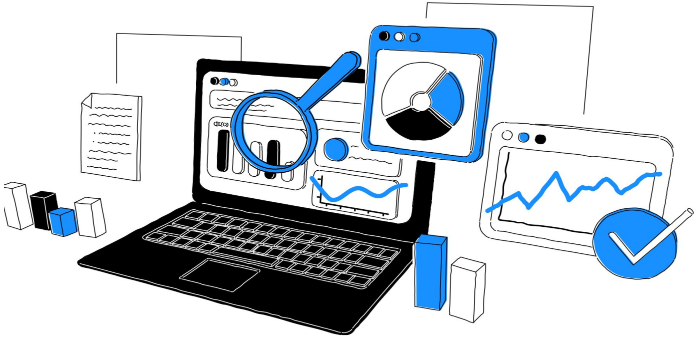
Lending Club Case Study



Group Members:
Anshika Joshi
Rudranil Gupta

Contents

- Objective/goal
- Data Description & Understanding
- Data Preparation/Cleaning
- Data Analysis Methodology
- Data Visualization
- Conclusion

Objective/Goal

- 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.

- To identify the risky loan applicants, which helps in cutting down the amount of credit loss by performing Exploratory Data Analysis.

- In other words, we need 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 Description & Understanding

- We are provided with Loan Dataset in .csv format along with Data_dictionary containing description of the variables provided in the data.

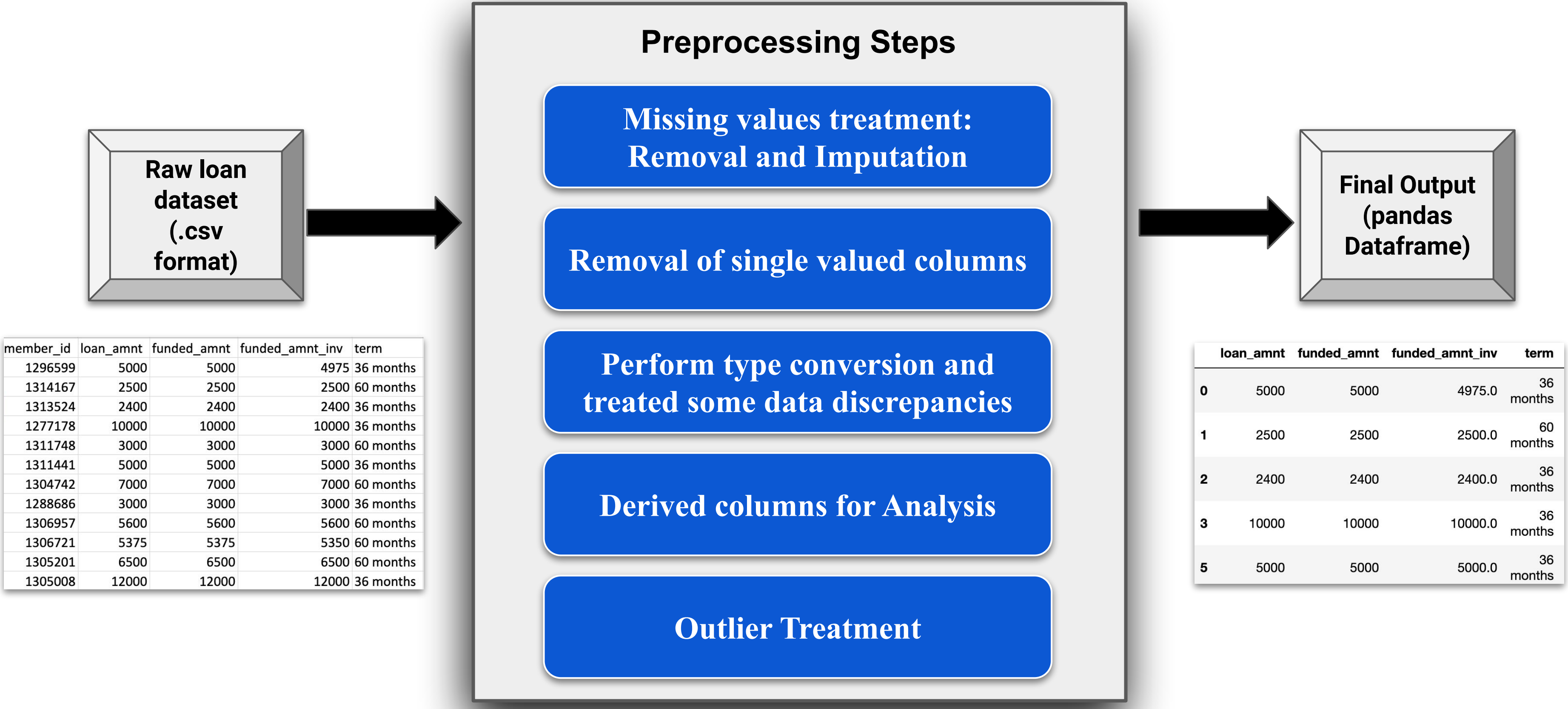
- The Loan Dataset is the historical data of past 5 years (2007-2011). The raw data contains 111 attributes as columns and 39717 records pertaining to each borrower.

- The attributes provided in the dataset are basically of 2 types Consumer Attributes and Loan Attributes. The consumer attributes gives us ample information about the borrower like their annual_income, home_ownership, purpose etc whereas loan attributes provides information like loan_amount, interest_rate, installment etc

- When a person applies for a loan, there are following two types of decisions that could be taken by the company,
    - Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
        - Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
        - Current: Applicant is in the process of paying the instalments.
        - 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
    - Loan rejected: *This data is not provided to us*

- We are focussing on those applicants having their Loan status as Fully paid and Charged-off, since the Current cannot state whether a borrower is a defaulter or not.

Data Preparation/Cleaning



Raw loan dataset (.csv format)

member_id	loan_amnt	funded_amnt	funded_amnt_inv	term
1296599	5000	5000	4975	36 months
1314167	2500	2500	2500	60 months
1313524	2400	2400	2400	36 months
1277178	10000	10000	10000	36 months
1311748	3000	3000	3000	60 months
1311441	5000	5000	5000	36 months
1304742	7000	7000	7000	60 months
1288686	3000	3000	3000	36 months
1306957	5600	5600	5600	60 months
1306721	5375	5375	5350	60 months
1305201	6500	6500	6500	60 months
1305008	12000	12000	12000	36 months

Preprocessing Steps

- Missing values treatment: Removal and Imputation
- Removal of single valued columns
- Perform type conversion and treated some data discrepancies
- Derived columns for Analysis
- Outlier Treatment

Final Output (pandas Dataframe)

	loan_amnt	funded_amnt	funded_amnt_inv	term
0	5000	5000	4975.0	36 months
1	2500	2500	2500.0	60 months
2	2400	2400	2400.0	36 months
3	10000	10000	10000.0	36 months
5	5000	5000	5000.0	36 months

Preprocessing Steps:

- Removal of unnecessary columns which does not impact our analysis. Example: id, member_id, employment title etc.

- Check the missing values present in each column, and removed those variables having missing values more than 60% of the data.

- Performed missing values imputation wherever required with mean,mode and median.

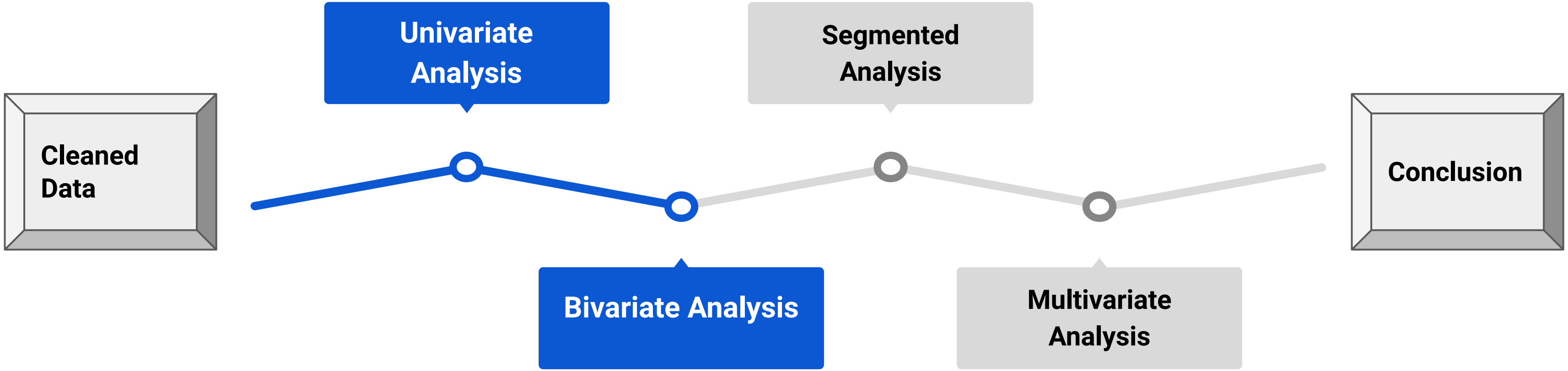
- Removed the columns having identical value.

- Perform type conversion of the variables like issue_date, and treated some data discrepancies like remove/replace special characters if present in data.

- Derived columns like “issue_month” & “issue_year” and filter the data of loan_status into subsets of “full_paid” & “charged_off”.

- Performed outlier treatment by checking their interquartile range with the help of boxplot and then removed those points which seems to be outlier from the column.

Data Analysis Methodology



Cleaned Data

Univariate Analysis

Bivariate Analysis

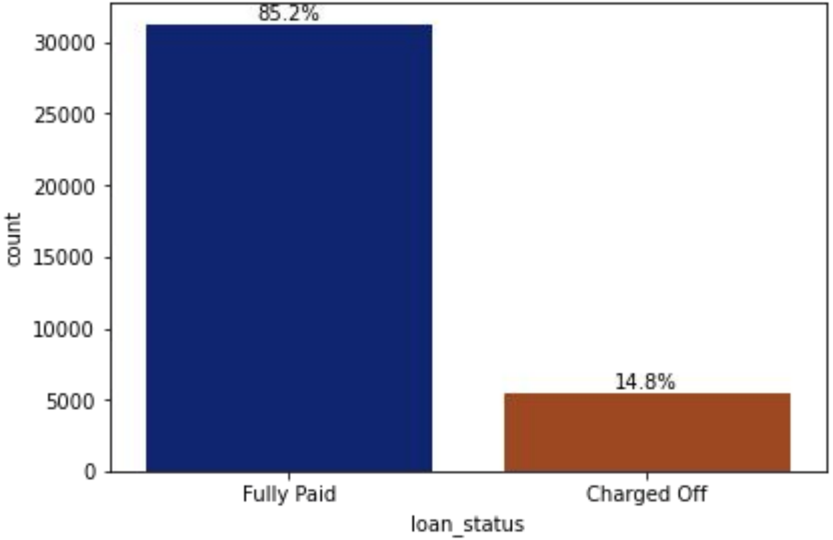
Segmented Analysis

Multivariate Analysis

Conclusion

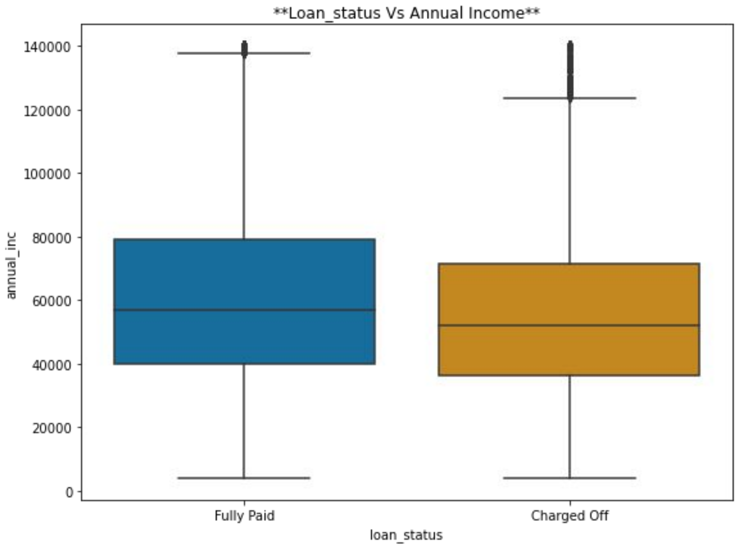
Data Visualization

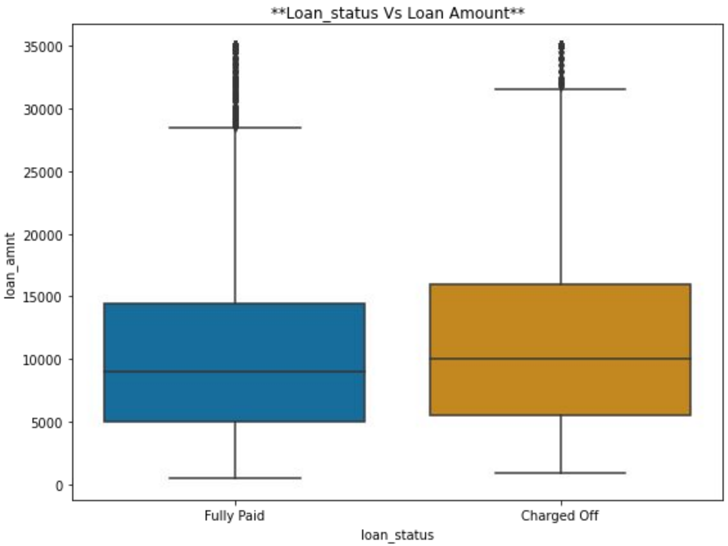
Visualize data distribution on basis of “fully-paid”/“good” and “charged-off”/“bad”/“defaulters” consumers.



- Here on X-axis we have loan_status variable there we can see out of total data, 85% data belongs to fully-paid consumers while we have only 14.8% data for bad/charged-off/defaulters.
- Since the data is imbalance for both the classes, we will consider percentage or proportion for further analysis of good & bad consumers.

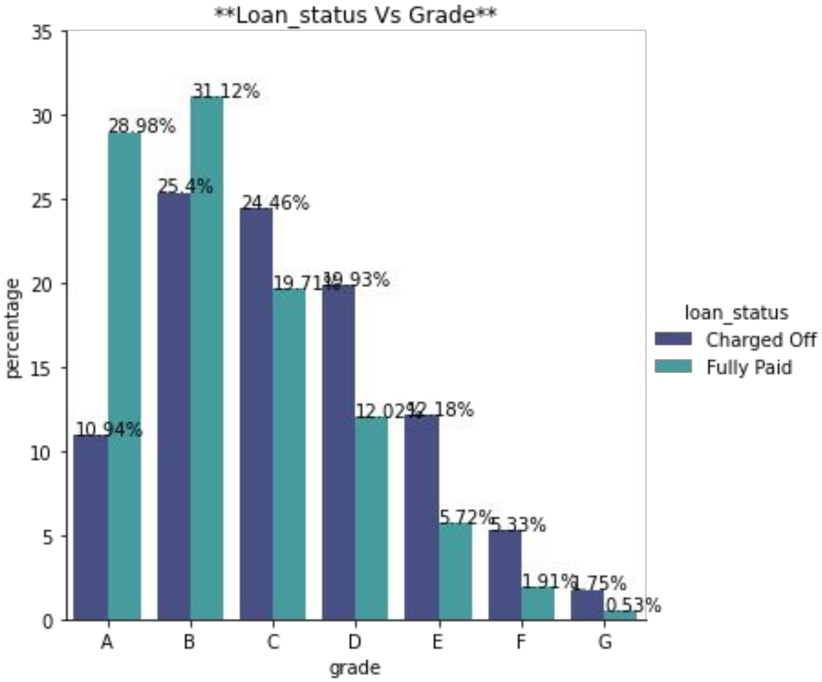
Visualize and analyse annual income and loan amount of good and bad consumers





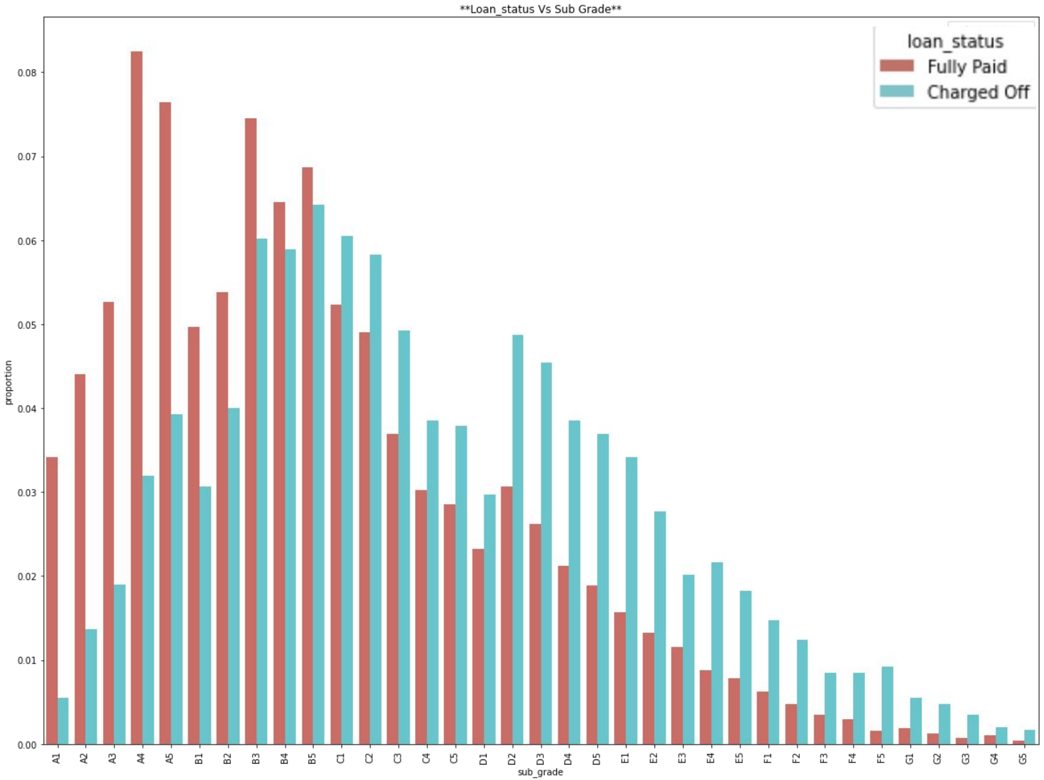
- Charged_off(defaulters) are having less annual income as compared to fully_paid consumers.
- But the loan amount range is more for charged-off , that means they demands for larger amount of loan as compared to fully paid.
- This can be potential reason why they are unable to pay loan amount

Visualize and Analyse grade of good and bad consumers



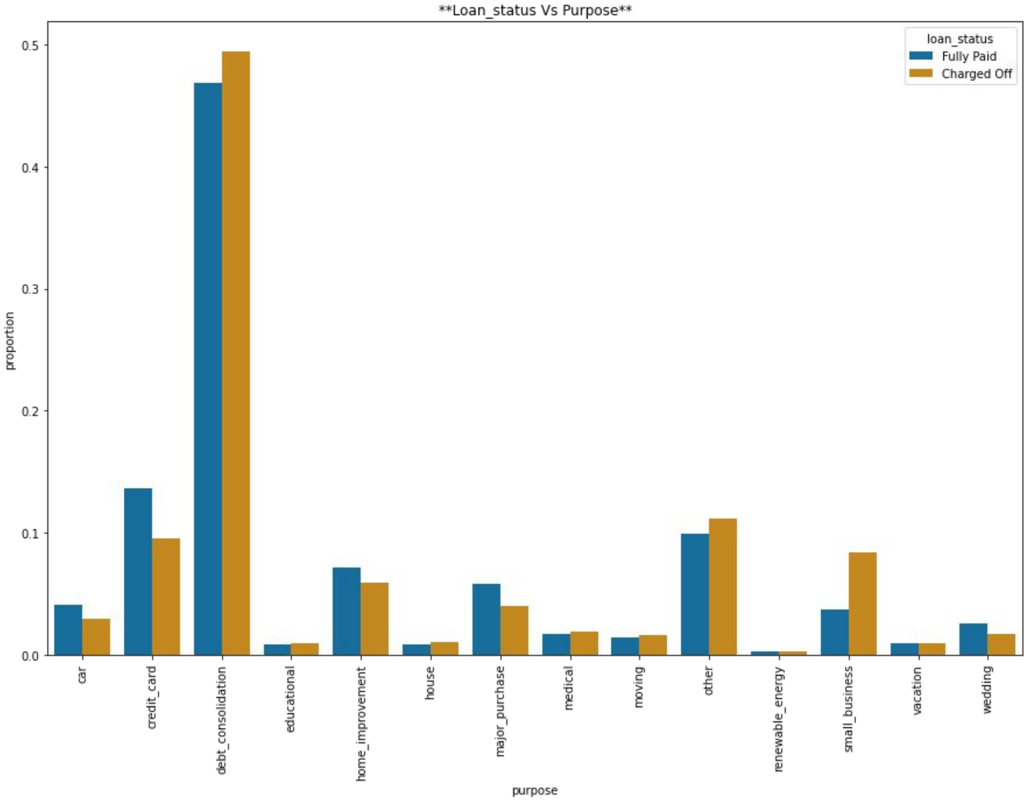
- The variable "grade" explains about the credit history of the consumer.
- Here we observe that charged_off consumers have more percentage of grade C, D, E, F & G as compared to fully_paid consumers.
- We can conclude that consumers with grade “A” & “B” can fully pay the loan while grade category C, D, E, F & G belongs to bad consumers.

Visualize and Analyse sub-grade of good and bad consumers



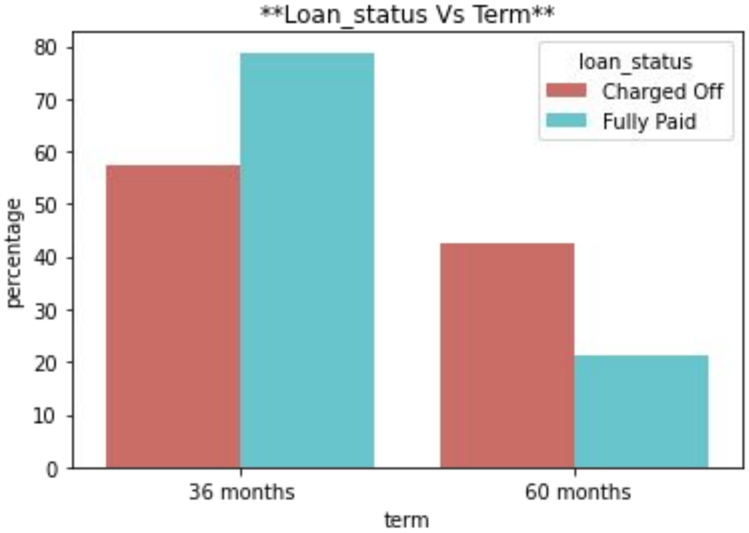
- Here we observe a trend that charged_off consumers have initially(A1,A2,A3,A4,A5,B1,B2,B3,B4, B5) lower proportion then fully_paid consumers.
- But gradually the proportion of charged_off increase and fully_paid decreases from subgrade C1 onwards.
- So we conclude that the consumers with subgrade C1 & so on are more likely to default.

Visualize and Analyse Purpose of taking for good and bad consumers



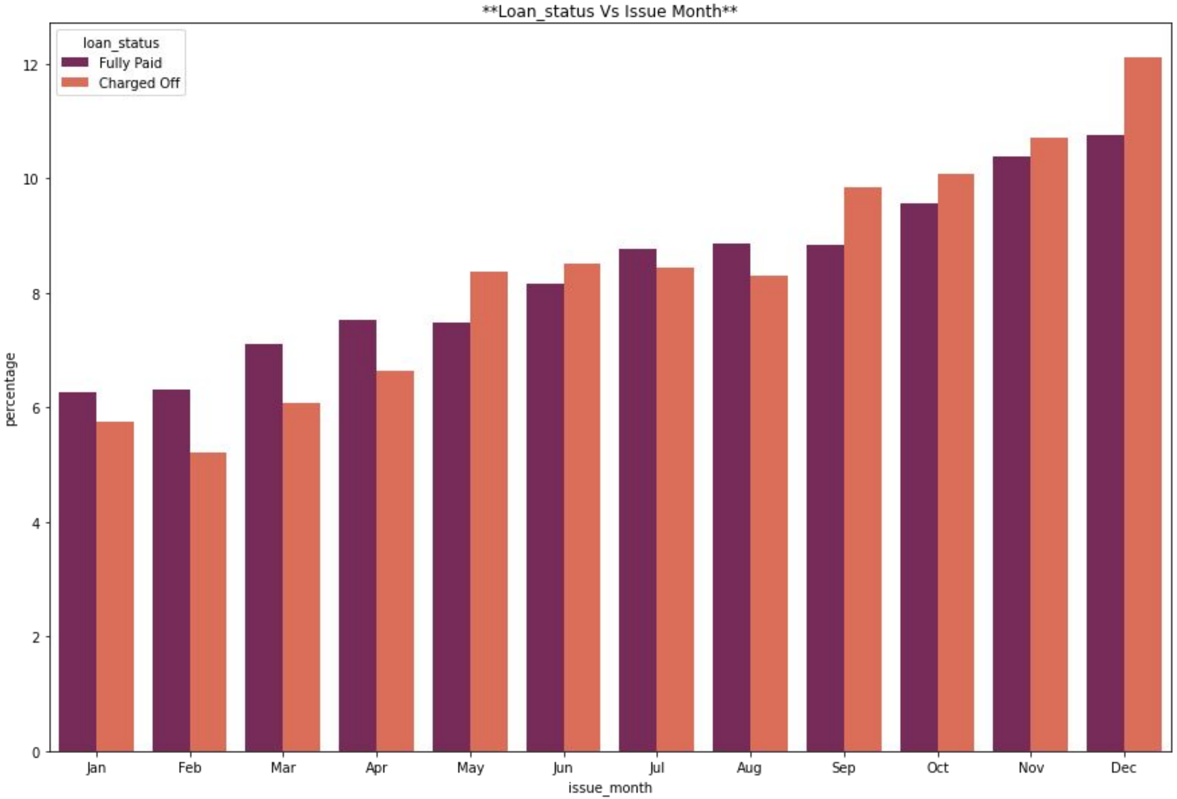
- Here we observe that most people are taking loan for the purpose of “debt_consolidation”, among which charged_off consumers are more in proportion.
- Also for the purpose like "small_business" and "other" has comparatively more proportion of charged_off consumers then fully-paid.
- We can say that because charged-off consumers are already in debt, so they are unable to pay the loans.

Visualize and Analyse Term of loan for good and bad consumers



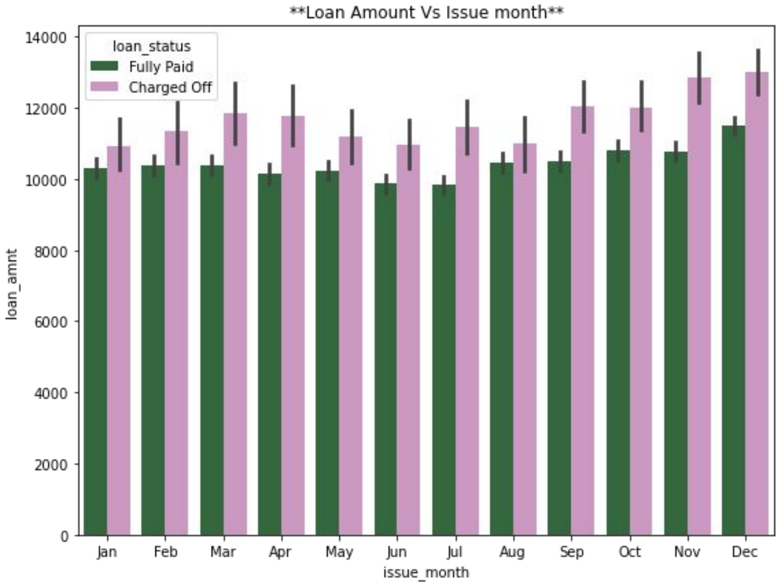
- We see that generally people tend to take loan for less duration(36 months).
- And mostly are fully-paid consumers.
- Among the people who take loan for long term(60 months) are the having more percentage of charged-off consumers.
- We can say that, more consumers who are opting for long term loan can default.

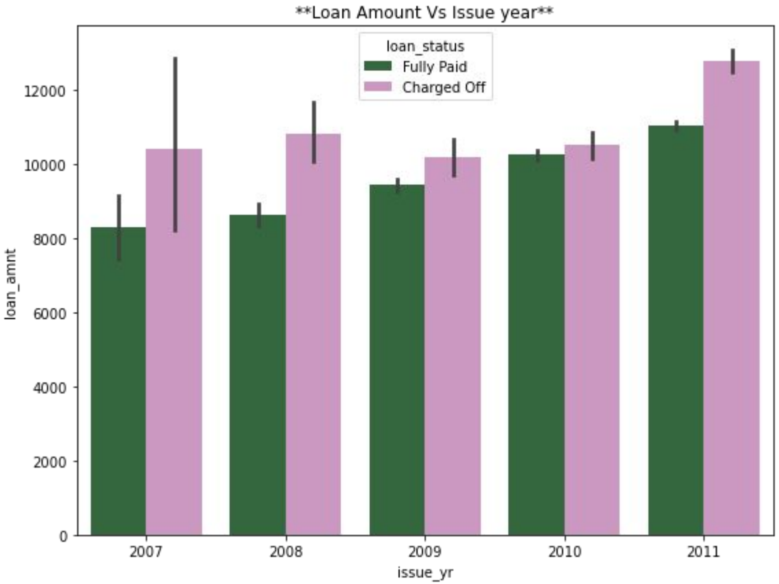
Visualize and Analyse Issue Month of loan for good and bad consumers



- We observe that, mostly consumers are taking loan in the year end.
- The graphs shows, more charged_off consumers take loan in the months May, Sept, Oct, Nov & Dec.
- We can conclude that mostly defaulters are taking loan in year end.

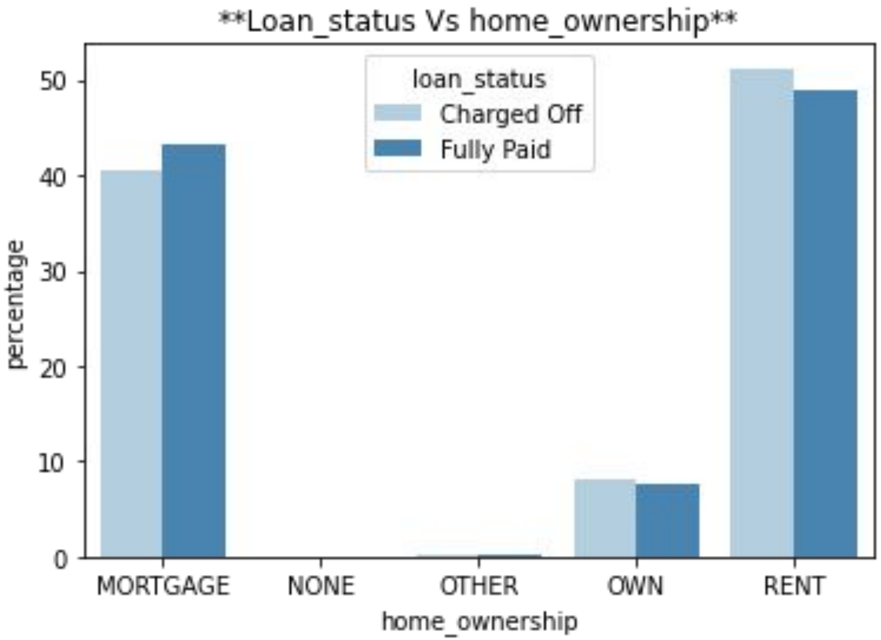
Visualize and Analyse Loan amount Issue Month & year of loan for good and bad consumers





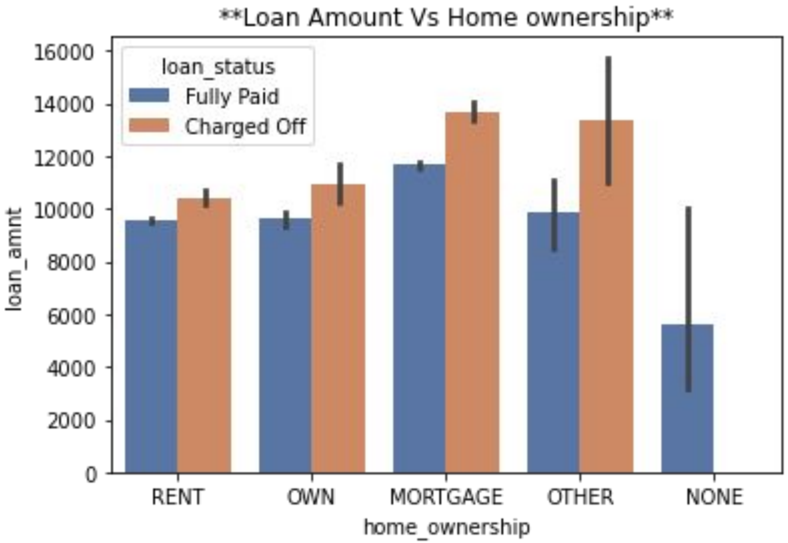
- The loan amount is more either in the beginning or in the end of the year for both good and bad consumers.
- As both the graphs signifies, that the loan amount is more for charged_off consumers as compared to fully_paid.
- There is an increasing trend loan amount in the course of 4 years historical data. Among all, charged-off consumers are always taking loan for larger amount.

Visualize and Analyse home ownership for good and bad consumers



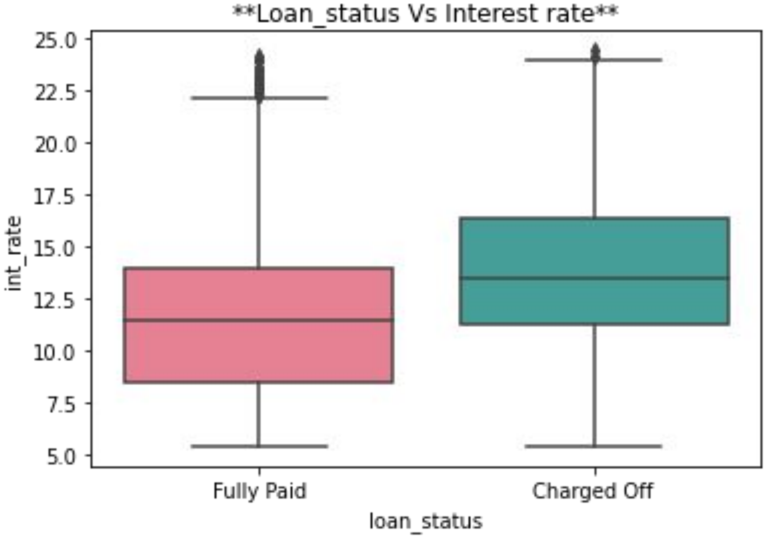
- More than 50% of the charged_off consumers are living with RENT
- And very least who OWNs a house are taking loan.
- We can conclude that mostly charged-off consumers are living with RENT.

Visualize and Analyse loan amount vs home ownership for good and bad consumers



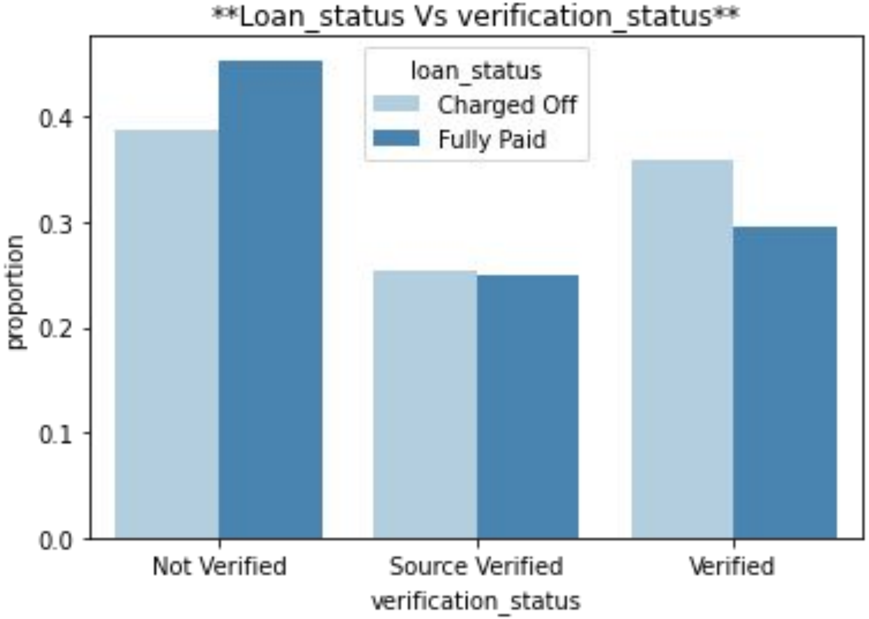
- Most consumers who has home ownership as mortgage do opt for large amount of loan.
- Among all, charged-off are taking loan for larger amount.

Visualize and Analyse Interest Rate for good and bad consumers



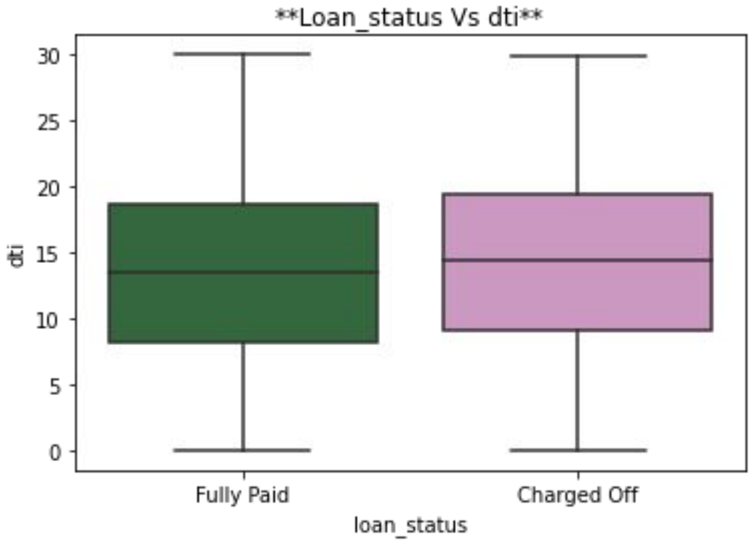
- Charged_off consumers are taking loans on comparatively higher interest rate.
- Here we can see 50% data of charged_off consumers are paying loan at the interest rate range 11.26-16.32, while fully_paid has a range 8.49-11.49

Visualize and Analyse Verification Status for good and bad consumers



- Here we can suspect some abnormality that is, among those consumers who are not verified, there we can find more proportion of fully_paid.
- while those who are verified are having more proportion of charged_off consumers.
- We can conclude that more not-verified consumers tend to fully pay the loan(good consumers) and those are verified, still can default.

Visualize and Analyse Debt-to-Income for good and bad consumers



- Charged_off consumers have bit more debt to income ratio than fully_paid consumers.
- A high DTI ratio can signifies that an individual has too much debt for the amount of income earned each month.

Conclusion

We can draw following conclusions by Exploratory Data Analysis:

- Defaulters tend to take loan of larger amount and have lower annual income as compared to fully-paid consumers.
- Fully-paid consumers comes under grade “A” & “B” while defaulters are graded as “C”, “D”, “E”, “F” & “G”
- Likewise mostly fully-paid consumers are having sub-grade (A1,A2,A3,A4,A5,B1,B2,B3,B4,B5) while charged-off consumers are more tend to have sub-grade C or below.
- Defaulters has a purpose as ‘debt_consolidation’ for opting a loan.
- Generally fully-paid opt for short term but defaulters opt for long term loan.
- Mostly defaulters are taking loan in year end (sept,oct,nov & dec)
- Mostly charged-off consumers are living with RENT.
- Charged_off consumers are taking loans on comparatively higher interest rate.
- We also suspected some abnormalities that, the charged_off consumers are verified by LC.
- Charged_off consumers have bit more debt to income ratio than fully_paid consumers.