Lending Club Case Study

LendingClub

Key Insights of EDA

Author 1: Tejasvi Chandola

Author 2: Deepthi Shreeya

Date: 23-07-2024

Problem Statement:

1 Business Understanding:

You work for a consumer finance company which specialises in lending various types of loans to urban customers. When a person applies for a loan, there are two types of decisions that could be taken by the company:

A. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

1. Fully paid: Applicant has fully paid the loan
2. Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed.
3. 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

B. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

2 Business Objectives:

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.

Data Loading and Understanding:

A. Data Loading

Loaded the dataset using pandas.

B. Key Statistics

Number of rows: 39,717

Number of columns: 111

Dataset Used:

loan.csv

Data_Dictionary.xlsx

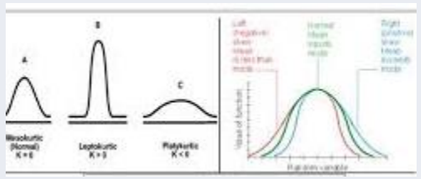
Modules of the Project:

1. Data Cleaning & Identifying Outliers

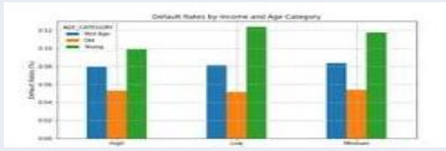
2. Analysing Data Imbalance

3. Creating Derived Metrics

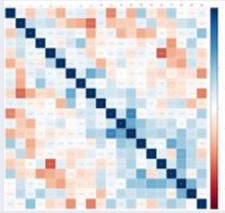
4. EDA-Univariate Analysis Insights



5. EDA-Segmented Univariate Analysis Insights



6. EDA-Bivariate Analysis Insights



8. Results

9.Recommendations

Data Cleaning & Identifying Outliers :

1. Initial Data Inspection- Displayed the first few rows of the dataset and observed various data types including floats, integers, and objects.

2. Missing Value Check- Checked for missing values in the dataset and identified columns with a significant number of missing values.

3. Column Removal- Removed those Columns which had more than 30% of missing values or were irrelevant to the problem statement (i.e. came into existence after loan approval or was not related to the borrower)

4. Row Removal- Duplicate rows were removed.

5. Missing Value Treatment- Null values were replaced with Median(For Numerical Columns) and Mode (For Categorical Columns)

6. Converted Data types- Converted the data types of Interest Rate and Revolving Line Utilization Rate Columns' values from String to float type using astype() function and removed the '%' using rstrip() function.

7. Converted date columns to datetime format- Converted date columns to datetime format using to_datetime( ) function.

8. Performed Sanity Check on the data- Outliers were removed from the Annual Income and the revolving balance Columns and the Column "Account Type" was removed as it has a single value throughout.

9. Removed some instances where categorical value counts are really low compared to total population-

(i) The data about the following states were removed from the dataset as their value counts were really low compared to the total population-

'MT', 'WY', 'AK', 'SD', 'VT', 'MS', 'TN', 'IN', 'ID', 'IA', 'NE', 'ME'

(ii) Some values which were 'NONE' and 'Other' under the Home Ownership Column were also removed.

EDA – Uni-variate Analysis Insights:

- Univariate Analysis was performed on the dataset
- Analysed Data Imbalance
- Created Derived Metrics like ‘delinquency_in_2years’, 'bankruptcies_occured’, 'month’, ‘Year’

Insights

1. More than 80% of the loans were fully paid.
2. Most of the loans were taken on month of November and least number of loans were taken in Feb.
3. Loans are almost getting doubled year after another 2007 having least percentage of loans whereas 2011 have highest volume. This implies that year after another this company is growing.
4. More than 95% of folks have no records of bankruptcies and nearly 90% have no previous delinquency.
5. Majority of folks applying for loan are from CA, NY.
6. Most of the loans are for paying other debts, after that loans taken are for home and business improvement.
7. More than 40% of the borrowers are not verified, some are verified and very few are sourced verified.
8. People living in rented and mortgage houses have majority of loans.

9. People working for more than 10 years have highest percentage of loan(~25%), People working for <1, 2 and 3 Years contribute for 33% of the loans. This may be because they are borrowing money for startups or career growth.

10. More than 73 percent borrowers takes loan on term of 36 months, this might be to save the interest on funded amount.

11. Loan Amount, Funded amount and Committed amount by investors follow a similar distribution, Mostly values of these three variables lies between $5000 to $15000, this may represent that investors are investing their money in this company and most of the amount Is fulfilled by investors only. Further insights will be seen in bivariate.

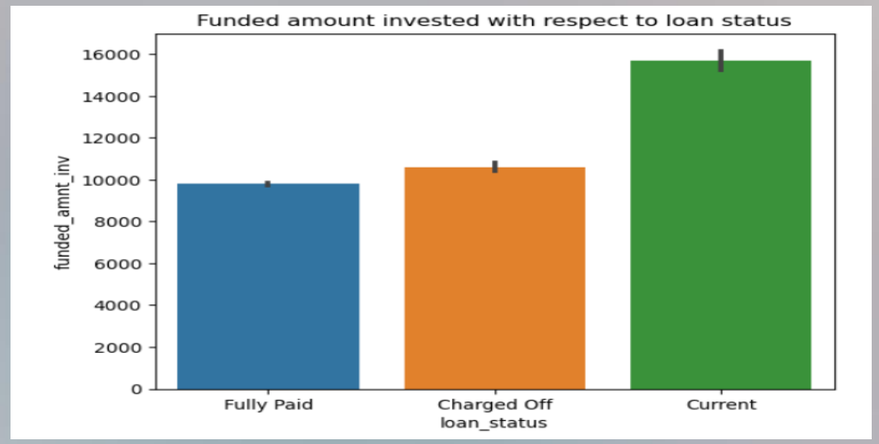
12. Most of the borrowers having an annual salary between $50000 and $70000 are coming to this organization for loan where peak lies around $50000, very few investors have annual salary greater than $200000

EDA –Segmented Uni-variate Analysis Insights:

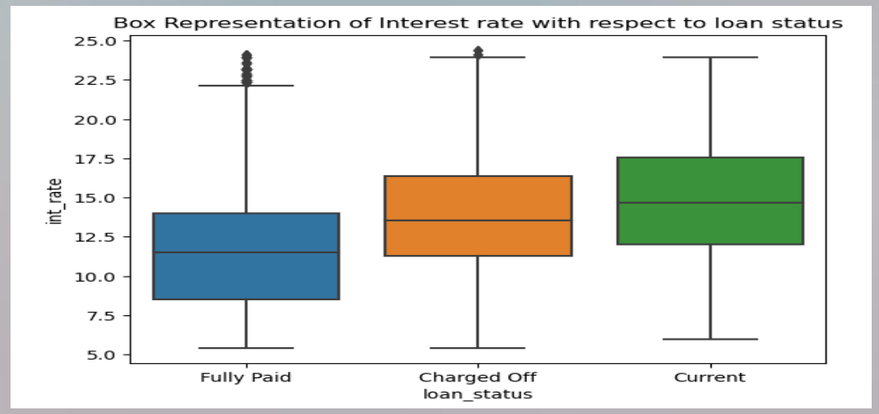
Insights

- Charged off borrowers have a higher funded amount, interest rates and installments compared to fully paid ones.

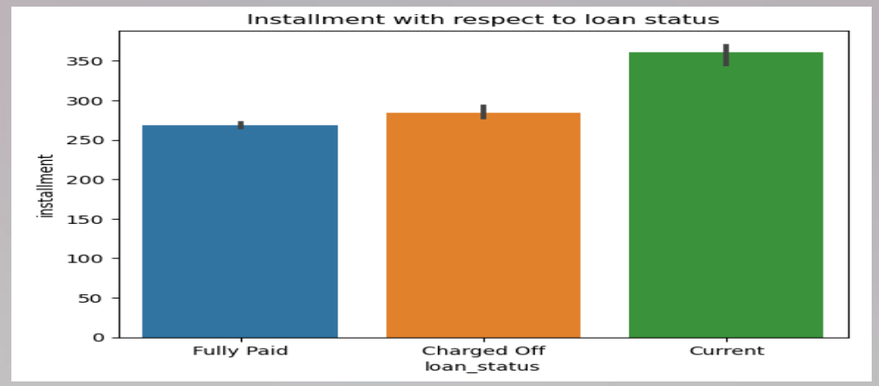
- Bar Plot of. Funded Amount by Investors w.r.t. Loan Status



- Box Plot of Interest Rates w.r.t. Loan Status.



- Bar Plot of Installment w.r.t. Loan Status.



EDA –Segmented Uni-variate Analysis Insights:

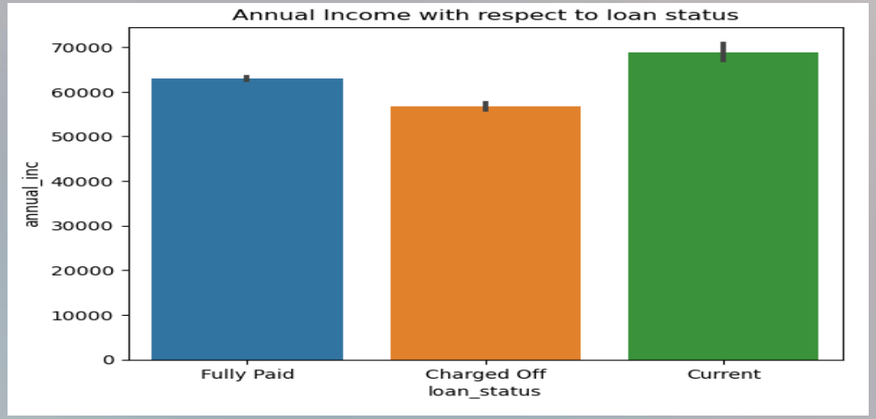
Insights

- Higher income depicts lower chances of defaulting .

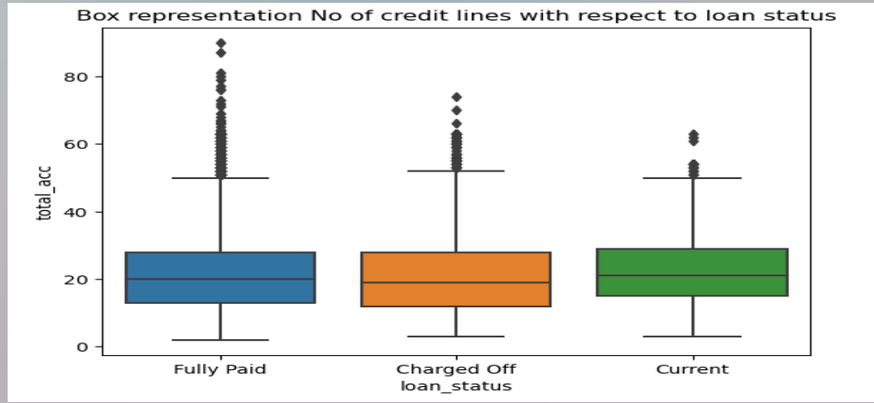
- Fully paid borrowers have a slightly higher number of credit lines in file compared to charged off borrowers. Also, the number of outliers are more in the case of Fully Paid borrowers. They might take more loan but pay their debt.

- Customers with Higher Debt-to-Income Ratio are more likely to be defaulters.

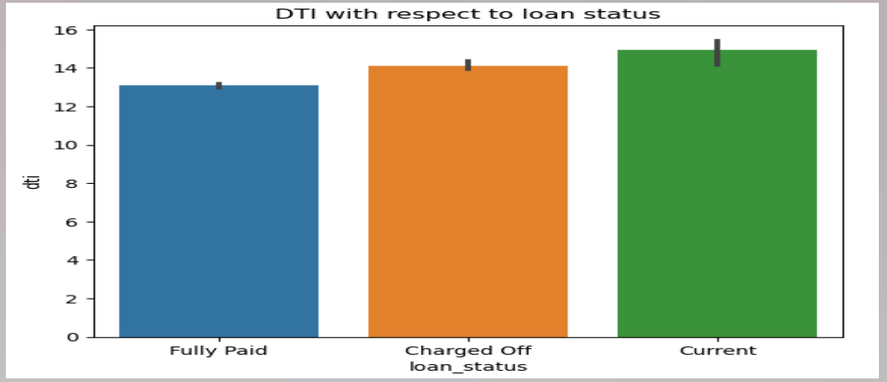
- Bar Plot of Annual Income w.r.t. Loan Status.



- Box Plot of Number of Credit Lines w.r.t. Loan Status.



- Bar Plot of Debt-to-Income Ratio w.r.t. Loan Status.



EDA –Segmented Uni-variate Analysis Insights:

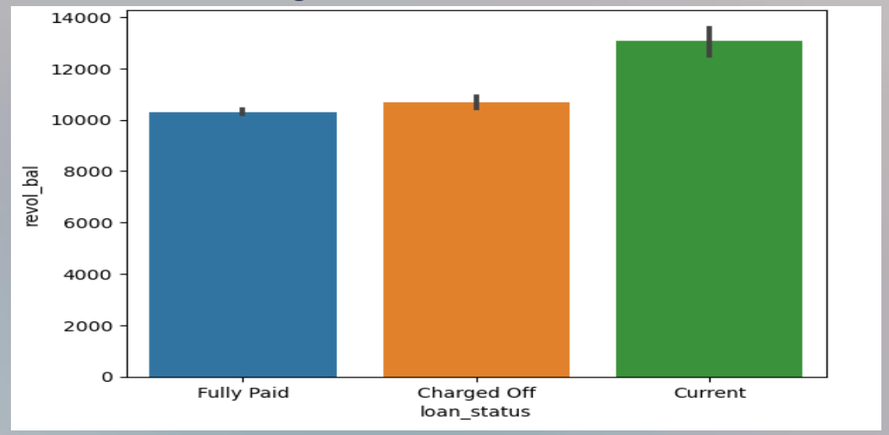
Insights

- Borrowers who are likely to default have slightly higher average revolving balance than fully paid ones.

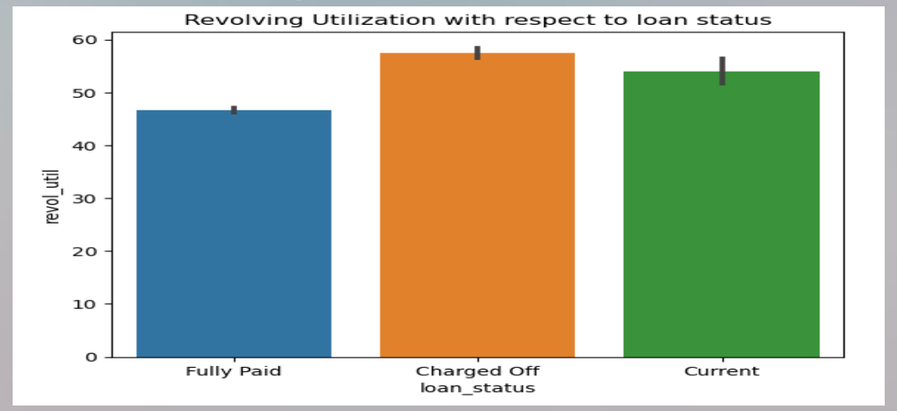
- Customers with Higher value Revolving Line Utilization rates are more likely to be defaulters.

- Customers with a term of 60 months are more likely to default than customers with a term of 36 months.

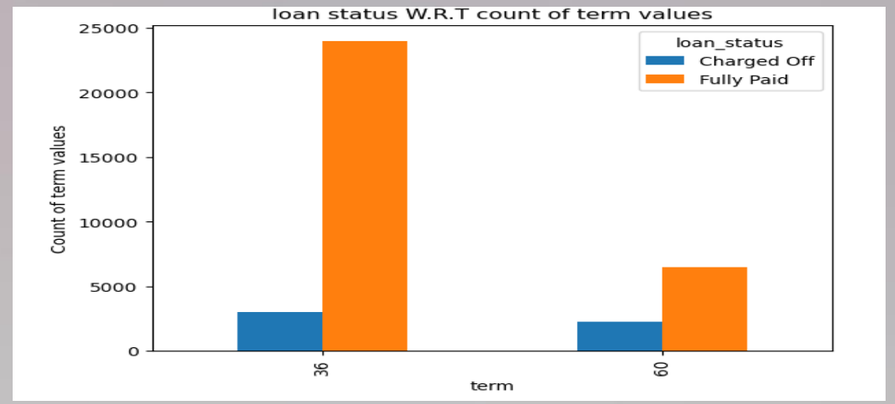
- Bar Plot of Revolving Balance w.r.t. Loan Status.



- Bar Plot of Revolving line utilization rates w.r.t. Loan Status.



- Bar Plot of Loan Status w.r.t Term.



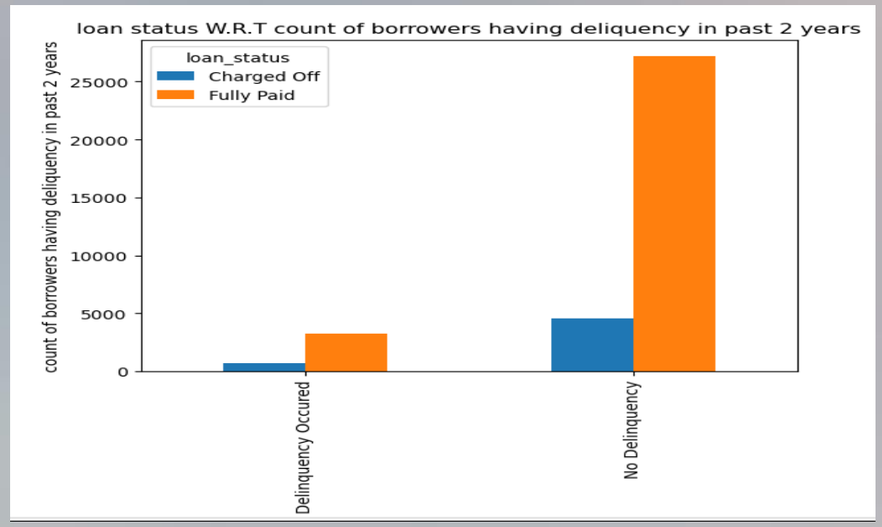
EDA –Segmented Uni-variate Analysis Insights:

Insights

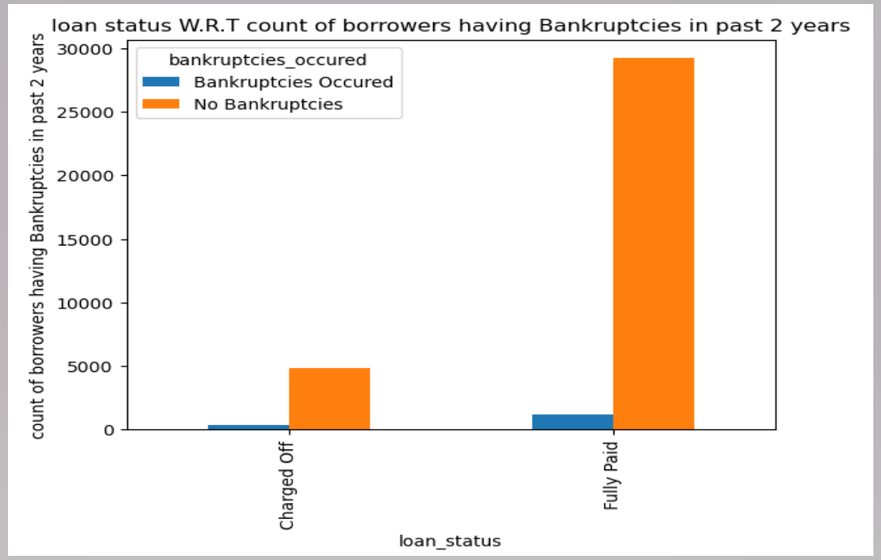
- People with past Delinquency has charged off to fully paid ratio of 20%, which is higher than that of non-delinquent borrowers with small margin.

- People with earlier Bankruptcies records are more like to be charged off that non-bankrupt borrowers with a small margin.

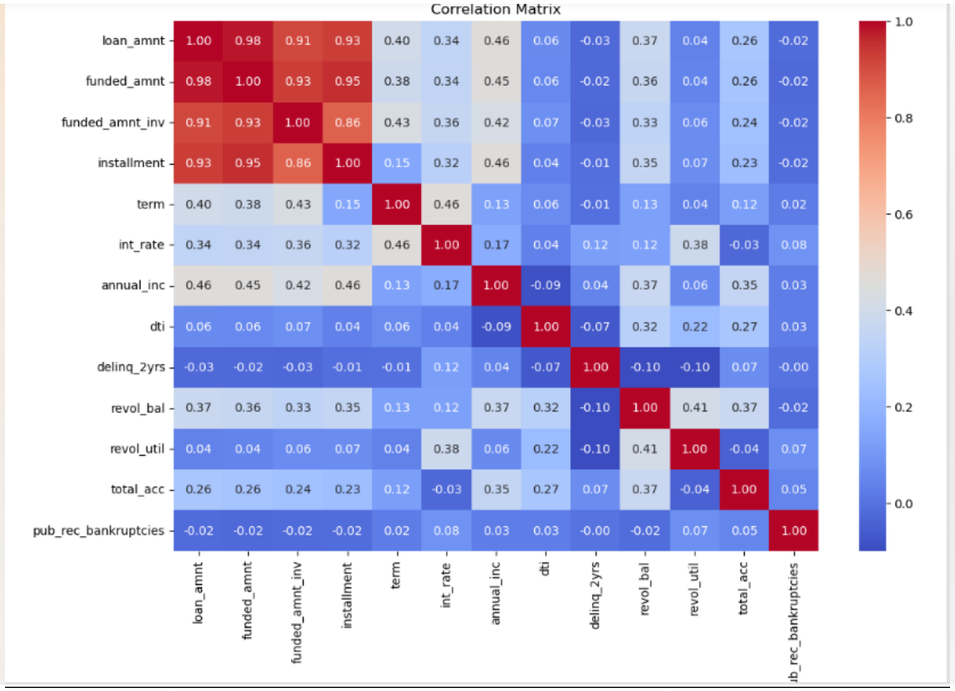
- Bar Plot of Loan Status w.r.t. Borrowers having delinquency in past 2 years



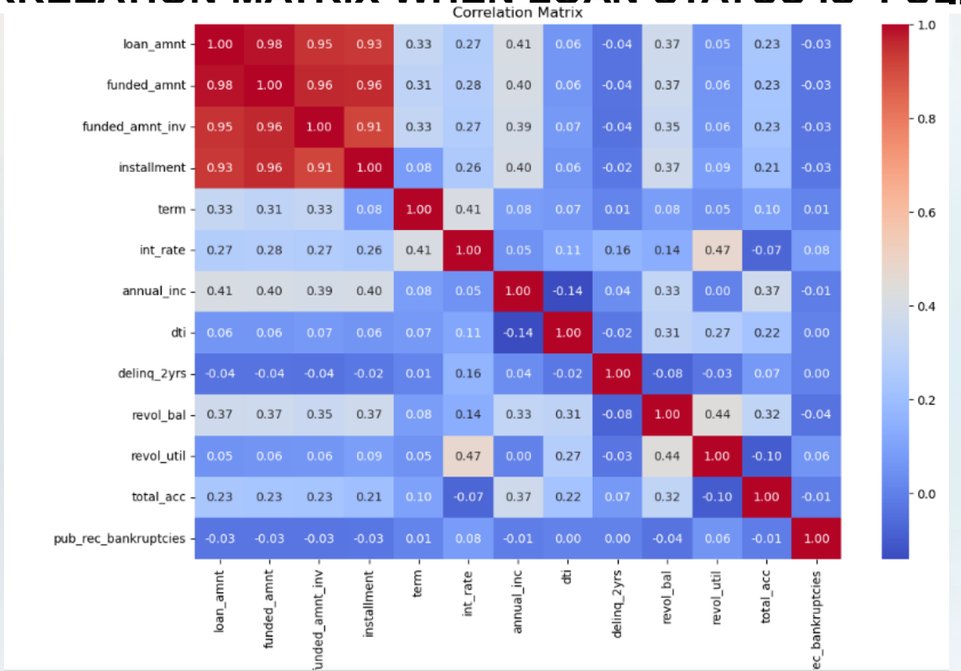
- Bar Plot of Loan Status w.r.t. Borrowers having bankruptcies in the past 2 years.



1) CORRELATION MATRIX WHEN LOAN STATUS IS 'CHARGED-OFF'



2) CORRELATION MATRIX WHEN LOAN STATUS IS 'FULLY-PAID'



EDA – Bi-variate Analysis Insights:

Key Comparisons:

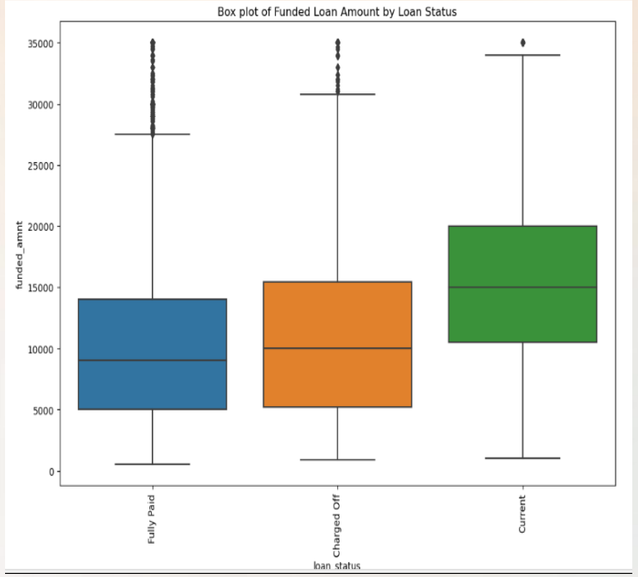
Consistency in High Correlations: Both fully paid and charged-off loans show consistent high correlations between loan amount, funded amount, funded amount by investors and instalment, indicating that the fundamental loan structure remains consistent regardless of loan status.

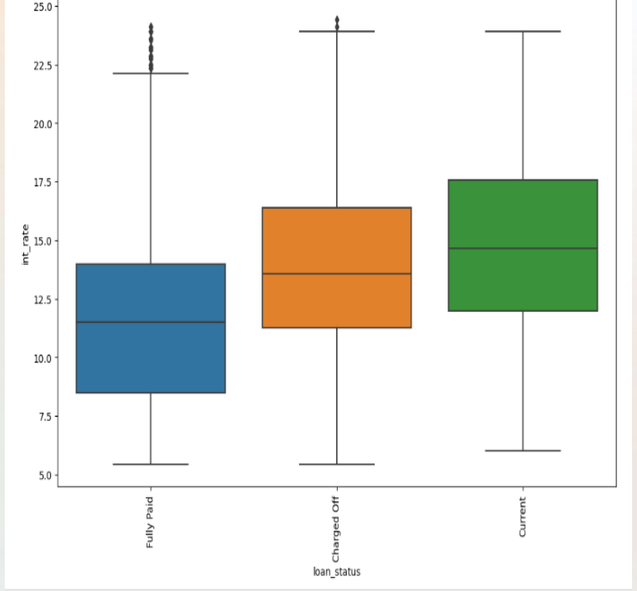
Term and Interest Rate: The correlation between term and interest rate is slightly higher in charged-off (0.46) compared to fully paid loans (0.41), suggesting that longer terms might be slightly more associated with higher interest rates in charged-off loans.

Revolving Balance and Utilization: The correlation is stronger in fully paid loans (0.44) than in charged-off loans (0.41), indicating that higher revolving balances are more critical in fully paid loans.

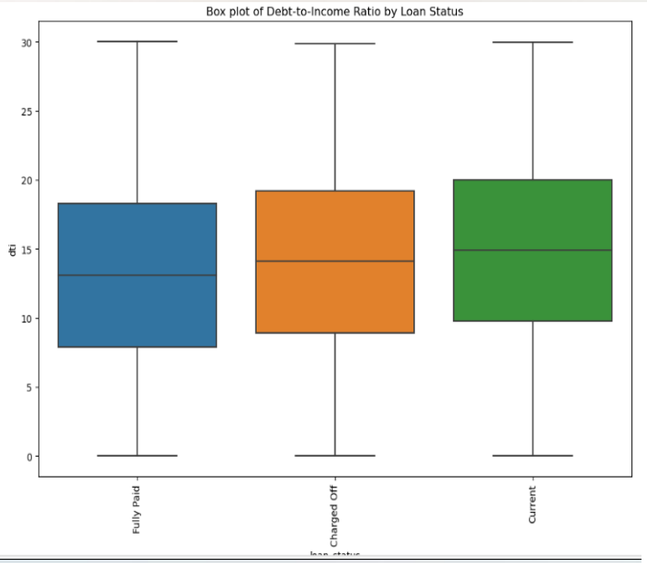
Annual Income: The moderate correlation between annual income and loan amount is stronger in charged-off loans (0.46) compared to fully paid loans (0.41), suggesting that higher income borrowers tend to take larger loans and are more likely to not repay them.

1)Box Plot of Funded Amt. by Loan Status

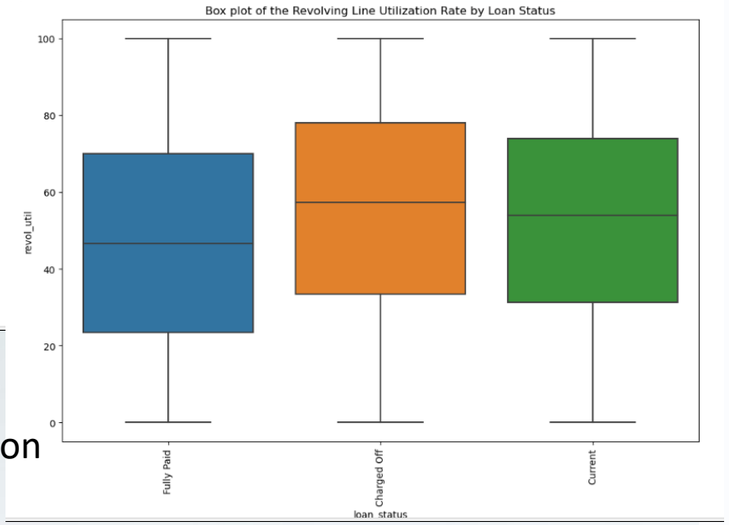




2) Box Plot of Interest Rate by Loan Status



3) Box Plot of Debt-to-Income Ratio by Loan Status



4) Box Plot of Revolving Utilization Rates by Loan Status

EDA – Bi-variate Analysis Insights:

1) "Charged Off" loans received ***more funded loan amounts*** compared to "Fully Paid" Loans and the ***variability in funded loan amounts is also more*** in "Charged Off" loans as compared to "Fully Paid" Loans. The presence of ***more frequent and higher outliers*** in "Fully Paid" Loans indicate a significant shift in lending practices or borrower behaviour for "Fully-Paid" Loans.

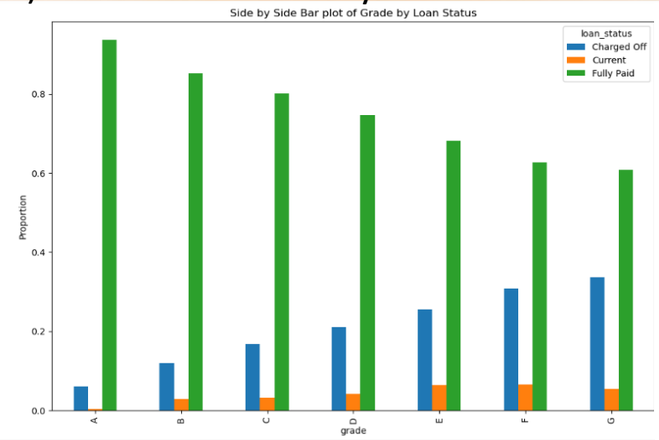
2) "Charged Off" loans had ***higher interest rates*** compared to "Fully Paid" Loans.

3) "Charged Off" loans had ***slightly higher Debt-to-Income Ratio*** compared to "Fully Paid" Loans.

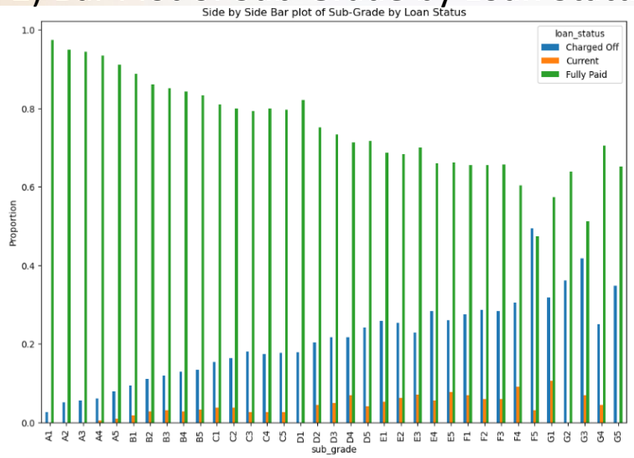
4) "Charged Off" loans had ***higher Revolving Line Utilization Rates*** compared to "Fully Paid" Loans.

EDA – Bi-variate Analysis Insights:

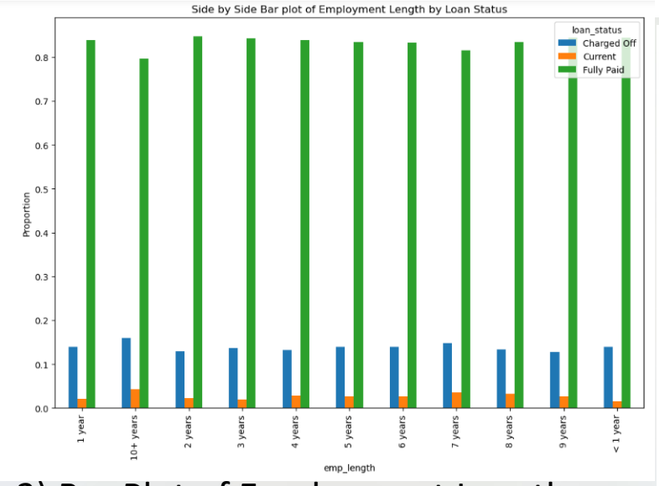
1) Bar Plot of Grade by Loan Status



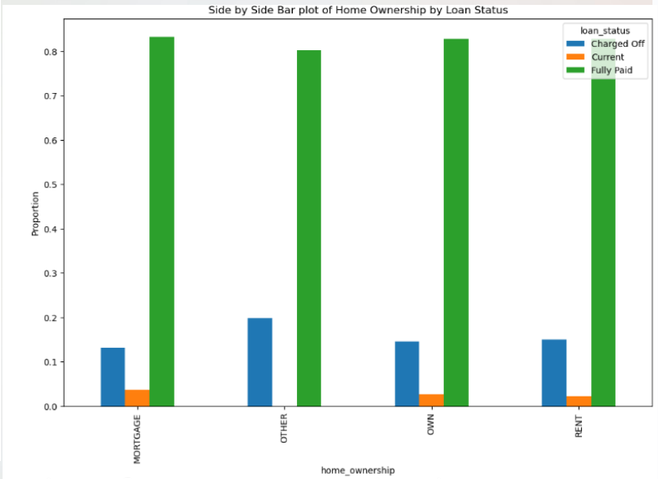
2) Bar Plot of Sub-Grade by Loan Status



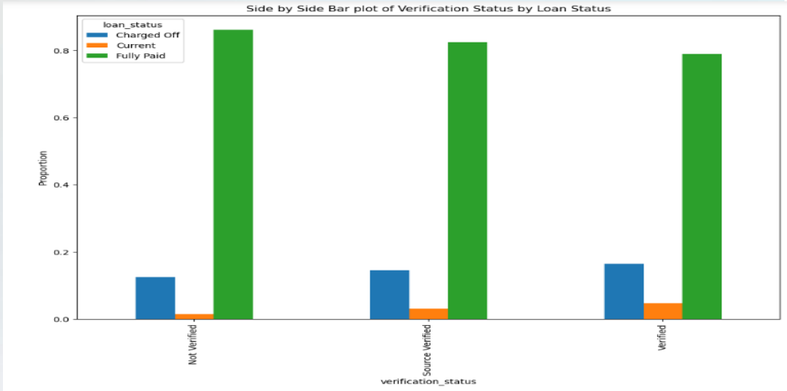
3) Bar Plot of Employment Length by Loan Status



4) Bar Plot of Home Ownership by Loan Status



5) Side-by-side Bar Plot of Verification Status by Loan Status



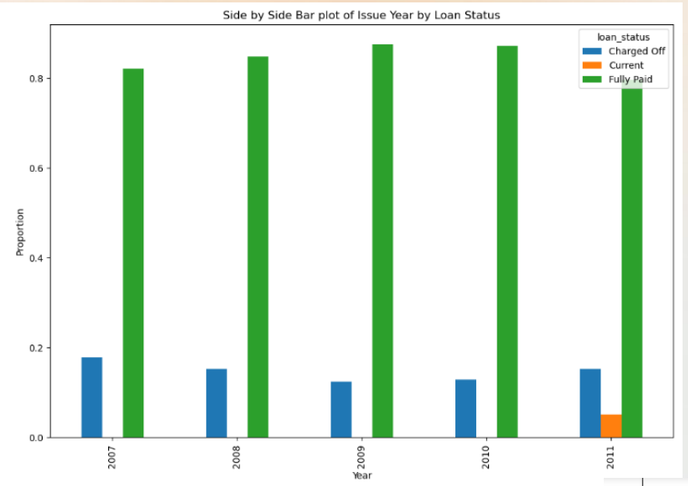
1) Overall loan grades are predictive of performance, with ***higher grades*** indicating ***lower risk and better outcomes.***

2) Sub-grade is a strong predictor of loan performance, with ***higher sub-grades*** (e.g., A1, A2) correlating with ***better outcomes*** than lower sub-grades (e.g., G5, F5)

3) While employment stability generally correlates with better loan performance, those ***at the extremes of employment length*** (Very new like 1 year or very long-term like 10+ years) might present ***higher risks***.

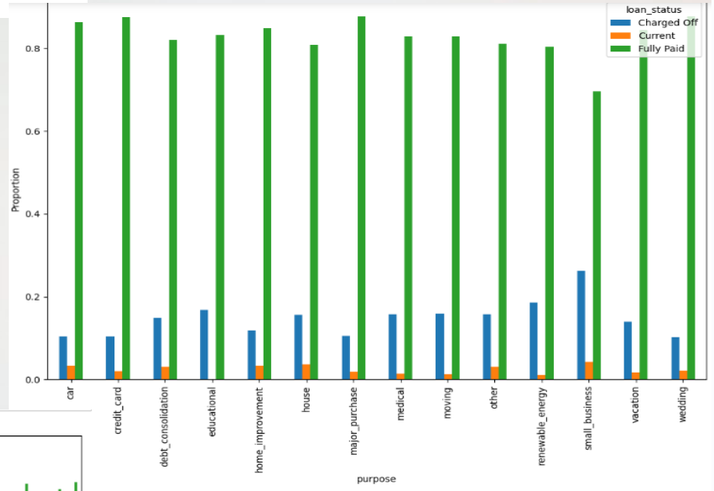
4) Homeownership status impacts loan performance, with those ***owning homes or having mortgages*** performing ***better than*** others.

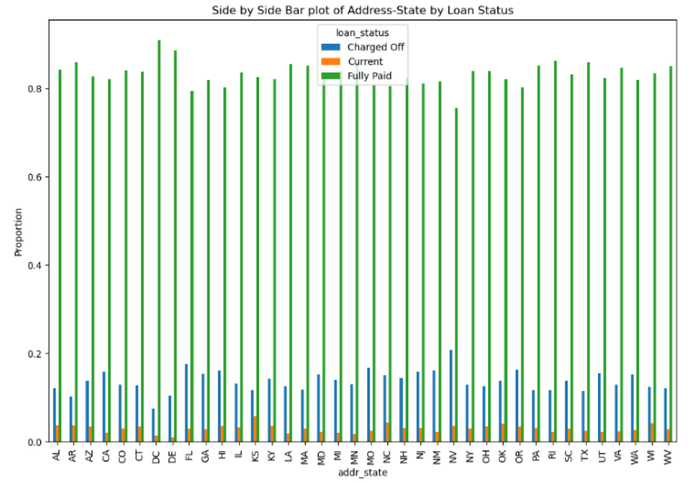
5) Verification status is a significant factor in loan performance, with ***verified loans*** showing ***better*** outcomes.



6)Side-by Side Bar Plot of Issue Year by Loan Status

7)Side-by Side Bar Plot of Purpose by Loan Status





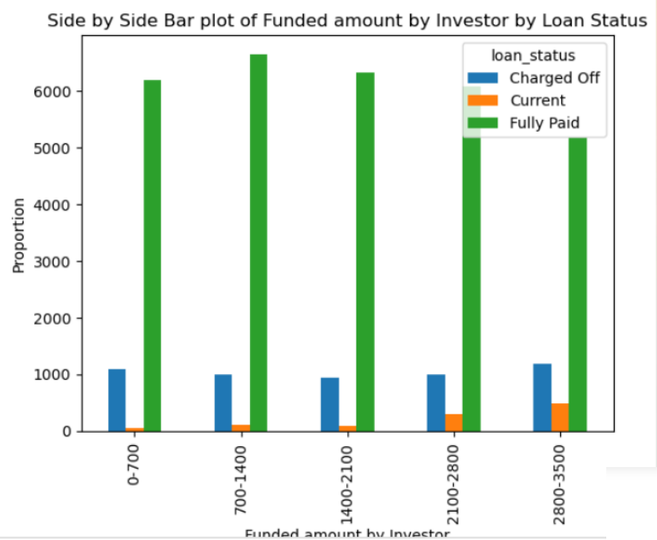
8)Side-by Side Bar Plot of Address-State by Loan Status

EDA – Bi-variate Analysis Insights:

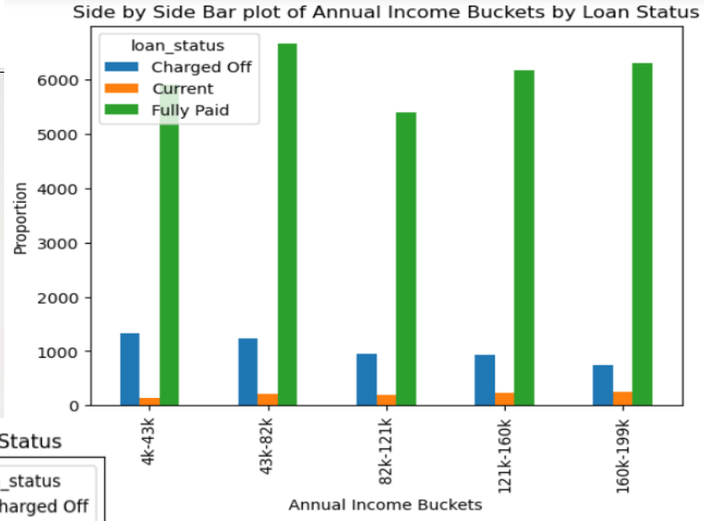
6) Loans issued in ***2007*** and ***2011*** have a ***higher proportion of charged-off loans*** compared to other years. This could reflect broader economic conditions during those years, such as the financial crisis around 2007-2008 and its lingering effects.

7) Loans for ***renewable energy*** and ***small business purposes*** have a relatively ***higher proportion of charged-off status*** compared to others like car or credit card loans, which have a higher proportion of fully paid statuses.

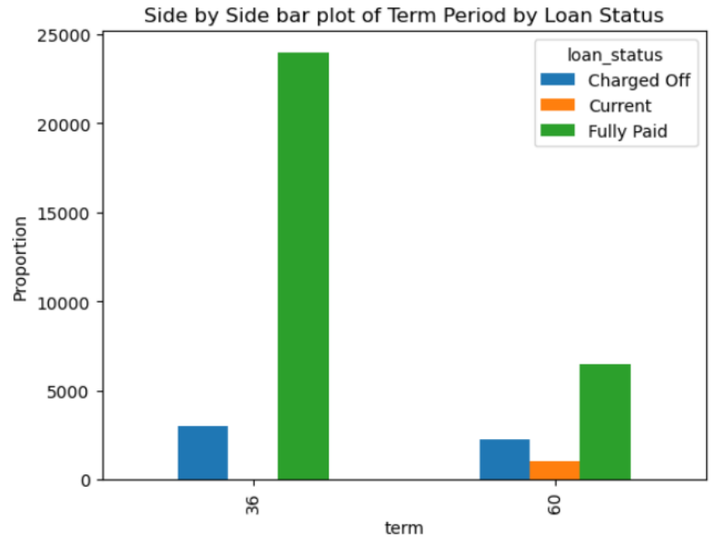
8) States like ***NV and FL*** have a ***higher proportion of charged-off loans*** compared to others, indicating potential economic stress or higher risk in these states.



9)Side-by Side Bar Plot of Funded Amount by Investors by Loan Status



10)Side-by Side Bar Plot of Annual Income by Loan Status



11)Side-by Side Bar Plot of Term Period by Loan Status

EDA – Bi-variate Analysis Insights:

9) Borrowers with Funded amount invested in the range of ***$2800-$3500*** are ***most likely to default*** having charged off to fully paid ratio of 22%. Other buckets have a ratio between 14-17%,where ***$700-$1400*** bucket has the ***least probability to default*** of all by a small margin.

10) As the ***annual income increases***, chances of ***defaulting decreases***. Income range ***$4k to $43k*** has a ***charged off / fully paid ratio*** of ***22%***, where as for income ***$160k to $199k*** the ratio is ***11%***.

11) For the term period of ***36 months***, the ***charged off to fully paid ratio is 1/10*** but in case of ***60 months***, the ***ratio is 1/3***. This indicates that customers with a term period of ***60 months*** are ***very likely to default***.

Results:

REPAYORS:

- There were more loan repayors than defaulters among *higher annual income people*.
- *Home owners* are more likely to be repayors than those who rented their homes.
- There were more loan repayors than defaulters among people with *higher number of credit lines in file.*
- Borrowers *with verified loans* are more likely to be repayors than borrowers with unverified or source verified loans.

DEFAULTERS:

- There were more loan defaulters than loan repayors among those borrowers who have a *higher loan amount, interest rate and instalments.*
- There were more defaulters than repayors among people with *higher Debt-to-Income Ratio.*
- There were more defaulters than repayors among people with *higher average revolving balance and revolving line utilization rates.*
- There were more defaulters than repayors among customers with *a term of 60 months* than customers with a term of 36 months.
- There were more defaulters than repayors among borrowers *with more than 10 years of employment.*
- As the *grade and sub-grade decreases*(From A1,A2 to F5,G5) the borrower is most like to be a defaulter.
- People with *past Delinquency* were more likely to be defaulters than non-delinquent borrowers.
- People with *earlier Bankruptcies records* are more like to be defaulters than non-bankrupt borrowers.
- There were more loan defaulters than repayors among borrowers who were issued loans in *2007* and *2011*
- There were more loan defaulters than repayors among borrowers who took loans for *renewable energy* and *small business purposes* compared to those who took *car or credit card loans*.
- There were more loan defaulters than repayors among borrowers who belonged to States like *NV and FL* compared to those from other states.

Recommendations:

As a result of our analysis, we are now able to predict whether a Client will repay the loan or not. The following are our recommendations:-

1) ***Interest rate, funded amount and installments*** **where interest rate have a** ***higher dependency on loan status*****. To avoid defaulters loan with** ***high interest*** **and** ***high amount should not*** **be given to** ***low annual income*** **personal with** ***income range $4000 to $43000.***

2) ***Grade G and F Loans*** **should be avoided for** ***low income*** **people who opt** ***for high amount loans*** **as they have really** ***high risk*** **of getting** ***charged off.***

3) People who opted ***for lower sub-grade(F5, G5) loans*****, have a** ***higher risk*** **of getting** ***charged off.***

4) It is seen that loan taken by borrowers who have ***more credit lines*** **are** ***more likely to repay the loan back*** **maybe from getting another loan.**

5) People taking loans for ***small business are really risky*****, a thorough** ***background check must be done*** **for them before approving loan.**

6) People asking for ***term period of 60 months,*** **have** ***high probability of getting charged off*****, hence** ***if loan amount/ Interest is low*** **then it is** ***much better*** **or they can be asked** ***to repay*** **the loan within a** ***term period of 36 months*****.**

7) People having ***employment history*** **of** ***more than 10 years*****, have a** ***higher risk*** **of getting** ***charged off.***

8) People having ***higher Debt-to-Income Ratio*****, have a** ***higher risk*** **of getting** ***charged off.***

9) People having ***higher average revolving balance and revolving line utilization rates*****, have a** ***higher risk*** **of getting** ***charged off.***

10) People who had ***Delinquency in the past 2 years,*** **have** ***a higher risk*** **of getting** ***charged off.***

11) People who are ***present*** **in** ***the earlier Bankruptcies records,*** **have a** ***higher risk*** **of getting** ***charged off.***

12) People who applied for a loan in ***2007*** **or** ***2011,*** **have a** ***higher risk*** **of getting** ***charged off.***

13) People who belong ***to States*** **like** ***NV*** **and** ***FL,*** **have** ***a higher risk*** **of getting** ***charged off.***

Conclusion:

Data Exploration:

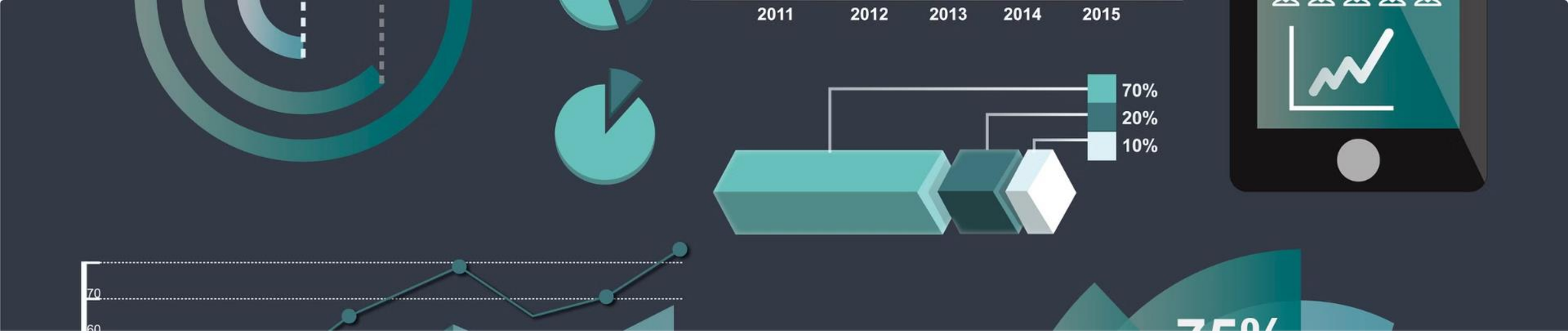
Explored the data about the Lending Club’s loan offers and the factors impacting the decision to approve or deny the loan application.

Derived Insights:

Identified trends and patterns related to loan offers for further analysis and decision-making.

Recommendations:

Provided recommendations to minimize the possibility of financial losses in the approval or rejection of loans.



THANK YOU!