Lending Club Case Study

Upgrad

Naga Shanmukha Kumar Atte
&
Soumojit Ash

Problem Statement

Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

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.

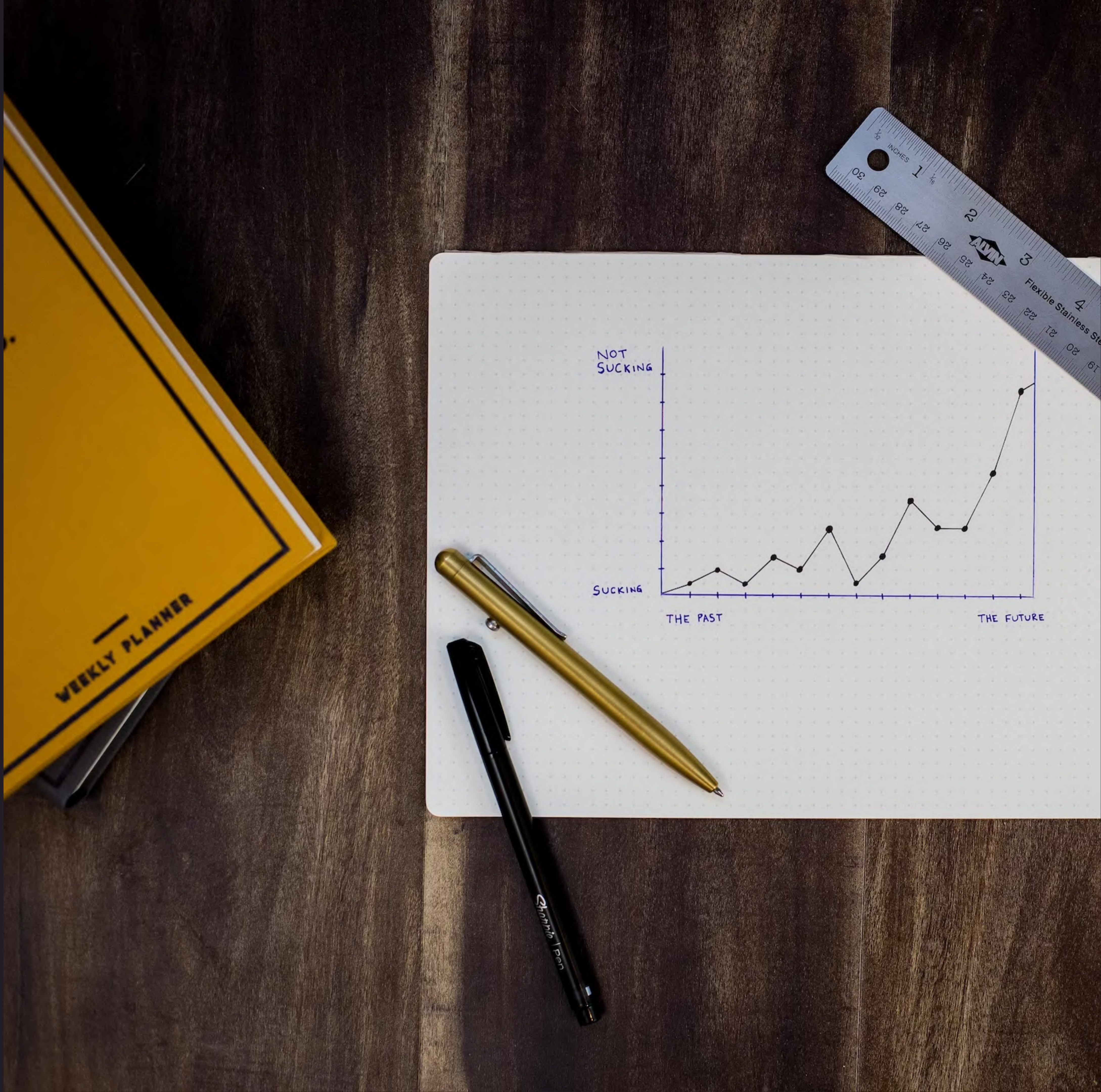
Business Objectives

- Use EDA to understand how consumer attributes and loan attributes influence the tendency of default
- Aim of this case study is to use EDA to identify such applicants.
- Understand the driving factors behind loan default i.e. the variables which are strong indicators of default

Analysis Approach

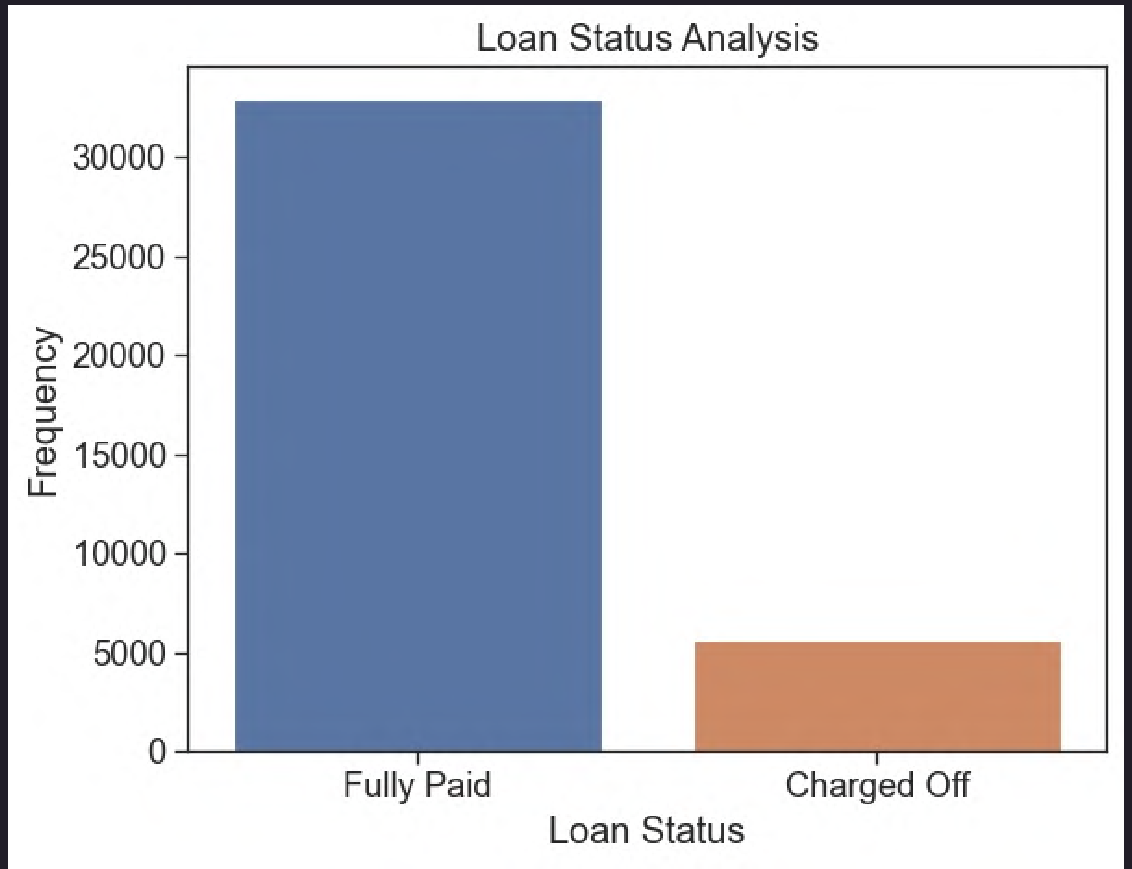
Approach followed

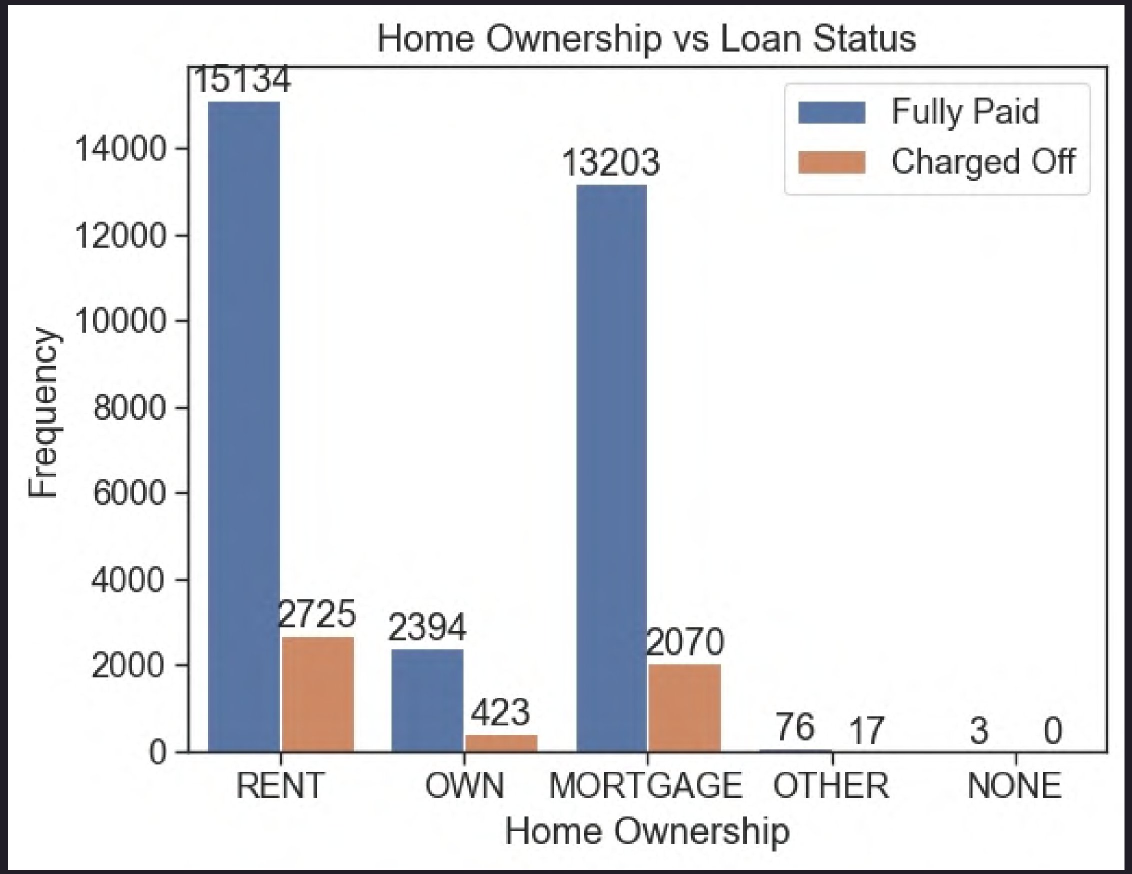
1. Data Handling & Cleaning & Sanity Checks
    a. Treating missing & unique values
    b. Treating incorrect / invalid data
2. Deriving additional columns
3. Univariate Analysis
    a. Unordered Categorical Variable Univariate Analysis
    b. Ordered Categorical Variable Univariate Analysis
    c. Continuous / Numeric Variable Univariate Analysis
4. Bivariate Analysis

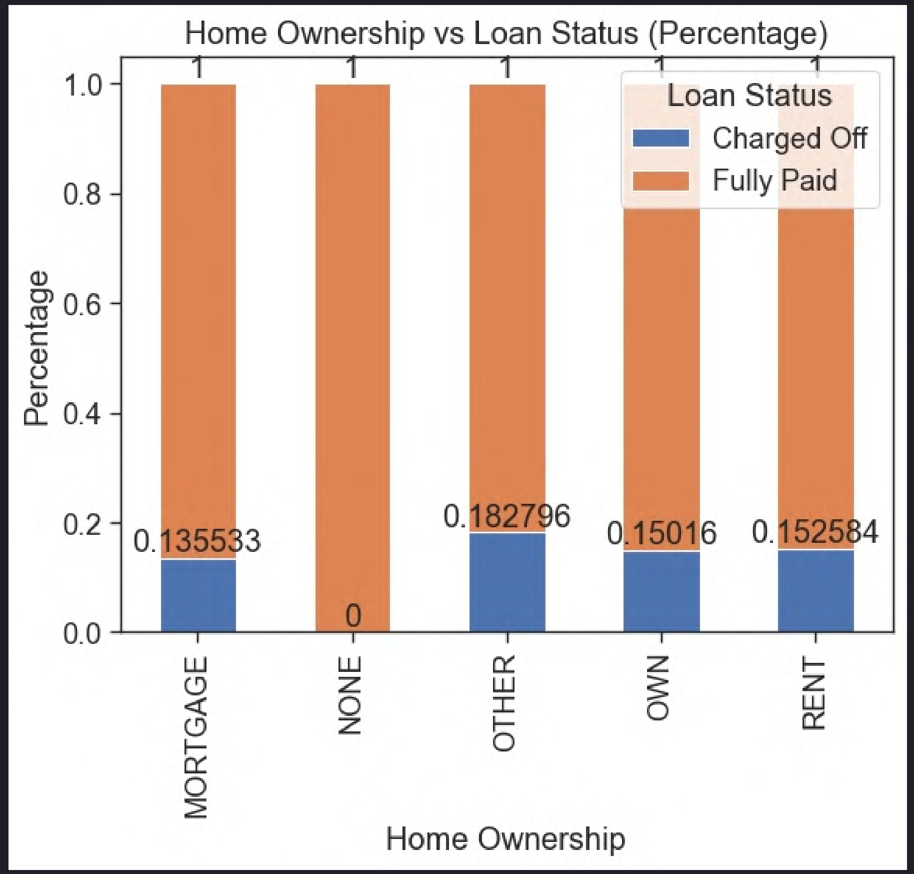


Univariate - Unordered Categorical Variables

Univariate Analysis - Unordered Categorical Variables - Key Observations







Loan Status Analysis

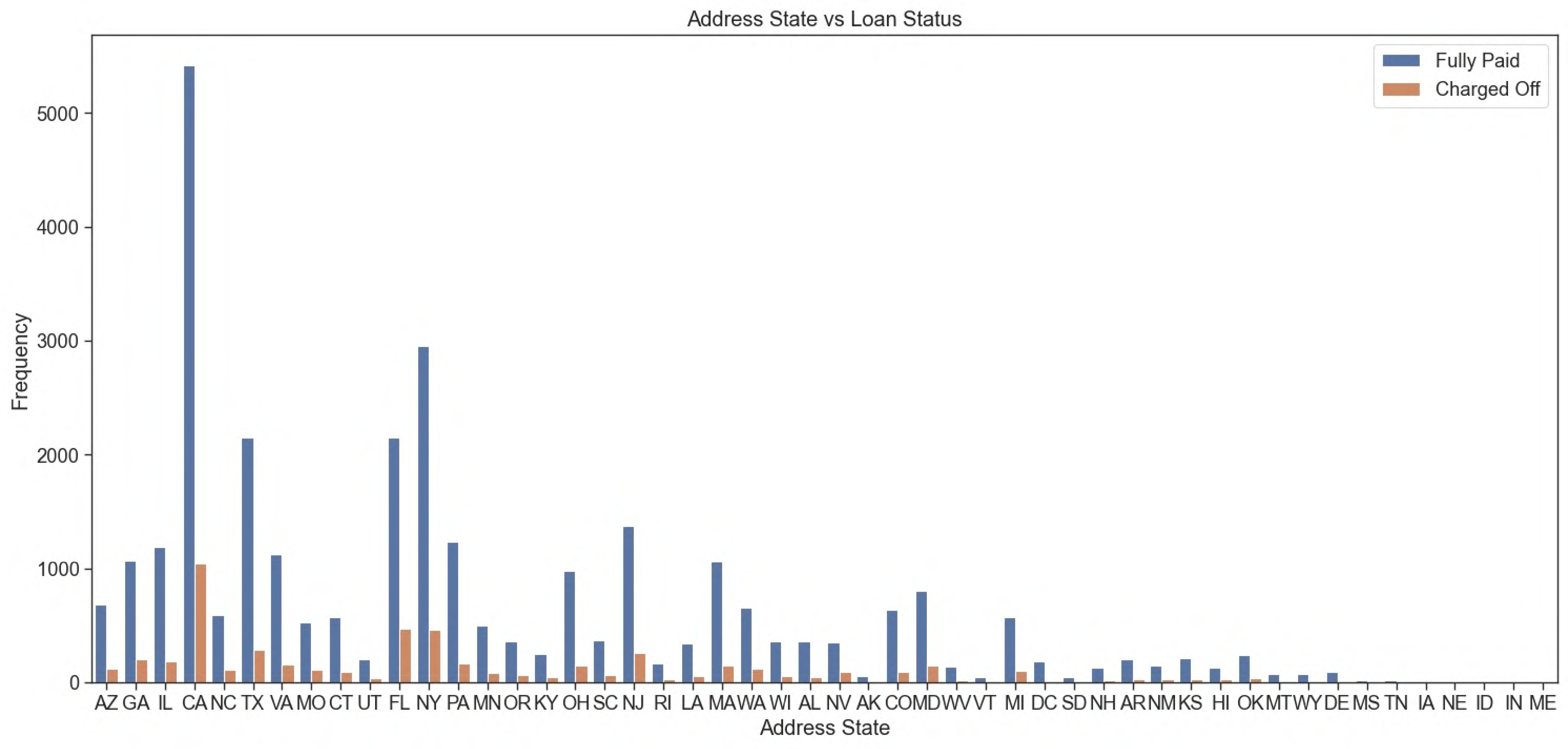
85% of the loan applicants data provided is fully paid

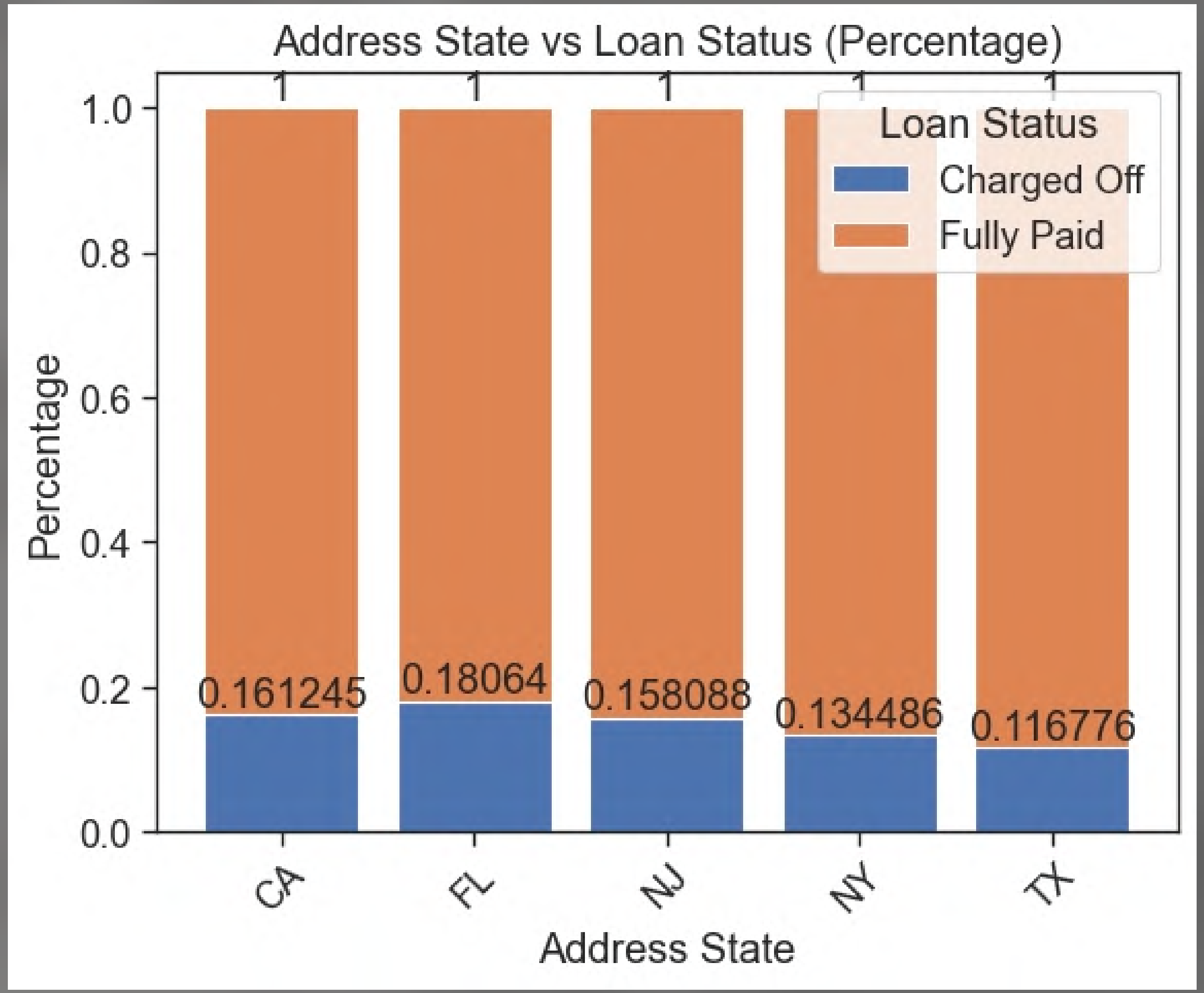
Home Ownership Analysis across loan statuses

1. More number of loan applicants have Home Ownership as 'RENT' followed by 'MORTGAGEE'.
2. However when we observe charged off data, proportion of 'RENT' is slightly higher

Try Pitch

Univariate Analysis - Unordered Categorical Variables - Key Observations

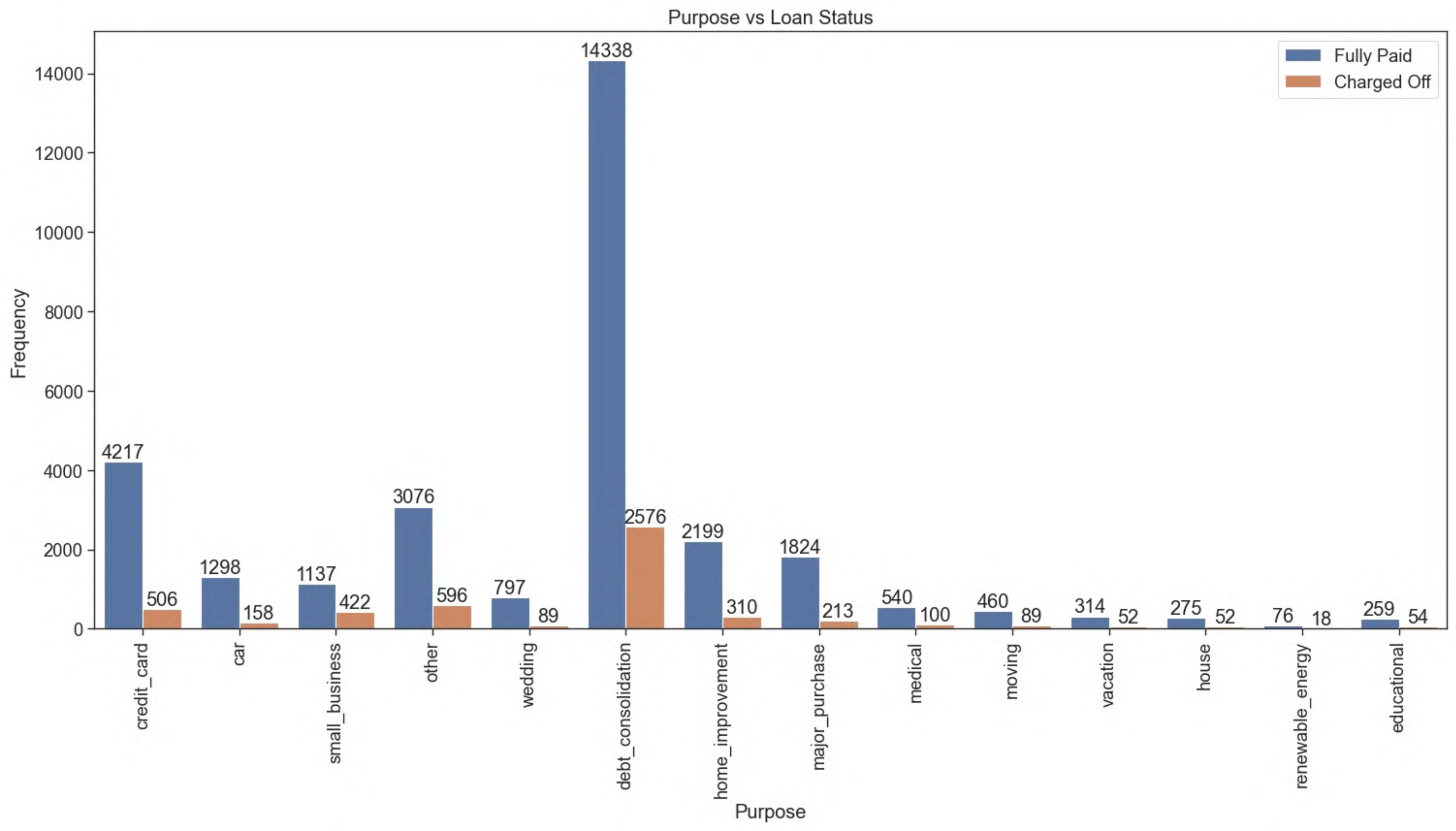


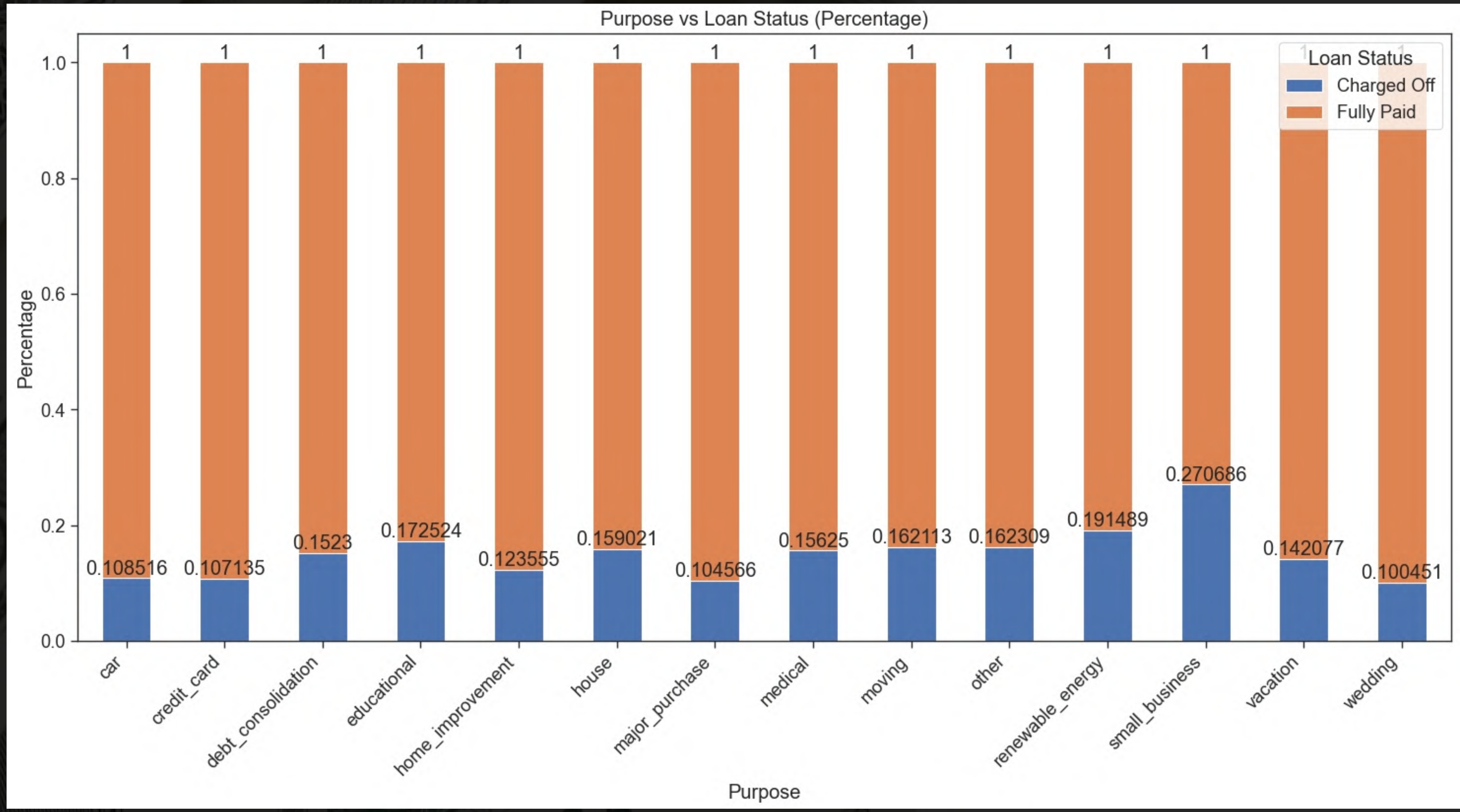


Address state across loan statuses

1. Address states wise, data is mainly coming from CA followed by NY, TX, FL and NJ
2. However proportion of charged off data is more in FL followed by CA

Univariate Analysis - Unordered Categorical Variables - Key Observations



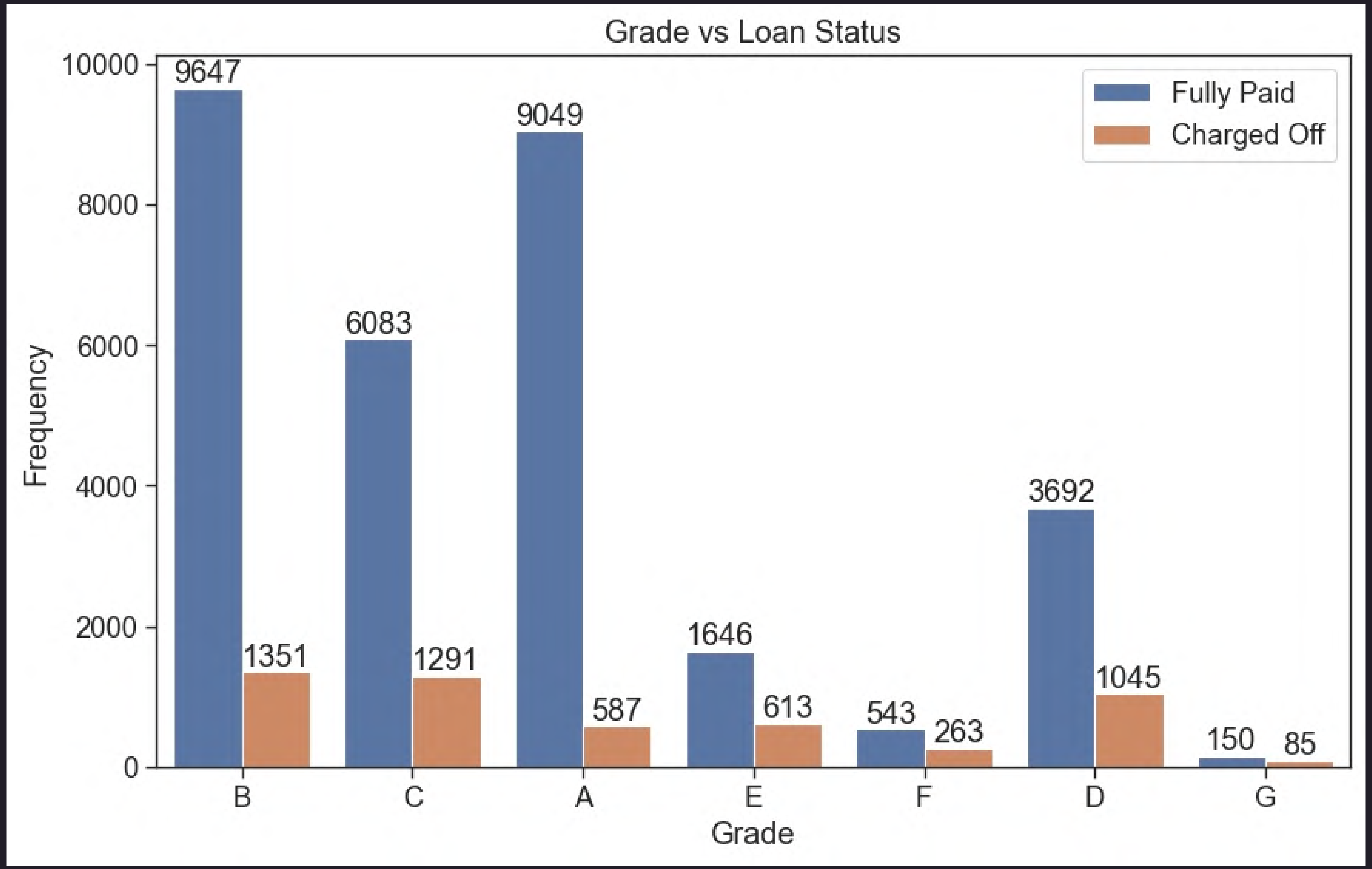


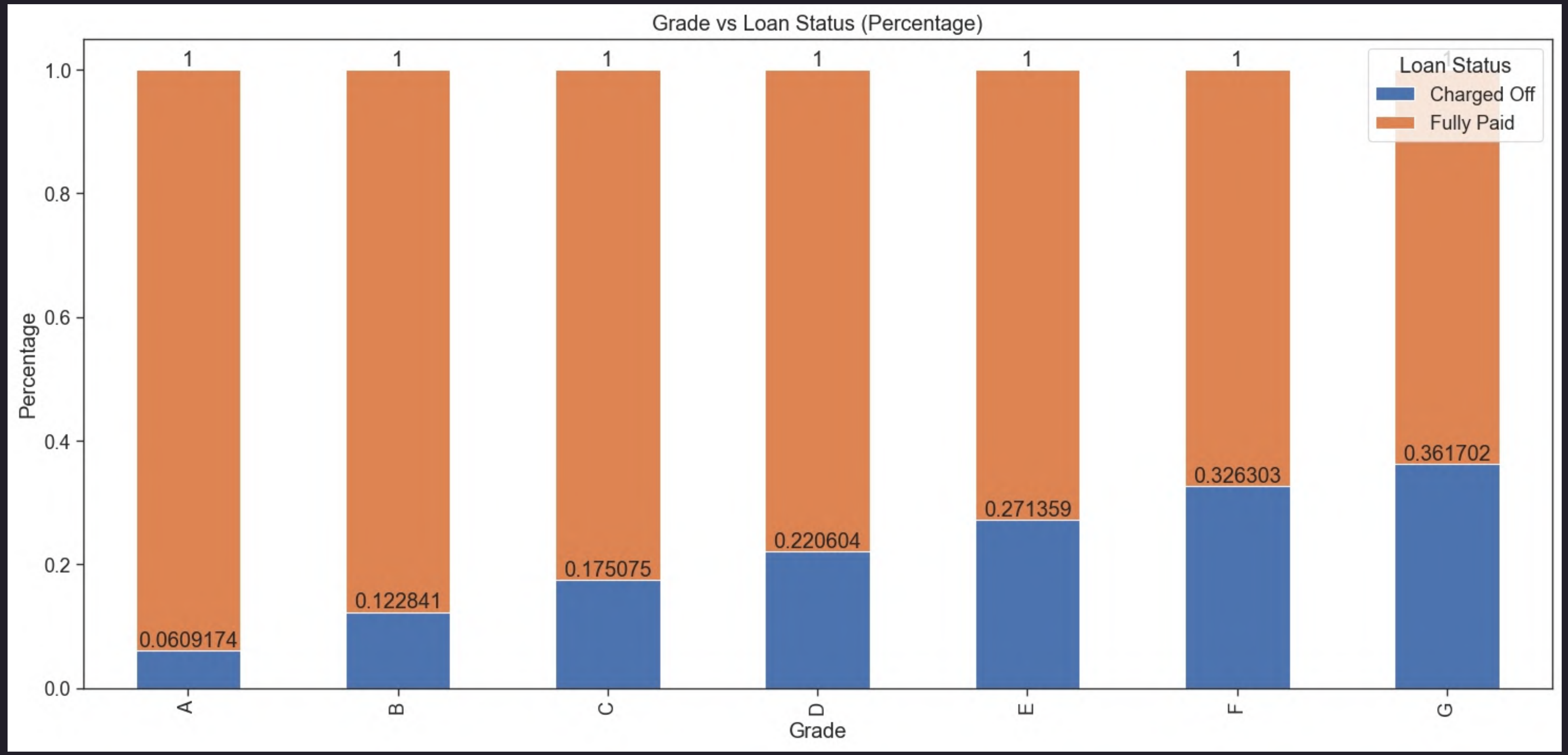
Purpose across loan statuses

1. Most of the loan applicants have given the purpose as 'debt consolidation' distantly followed by 'credit card', 'other' etc.
2. Even though the data for 'small business' is small when compared with 'debt consolidation', close to 40% of the loan applications with purpose as 'small business' have been charged off.

Univariate Analysis - Ordered Categorical Variables

Univariate Analysis - Ordered Categorical Variables - Key Observations





Grade across loan statuses

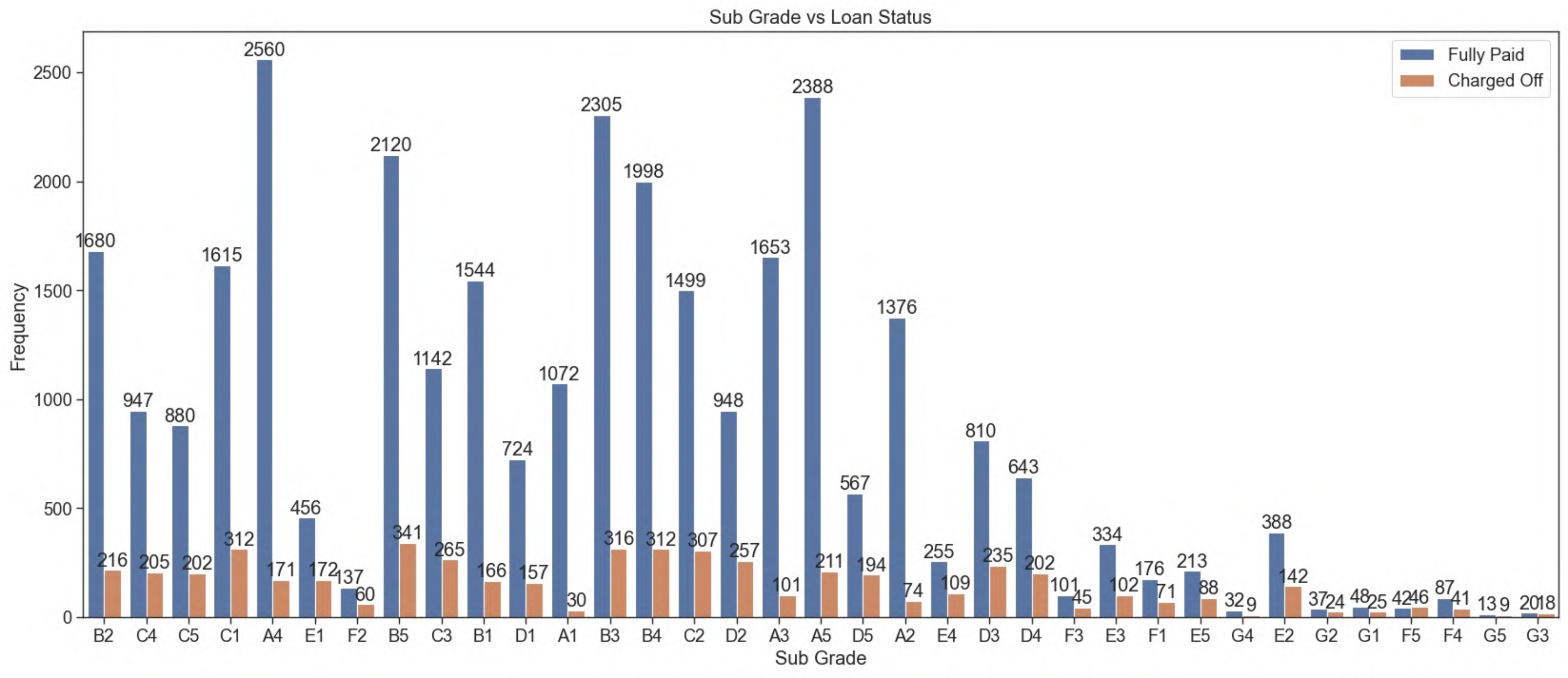
1. More number of applicants have grade as 'B' closely followed by 'A' and 'C'

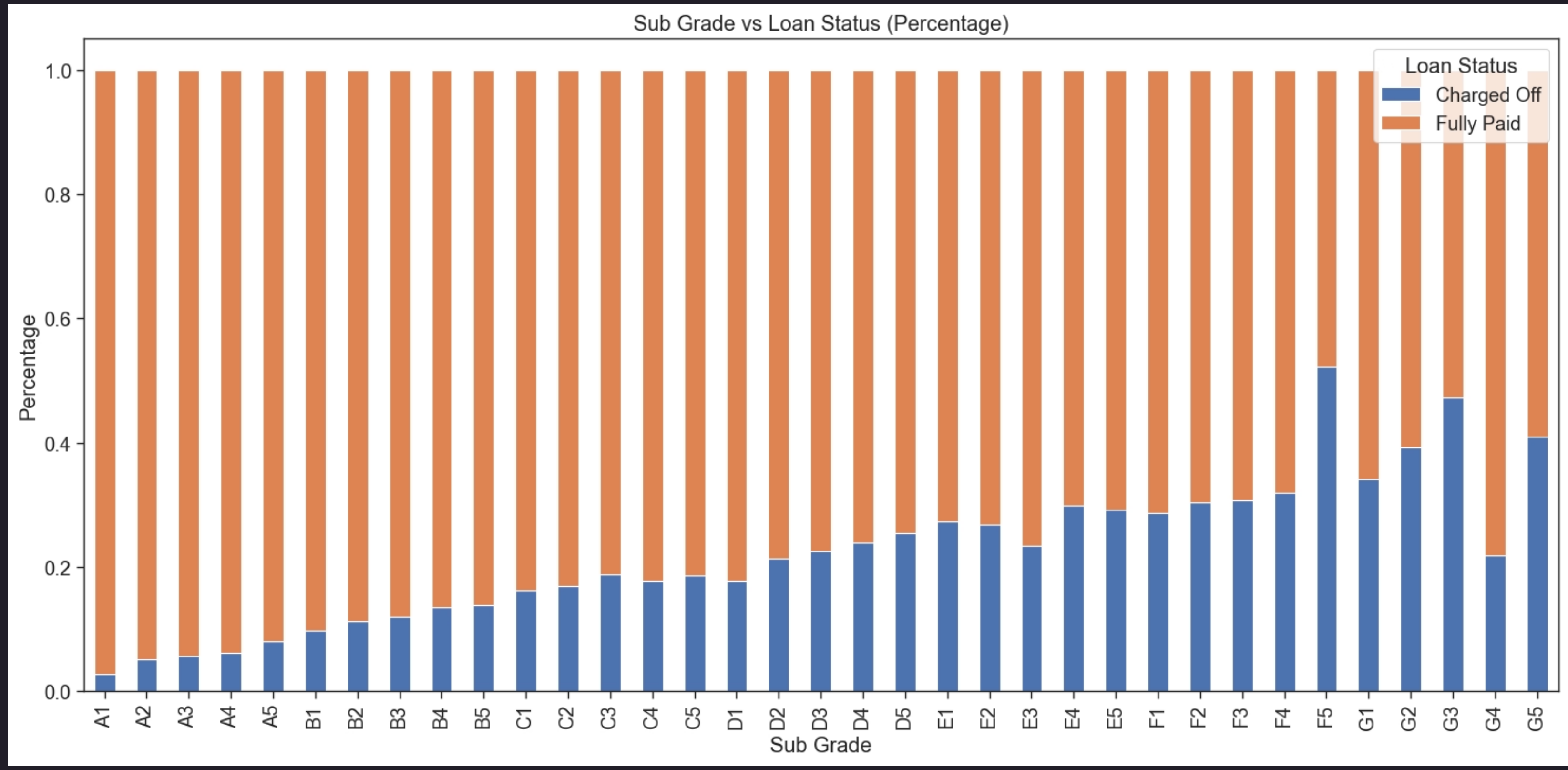
2. Even though the data is small for grades 'E', 'F' and 'G', proportion wise it is increasing as we move from A to G

3. If we look at only A, B and C as more data is from those grades, it is applicants with 'C' grade who are more defaulting followed by 'B' and then 'A'

4. As the Grade is an ordered categorical variable, and risk of the applicant increases from A to G, our results also indicate the same.

Univariate Analysis - Ordered Categorical Variables - Key Observations





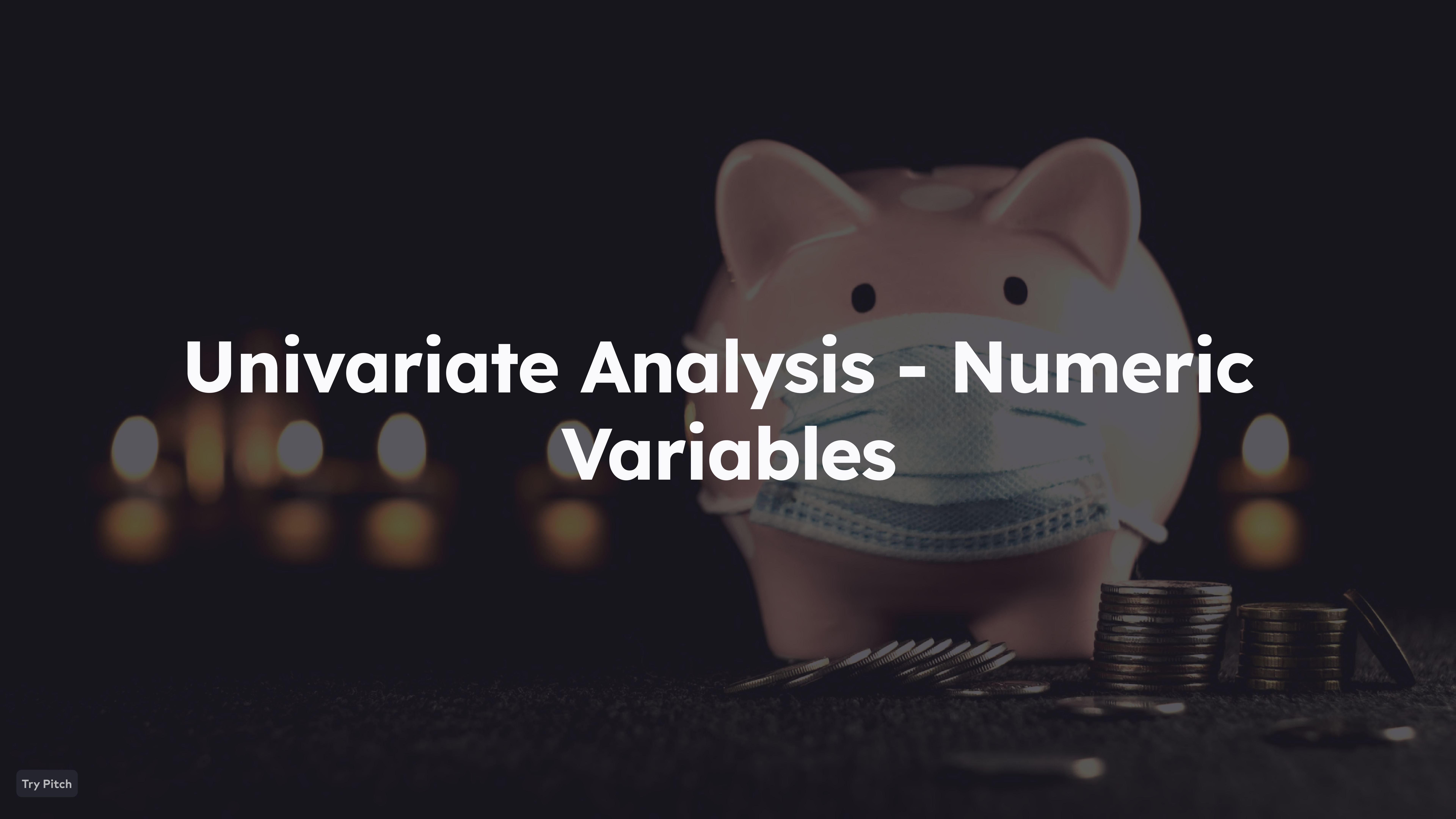
Sub Grade across loan statuses

1. More number of applicants have sub grade as 'A4', 'A5' followed by 'B3' and 'B5'

2. Even though the data is small for grades 'E', 'F' and 'G', proportion wise it is increasing as we move from A to G

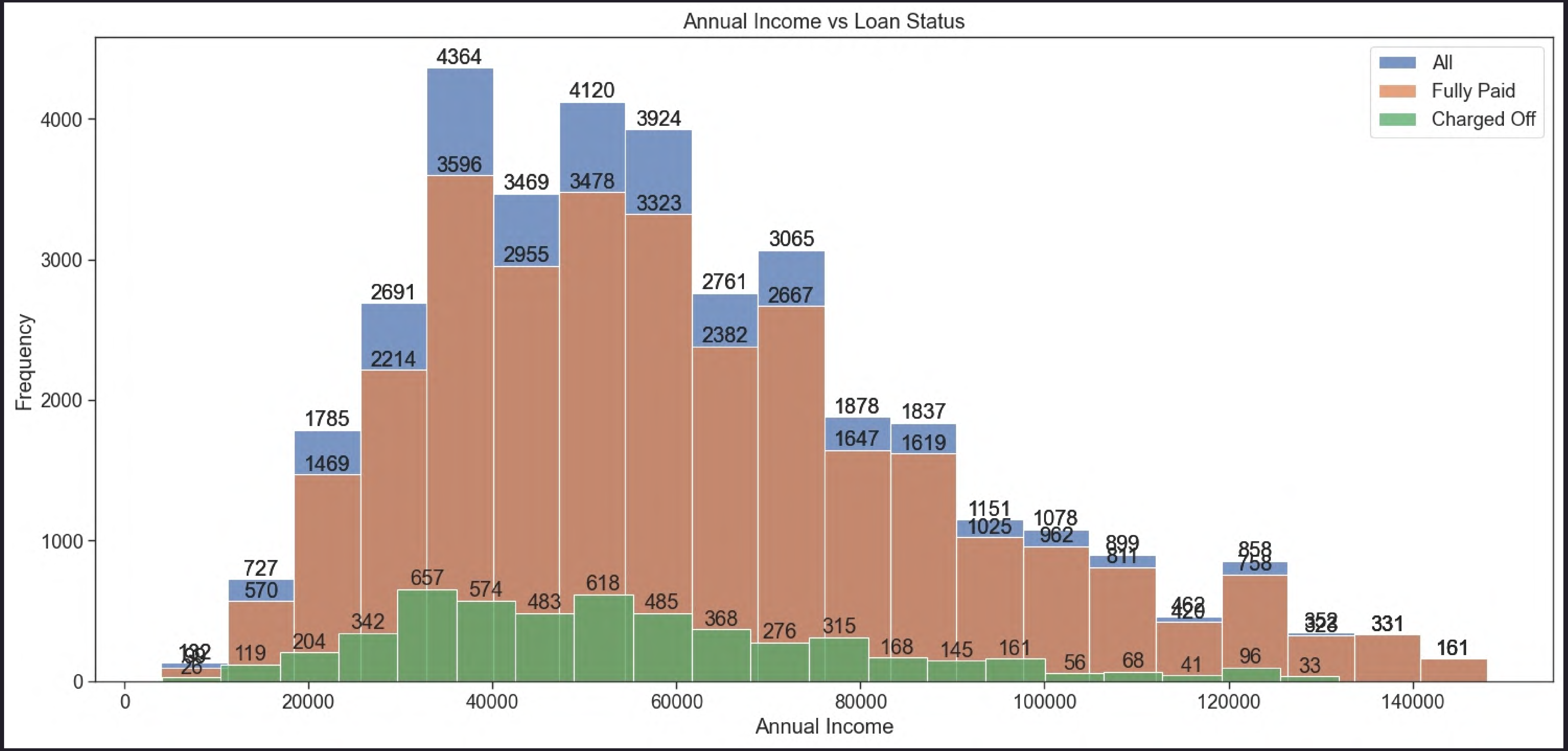
3. If we look at only A, B and C as more data is from those grades, it is applicants with 'C3' sub grade who are more defaulting followed by 'C4', 'C5'

4. Even these results also vindicate the sub grade significance



Univariate Analysis - Numeric Variables

Univariate Analysis - Continuous / Numeric Variables - Key Observations

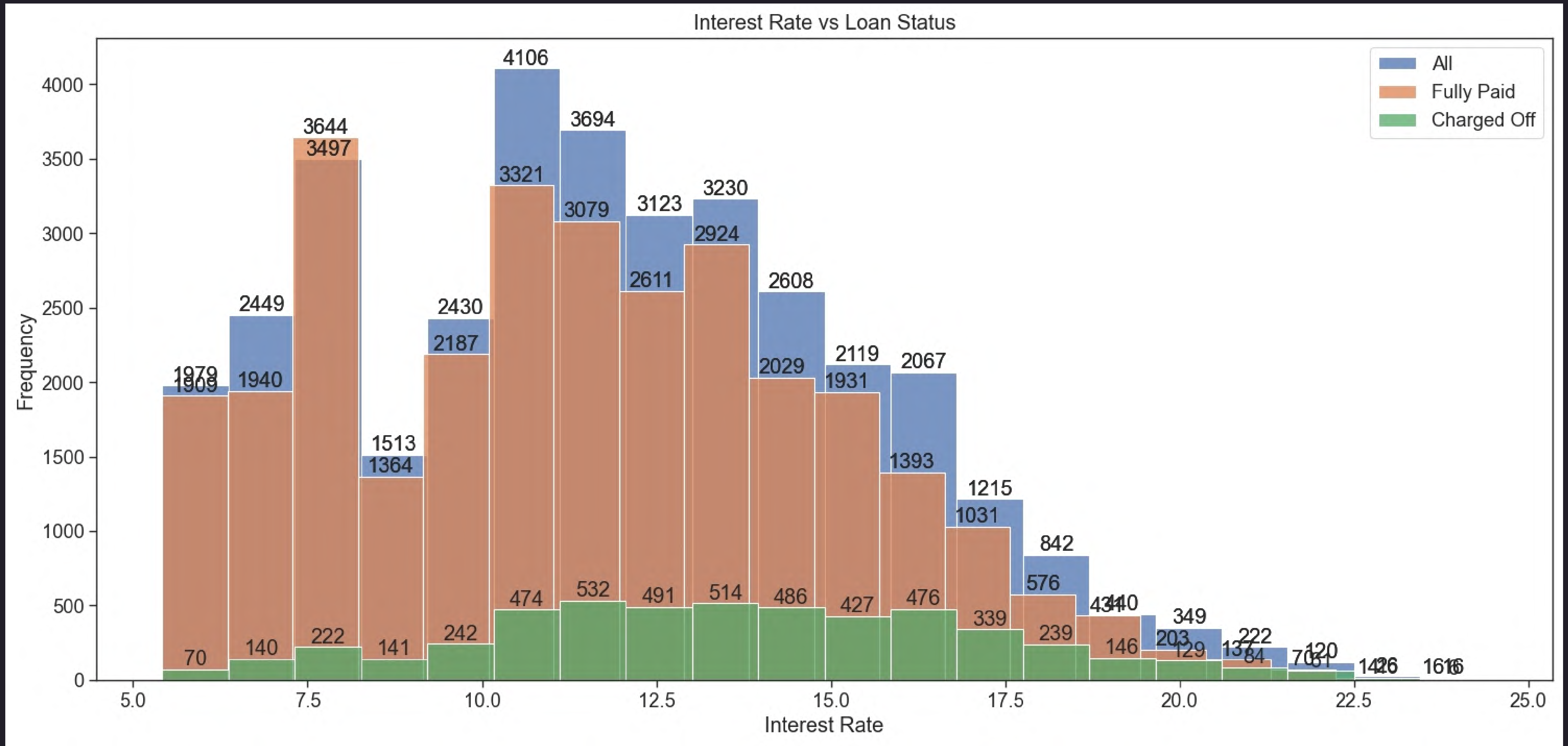


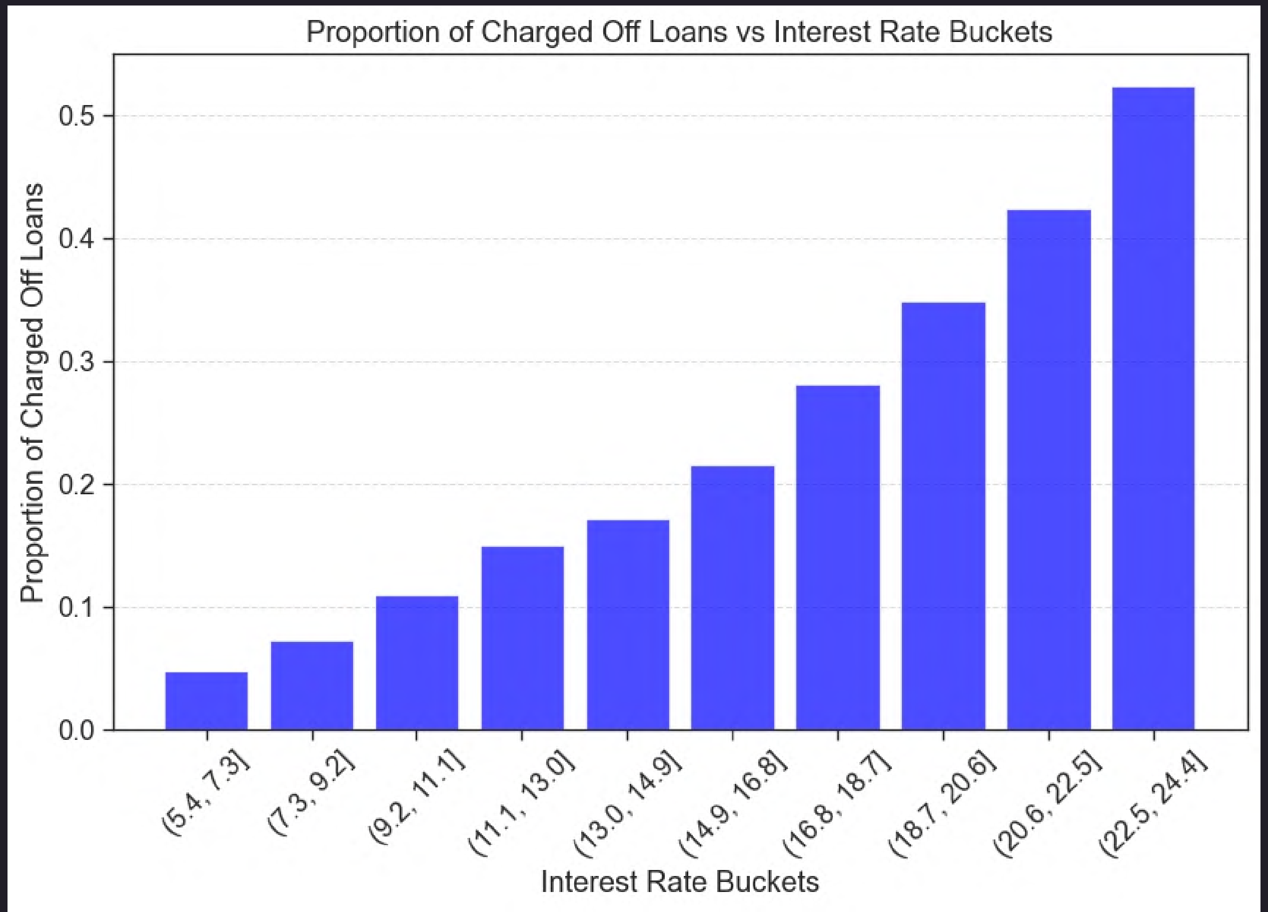
Annual Income across loan statuses

1. We see that highest number of loan applicants are from 30,000 to 40,000 followed by 40000 to 50000

2. Proportion wise, we see that majority of loan write offs are for loan applicants having annual income between 35000 and 50000

Univariate Analysis - Continuous / Numeric Variables - Key Observations

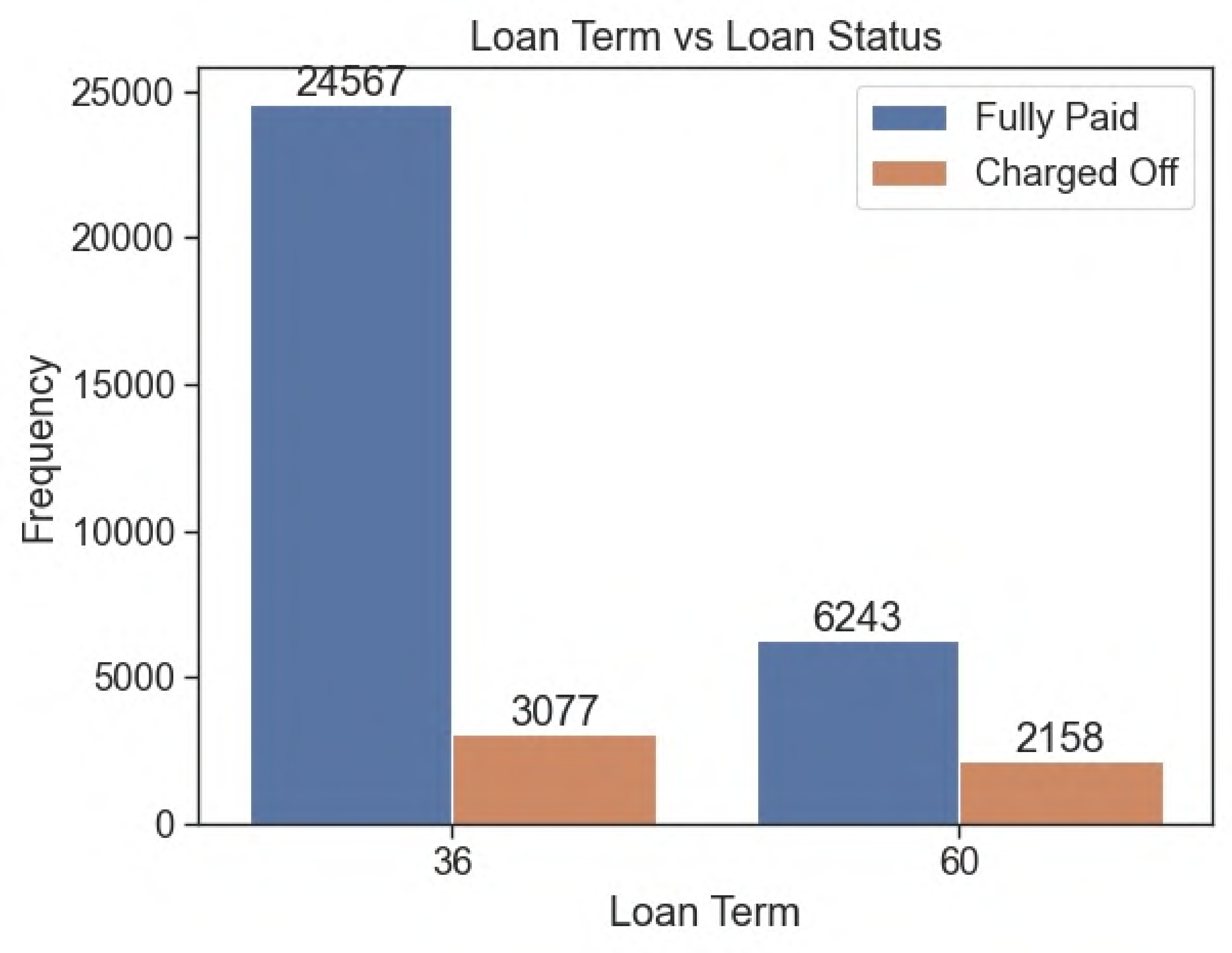


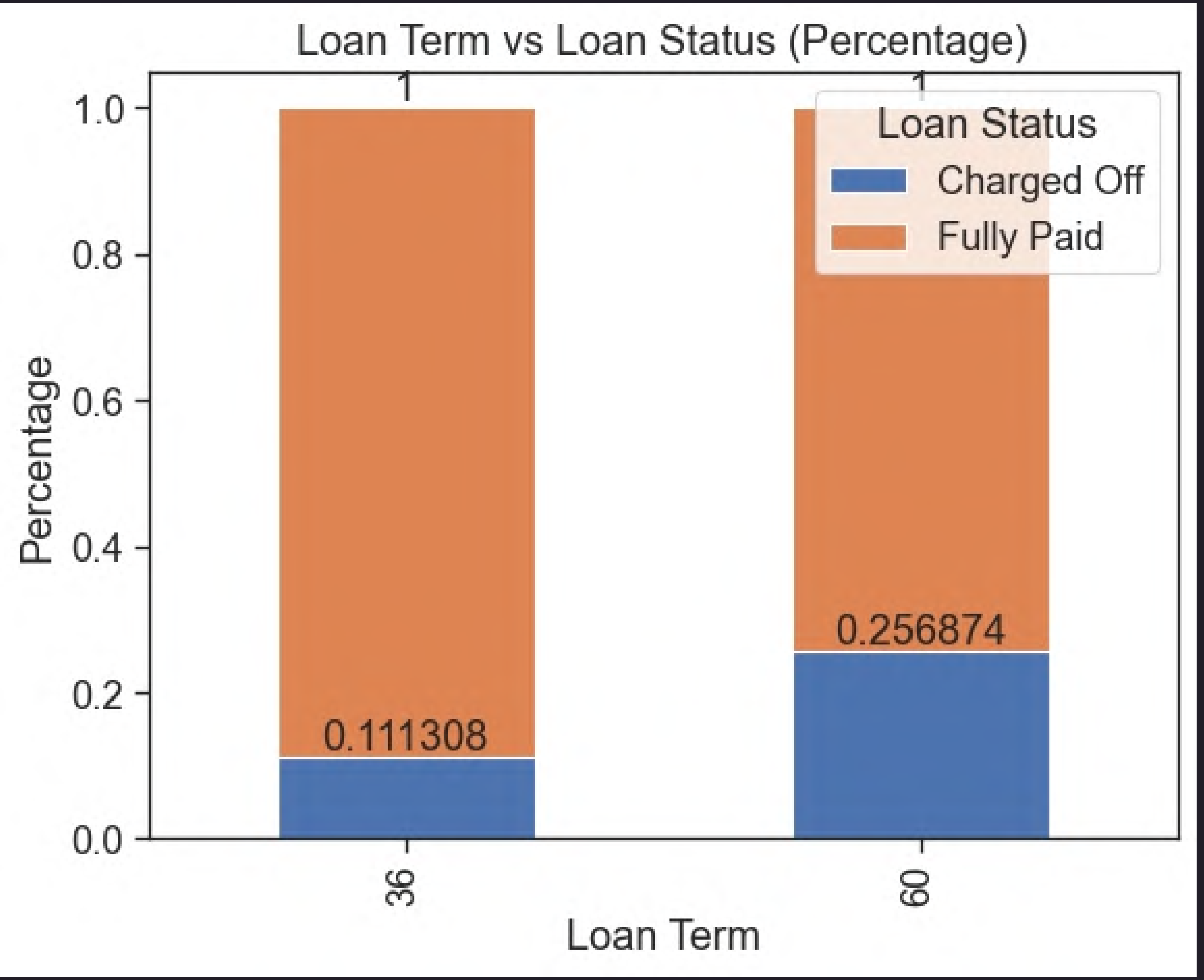


Interest rate across loan statuses

1. Majority of the loan applicants are having interest rate between 10 and 15
2. Even though the data is small for applicants with high interest rate, proportion wise we are able to see that charged offs are increasing as the interest rate grows

Univariate Analysis - Continuous / Numeric Variables - Key Observations

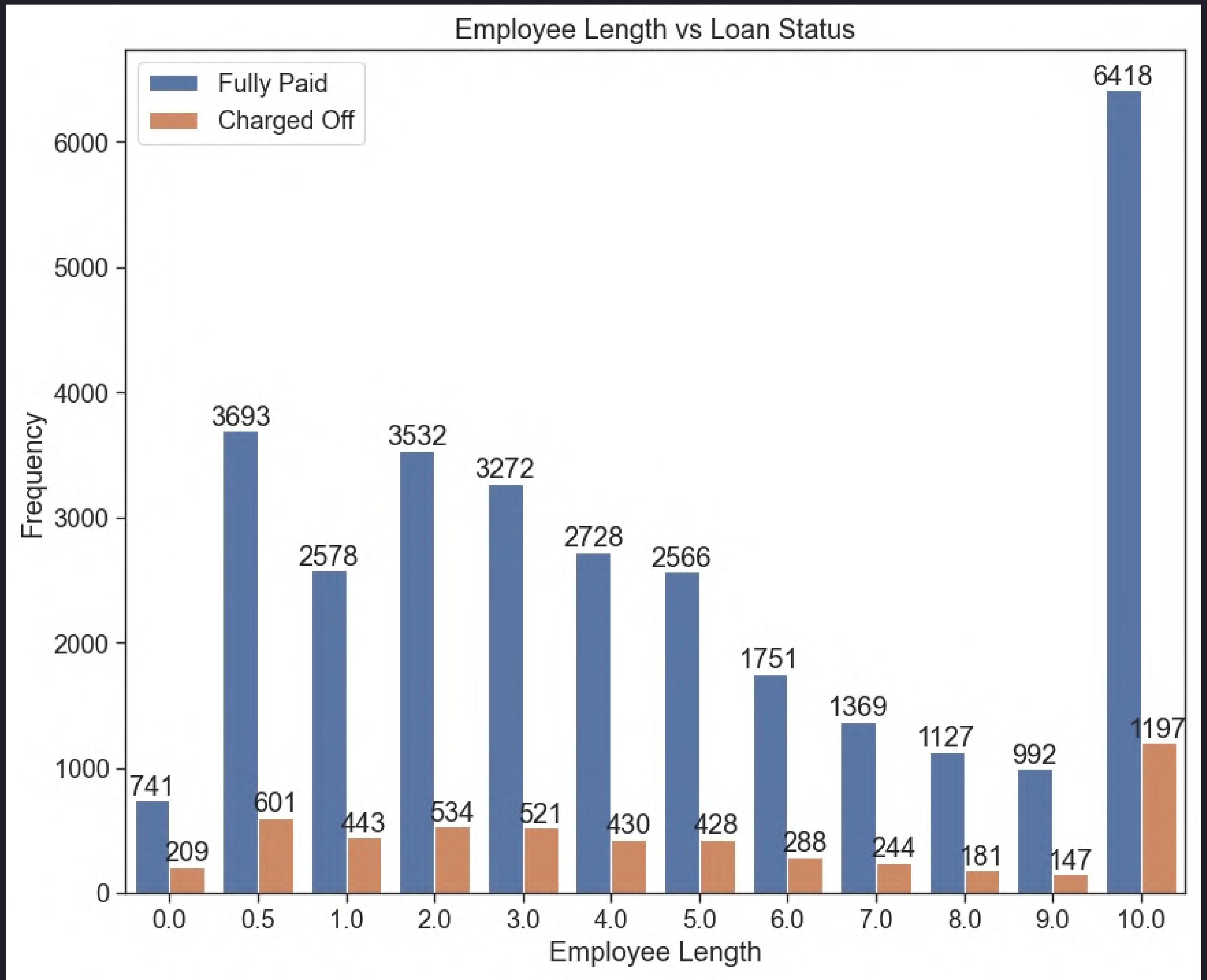


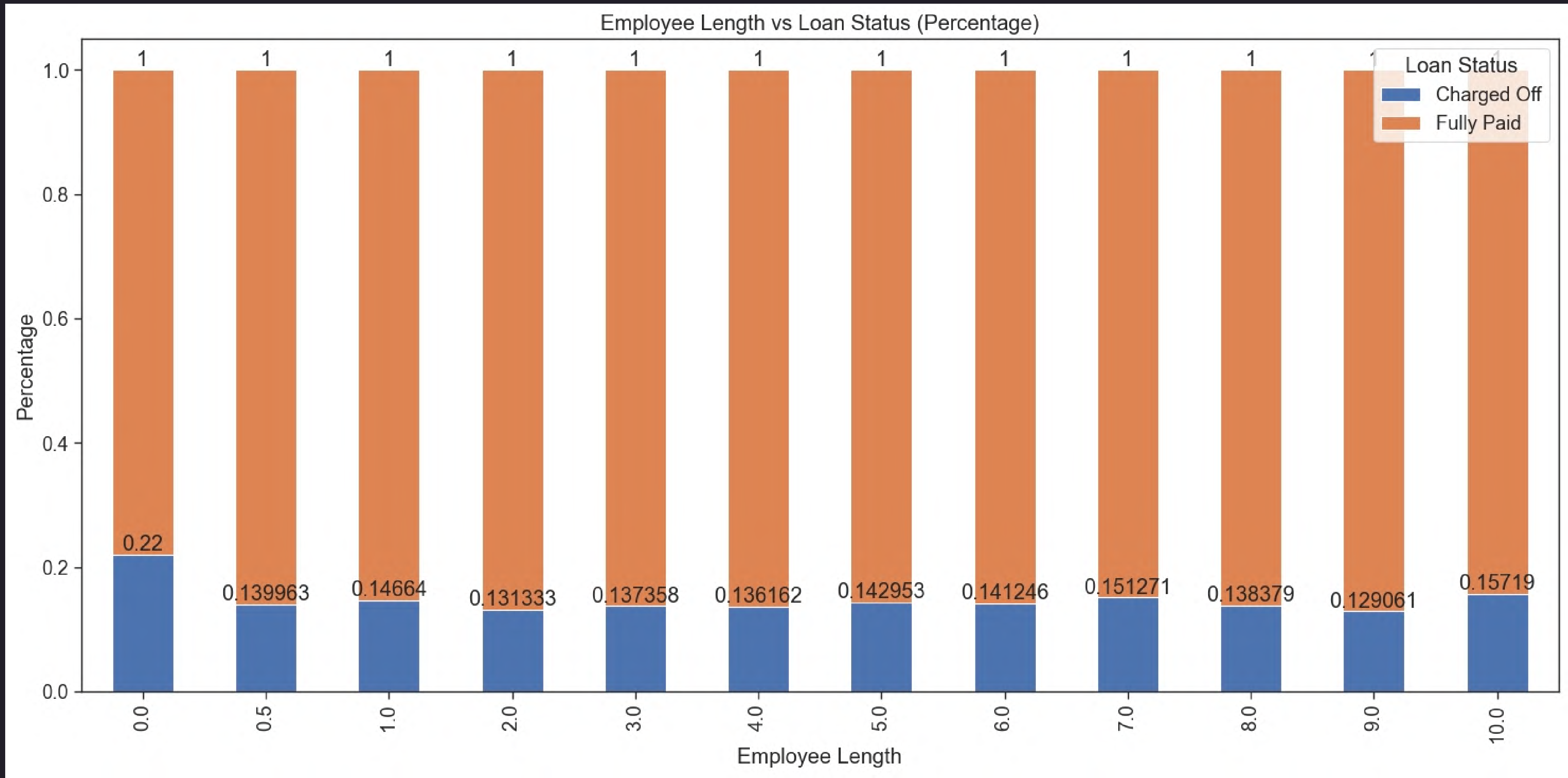


Loan Term across loan statuses

1. Majority of the loan applicants have loan term as 36 months
2. However proportion wise, charged off data is significantly high in applicants who have loan term as 60 months

Univariate Analysis - Continuous / Numeric Variables - Key Observations

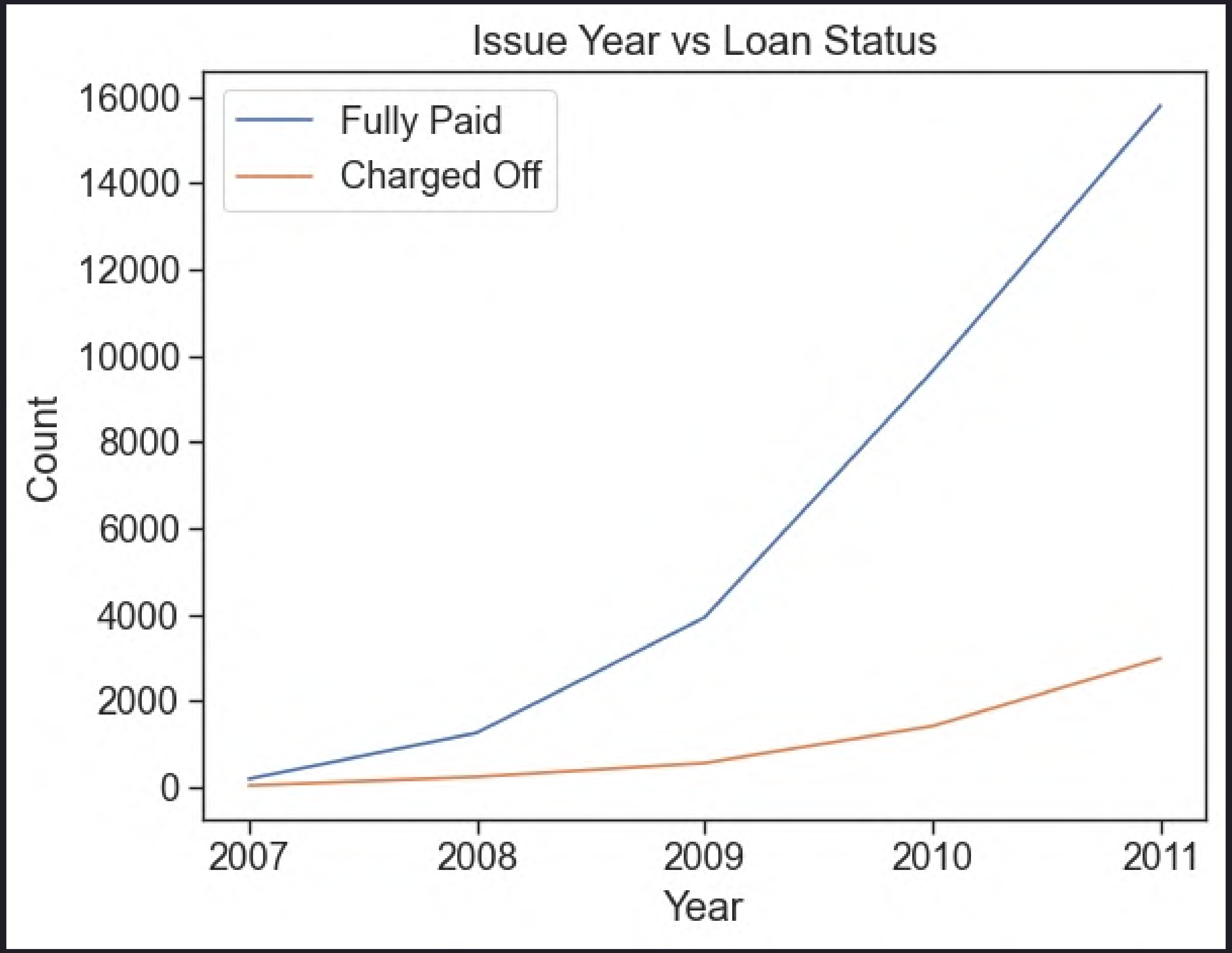


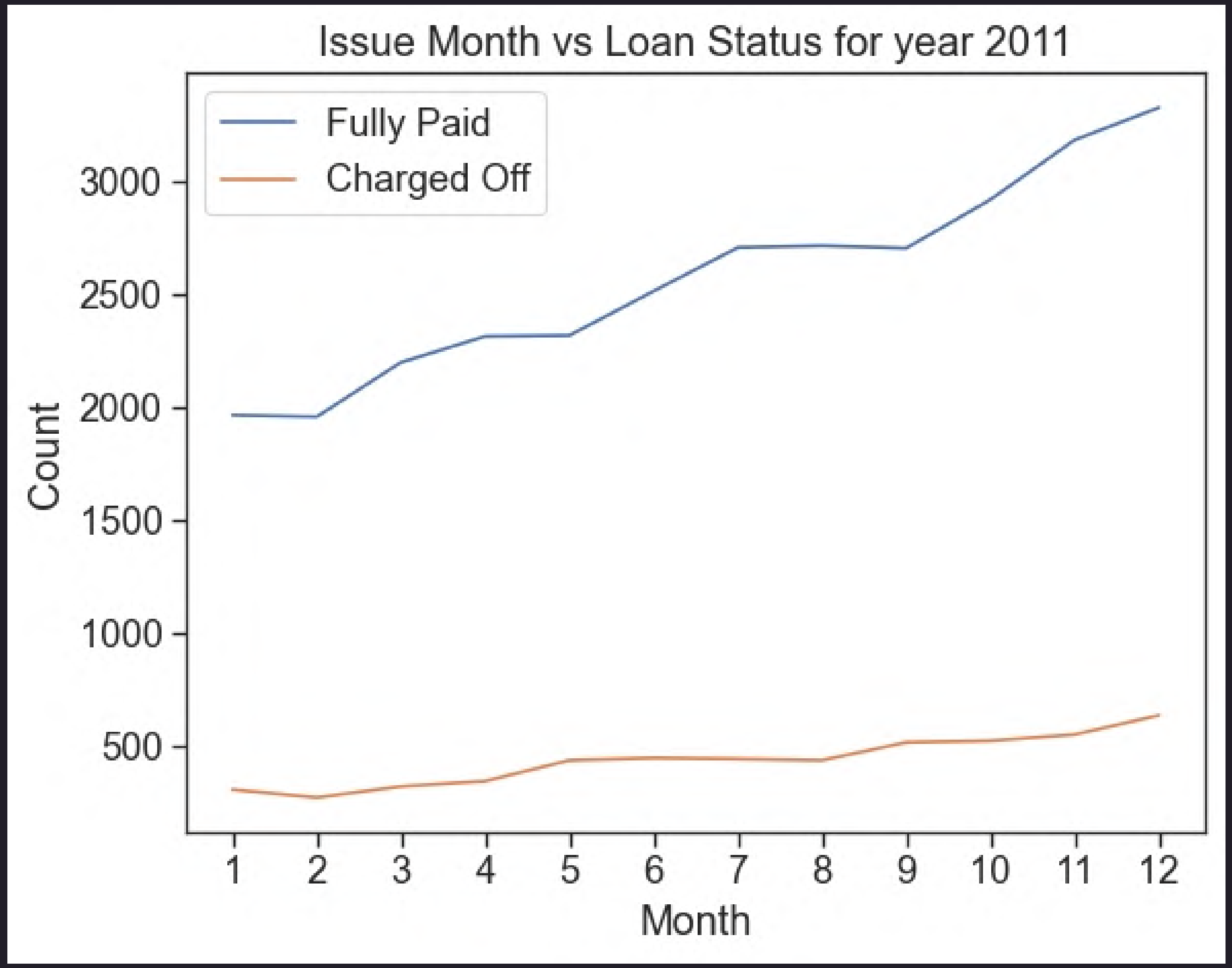


Employee Length across loan statuses

1. More number of loan applicants have employee length as 10+ years
2. Proportion wise, we notice almost the same pattern across all employee lengths

Univariate Analysis - Continuous / Numeric Variables - Key Observations





Issue Year

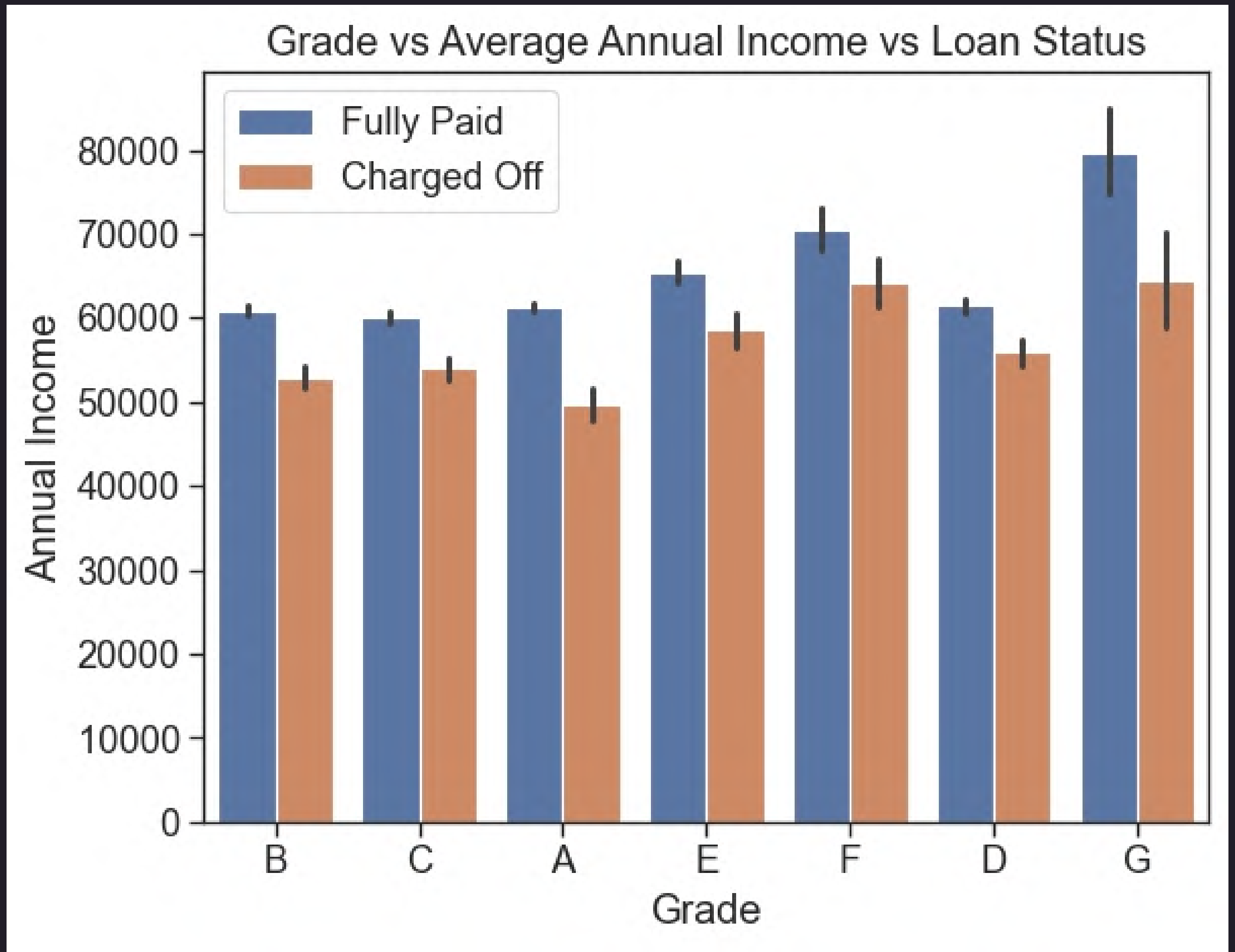
1. Loan applicants have significantly increased in 2011, there is year on year exponential increase in terms of overall loan applications
2. With respect to charged off data, it is almost same for the years 2007 to 2009. For years 2010 and 2011 it has increased significantly

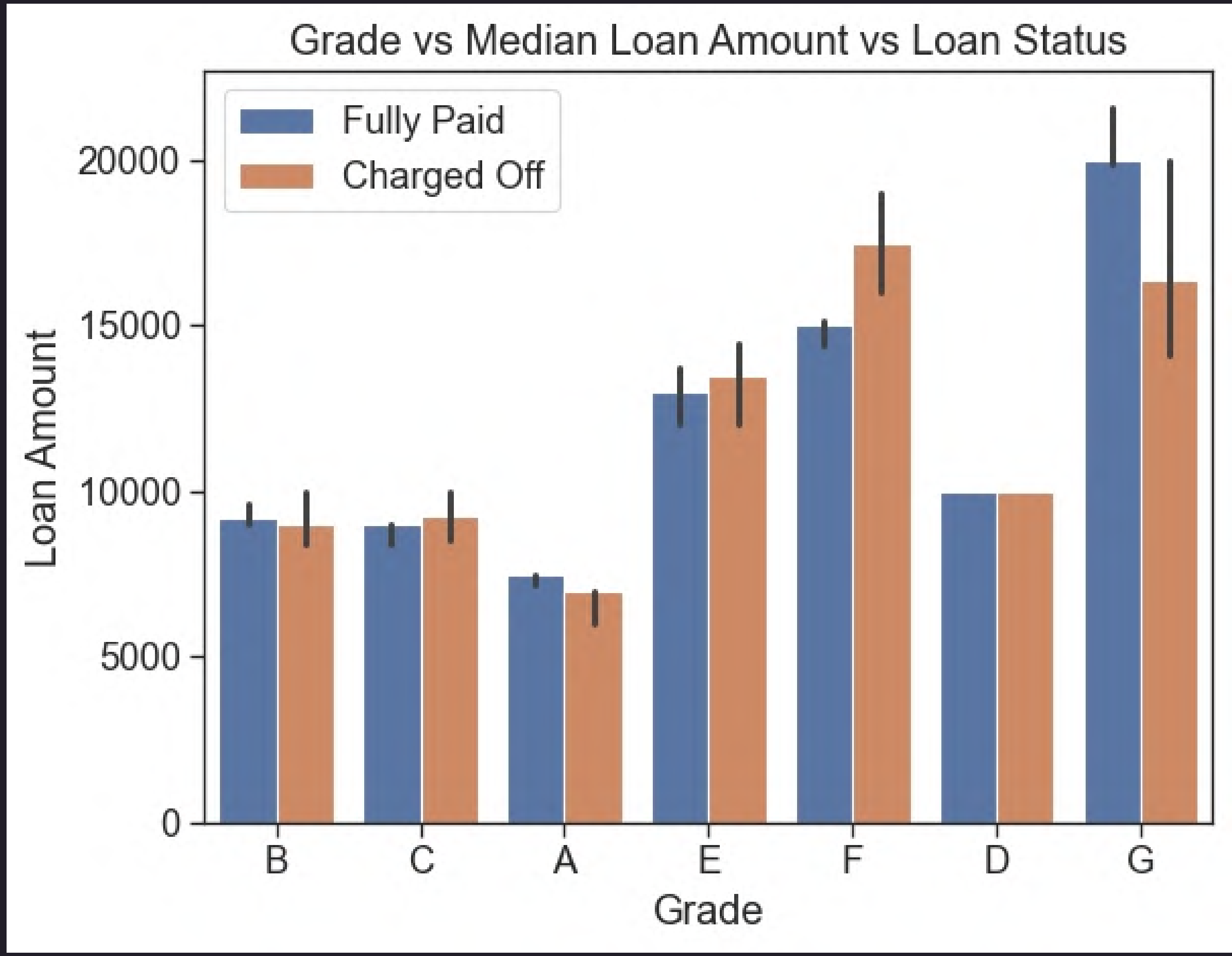
Issue Month for the year 2011

1. Loan applications data increased as the months are passing and it is significantly high by the end of the year
2. However charged off data is almost similar except that it is little high for the loans issued in December. Overall there is no much impact of the loan issued month

Bivariate Analysis

Bivariate Analysis - Key Observations





Grade vs Annual Income across loan statuses

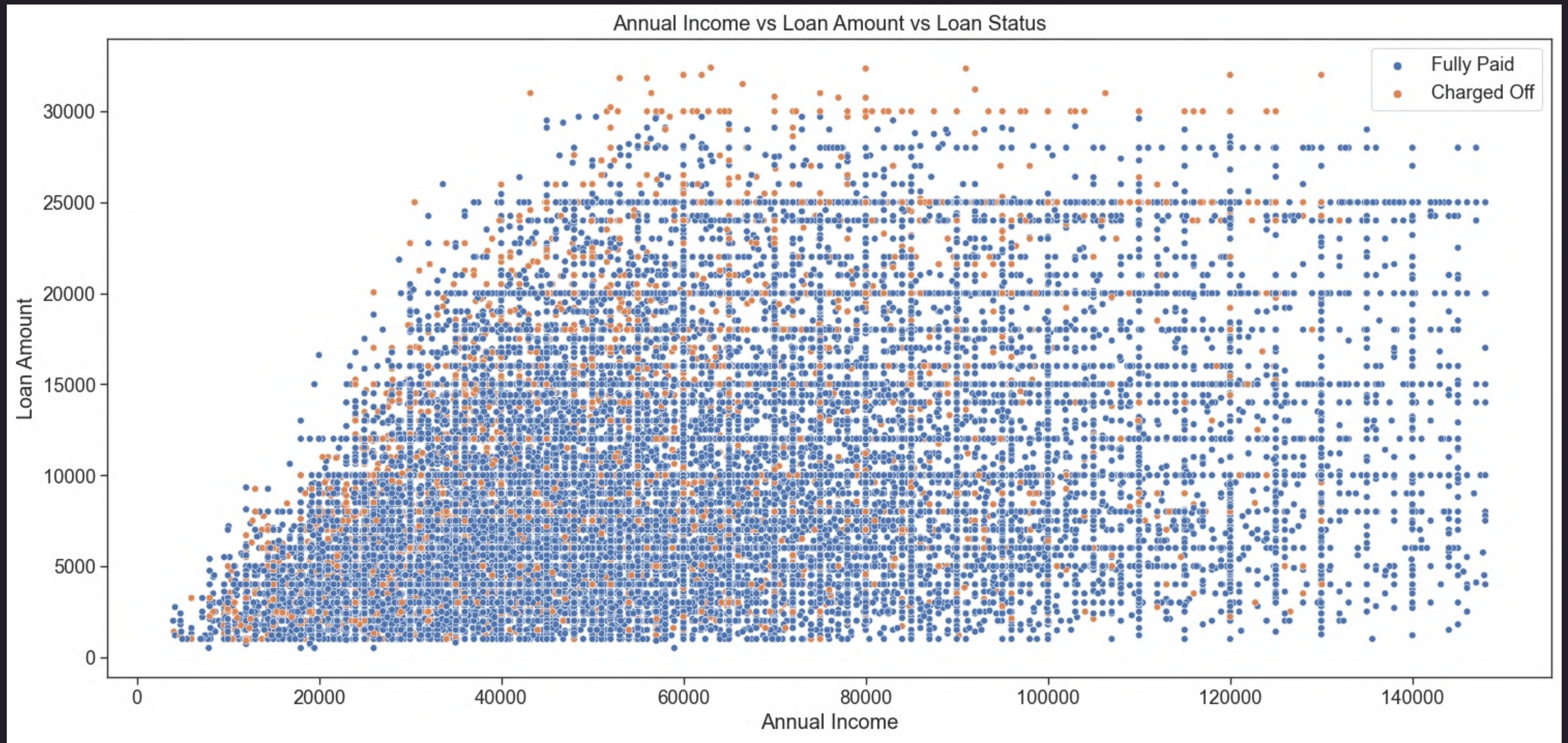
1. Average annual income is higher for Grade 'G'

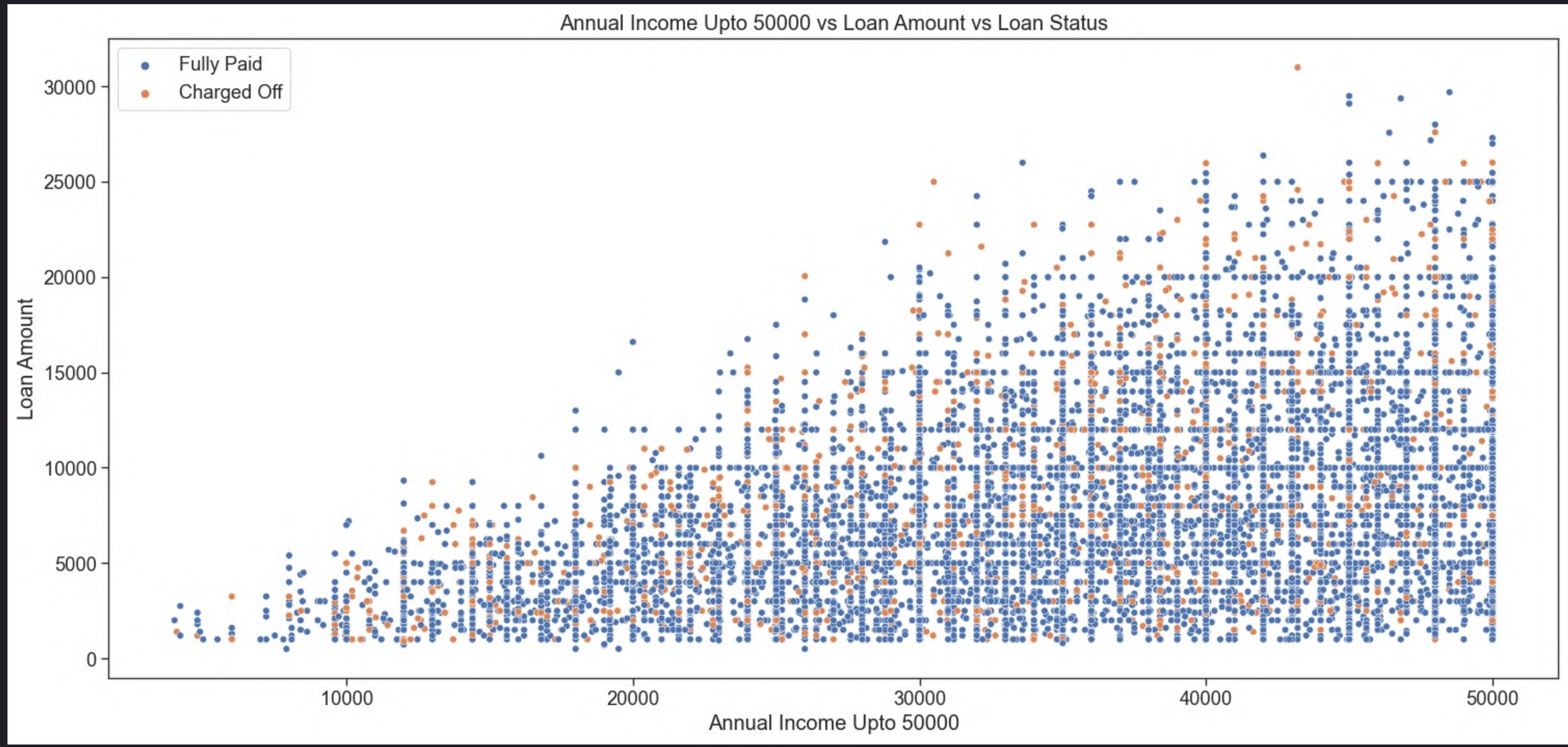
2. This makes sense as the risk increases from A to G, we tend to look for high annual income

Grade vs Loan Amount across loan statuses

1. Median loan amounts are high in 'F' followed by 'G' and 'E'

Bivariate Analysis - Key Observations

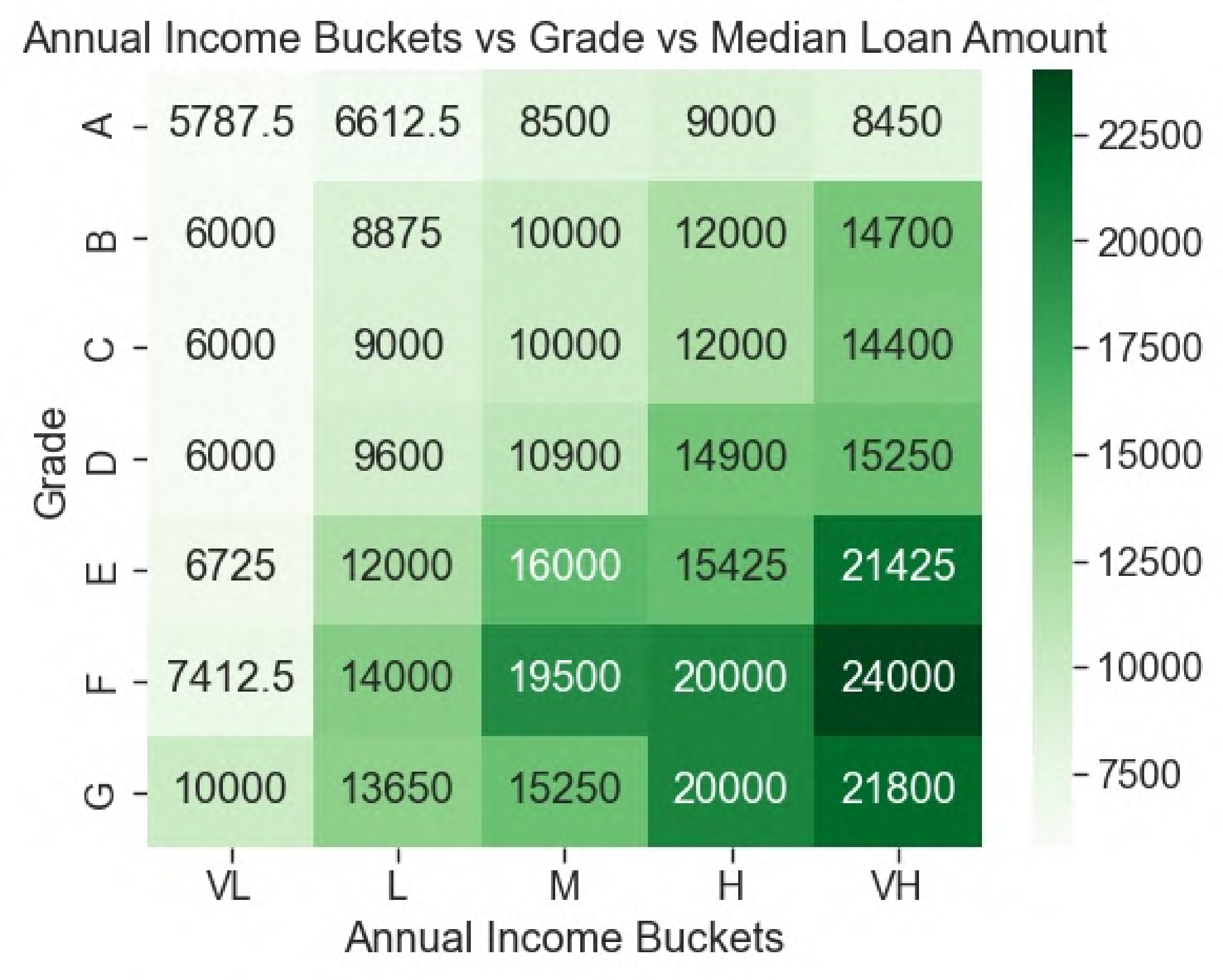


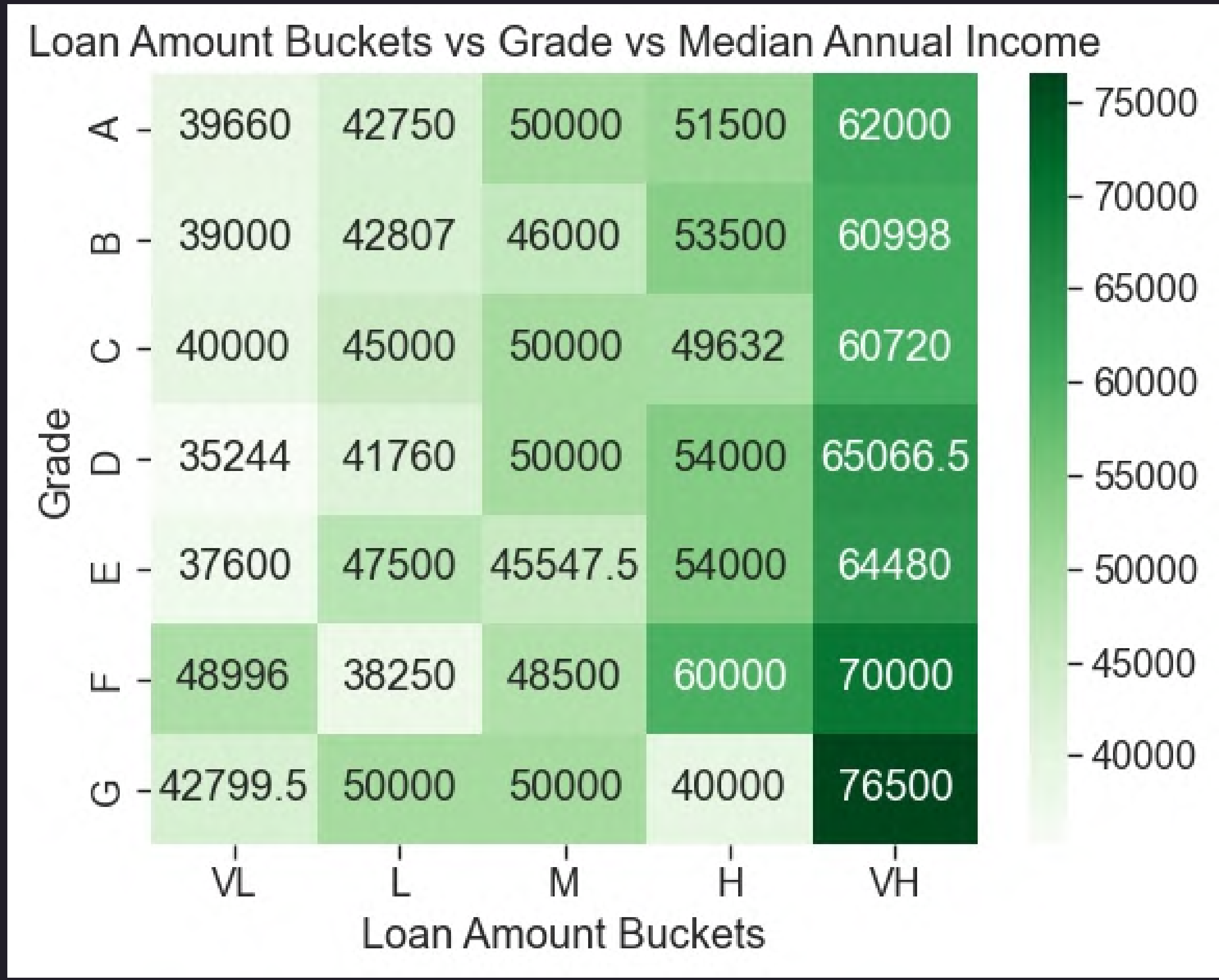


Annual Income vs Loan Amount

1. We see that majority of charged off data is present in the lower annual income range
2. Loan Amount is increasing as the annual income increases

Bivariate Analysis - Key Observations





Loan Amount across Grade & Annual Income Categories

1. We see median loan amounts for the charged off data across different annual income buckets and different grades.
2. We see more density in charged off data for the loan amounts 19,000 to 24, 000 across three different annual income buckets and that is for grades 'E', 'F' and 'G'

Annual Income across Grade & Loan Amount Categories

1. We see that median annual income is increasing with Grade and as the loan amount increases

Summary

Key Observations from our analysis

1. Grade & Sub Grade - As the loan applicant's Grade / sub grade increase from A to G, we see that proportion of charged off data is increasing.
2. Loan Term - More charged off data is having loan term as 60 months
3. Interest Rate - As the interest rate increases, the tendency of charged off is also increasing.
4. Annual Income - Majority of loan write offs are for loan applicants having annual income between 35000 and 50000
5. Loan Amount increases as the annual income increases.
6. We see more density in charged off data for the loan amounts 19,000 to 24, 000 across three different annual income buckets and that is for grades 'E', 'F' and 'G'

Thank You