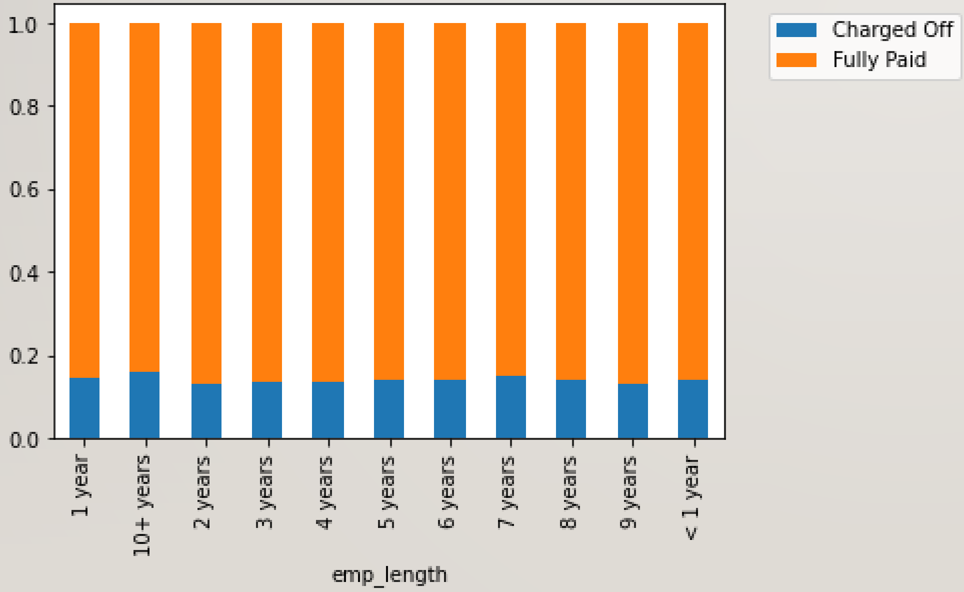
- From bar chart, we can see clearly see a spike in the number of applications from the state California.



LCCS: BIVARIATE ANALYSIS
IMPACT OF THE GRADE ON LOAN STATUS

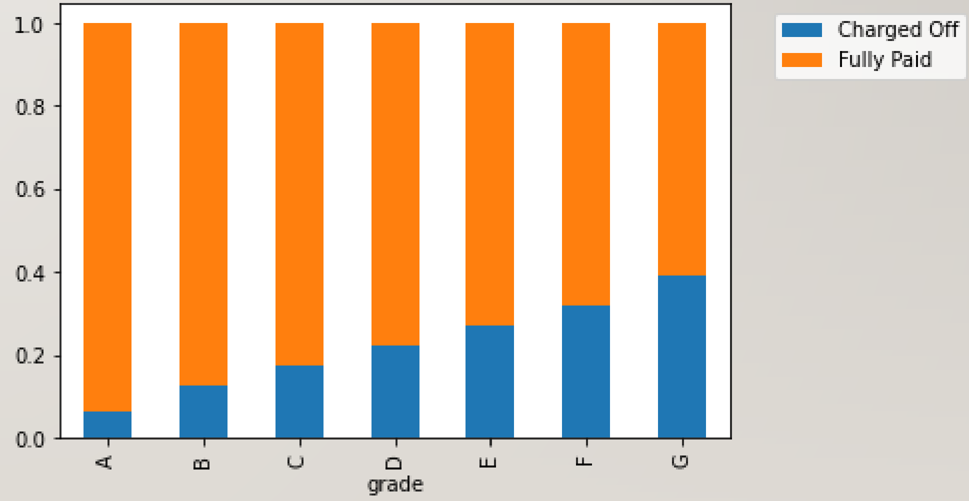
- The number of applications are more from grade B which also has the highest charged off count.

Comparison of Employee length against loan status(normalized)



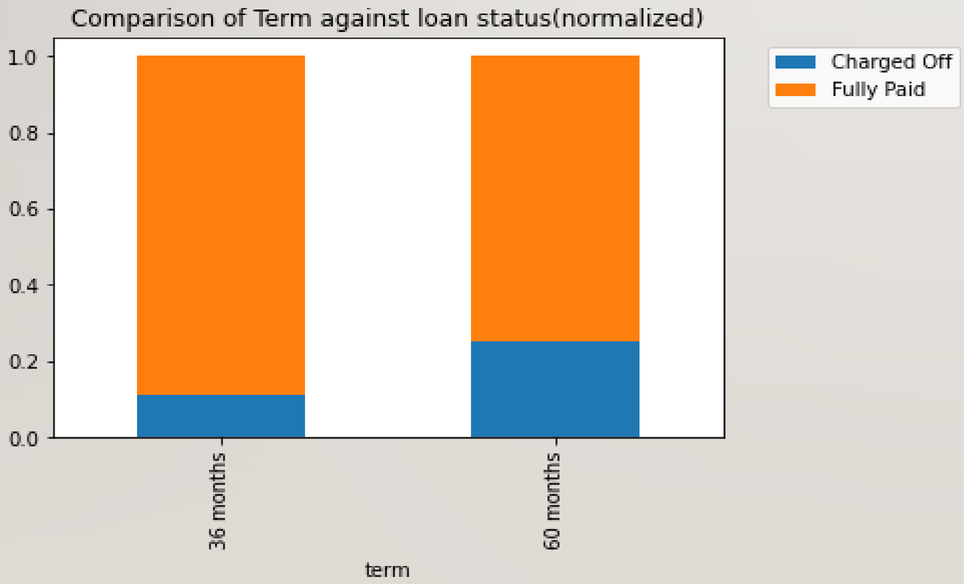
- However, when the data is normalized , the chart indicates that the percentage of applicants likely to charge off is more for grade G followed by F, E, D, C, B, A.

Comparison of Grade against loan status(normalized)

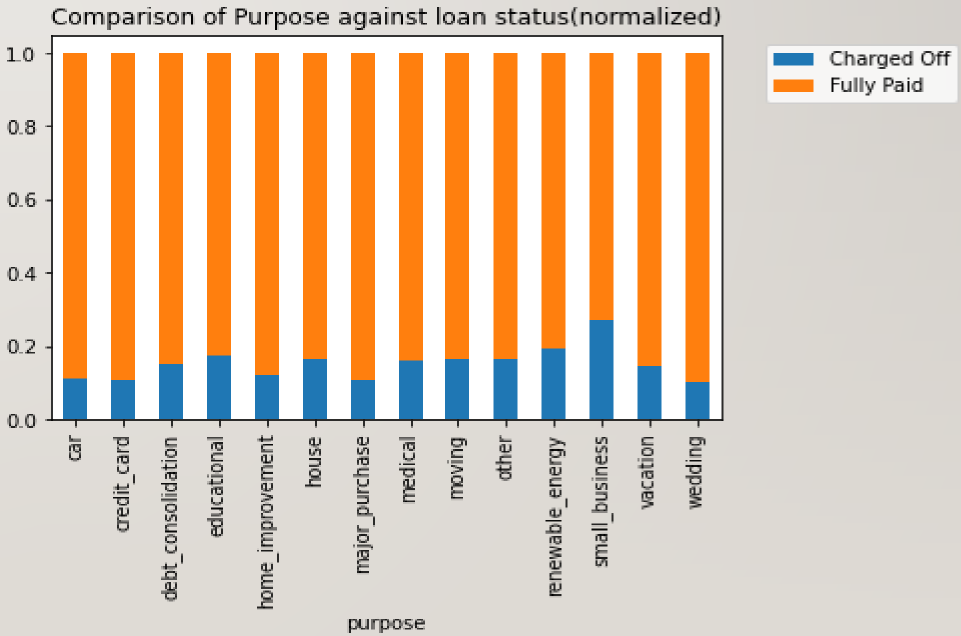


LCCS: BIVARIATE ANALYSIS
IMPACT OF THE TERM AND PURPOSE ON LOAN STATUS

- The bar chart indicates that when the data is normalized, term or duration of loan is less applicants are less likely to default.

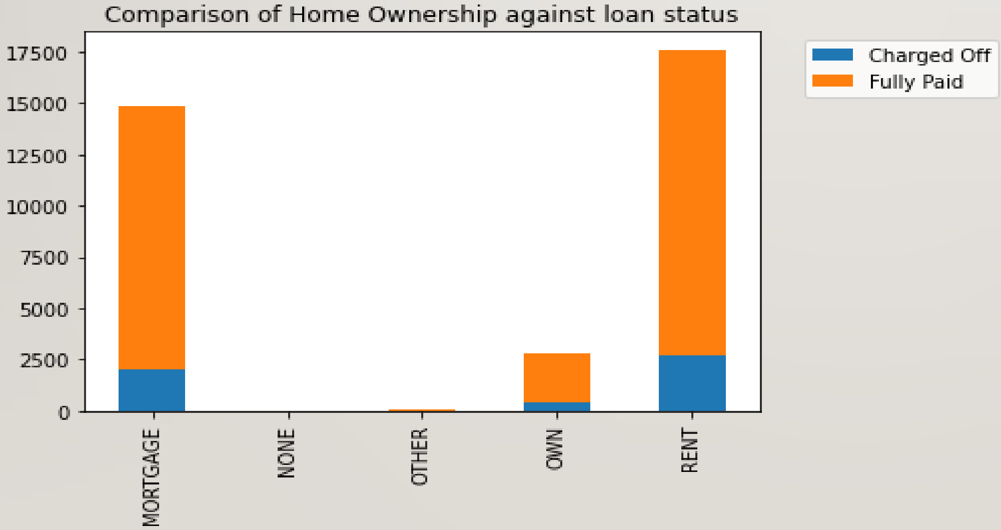


- The bar chart indicates that when the data is normalized, small business applicants are most likely to default.

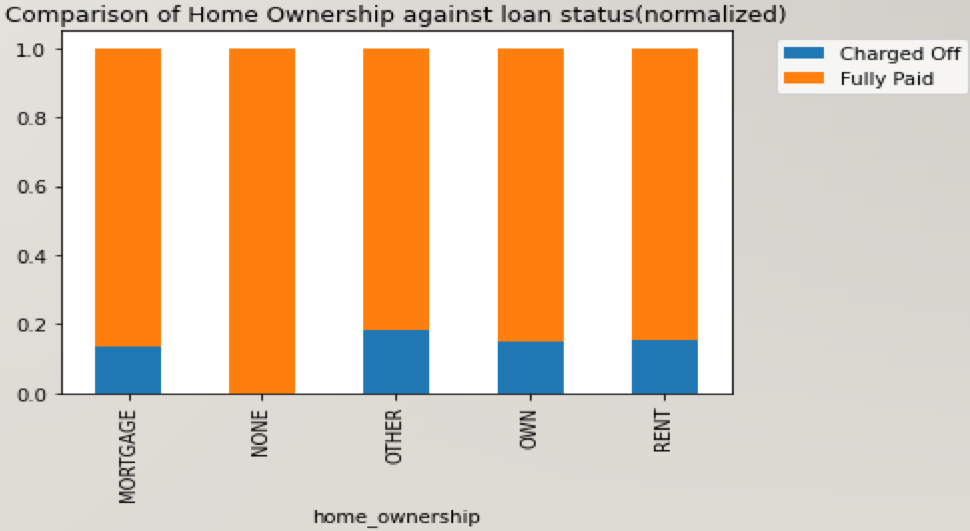


LCCS: BIVARIATE ANALYSIS
IMPACT OF THE HOME OWNERSHIP ON LOAN STATUS

- The bar chart specifies that applicants who rented or mortgaged properties are more likely to default.



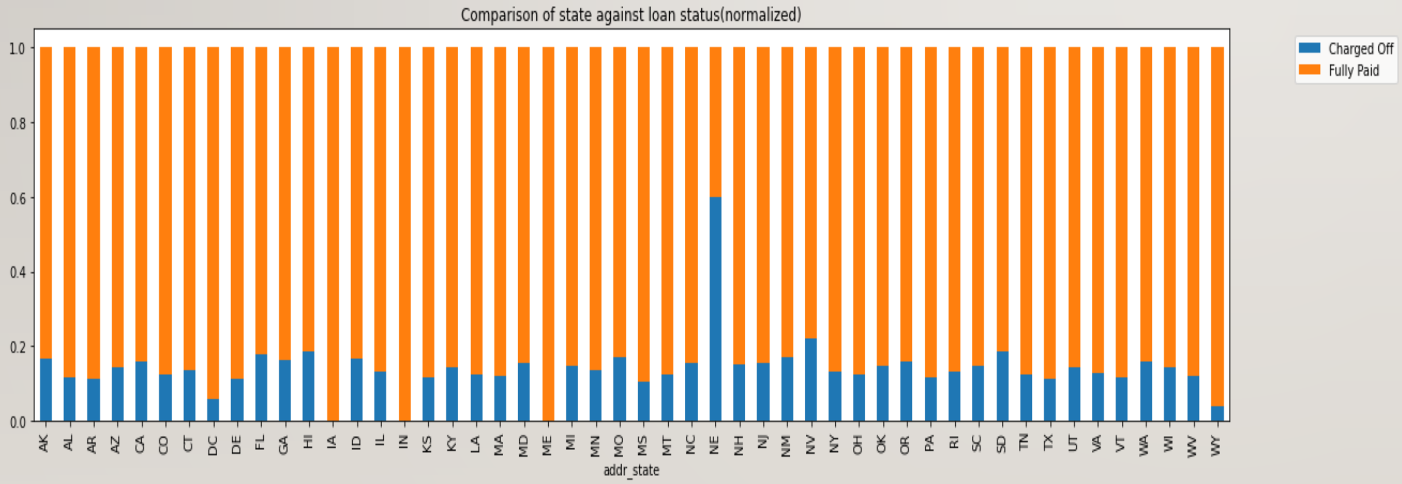
- The bar chart indicates that when the data is normalized, specifies that applicants who have ownership of "others" are most likely to default followed by applicants who rent and own house. Applicants who have no home ownership are least likely to default.



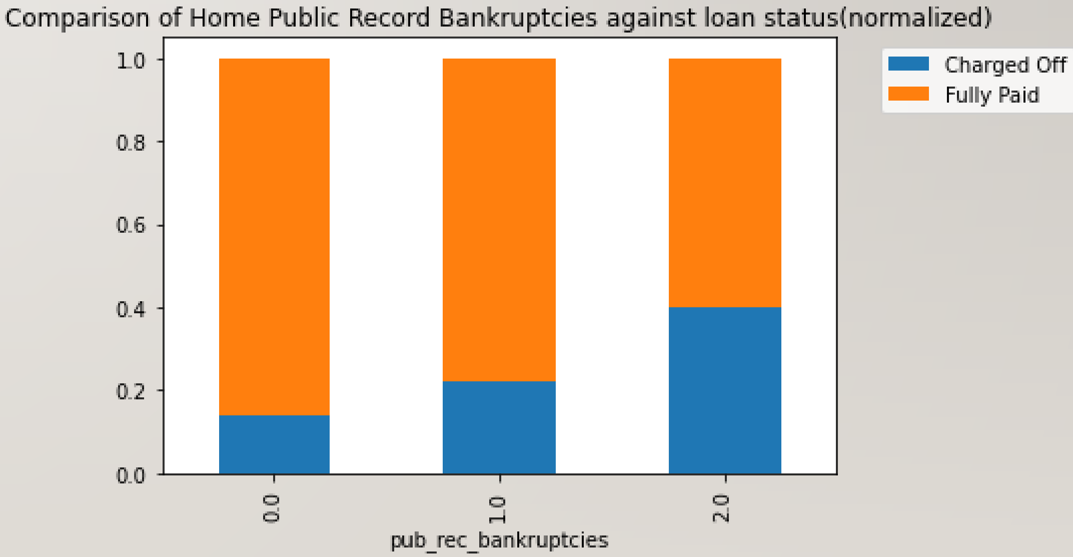
LCCS: BIVARIATE ANALYSIS

IMPACT OF THE STATE AND BANKRUPTCIES ON LOAN STATUS

- The bar chart suggest that most applicants from the state of Nebraska(NE) are most likely to default.



- The bar chart indicates that applicants with 2 bankruptcies are more likely to default. There is also a trend that shows that the number of bankruptcies is positively correlated to loans being charged off.

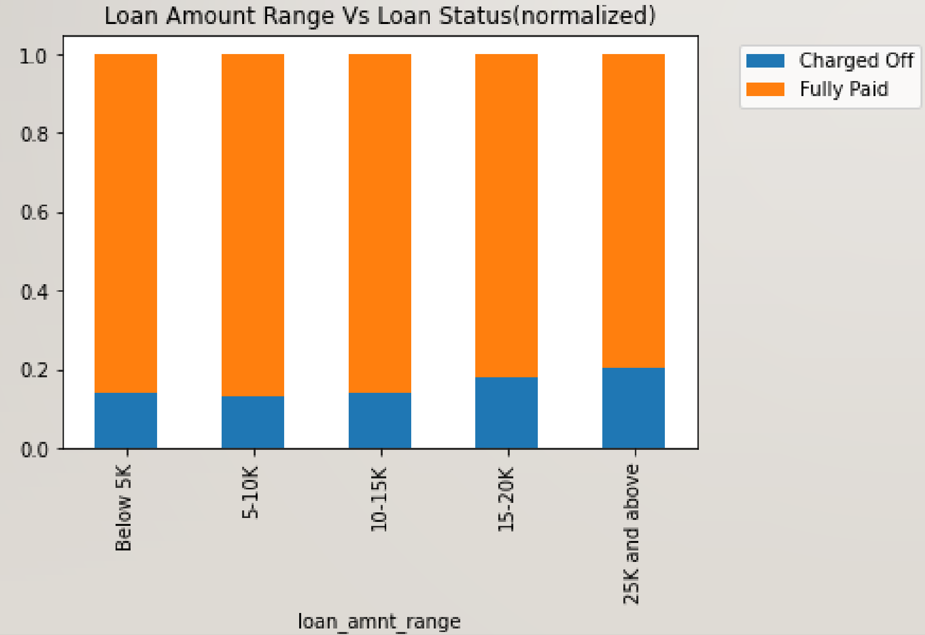


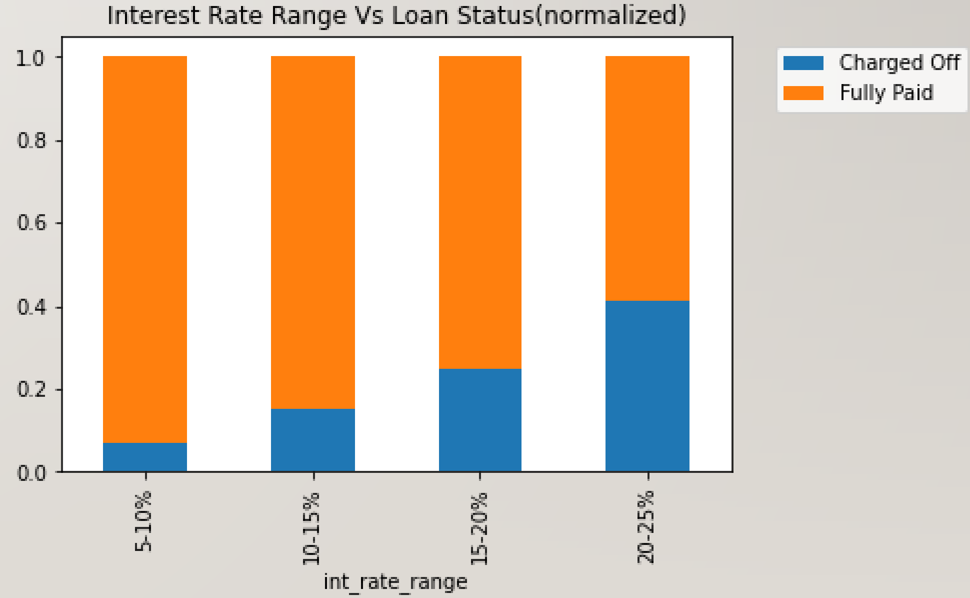
LCCS: SEGMENTED BIVARIATE ANALYSIS

IMPACT OF LOAN AMOUNT RANGE AND INTEREST RATE RANGE ON LOAN STATUS

- The bar chart indicates that most charged off loans lie in the range of loan amount which are above 25K.

- The bar chart indicates that most charged off loans lie in the range of interest which lie between 20-25%.



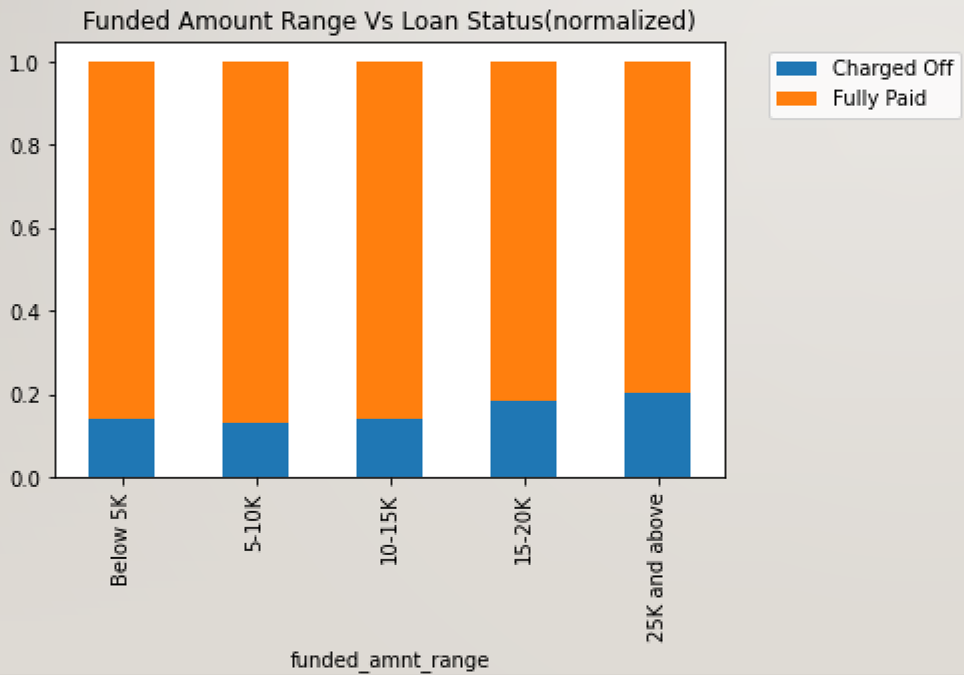


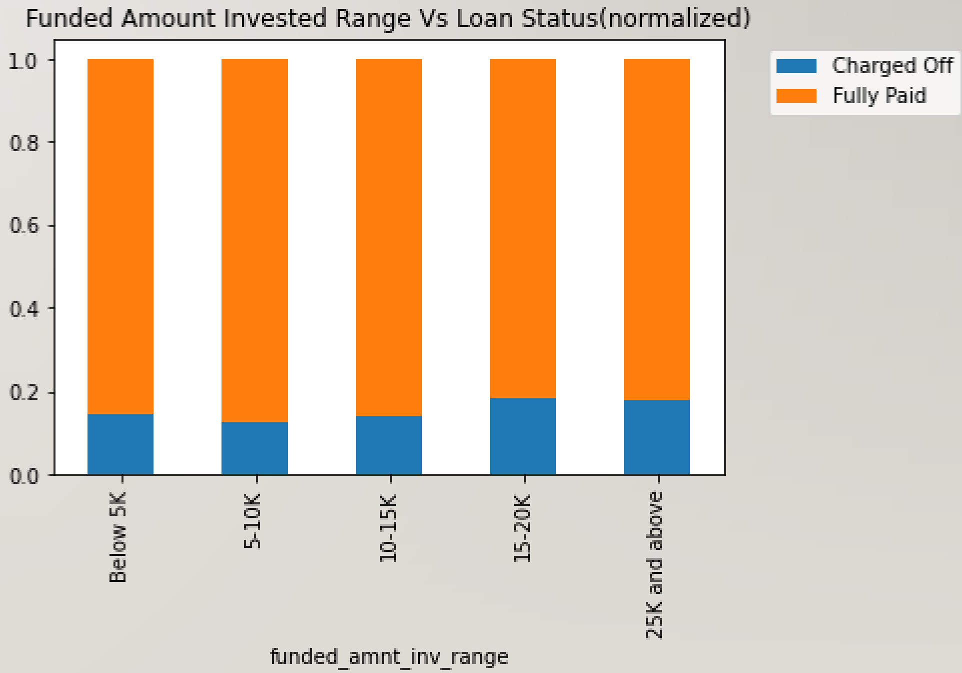
LCCS: SEGMENTED BIVARIATE ANALYSIS

IMPACT OF FUNDED AMOUNT RANGE AND FUNDED AMOUNT INVESTED RANGE ON LOAN STATUS

- The bar chart indicates that that most charged off loans lie in the range of funded amount which are above 25K.

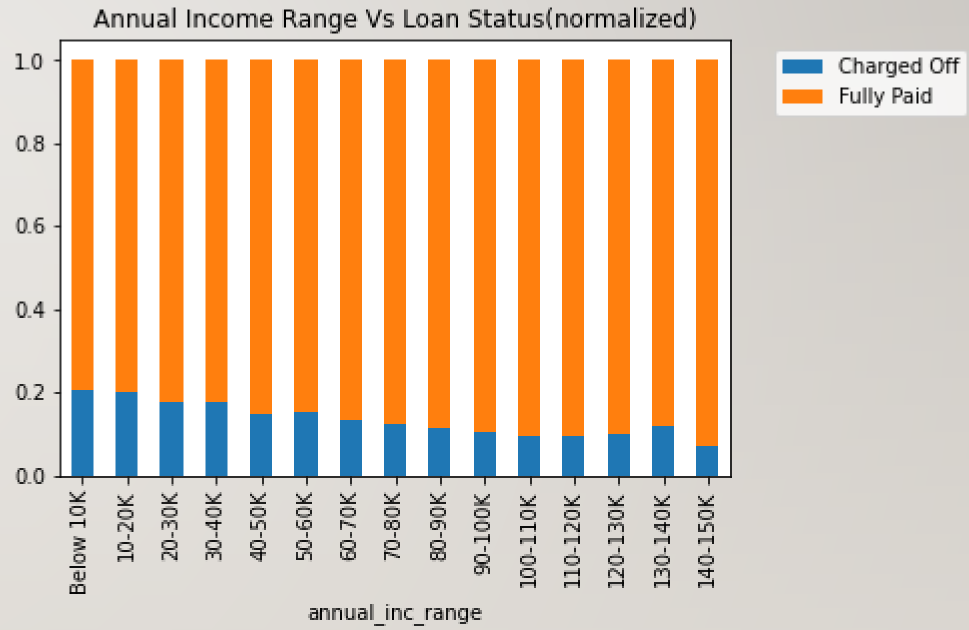
- The bar chart that most charged off loans lie in the range of funded amount invested which are above 25K.





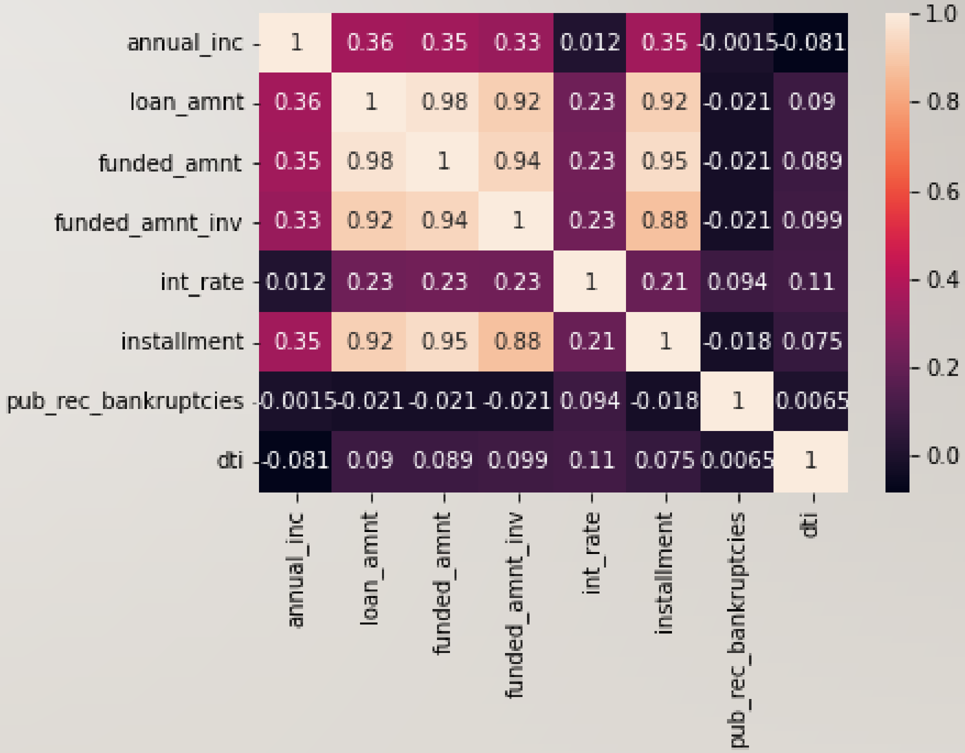
LCCS: SEGMENTED BIVARIATE ANALYSIS
IMPACT OF ANNUAL INCOME RANGE ON LOAN STATUS

- The bar chart indicates that most charged off loans are similar and lie in the range of annual income which are below 10K and 10K-20K.
- Also, applicants with higher annual income have lower chances of defaulting



LCCS: CORRELATION HEATMAP

- A correlation matrix is a table showing correlation coefficients between variables.
- The values can range between -1 and 1, the higher the value the more closely the variable are related.
- Heatmap is graphical representation of correlation matrix.
- There is a high positive correlation between loan_amnt and funded_amnt
- pub_rec_bankruptcies have a negative impact on loan_amnt



	annual_inc	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	pub_rec_bankruptcies	dti
annual_inc	1	0.36	0.35	0.33	0.012	0.35	-0.0015	-0.081
loan_amnt	0.36	1	0.98	0.92	0.23	0.92	-0.021	0.09
funded_amnt	0.35	0.98	1	0.94	0.23	0.95	-0.021	0.089
funded_amnt_inv	0.33	0.92	0.94	1	0.23	0.88	-0.021	0.099
int_rate	0.012	0.23	0.23	0.23	1	0.21	0.094	0.11
installment	0.35	0.92	0.95	0.88	0.21	1	-0.018	0.075
pub_rec_bankruptcies	-0.0015	-0.021	-0.021	-0.021	0.094	-0.018	1	0.0065
dti	-0.081	0.09	0.089	0.099	0.11	0.075	0.0065	1

LCCS: RECOMENDATIONS

- Applicatants who fall under grade ’G’ have a higher chance of defaulting which can be attributed to high interest rate.
- Small business applicants are more prone to defaulting
- Even if the number of application are more for 36 months, there seems to be a higher percentage of defaulters in 60 months tenure. Thus, higher the loan term, higher the chances of defaulting.
- There seems to be a larger number of charged off applicants who rent a house against other values but there is a higher percentage of charged off applicants who have ‘other’ ownership.
- Loan amount, funded amount and funded amount invested are closely related. Hence, the risk increases if the value for these variables increase.
- Annual income in inversely proportional to loan status indicating that applicants with higher annual income are less likely to default.
- Applicants who have higher records of bankruptcies tend to not pay the loan in full.

THANK YOU!