LENDING CLUB CASE STUDY EDA ASSIGNMENT SUBMISSION

Submitted By-

Poolakit Bharadwaj

Juhi Ramzai

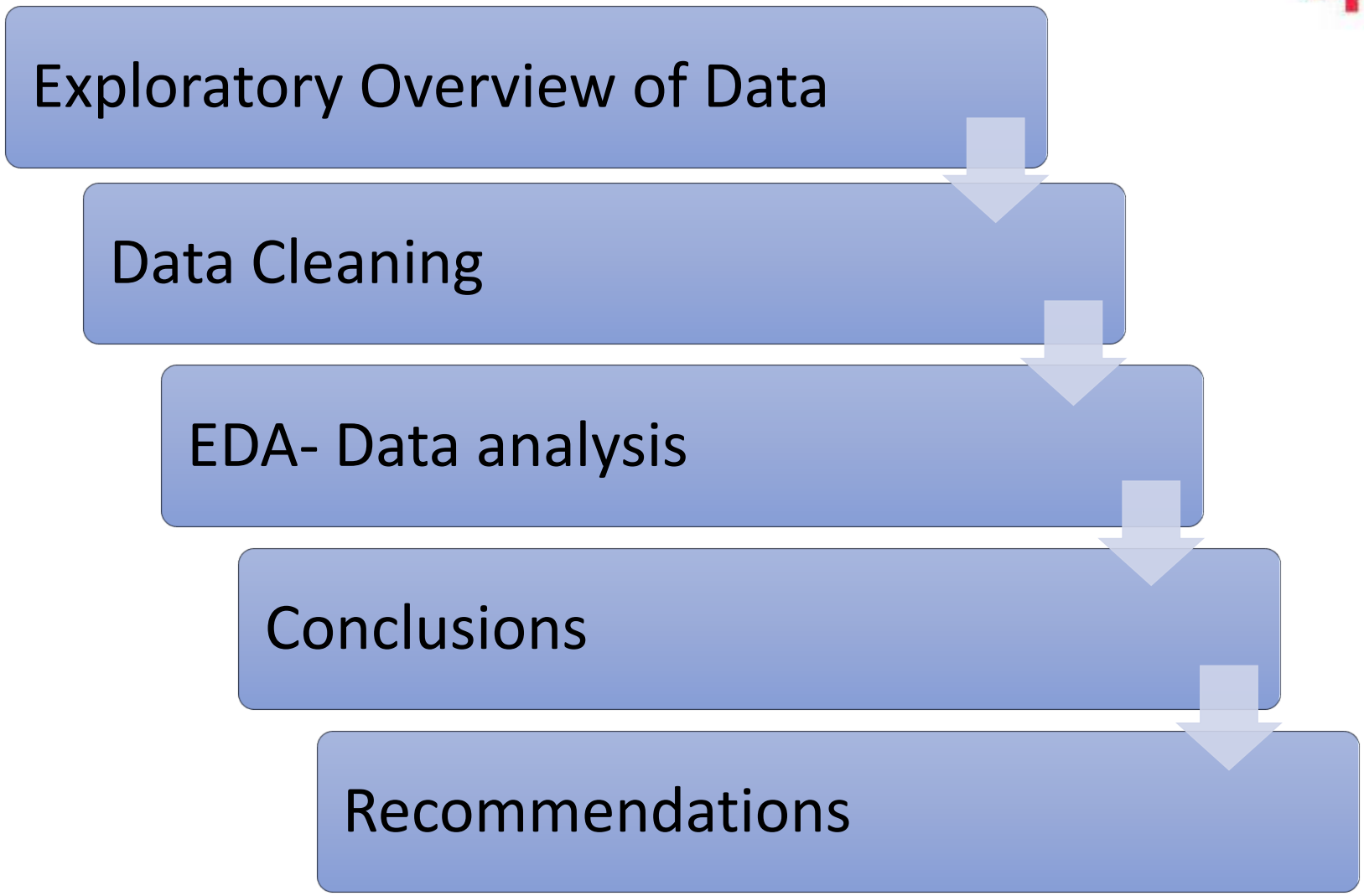
PROBLEM STATEMENT

The business problem is for a consumer finance company which specialises in giving out various types of loans to urban customers. When the company receives a loan application, it has to make a decision for loan approval based on the applicant's profile.

We have been provided the historical data for loan applications that were approved by the company after analysing the associated business risks and whether the borrowers defaulted or not.

Our aim is to identify, using EDA, the strong variables which indicate whether a borrower is likely to default. This will be used for taking actions such as denying the loan, reducing the amount of loan, lending at competitive interest rates or at a higher interest rate (for risky applications), etc. and hence reduce credit loss for the company.

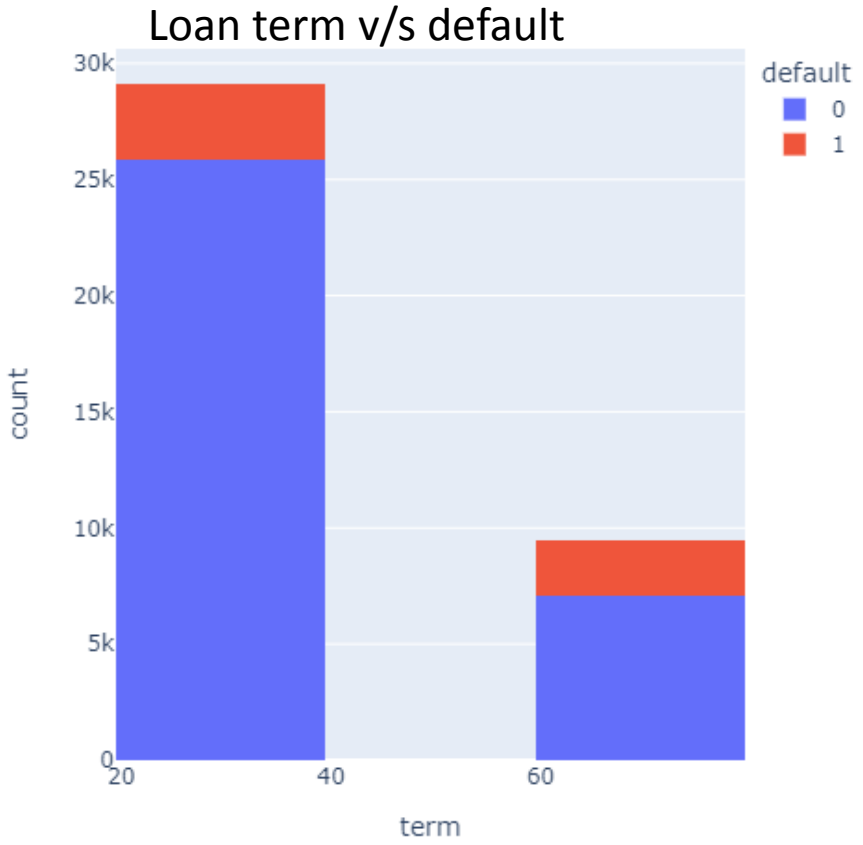
ANALYSIS WORKFLOW



1. Analysis of ‘Term’

- **Hypothesis:**
  - Longer loan terms increase the risk of a borrower defaulting, thus will lead to higher average default rates
- **Univariate Analysis:**
  - **75%** of the loans have a repayment term of **36** months
  - **25%** of the loans have a repayment term of **60** months
- **Bivariate Analysis (vs Default/Charged-Off)**
  - For **36** months loan terms, average default rate is **11%**
  - For **60** months loan terms, average default rate is **25%**

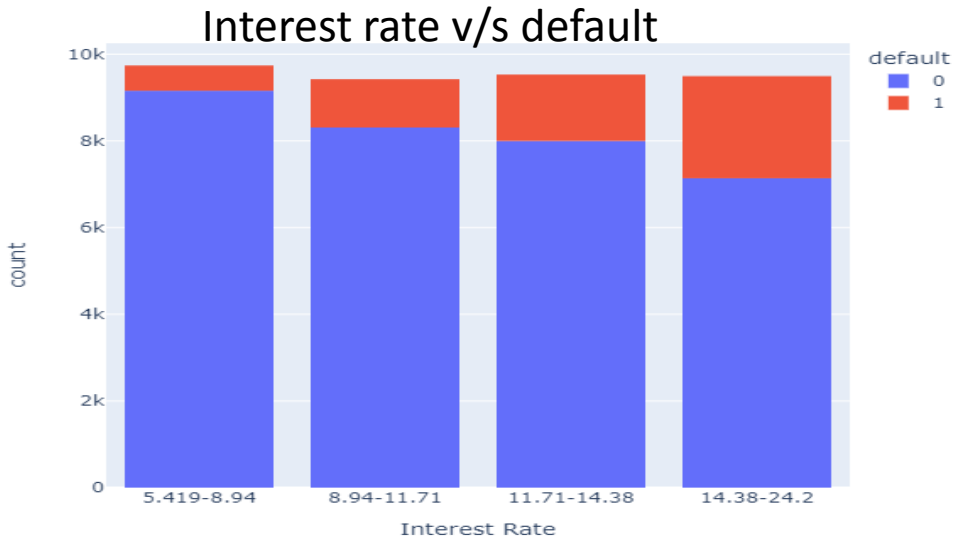
Hypothesis Confirmed



2. Analysis of ‘Interest Rate’

- **Hypothesis:**
  - Higher interest rates are generally offered to riskier borrowers, thus an increase in interest rate should lead to an increase in default rates as well
- **Univariate Analysis:**
  - We have divided this variable into 4 categories of interest rate offered by LC for conducting the further bivariate analysis
- **Bivariate Analysis (vs Default/Charged-Off)**
  - There is a steady increase in default rate with increasing loan interest rate ranging from **5% to 24%**

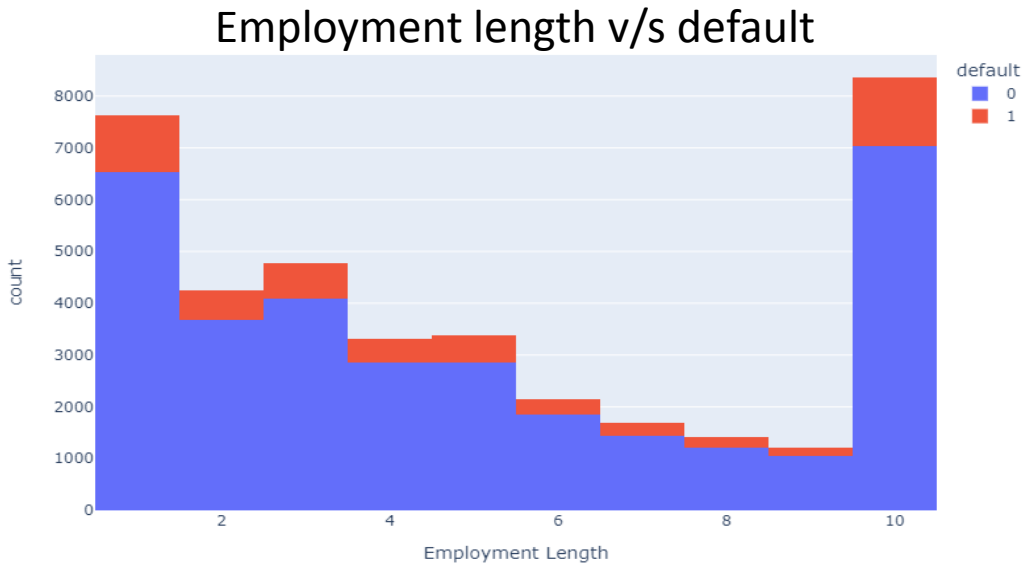
Hypothesis Confirmed



3. Analysis of ‘Employment Length’

- **Hypothesis:**
  - Employment length should not substantially impact a borrower’s probability to default as the employment length doesn’t necessarily indicate financial well being
- **Univariate Analysis:**
  - Highest number of loans have been approved for borrowers with **10+** years of employment length, followed by borrowers with **1** year of employment length
- **Bivariate Analysis (vs Default/Charged-Off)**
  - The default rate for all loan applications ranges from **13-15%** which shows that there is no significant impact of the borrower’s employment length

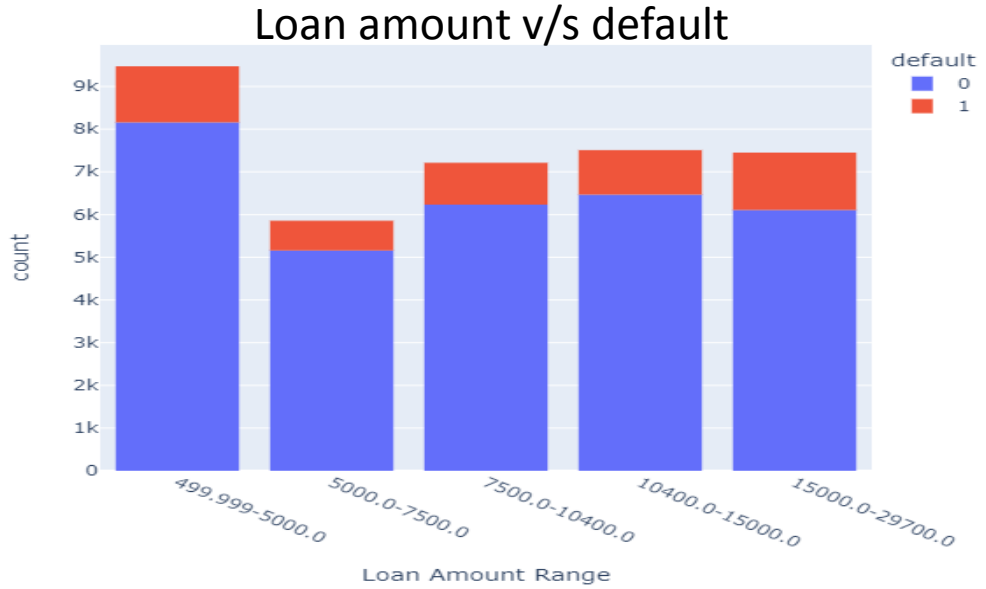
Hypothesis Confirmed



8. Analysis of ‘Loan Amount’

- **Hypothesis:**
  - Higher the loan amount, higher the risk of default, thus will lead to higher average default rates
- **Univariate Analysis:**
  - We have divided this variable into 5 categories of loan amount offered by LC for conducting the further bivariate analysis
- **Bivariate Analysis (vs Default/Charged-Off)**
  - Loan amount category **15,000-29,700** has a default rate of **~18%**, significantly higher than other categories. There is a weak increasing trend of default

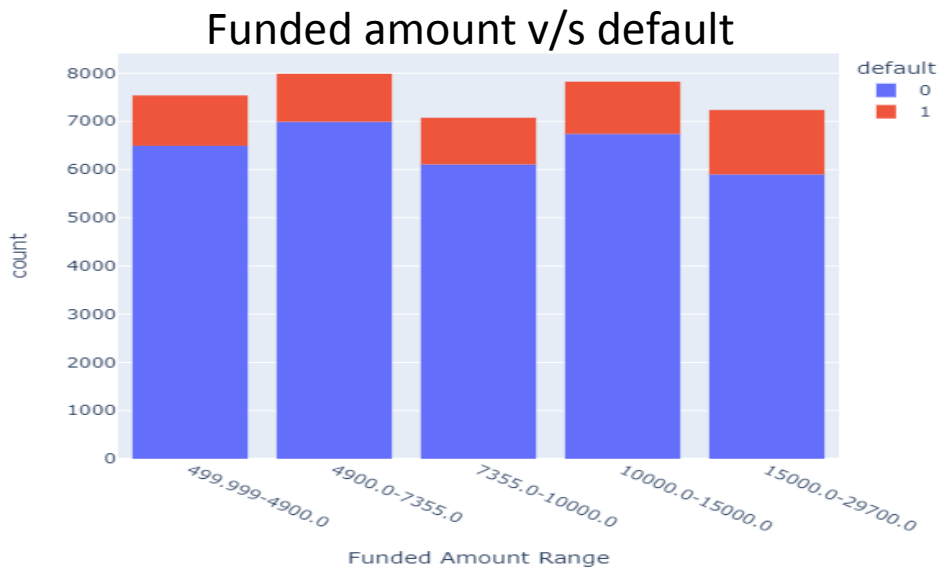
Hypothesis Confirmed

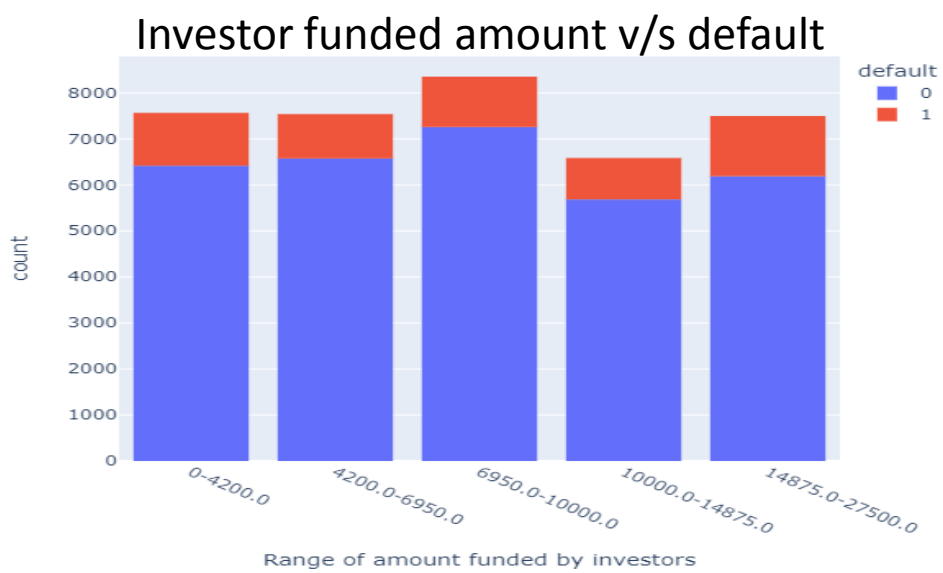


9. Analysis of ‘Funded amount’ and Funded amount inv

- **Hypothesis:**
  - Higher the funded amount, higher the risk of default, thus will lead to higher average default rates
- **Univariate Analysis:**
  - We have divided these variables into 5 categories as per available data for conducting the further bivariate analysis
- **Bivariate Analysis (vs Default/Charged-Off)**
  - Funded amount category **15,000-27,500** has a default rate of **~18%**, significantly higher than other categories
  - The default rate is highest at **~17%** where the amount funded by investors is **14,875-27,500**

Hypothesis Confirmed

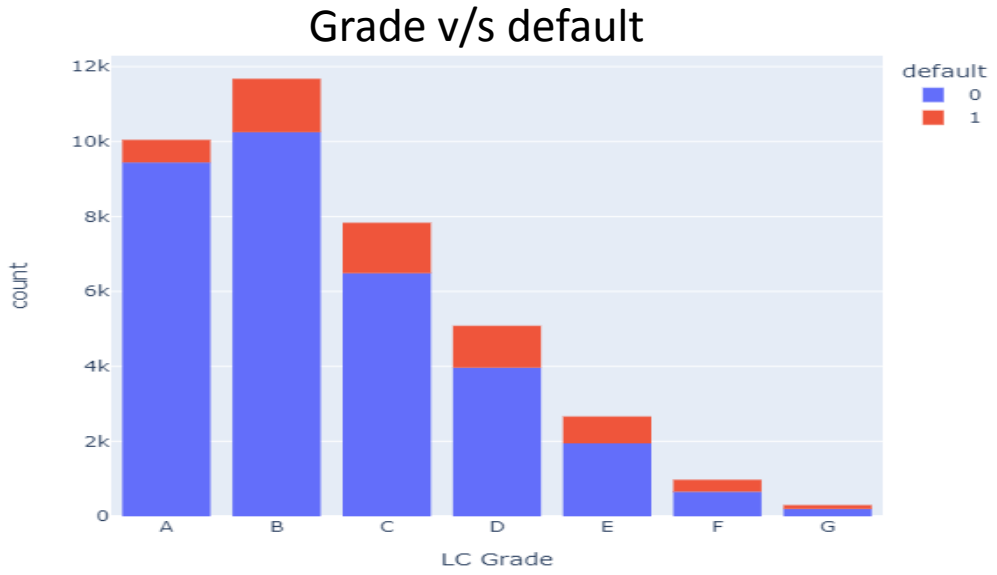




11. Analysis of ‘Grade’

- **Hypothesis:**
  - Lower the LS assigned grade, higher the risk of default, thus will lead to higher default rates
- **Univariate Analysis:**
  - **~56%** of the loans have been given out to grades A and B, **<1%** to grade G
- **Bivariate Analysis (vs Default/Charged-Off)**
  - Only **5%** default in grade A
  - There is an increasing trend with **33%** default in grade G

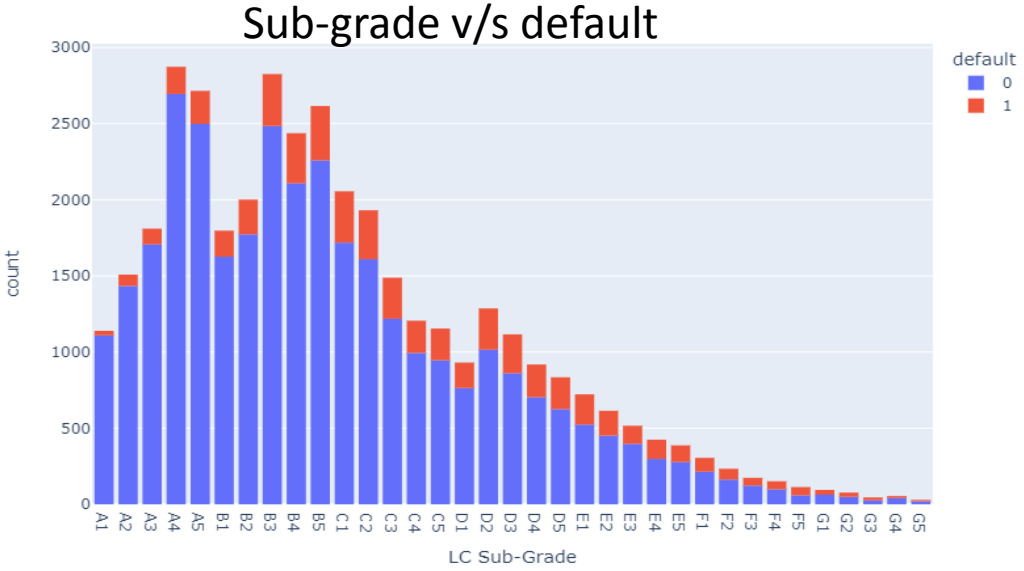
Hypothesis Confirmed



12. Analysis of ‘Sub-Grade’

- **Hypothesis:**
  - Lower the LS assigned sub-grade, higher the risk of default, thus will lead to higher default rates
- **Univariate Analysis:**
  - **~56%** of the loans have been given out to sub-grades of A and B grades
- **Bivariate Analysis (vs Default/Charged-Off)**
  - **2-7%** of default in sub-grades of A grade
  - Steady increasing trend with **24-42%** default in G sub-grades

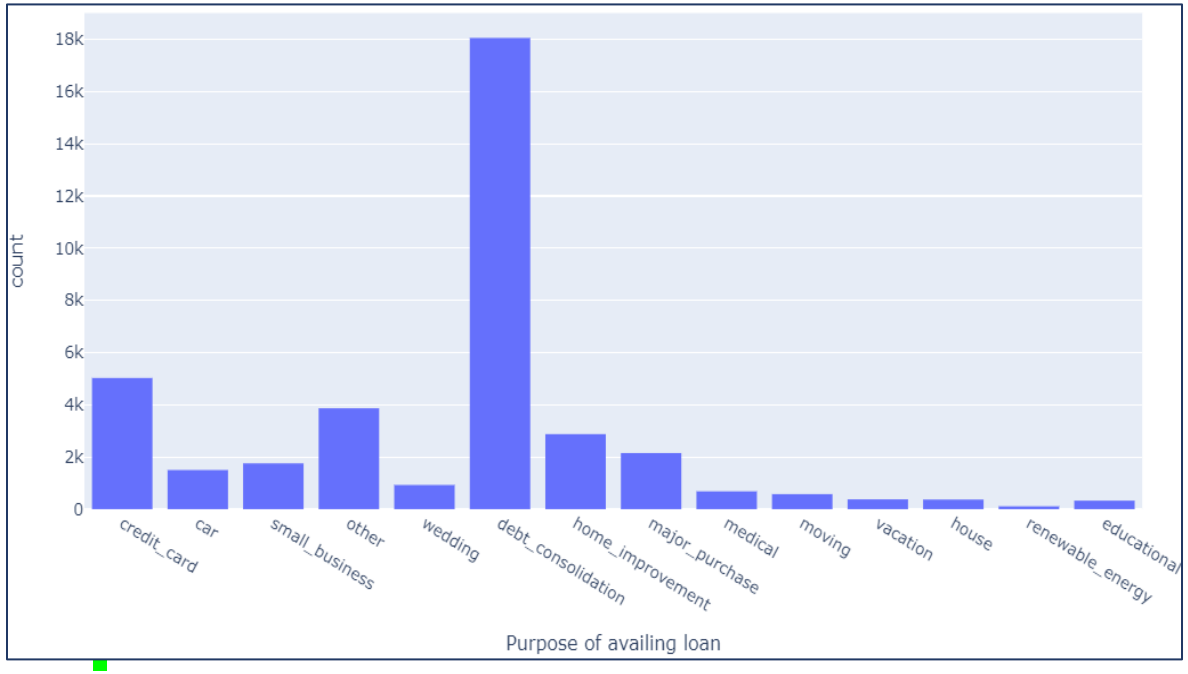
Hypothesis Confirmed



13. Analysis of ‘Purpose’

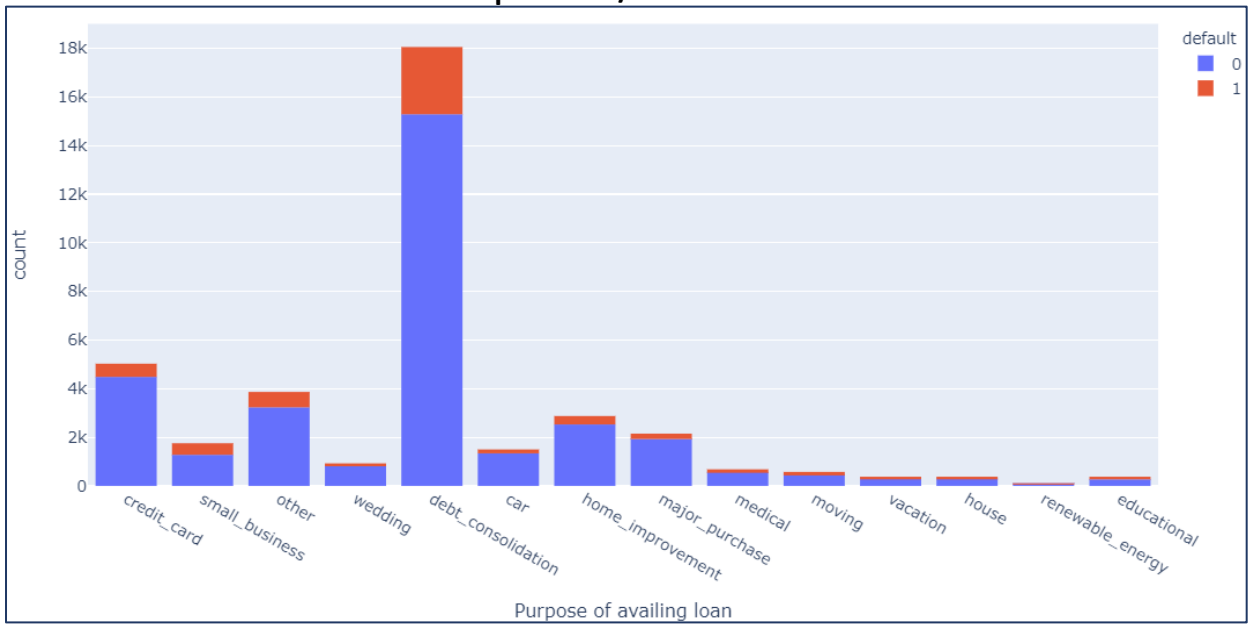
- **Hypothesis:**
  - There can be host of purposes which are more prevalent than the others. No hypotheses needed. This is a factual scenario.

- **Univariate Analysis:**
  - Debt consolidation is the most prevalent purpose across all the loan applications



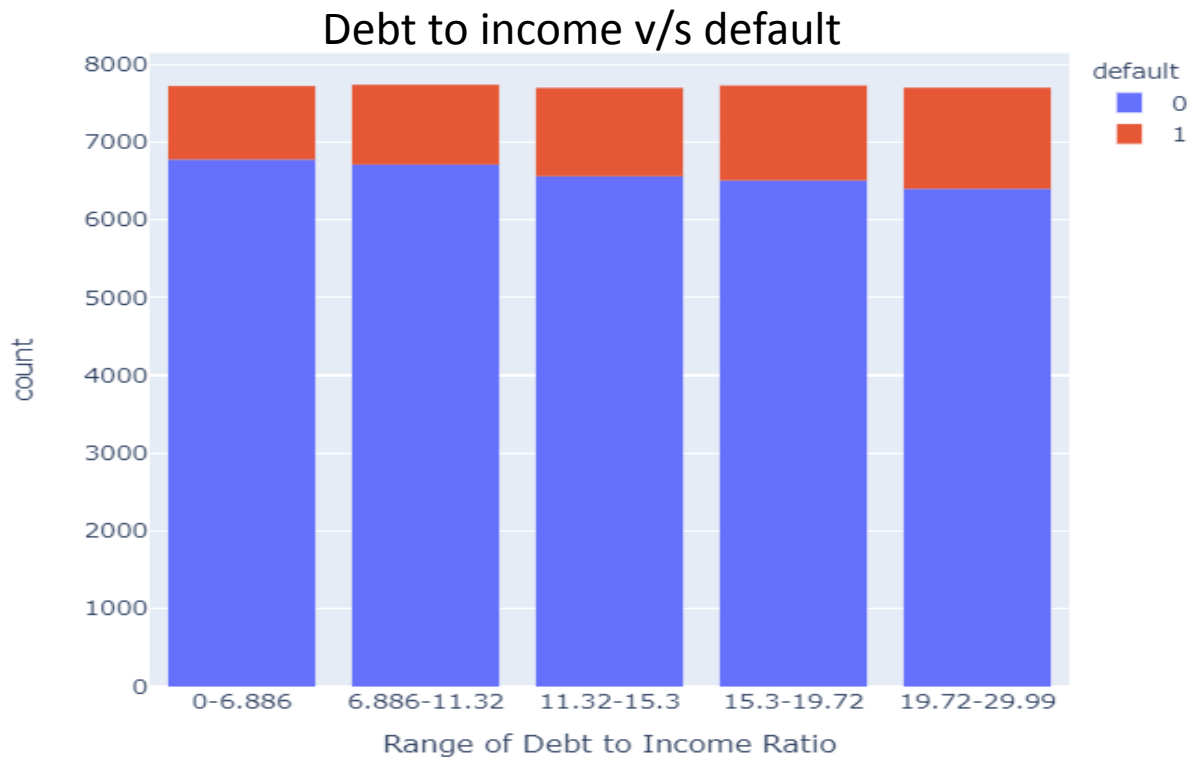
- **Bivariate Analysis (vs Charged-Off)**
  - Small business loans have the highest default rate at **~27%**, followed by renewable energy loans at **~19%**

Purpose v/s default



14. Analysis of ‘Debt to Income Ratio’

- **Hypothesis:**
  - Higher DTI ratios should yield higher default rates
- **Univariate and Bivariate Analysis:**
  - There is a steady increase in the default rate from **~12%** default to **~17%** with increase in the borrower's debt to income ratio

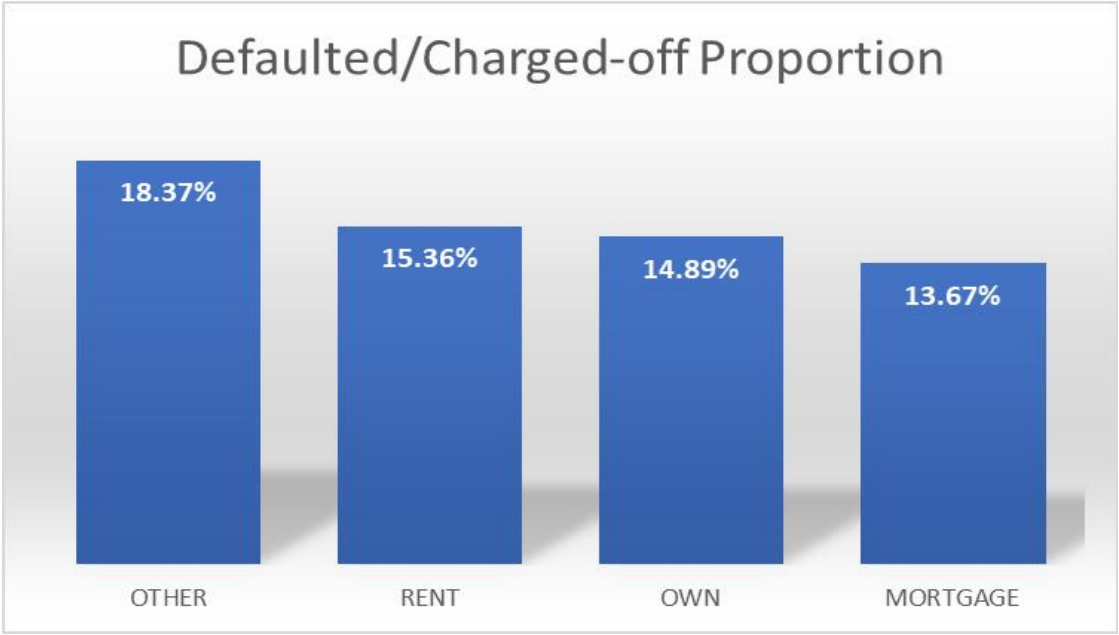


Hypothesis Confirmed

15. Analysis of ‘Home Ownership’

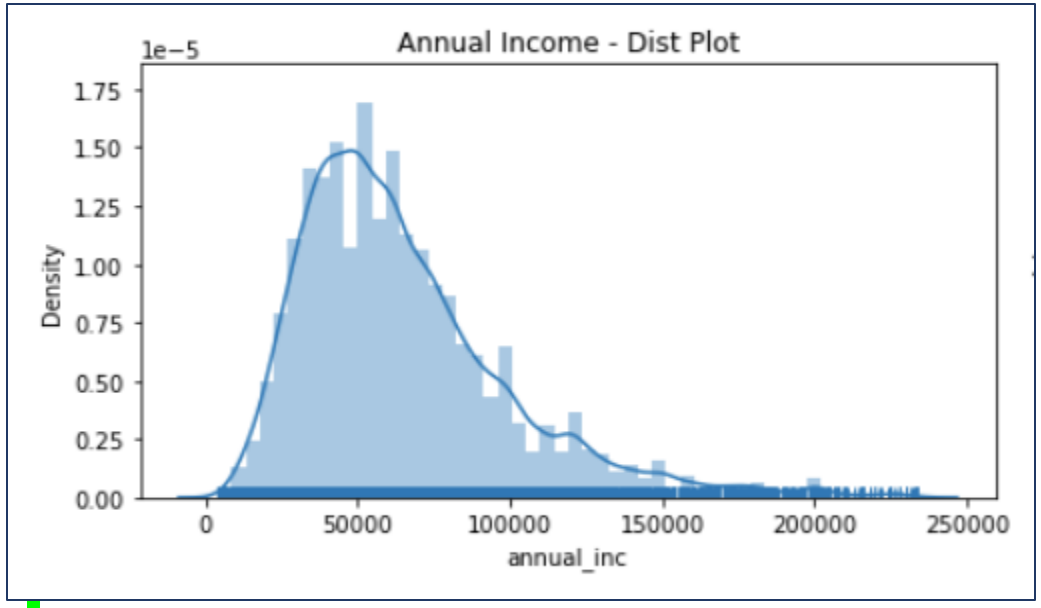
- **Hypothesis:**
  - Loan applications where borrower has his/her own home is less likely to default as compared to the applications where borrower lives in a rented or mortgaged accommodation.
- **Univariate Analysis:**
  - Out of the total loan applications, **92%** applications are from people who live in rented/mortgaged accommodations
  - Only **7.7%** loan applications are from people who live in their own houses
- **Bivariate Analysis (vs Default/Charged-Off)**
  - There is not a significant difference between the charged off proportions across different categories of Home ownership of the loan applicants.
  - No useful insights can be drawn from this variable to predict the repayment behavior of the applicants.

Hypothesis Not Confirmed

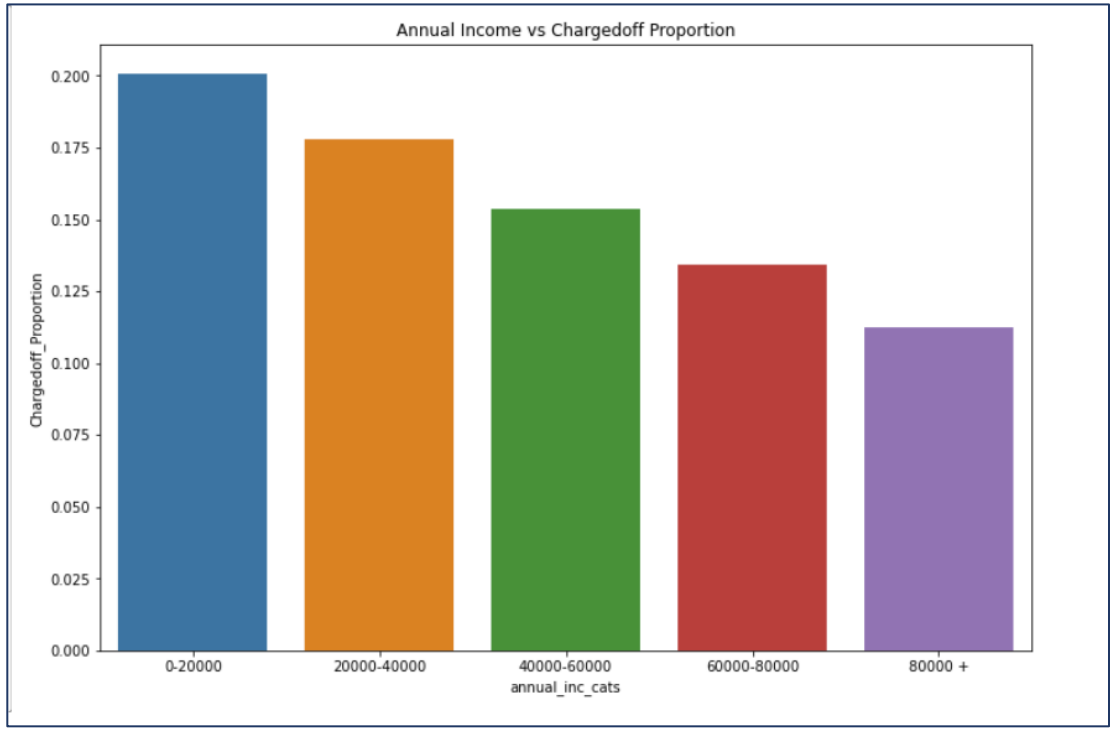


16. Analysis of ‘Annual Income’

- **Hypothesis:**
  - Higher Annual income should lead to better financial stability and positive repayment outlook, thus leading to lower default/charged-off rates. There should be an inverse proportional relationship.
- **Univariate Analysis:**
  - Most of the loan applications corresponds to borrowers with annual incomes between **40K-80K**
- **Bivariate Analysis (vs Charged-Off)**
  - There is a clear decreasing trend of charged off proportion (default rate) as the annual incomes are increasing. People in **0-20K** are most likely to default.



Hypothesis Confirmed



17. Analysis of ‘Verification Status’

- **Hypothesis:**
  - If the borrower’s income has been “Verified” or at least the source of the income has been verified (“Source Verified”) by the lending club then a particular loan application seems more credible and can be attributed to having lower default rates when compared to an application where borrower is marked as “Not Verified”
- **Univariate Analysis:**
  - **57.4%** loan applications are marked as either Verified or Source Verified
  - **42.6%** loan applications are marked as Not Verified
- **Bivariate Analysis (vs Default/Charged-Off)**
  - Loan applications with “Verified” status have an average default rate of **17%**, “Source Verified” correspond to **14%**
  - Loan applications which were “Not Verified” have an average status of **12.8%** which is contrary to the hypothesis
  - Not significant differences between percentages to draw any conclusion

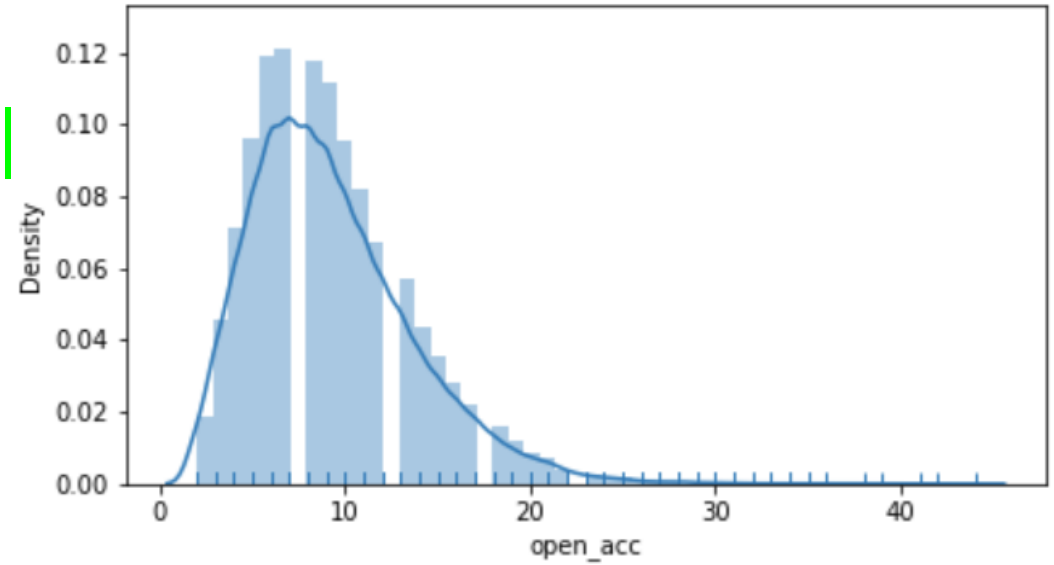
loan_status	verification_status	Charged Off	Fully Paid	Total	Chargedoff_Proportion
2	Verified	2032	9999	12031	0.168897
1	Source Verified	1422	8134	9556	0.148807
0	Not Verified	2130	14474	16604	0.128282

Hypothesis not confirmed

18. Analysis of ‘Open Credit Lines’

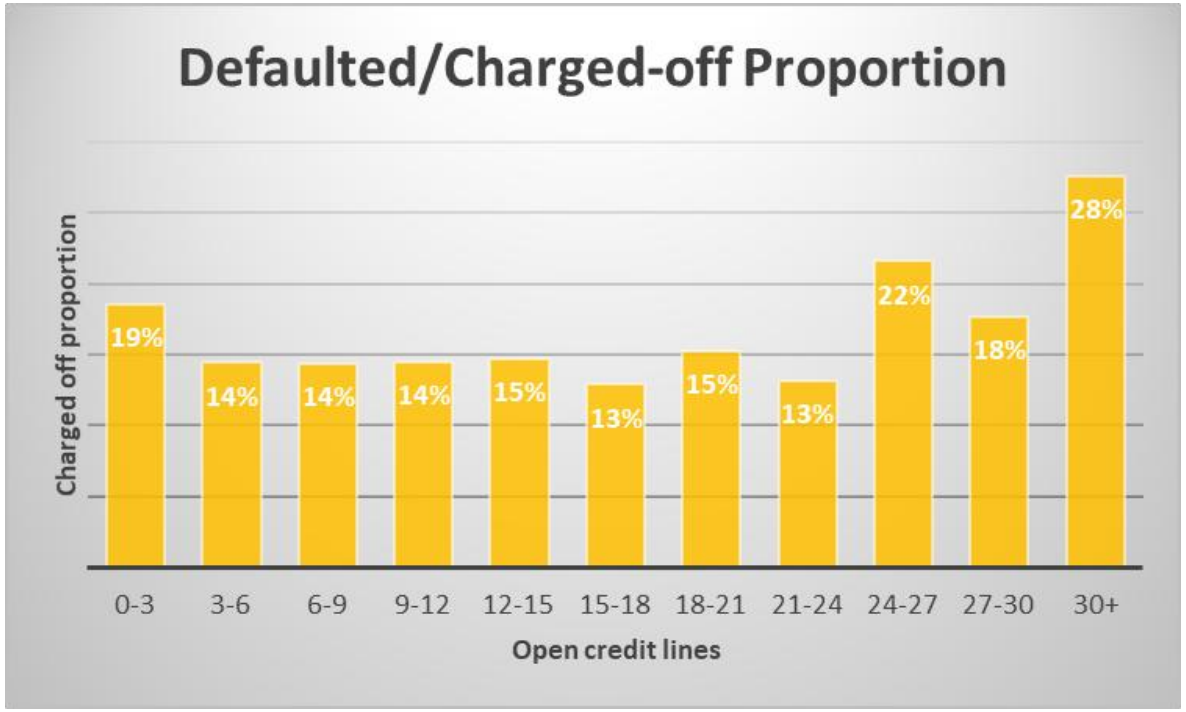
- **Hypothesis:**
  - A higher number of open credit lines currently in the borrower’s credit files can indicate increased load on a borrower but there is no clear behavior that can be attributed to default rates driven by Open credit Lines.

- **Univariate Analysis:**
  - Most of the borrowers have open credit lines in the range **5-15**



Hypothesis confirmed

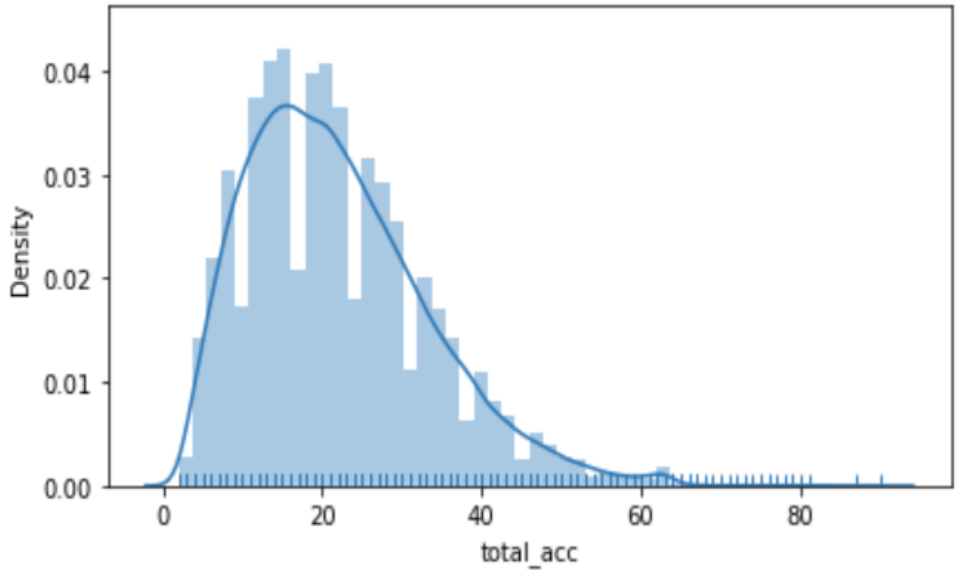
- **Bivariate Analysis (vs Charged-Off)**
  - There is a weak increasing trend of charged off proportion as the open credit lines increase



19. Analysis of ‘Total Credit Lines’

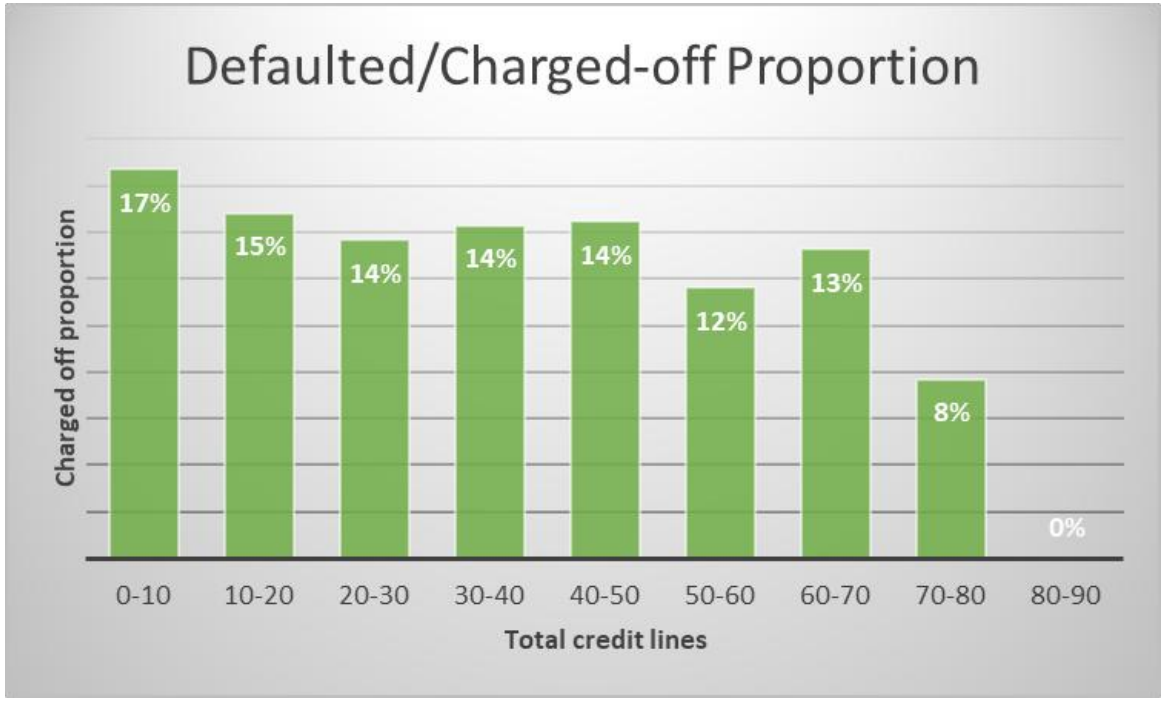
- **Hypothesis:**
  - A higher total number of credit lines currently in the borrower’s credit files can indicate bank’s confidence in the borrower’s ability to repay the loans and thus should be inversely proportional to default rates

- **Univariate Analysis:**
  - Most of the borrowers have total credit lines in the range **10-30**



Hypothesis Confirmed

- **Bivariate Analysis (vs Charged-Off)**
  - There is a weak decreasing trend of charged off proportion as the total credit lines increase



20. Analysis of ‘Derogatory Public Records’

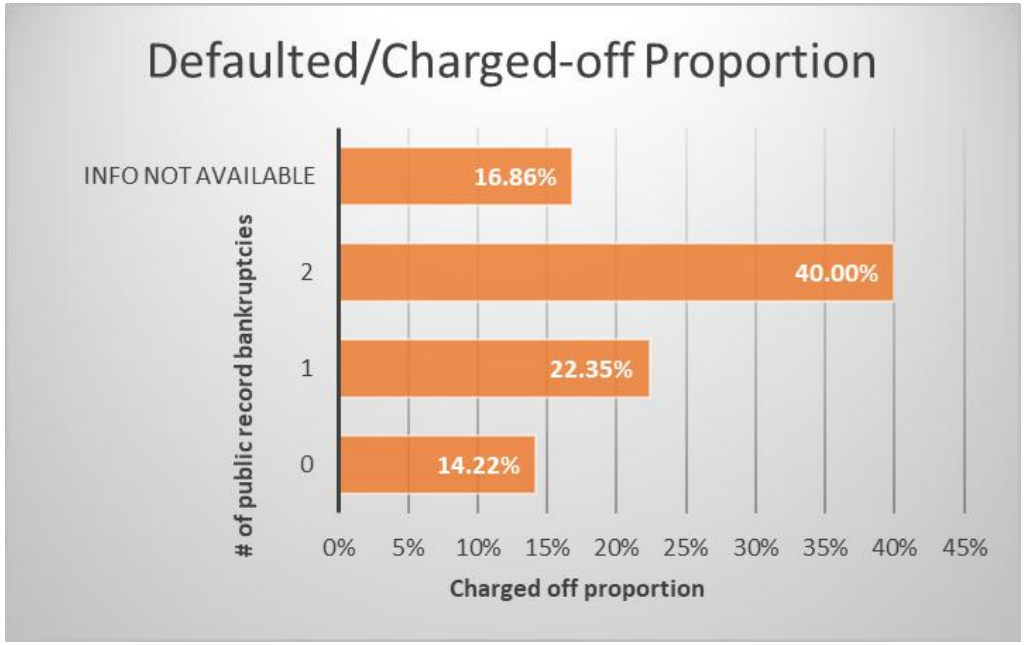
- **Hypothesis:**
  - A higher number of derogatory public records indicate towards riskier borrower characteristics which should lead to higher default/charged-off rates
- **Univariate Analysis:**
  - **95%** of loan applications are from borrowers having **0** derogatory public records
  - **5%** applications are from people having **1** derogatory public record
  - Only **0.15%** applications belong to people having more than **1** derogatory public records
- **Bivariate Analysis (vs Default/Charged-Off)**
  - Loan applications corresponding to **0** public derogatory records have an average default rate of **14%**
  - Loan applications corresponding to **1** public derogatory record have an average default rate of **22% (Increase of 8%)**, thus they are far more likely to default than borrowers with a clean record

Hypothesis Confirmed

21. Analysis of 'Public Record Bankruptcies'

- **Hypothesis:**
  - A higher number of public record bankruptcies indicate towards poor borrower repayment behavior which should lead to higher default/charged-off rates
- **Univariate Analysis:**
  - **94%** of loan applications have 0 public record bankruptcies against the applicants
  - **4.2%** of the loan applications are from applicants having **1** public record bankruptcy
  - **2%** loan applications have either No record or are from applicants having **more than 1** public record bankruptcy
- **Bivariate Analysis (vs Default/Charged-Off)**
  - Loan applications corresponding to **0** public record bankruptcies have an average default rate of **14%**
  - Loan applications corresponding to **1** public record bankruptcy have an increased average default rate of **22%**
  - We only have **5** loan applications from applicants having 2 public record bankruptcies, thus, there is not sufficient data to derive any conclusion though average default rate climbed upto **40%** for these **5** cases inline with expectations.

Hypothesis Confirmed



22. Analysis of rest of the variables

- **The EDA of the following variables did not yield significant insights**

  - **Issue date** - No significant relationship could be identified with default rate
  - **Earliest credit line** - This has no impact on the probability of a loan application defaulting or not. it ranges from **13-15%** implying that the default rate is not significantly affected by how much time has elapsed since the borrowers' first credit line
  - **Installment** - No significant change in default rate (**13-16%**) across various installment ranges
  - **State** -
    - State **CA** has the highest share of loan applications at **18%** followed by **FL** and **NY**
    - Top **7** states contribute to **~50%** of the total loan applications
    - State **NE** has the highest default rate at **60%** but it has very low volume of loan applications; **CA** has the highest number of defaulters
  - **Delinquency 2 years**
    - **89%** of the loan applications have no 30+ DPD's in the last 2 years
    - The default (charged-off) rate shows a steady increase with 30+ DPD's in the last 2 years increasing

Major features to predict default scenario

- Purpose of availing the loan
- Annual Income of the borrower
- Term of loan repayment
- Interest rate offered by the lender
- Public derogatory record or bankruptcy record
- Debt to income ratio
- Loan amount
- Funded amount and amount funded by investors
- LC assigned grade
- LC assigned sub-grade

Recommendations

Basis on the detailed exploratory data analysis done on loans.csv, we have concluded the following pointers:

- Loan applications for small businesses, renewable energy and educational purposes have a risk of defaulting. They should be given out cautiously after verification and also at higher interest rates.
- Annual income is a major factor that should be considered before approving/declining a loan application. Applicants having below **40K** annual income are prone to high default rates. Loan should be disbursed to them for shorter terms and higher interest rates.
- If a loan applicant has even one public derogatory record or public record of bankruptcy, then bank should be extremely cautious of disbursing loan to such an applicant. There are significantly higher chances of default in such cases.
- Debt to income ratio is also an important variable and should be as low as possible. It should be checked along with other variables such as already open credit lines. If there is an additional debt burden on the borrower with no increase in income, there is a high risk of default
- If the loan amount is extremely high (**15,000+**), the lender and investors should exercise caution and do enhanced due diligence on the customer as there is significantly higher risk of default than in case of lower amounts
- If the LC assigned grades and sub-grades are lower, then it is a red flag with high risk of default and such loans should be avoided or given out at higher interest rates