

Problem Statement

Lending Club Case Study:



You work for a consumer finance company which specialises 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

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tend

In other words, 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 Understanding

Data Understanding

The data set had 111 columns with a mix of customer behaviors & product features. The dataset has basic 3 categories of customers as mentioned below.

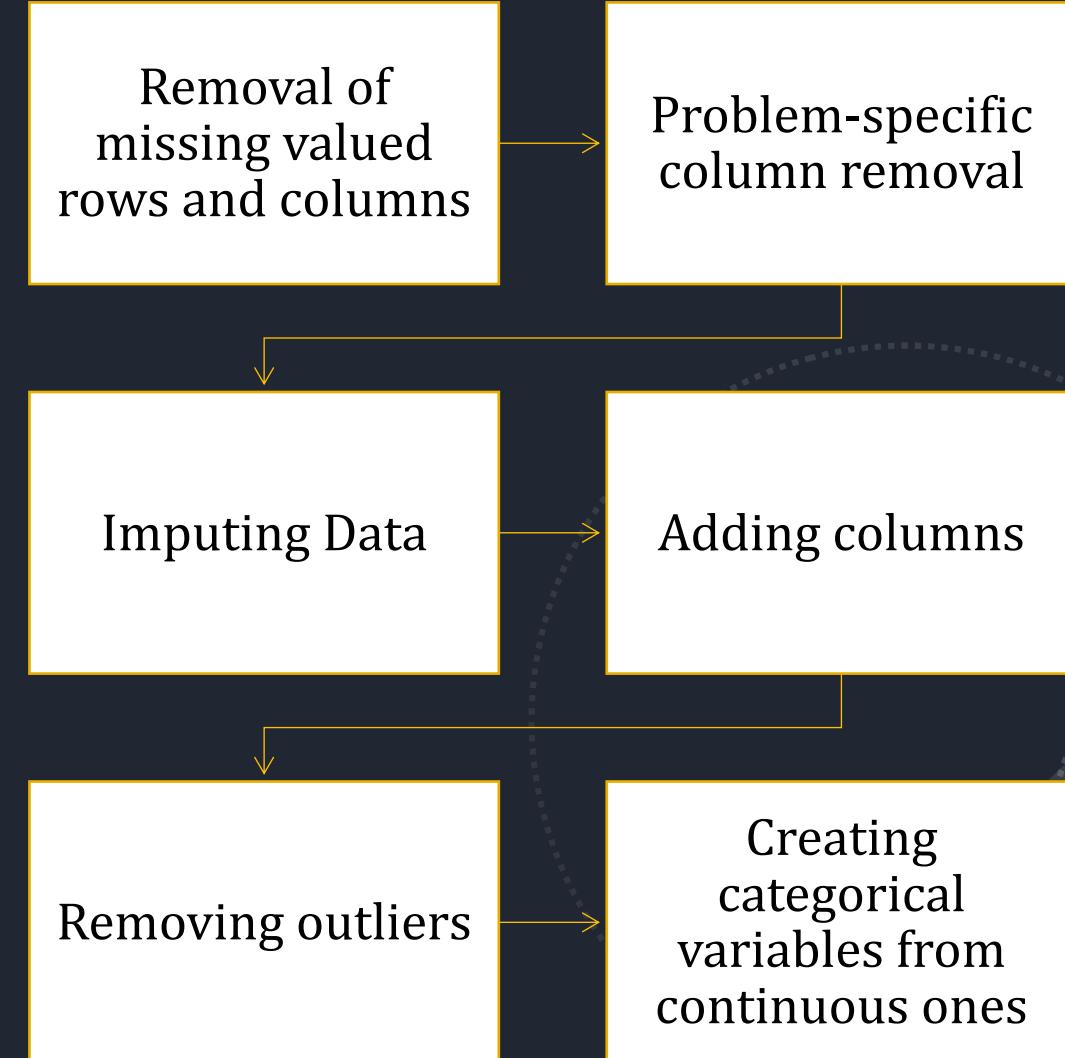
- 1- Fully Paid
- 2 – Charged Off
- 3 – Current

There were more than 50+ columns that has 100% NAN values & there was a mix of categorical, continuous, quantitative & ordered categorical types of variables in each columns.



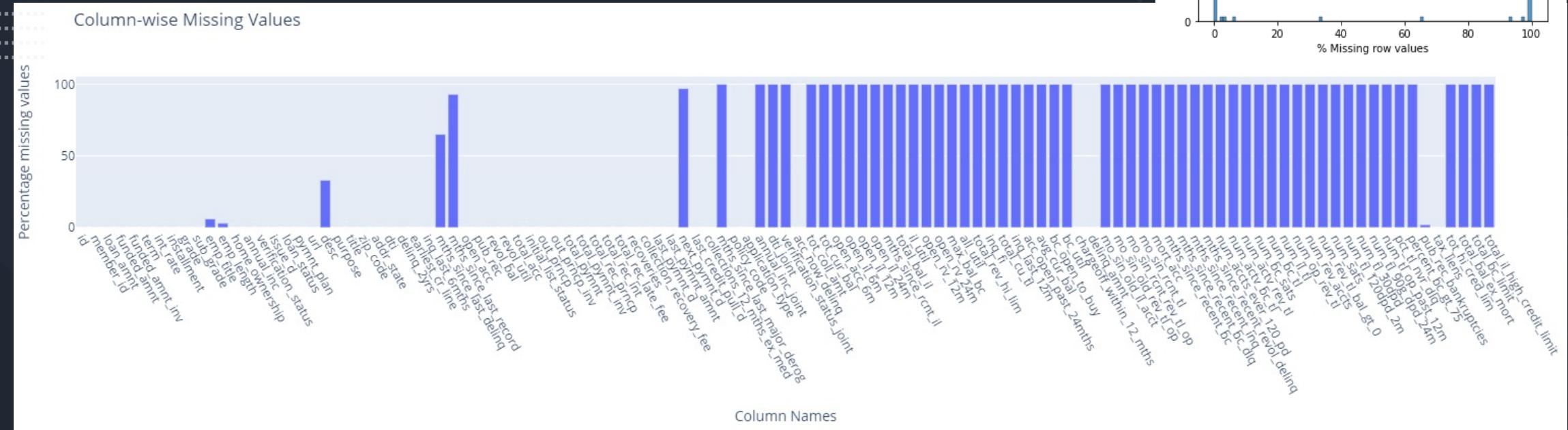
Dataset Cleaning

Methods Implemented



Removing Missing Values

- Column-wise analysis of missing values:
 - 54 columns with 100% missing values



Removing Missing Values

- Row-wise analysis of missing values
 - #Missing-value columns in a row
 - Minimum: 54 (48.65%)
 - Maximum: 62 (55.86%)

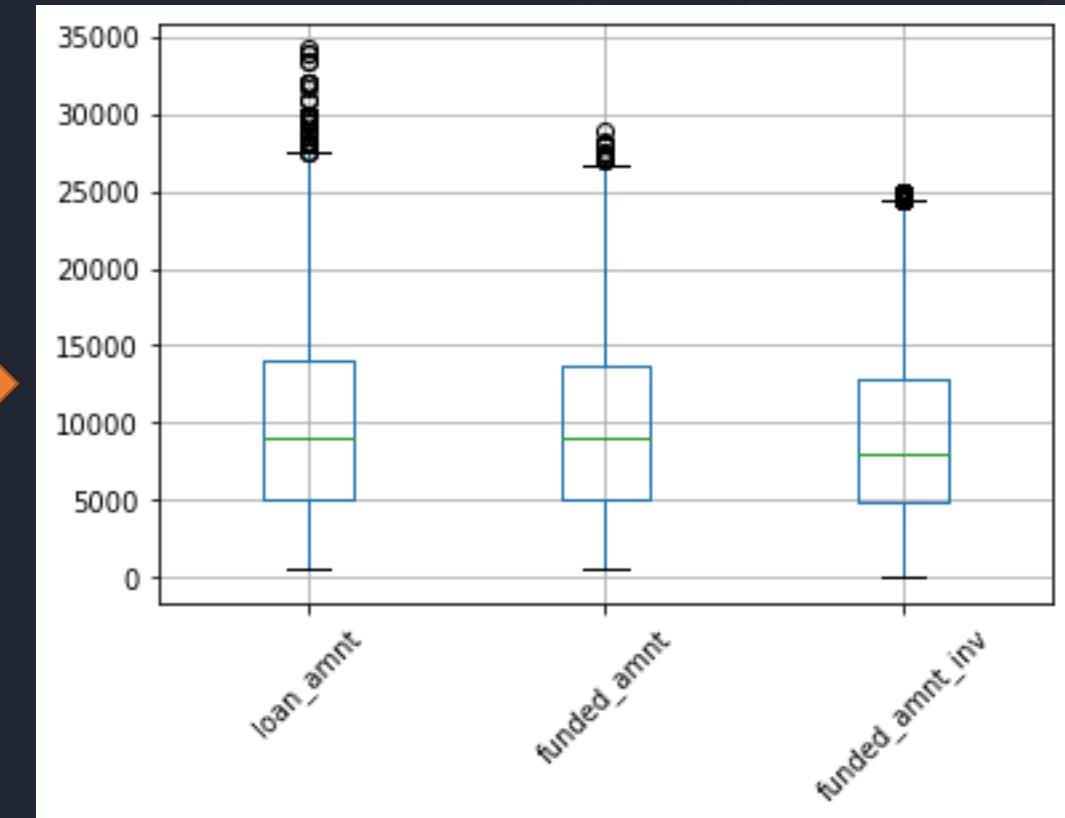
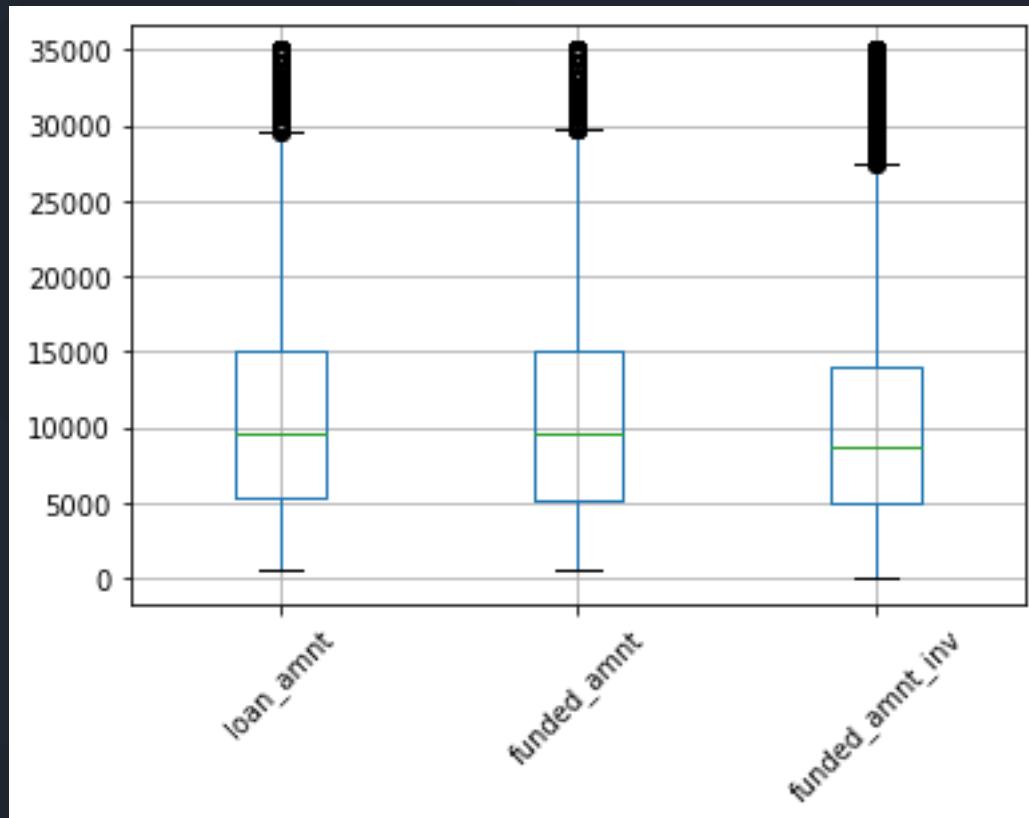


Case-specific data cleaning

- Remove:
 - Redundant columns
 - *url, desc*
 - No additional information
 - Useless columns
 - *id, member_id, emp_title, title*
 - Do not provide any reasoning to why loans were charged off
 - Columns that provide current customer behavior
 - Useless rows
 - *loan_status == Current*
 - Constant-valued columns
 - #Columns dropped: 9
- Fixing *dirty* data entries
 - Rounding off monetary amount numerical values
 - Converting issue date to month and year
 - Removing “%” from *int_rate* column
- Imputing data
 - Forward filling of *emp_length* and *pub_rec_bankruptcies*
- Adding columns
 - *funding_status*
 - *equal* if *loan_amnt* equals *funded_amnt_inv*
 - *lesser* otherwise

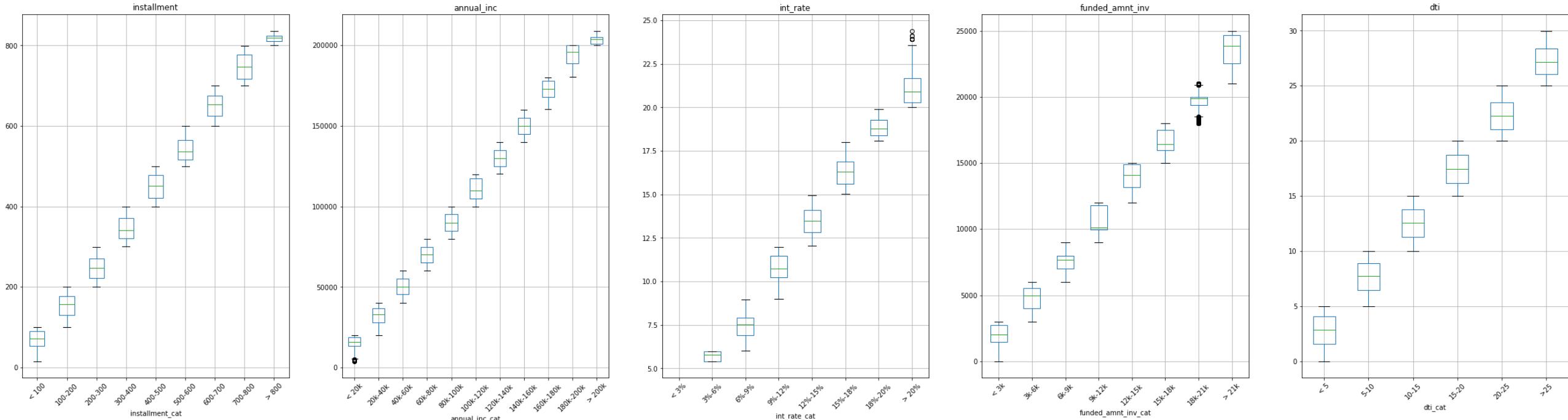
Case-specific data cleaning

- Removing outliers
 - Removed data beyond the 90 percentile
 - *loan_amnt, funded_amnt, funded_amnt_inv*
 - *installment, annual_inc*



Case-specific data cleaning

- Creating new columns (categorical variables)
 - installment -> installment_cat
 - annual_inc -> annual_inc_cat
 - int_rate -> int_rate_cat
 - funded_amnt_inv -> funded_amnt_inv_cat
 - dti -> dti_cat



Univariate Analysis



Analysis of Variable Types

Unordered Categorical

Ordered Categorical

Quantitative

Unordered Categorical Variables

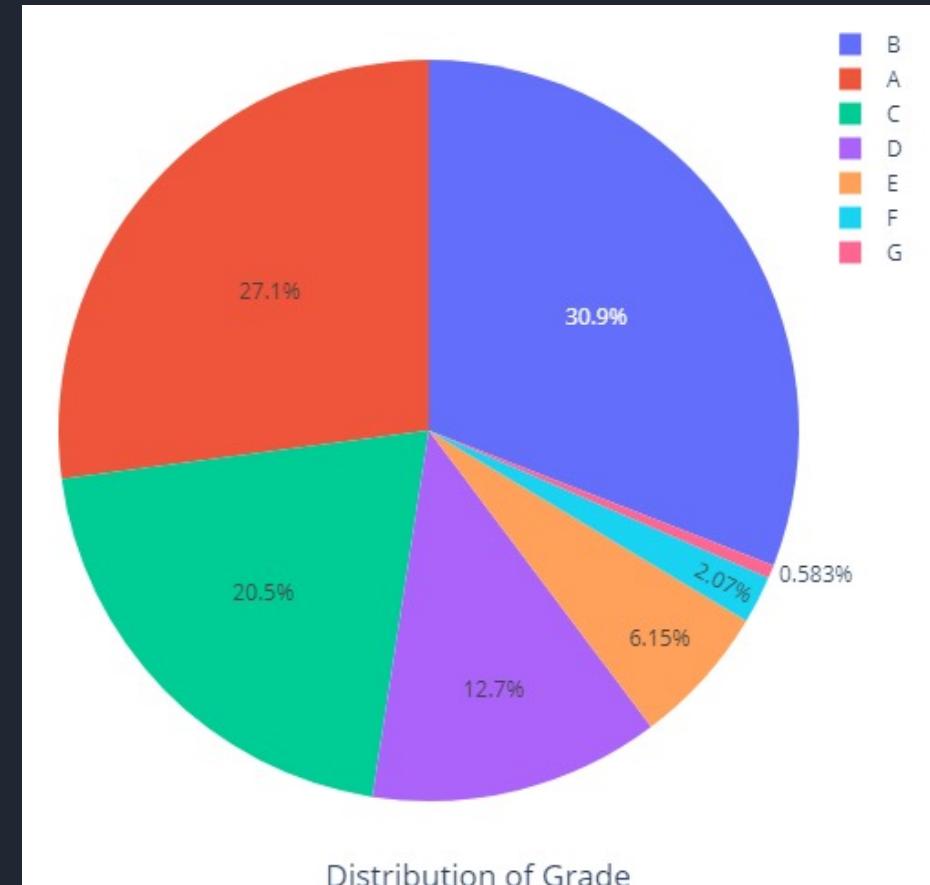
We brought all the columns falling under Unordered Categorical Variables as mentioned below and performed Univariate Analysis.

Below slides have a few pictorial representation of the Univariate Analysis

- Grade
- Sub-grade
- Home Ownership
- Verification Status
- Purpose
- Address State
- Loan Status

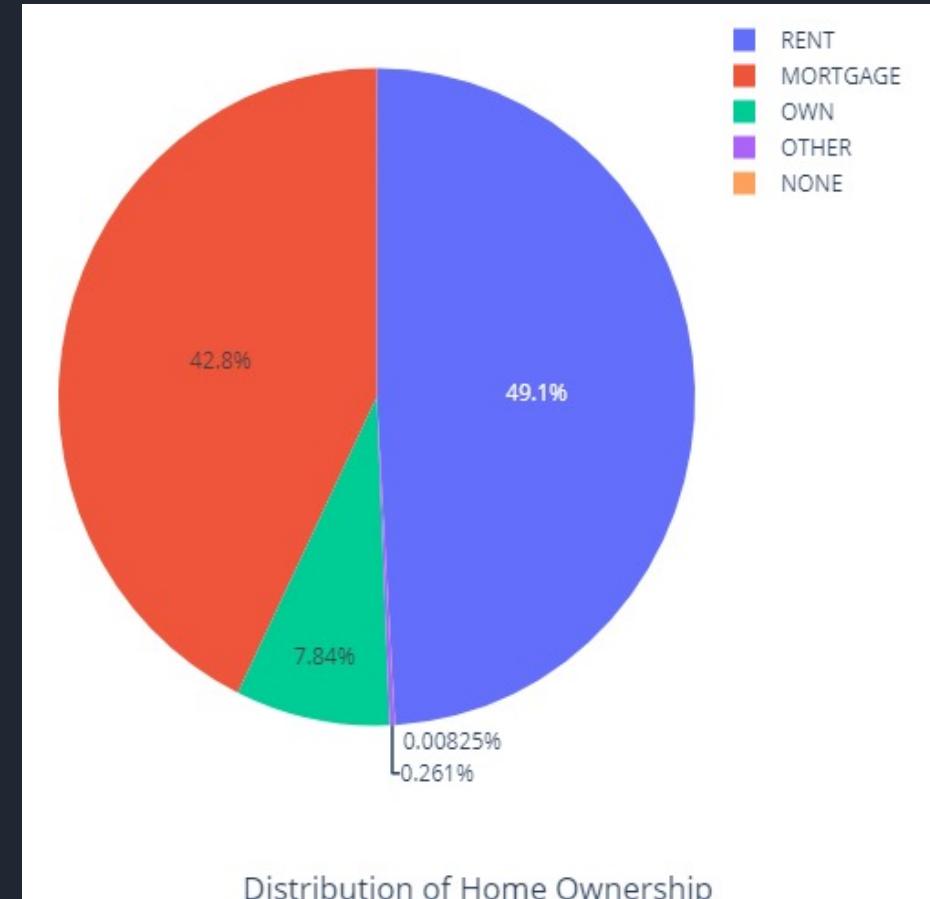
Univariate Analysis: *Grade*

- Takeaway:
- Grade B – Had highest frequency
- Grade G - Had lowest frequency



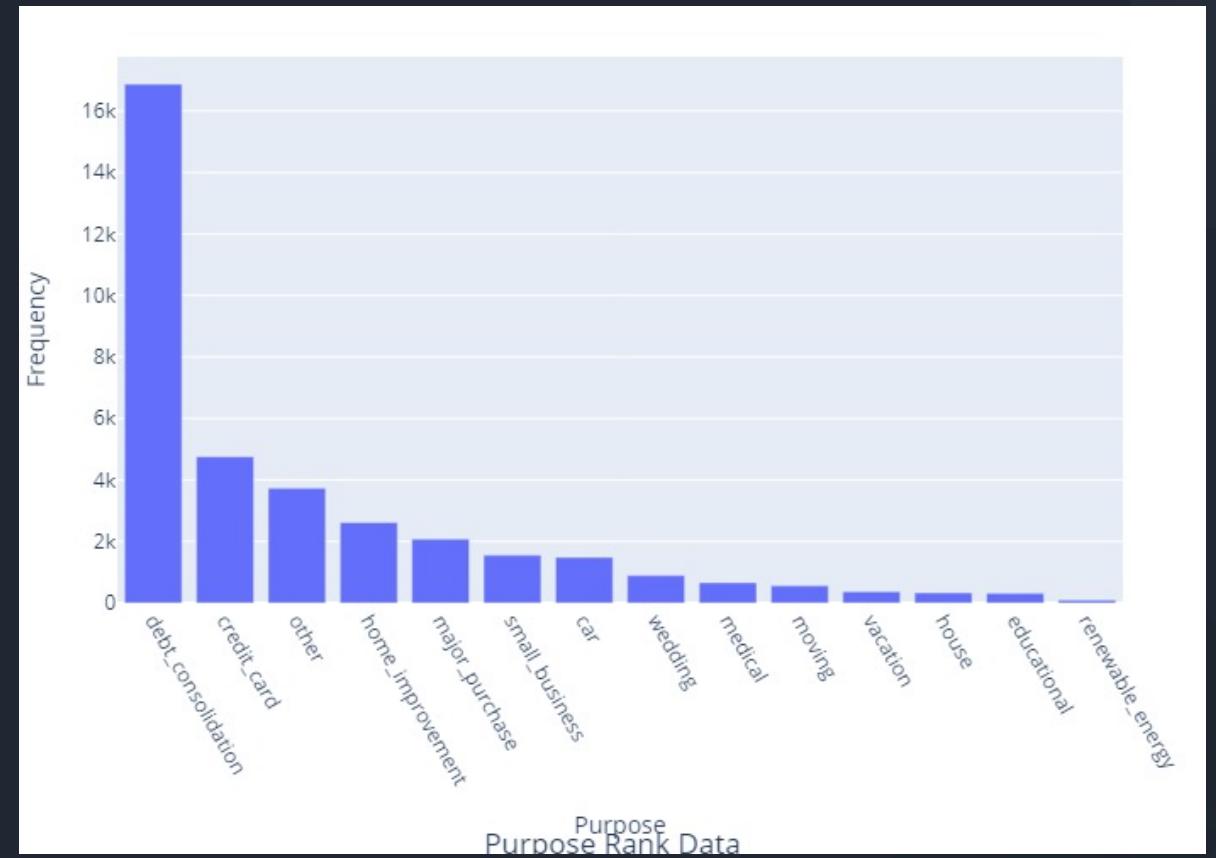
Univariate Analysis: Home Ownership

- Takeaway
- Rented - 49.1%
- Mortgage - 42.8%



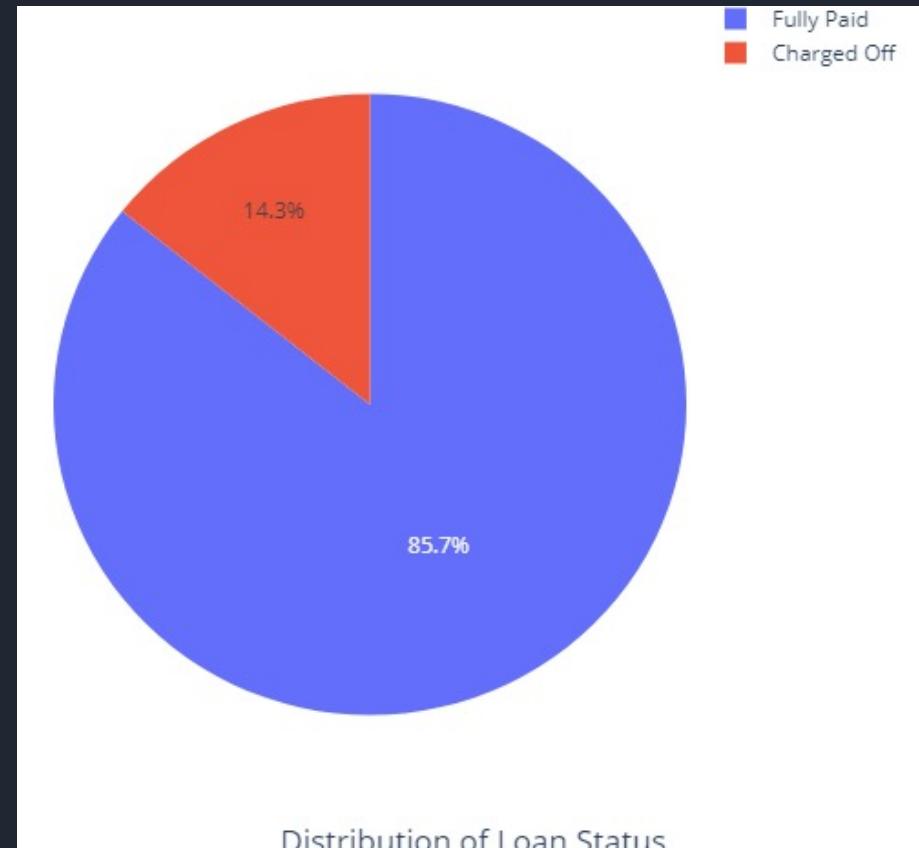
Univariate Analysis: *Loan Purpose*

- Takeaway:
- Debt Consolidation happened to be the most frequent purpose for taking a loan, while renewable energy was the least frequent purpose for taking loan.



Univariate Analysis: *Loan Status*

- Takeaway:
- After the data cleaning exercise, we found that 85.7% customer had fully paid the loan back while 14.3% resulted in Charge-Off for the Lending Club.



Ordered Categorical Variables

We brought all the columns falling under Ordered Categorical Variables as mentioned below and performed Univariate Analysis.

Below slides have a few pictorial representation of the Univariate Analysis

- Loan Term
- Employment Length
- Issue Month
- Issue Year
- Installment
- Annual Income
- Interest Rate
- Funded Amount Invested
- Debt-to-Income Ratio

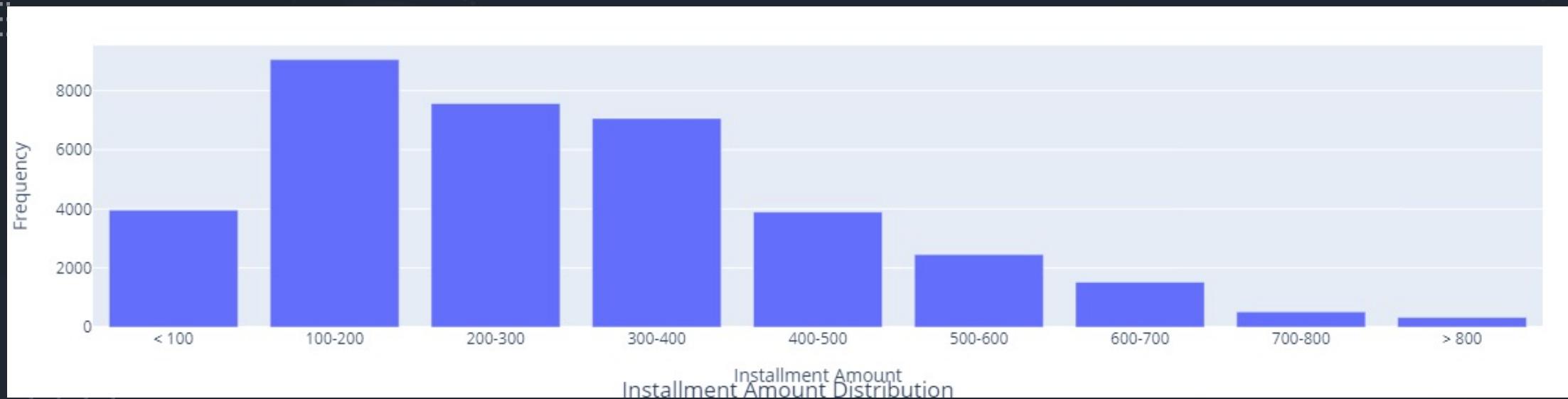
Univariate Analysis: *Employment Length*

- Takeaway:
- Customer with 10+ years of experience had the highest frequency of taking loans from Lending Club.



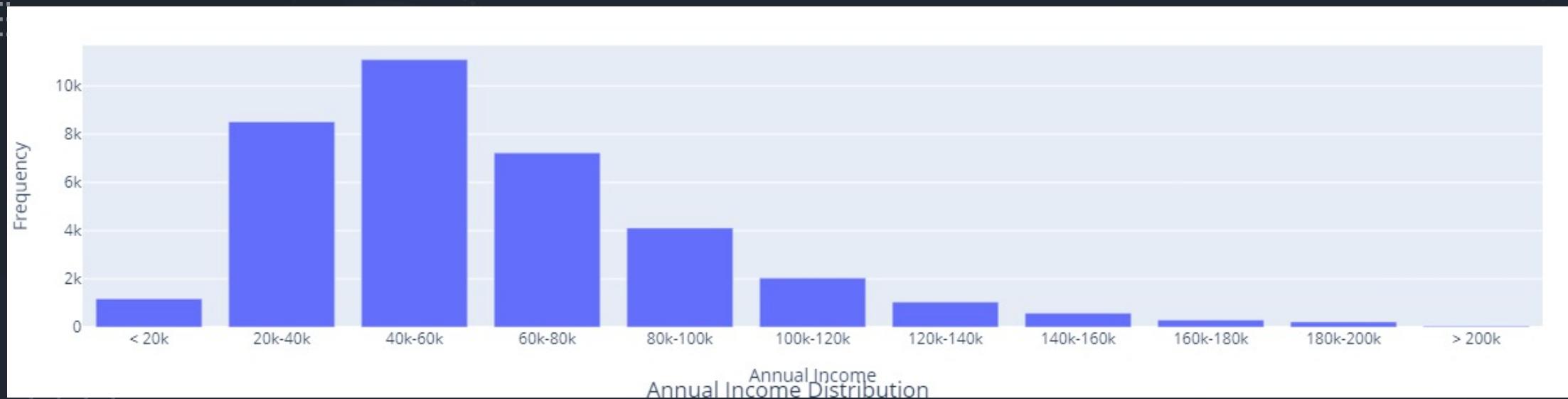
Univariate Analysis: *Installment Amount*

- Takeaway:
- The highest frequency came in the group of customer who were paying an Instalment of \$100-\$200/month.



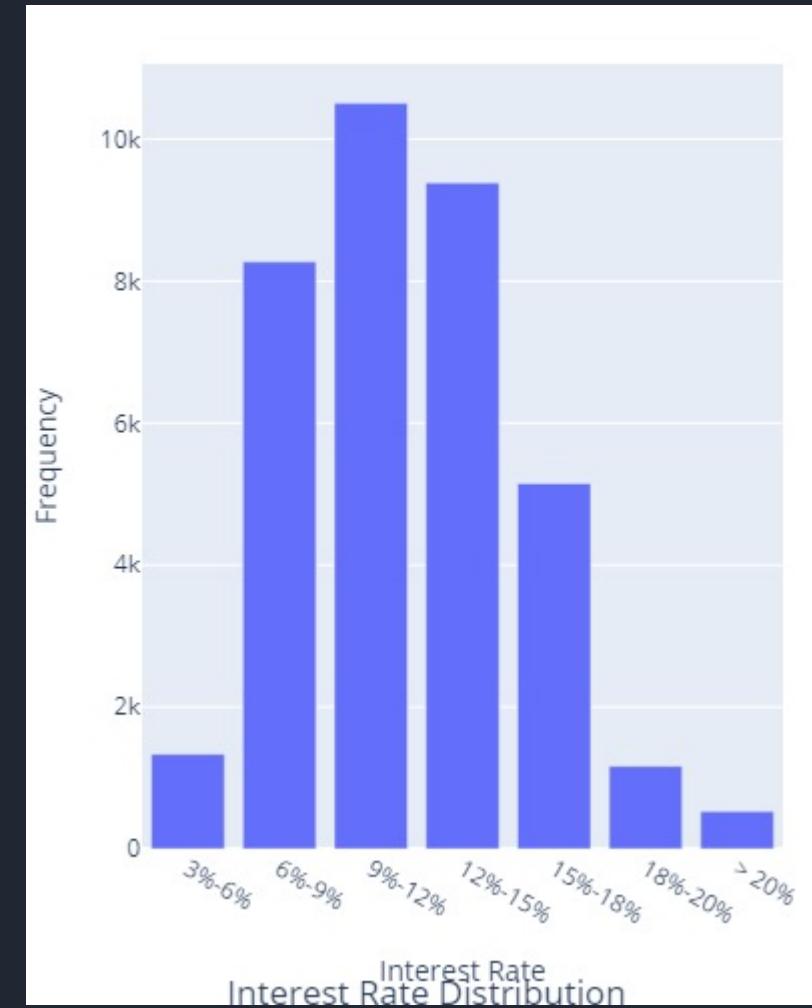
Univariate Analysis: *Annual income*

- Takeaway:
- Customer with 40-60k Annual income had the highest frequency of taking loans from Lending Club.



Univariate Analysis: *Interest Rate*

- Takeaway:
- Maximum customers where paying an interest rate ranging from 9-12% on the loans that they had taken from Lending Club.



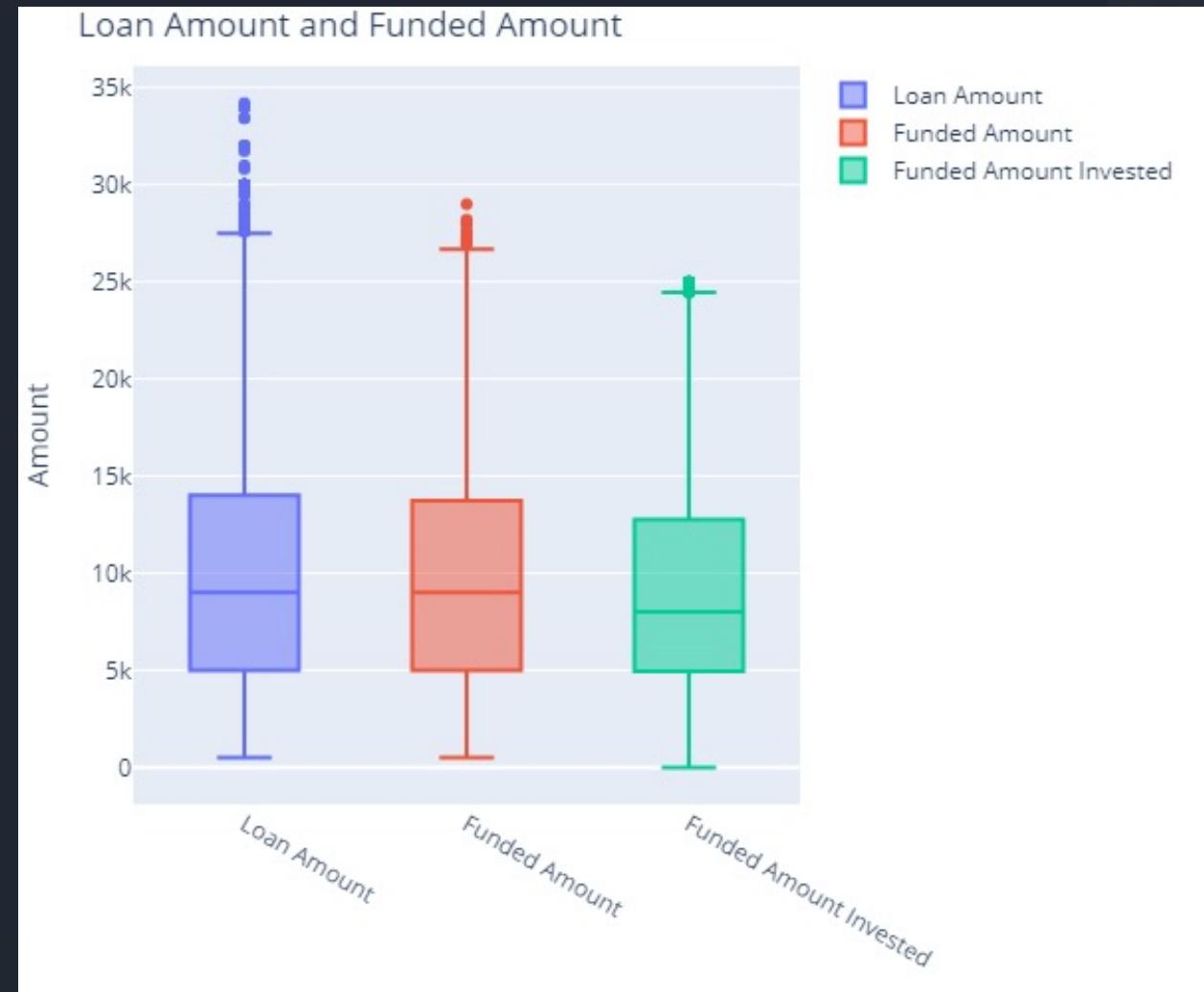
Quantitative Variables

We brought all the columns falling under Quantitative Variables as mentioned below and performed Univariate Analysis.

Below slides have a few pictorial representation of the Univariate Analysis

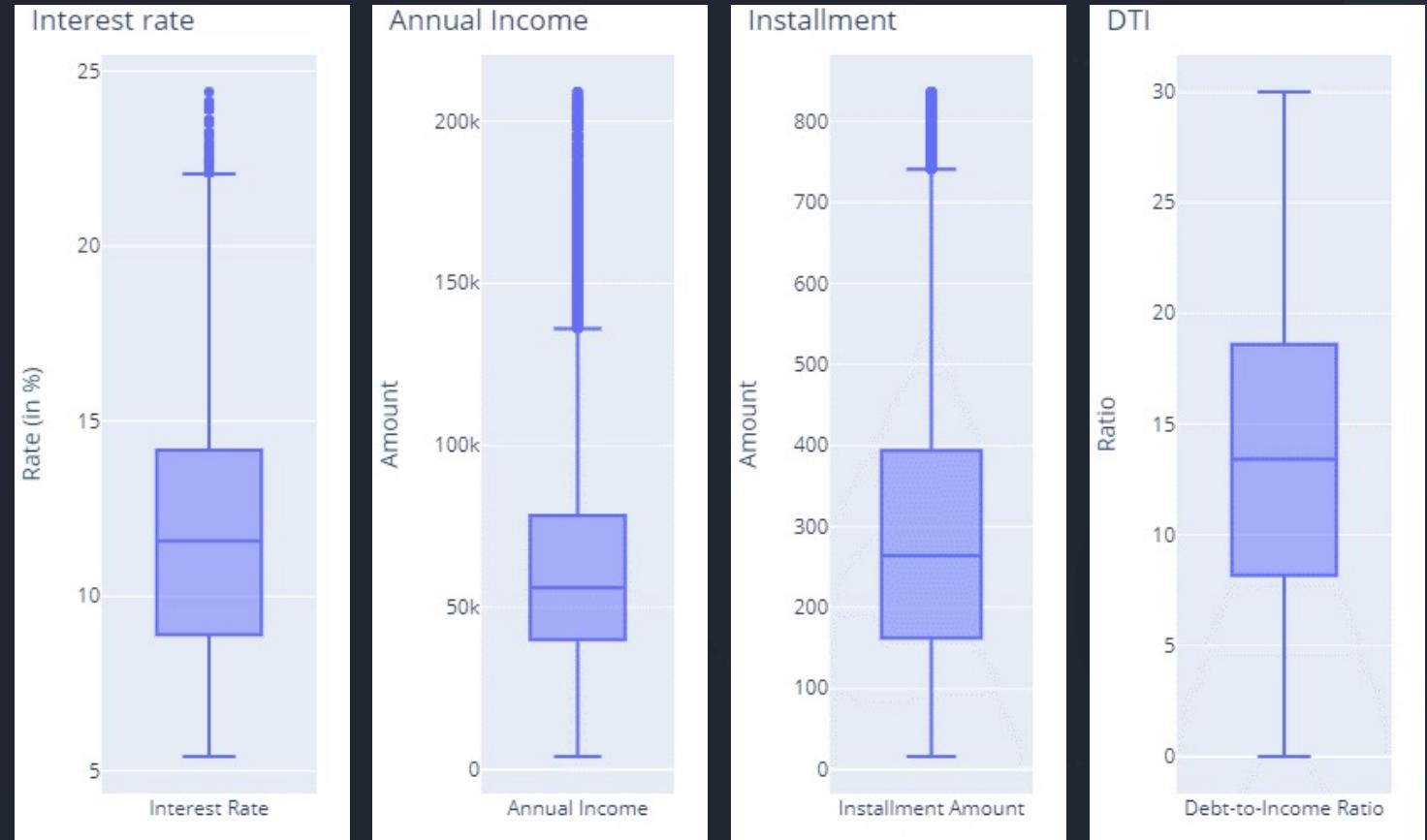
- Loan Amount
- Funded Amount
- Funded Amount Invested
- Interest Rate
- Installment
- Annual Income
- Debt-to-Income Ratio

Univariate Analysis: *Loan Amounts*



Univariate Analysis: *Others*

- Interest Rate
- Annual Income
- Installment Amount
- DTI Ratio



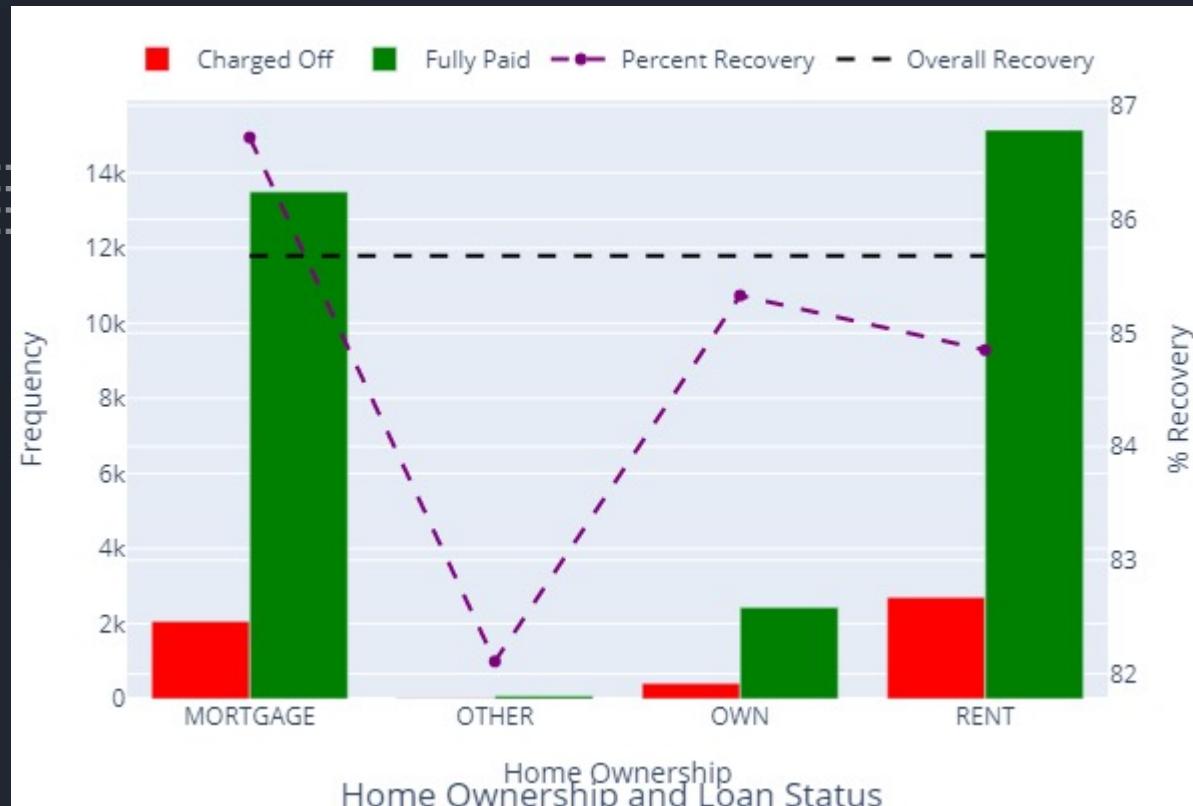
Segmented Univariate Analysis



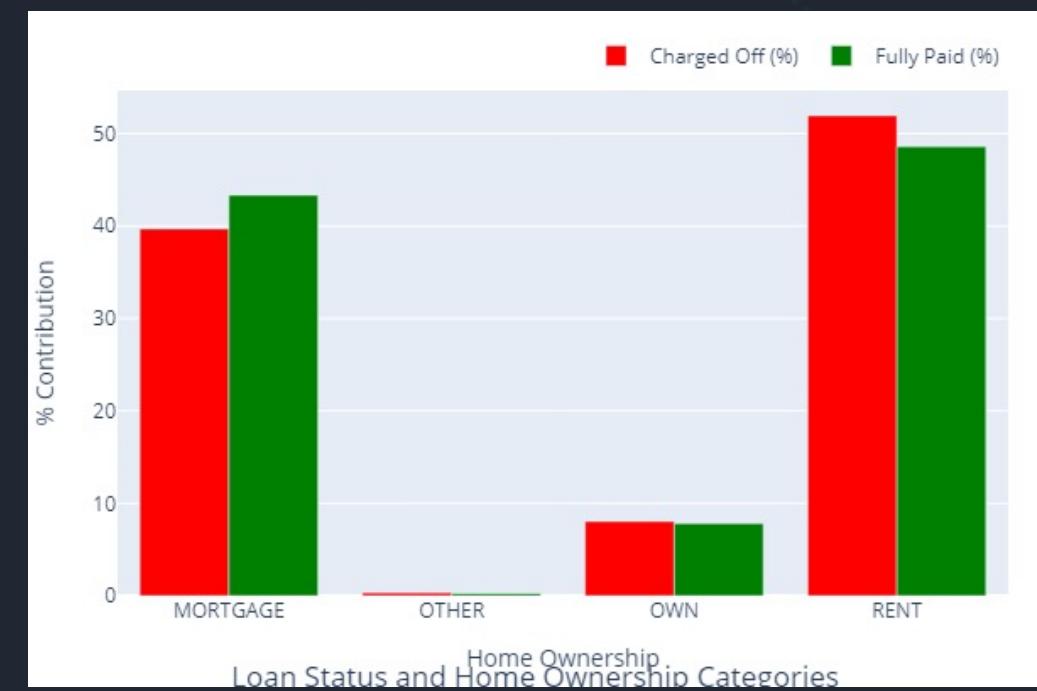
Categorical Variables for Analysis

- Loan Term
- Grade
- Sub-grade
- Home Ownership
- Verification Status
- Purpose
- Address State
- Bankruptcies
- Issue Month
- Issue Year
- Employment Length
- Funding Status
- Funded Amount Invested (Categorical)
- Interest Rate (Categorical)
- Installment (Categorical)
- Annual Income (Categorical)
- Debt-to-Income Ratio (Categorical)

Segmented Univariate Analysis: *Home ownership*

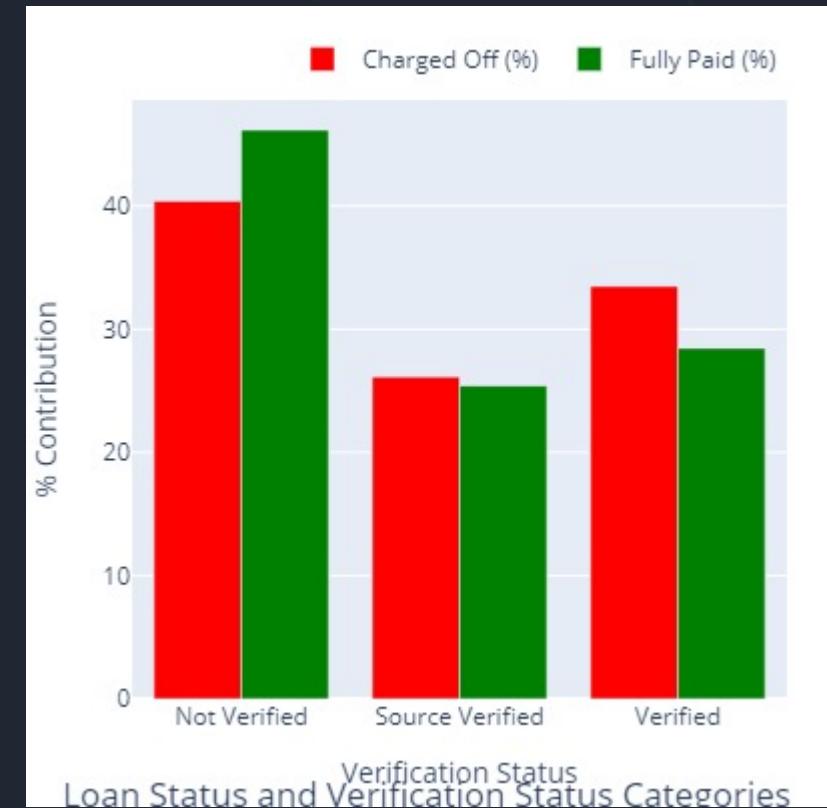
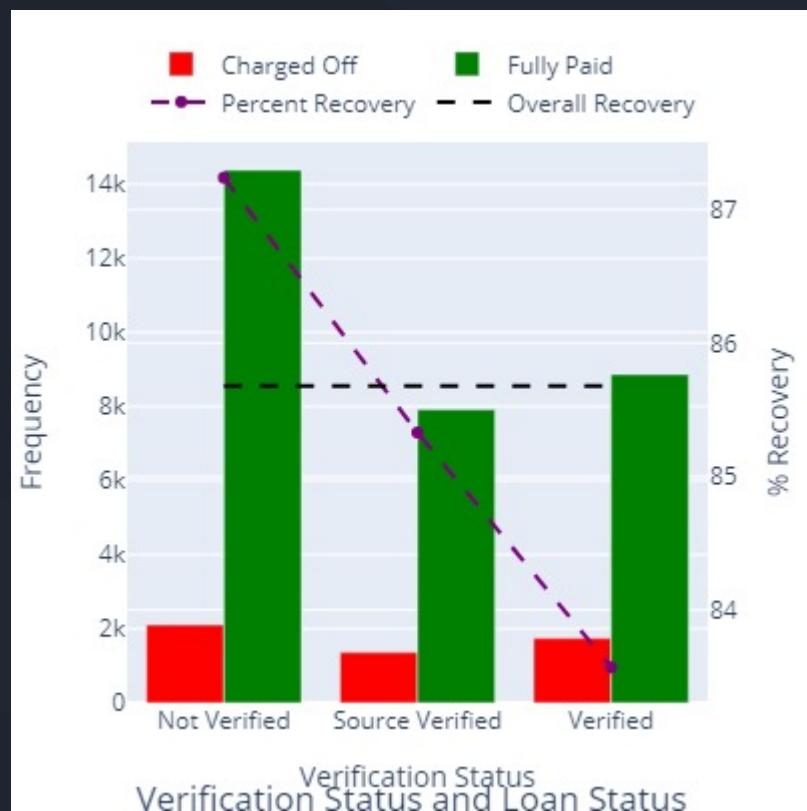


- Charged Off
 - Maximum contribution: RENT (51.94%)
 - Minimum contribution: OTHER (0.33%)
- Fully paid
 - Maximum contribution: RENT (48.61%)
 - Minimum contribution: OTHER (0.25%)



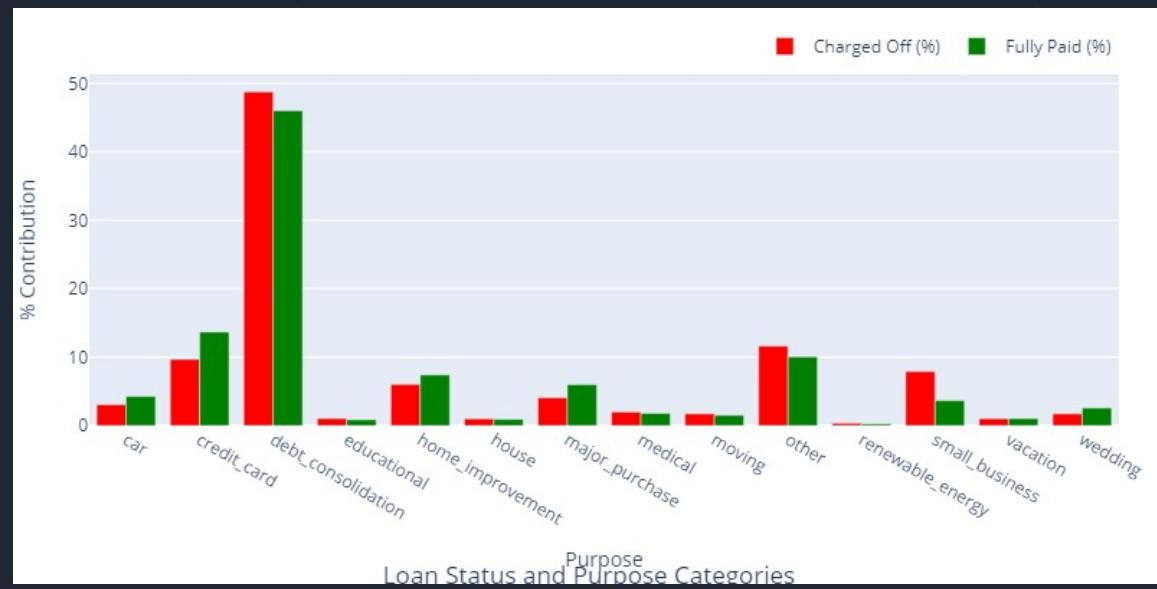
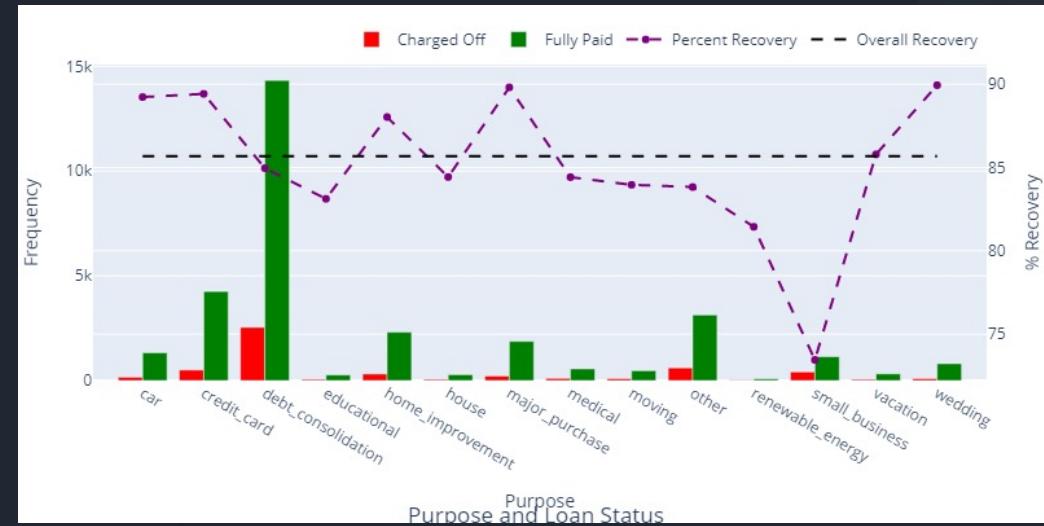
Segmented Univariate Analysis: *Verification*

- Charged Off
 - Maximum contribution: Not Verified (40.39%)
 - Minimum contribution: Source Verified (26.13%)
- Fully paid
 - Maximum contribution: Not Verified (46.15%)
 - Minimum contribution: Source Verified (25.4%)

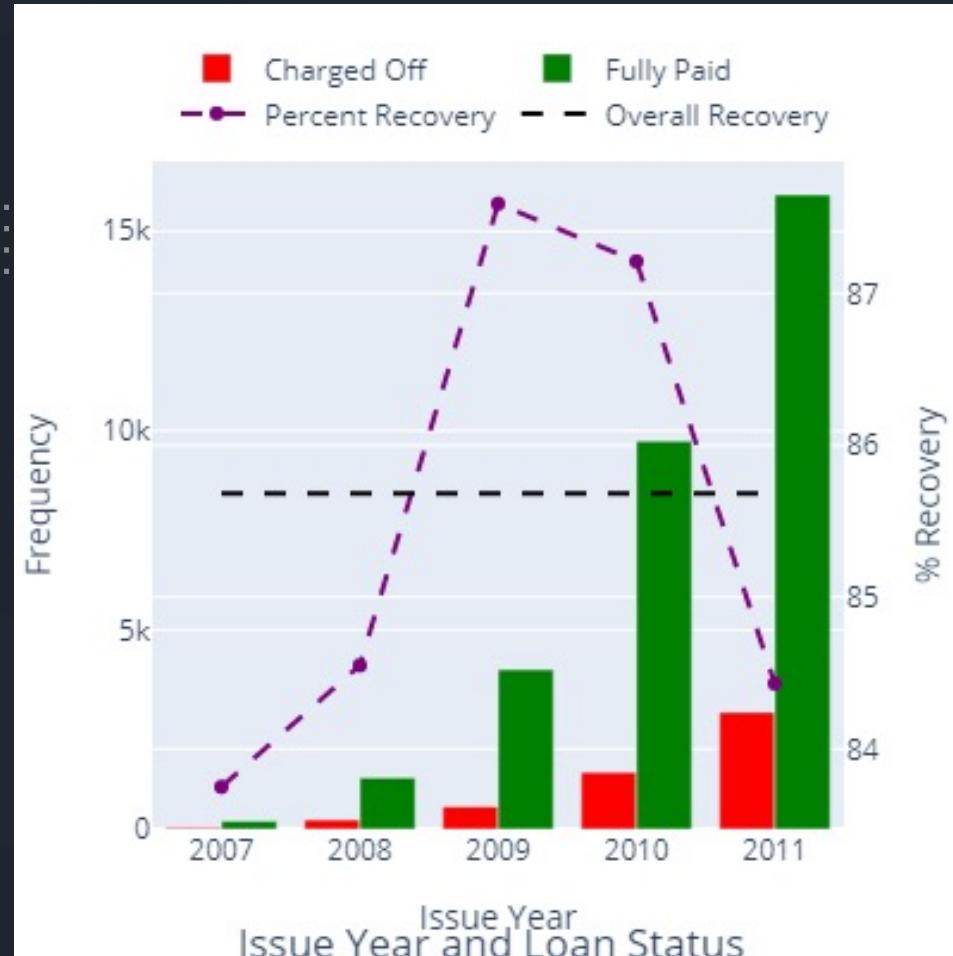


Segmented Univariate Analysis: *Loan Purpose*

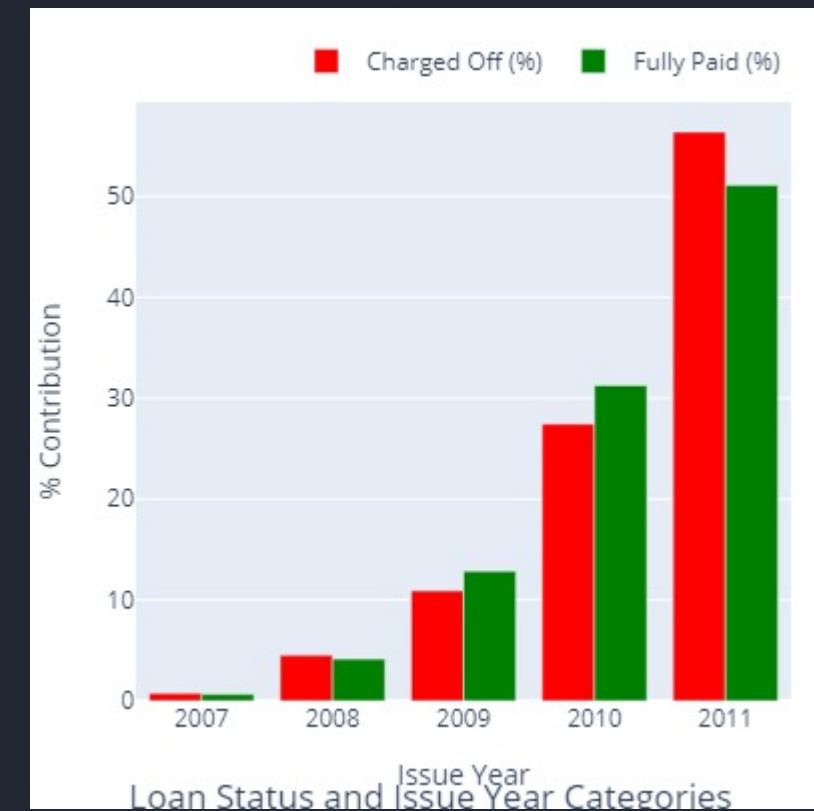
- Charged Off
 - Maximum contribution: debt consolidation (48.77%)
 - Minimum contribution: renewable energy (0.35%)
- Fully paid
 - Maximum contribution: debt consolidation (46.02%)
 - Minimum contribution: renewable energy (0.25%)



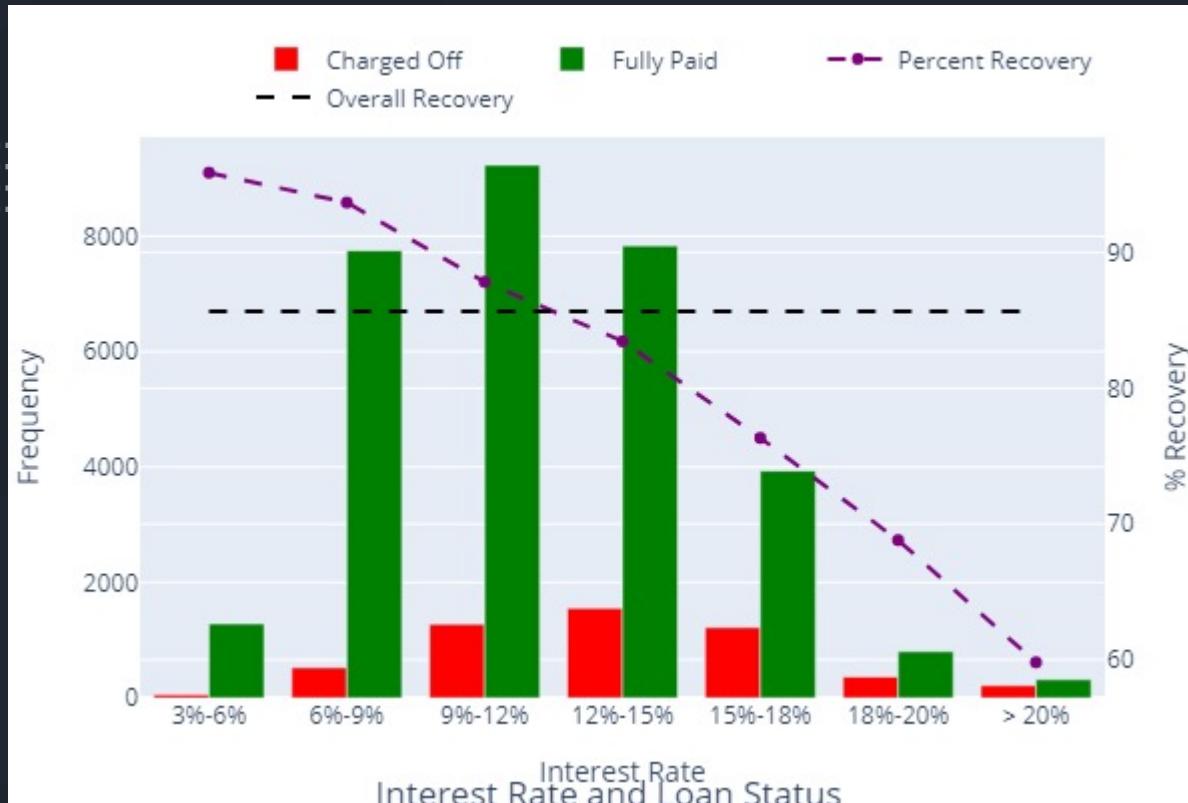
Segmented Univariate Analysis: Issue Year



- Charged Off
 - Maximum contribution: 2011 (56.38%)
 - Minimum contribution: 2007 (0.75%)
- Fully paid
 - Maximum contribution: 2011 (51.1%)
 - Minimum contribution: 2007 (0.65%)



Segmented Univariate Analysis: *Interest Rate*

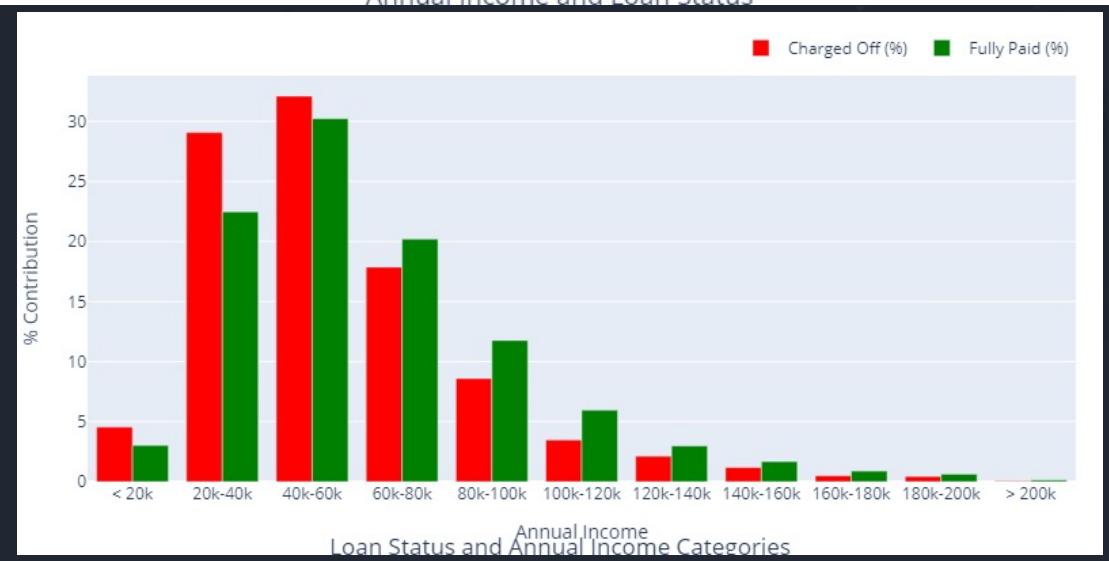


- Charged Off
 - Maximum contribution: 12%-15% (29.82%)
 - Minimum contribution: 3%-6% (1.06%)
- Fully paid
 - Maximum contribution: 9%-12% (29.63%)
 - Minimum contribution: > 20% (1.02%)



Segmented Univariate Analysis: *Annual Income*

- Charged Off
 - Maximum contribution: 40k-60k (32.11%)
 - Minimum contribution: > 200k (0.06%)
- Fully paid
 - Maximum contribution: 40k-60k (30.24%)
 - Minimum contribution: > 200k (0.14%)



Bivariate Analysis

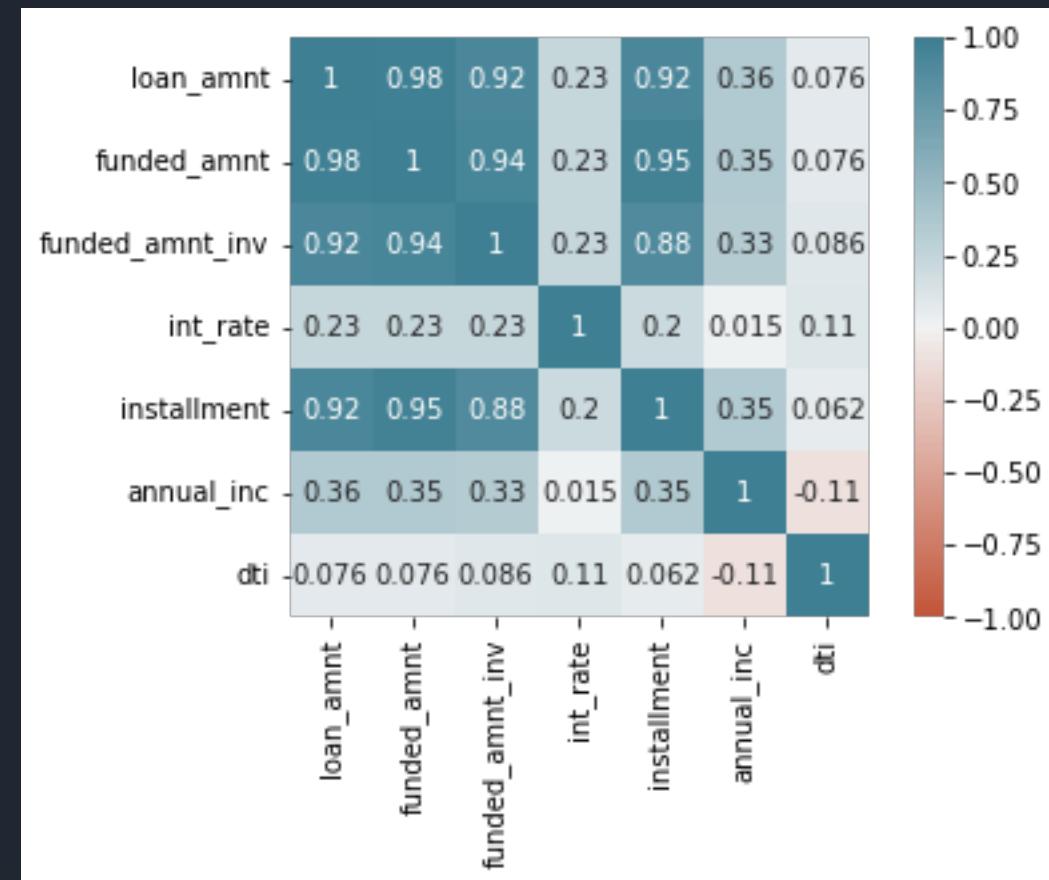
Analysis of Variable Types

Continuous

Categorical

Bivariate Analysis: *Continuous Variables* – Correlation Matrix

- Takeaway:
- Correlation Matrix suggests that:
 - Positive Correlation was seen between loan_amnt, funded_amnt, funded_amnt_inv & annual income
 - Negative Correlation was observed between annual_inc & dti



Bivariate Analysis: *Home ownership and Address State*

- Percent Recovery
 - Maximum: [MORTGAGE & DC] (97.92%)
 - Minimum: [MORTGAGE & NE] (33.33%)

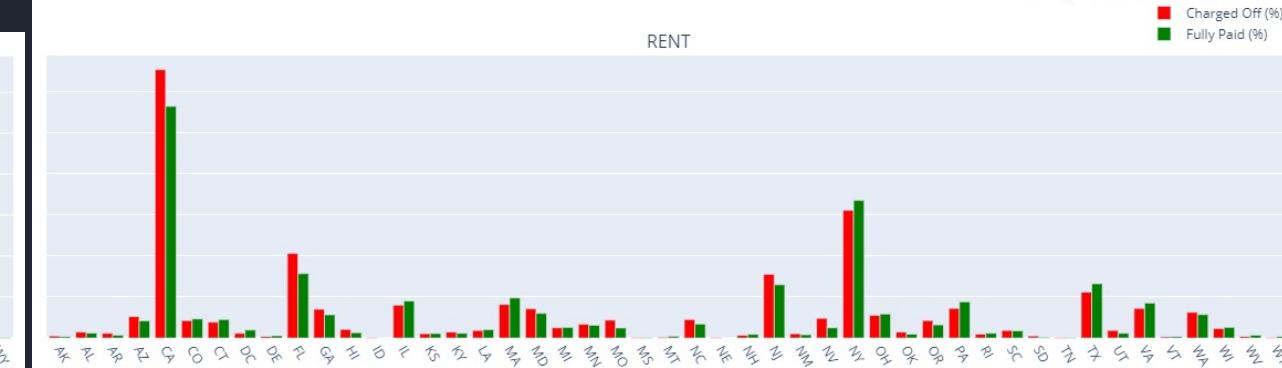


Bivariate Analysis: *Home ownership and Address State*

- Charged Off
 - Maximum contribution: [RENT & CA] (13.04%)
 - Minimum contribution: [MORTGAGE & WY] (0.02%)
- Fully paid
 - Maximum contribution: [RENT & CA] (11.26%)
 - Minimum contribution: [MORTGAGE & NE] (0.0%)

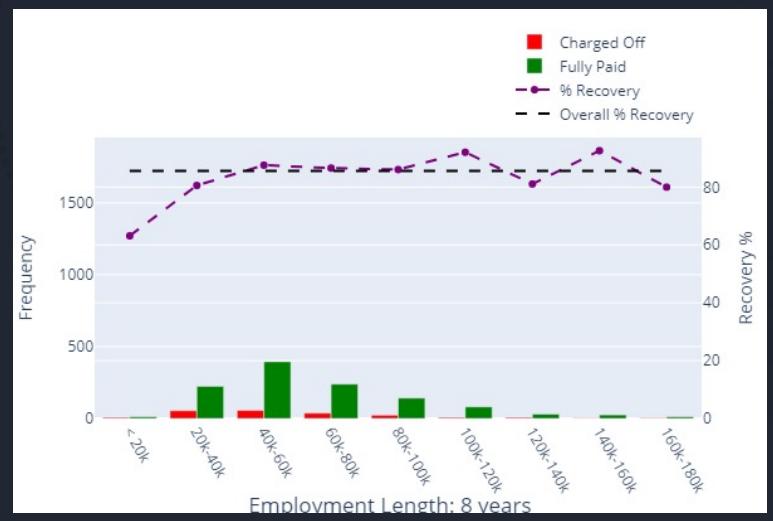
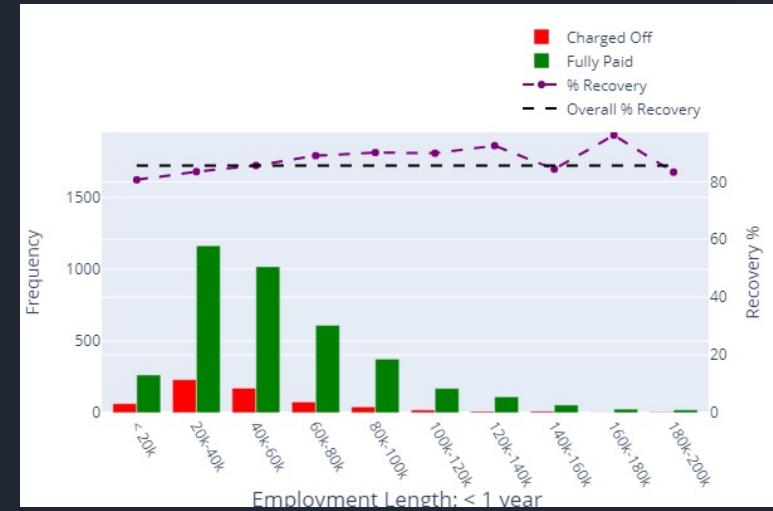


OWN



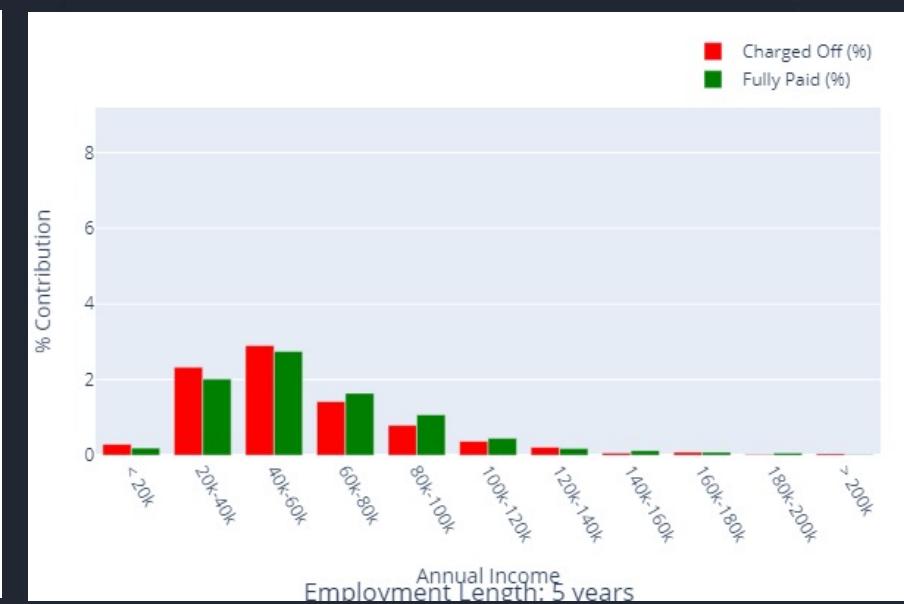
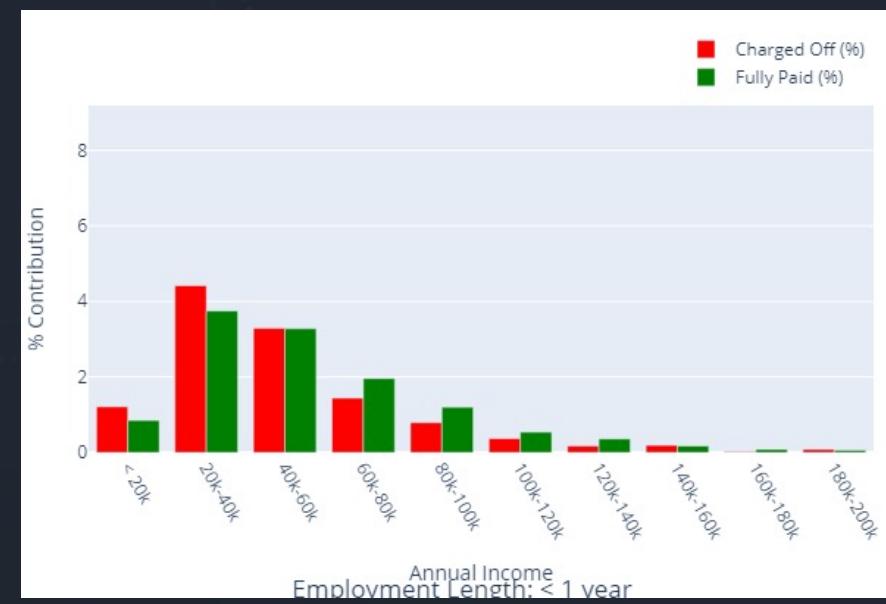
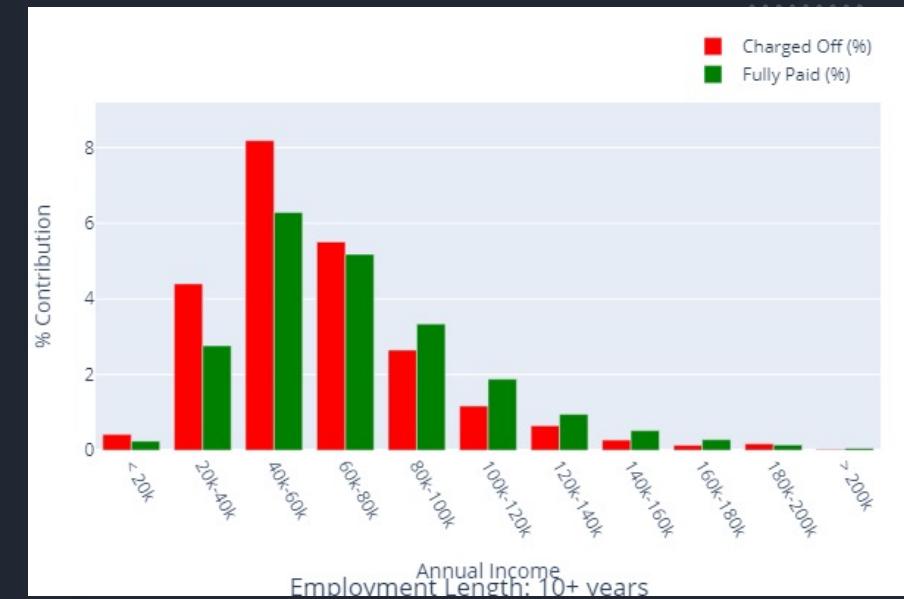
Bivariate Analysis: *Employment Length and Annual Income*

- Percent Recovery
 - Maximum: [< 1 year & 160k-180k] (96.15%)
 - Minimum: [8 years & < 20k] (63.16%)



Bivariate Analysis: Home ownership and Address State

- Charged Off
 - Maximum contribution: [10+ years & 40k-60k] (8.19%)
 - Minimum contribution: [< 1 year & 160k-180k] (0.02%)
- Fully paid
 - Maximum contribution: [10+ years & 40k-60k] (6.29%)
 - Minimum contribution: [5 years & > 200k] (0.01%)



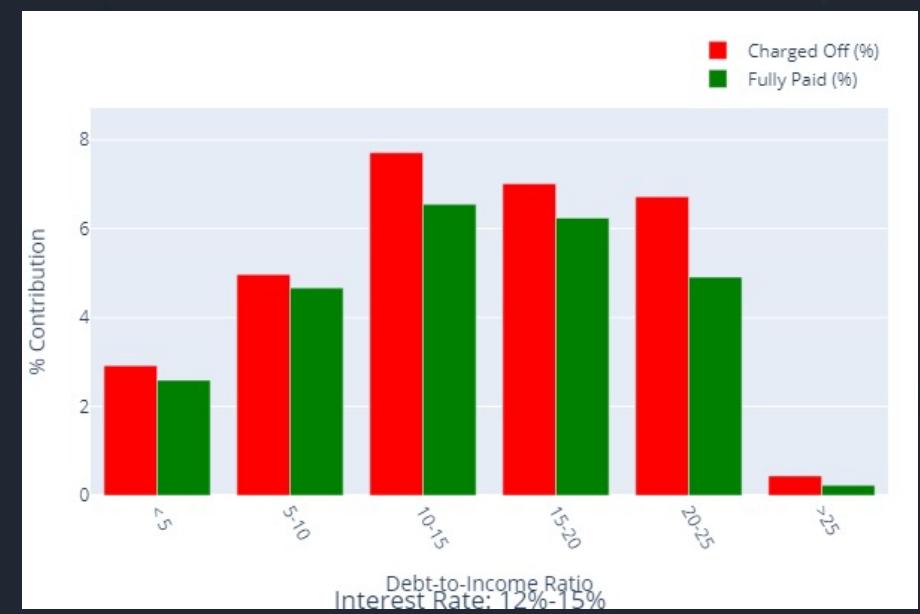
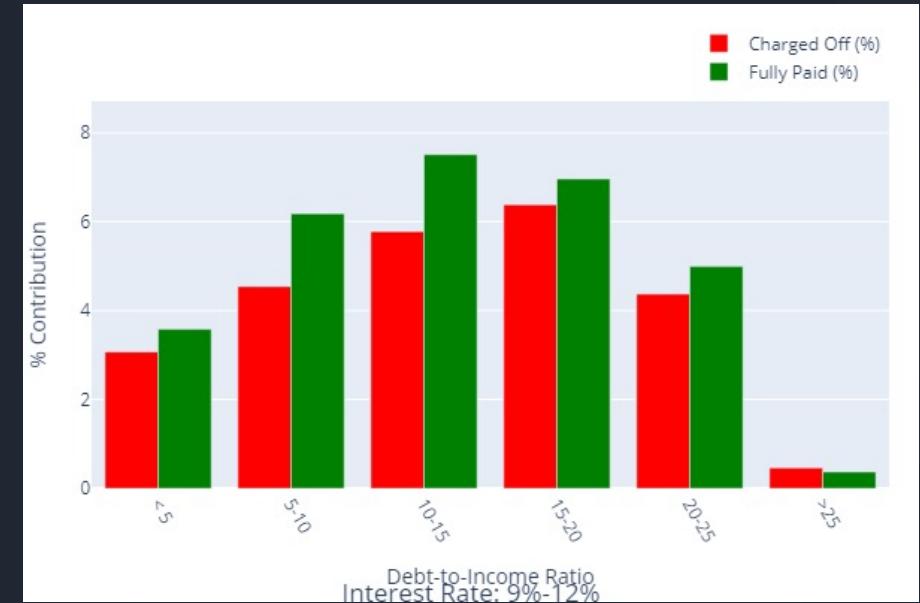
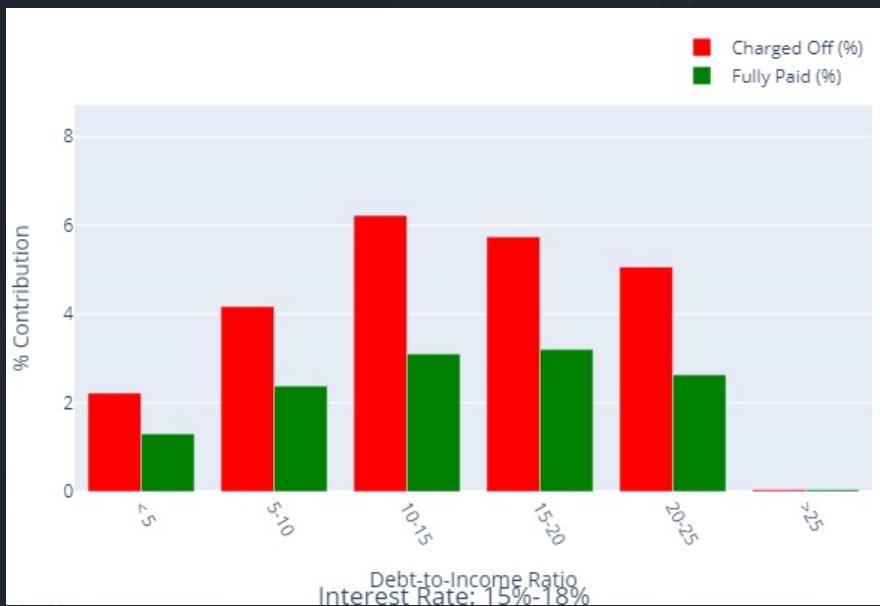
Bivariate Analysis: *Interest Rates and DTI*

- Percent Recovery
 - Maximum: [3%-6% & 20-25] (97.08%)
 - Minimum: [> 20% & 10-15] (55.12%)



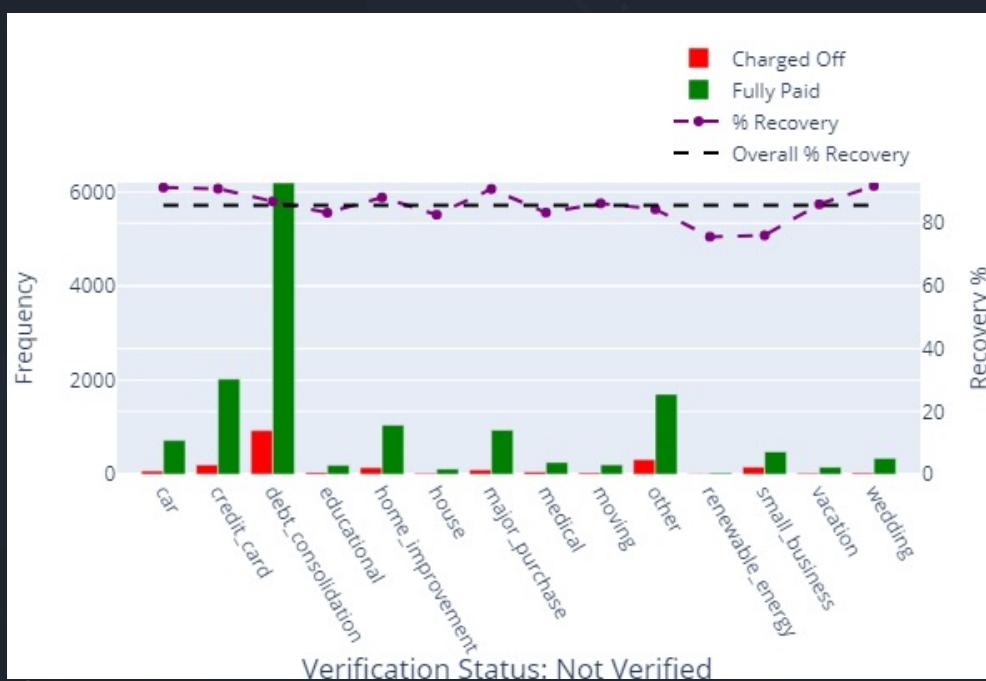
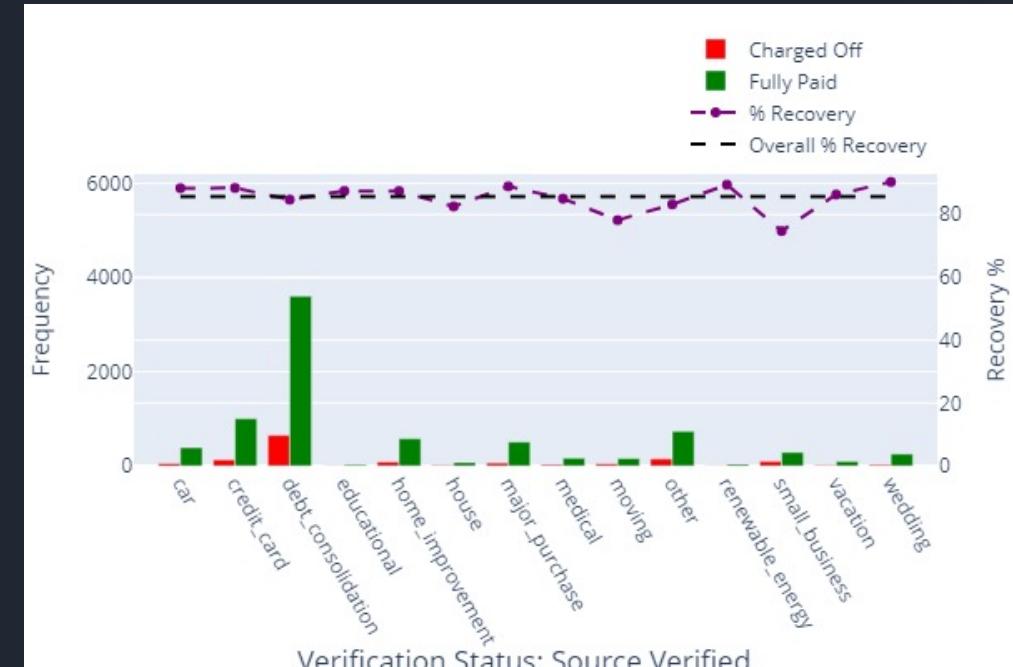
Bivariate Analysis: *Interest Rates and DTI*

- Charged Off
 - Maximum contribution: [12%-15% & 10-15] (7.71%)
 - Minimum contribution: [15%-18% & >25] (0.04%)
- Fully paid
 - Maximum contribution: [9%-12% & 10-15] (7.51%)
 - Minimum contribution: [15%-18% & >25] (0.04%)



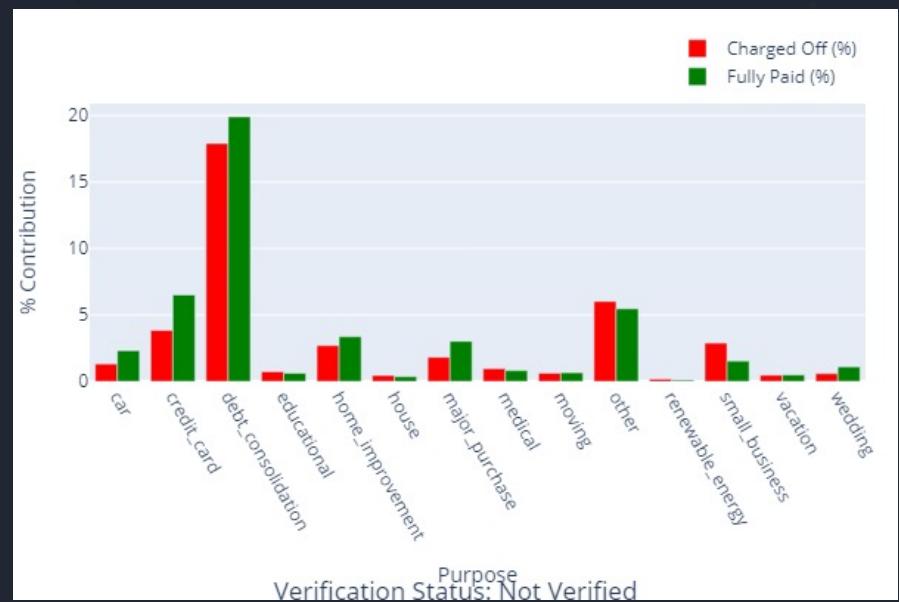
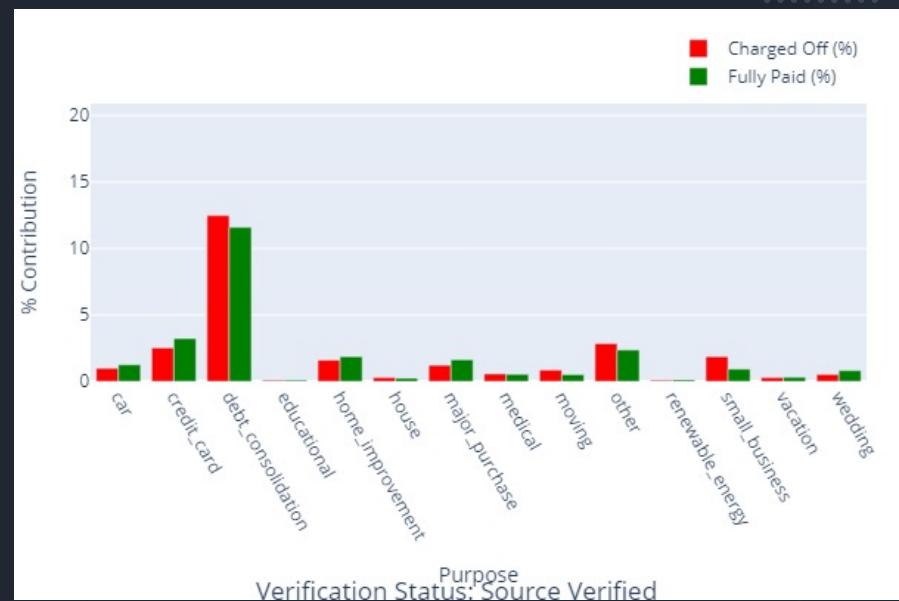
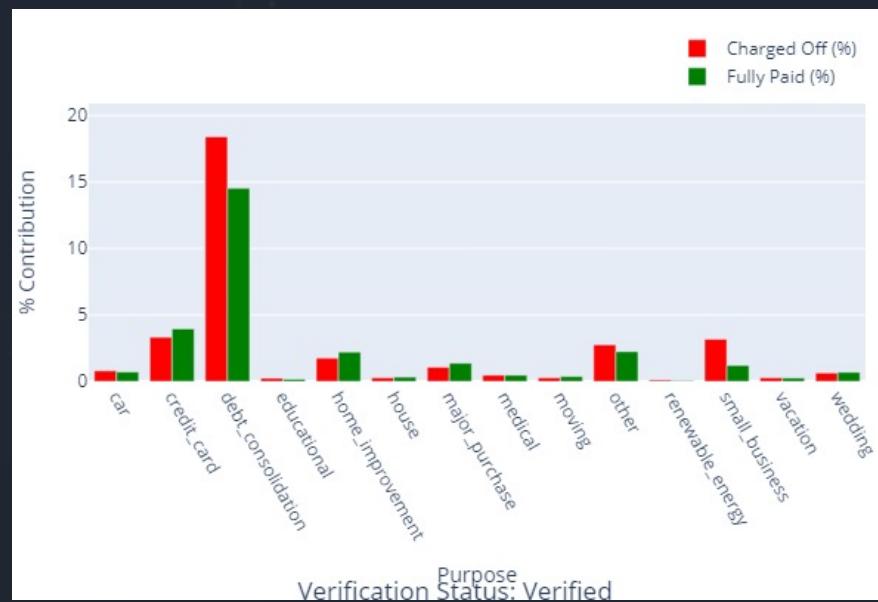
Bivariate Analysis: *Loan Purpose and Verification Status*

- Percent Recovery
 - Maximum: [Not Verified & wedding] (91.87%)
 - Minimum: [Verified & small_business] (69.44%)



Bivariate Analysis: *Loan Purpose and Verification Status*

- Charged Off
 - Maximum contribution: [Verified & debt_consolidation] (18.41%)
 - Minimum contribution: [Source Verified & renewable_energy] (0.08%)
- Fully paid
 - Maximum contribution: [Not Verified & debt_consolidation] (19.9%)
 - Minimum contribution: [Verified & renewable_energy] (0.05%)





Recommendations

Recommendations

- Whenever the loan funding amount was lesser than the loan requested amount the possibility of Charge off was lesser.
- Shorter term of 36 Months had higher possibly of Chargeback compared to 60months term.
- Installments ranging from \$100 - \$200 had more Charge Off's compared to other range of instalments
- Annual Income Ranges of \$20K -\$60K had highest amount of Charge Off.
- Interest Ranges of 12-15% had the highest amount of Charge Off's
- DTI Ranging from 10-10% had the highest amount of Charge Off's
- Customers with Employment length of 10+ years contributed to highest amount of Charge Off
- Loans issued in the month of December saw the highest amount of Charge off's
- People who have at least 1 Or 2 past bankruptcies contributed to less Charge Off's while customers with no bankruptcies caused 80% of the charge backs.
- CA-California State contributed to highest amount of Charge Off, followed by FL & MA
- Customer seeking loan for debt consolidation contributed to highest amount of charge off followed by small business.
- Not Verified customers contributed to highest amount of charge off.
- Customers staying in rented houses contributed highest amount of charge off's
- Grade B & B5 applicants contributed to highest amount of Charge Off.