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

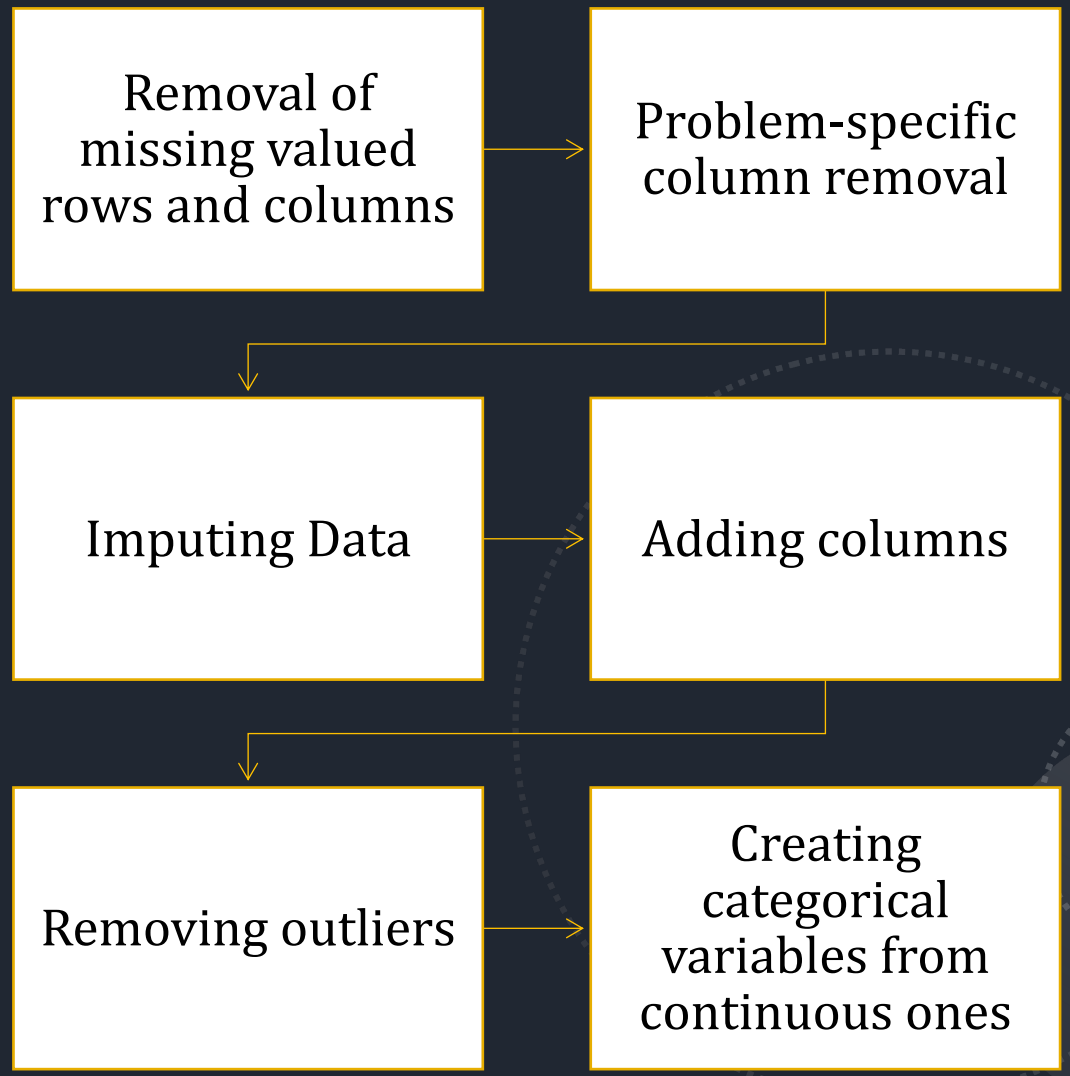
2 – Charged Off

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.

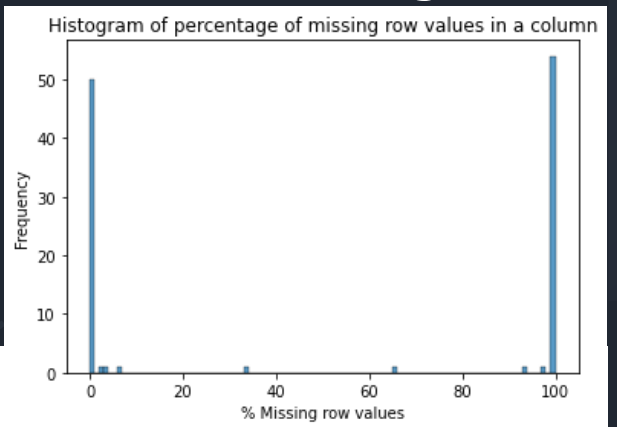
Dataset Cleaning

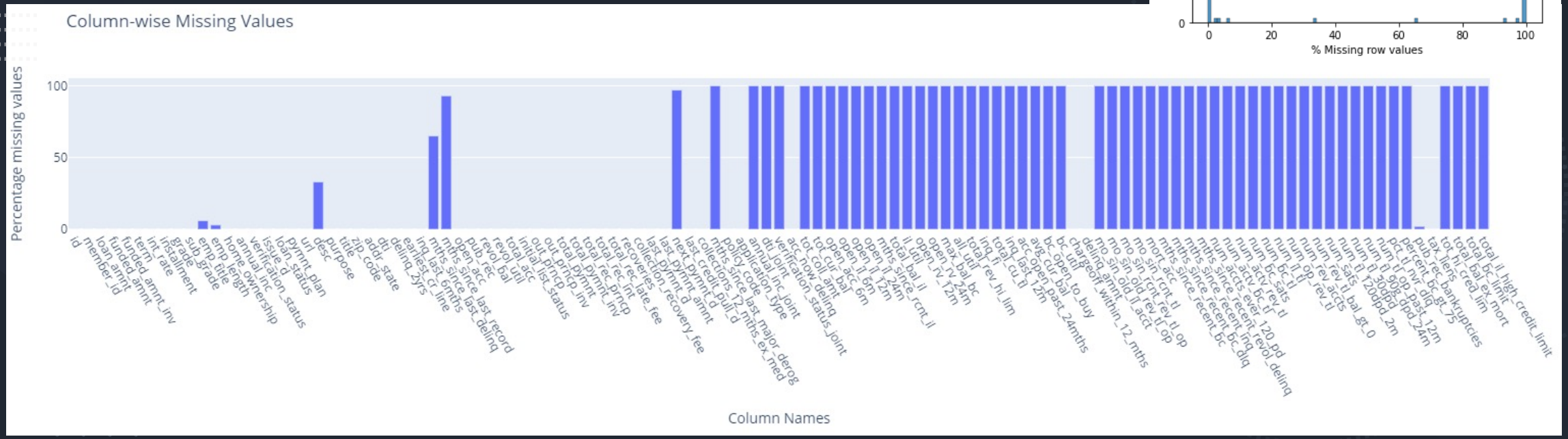
Methods Implemented



Removing Missing Values

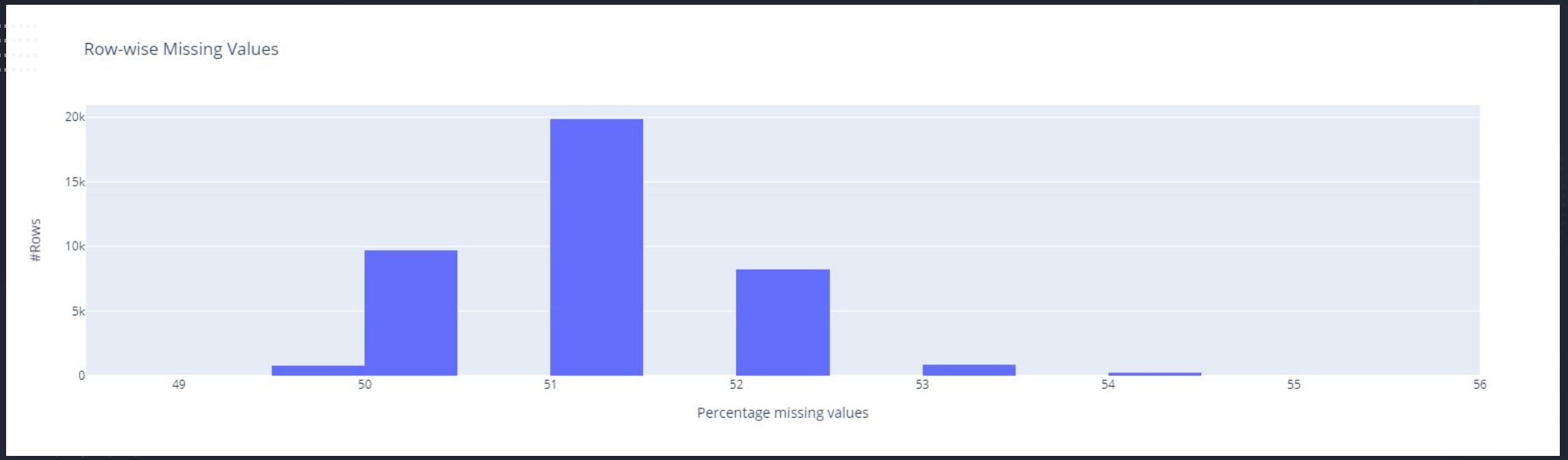
- Column-wise analysis of missing values:
  - 54 columns with 100% missing values





Removing Missing Values

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



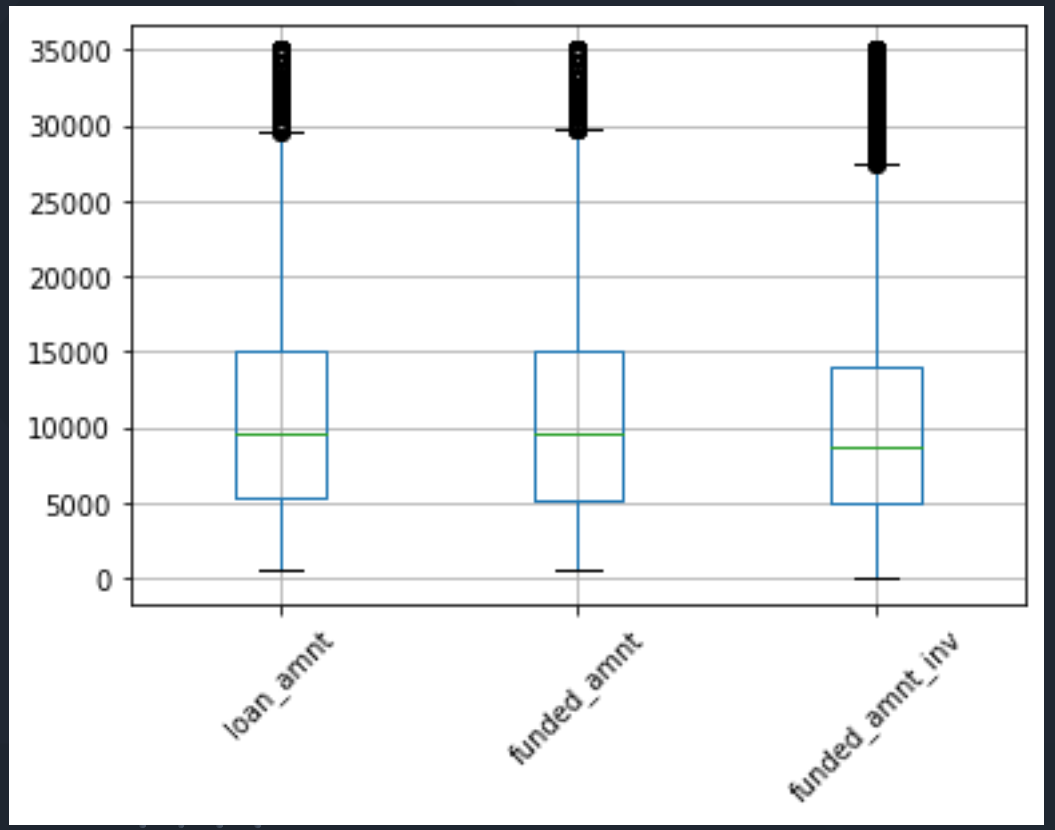
Case-specific data cleaning

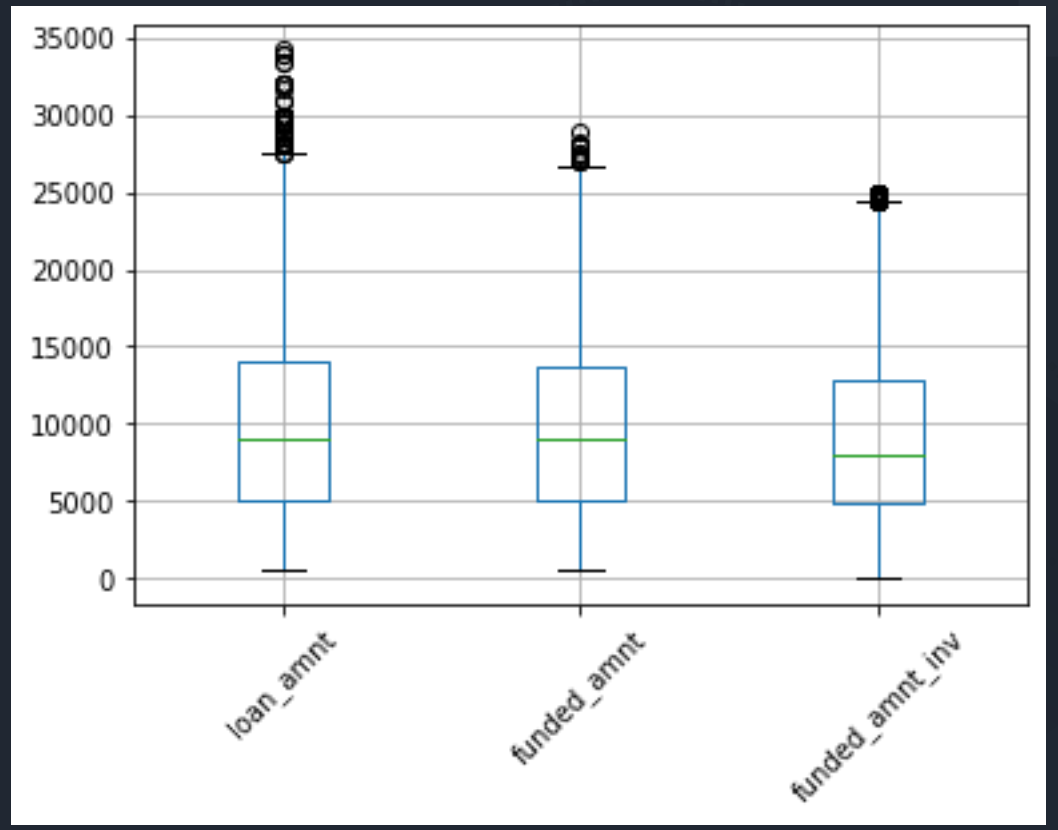
- Remove:
  - Redundant columns
    - *url, desc*
      - No additional information
  - Useless columns
    - *id, member_id, emp_title, title*
      - Do not provide any reasoning to why loans were charged off
    - Columns that provide current customer behavior
  - Useless rows
    - *loan_status == 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

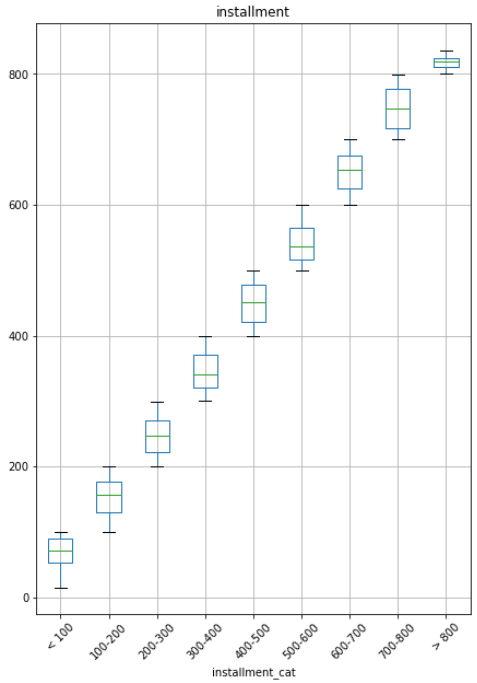
- Removing outliers
  - Removed data beyond the 90 percentile
    - *loan_amnt, funded_amnt, funded_amnt_inv*
    - *installment, annual_inc*

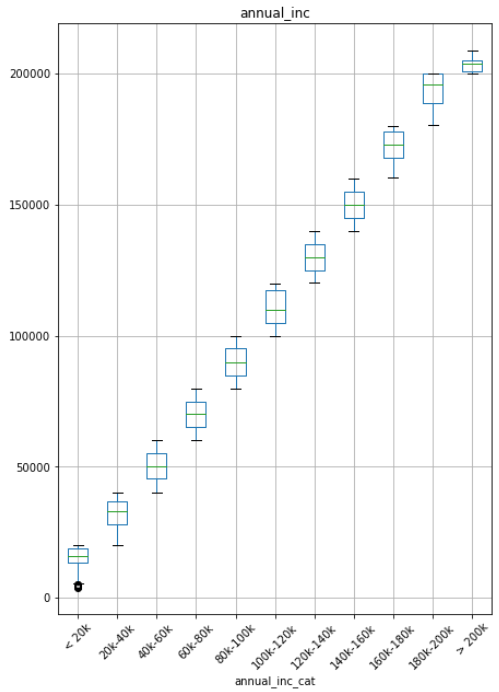


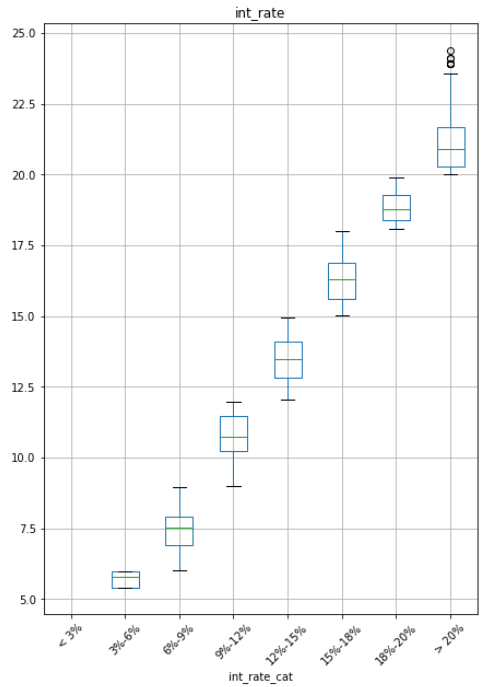


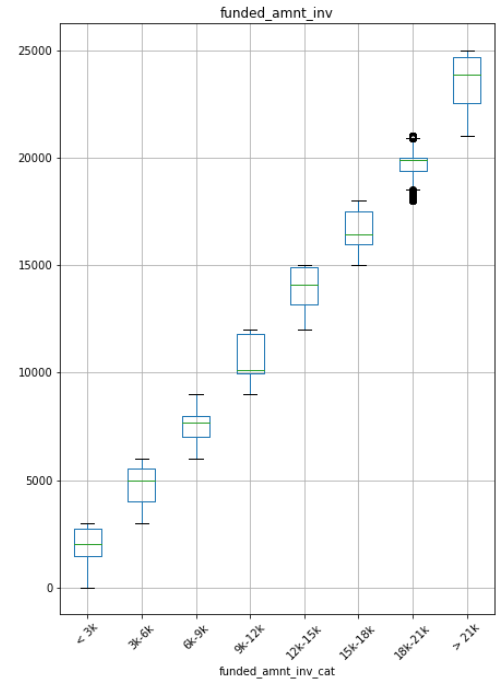
Case-specific data cleaning

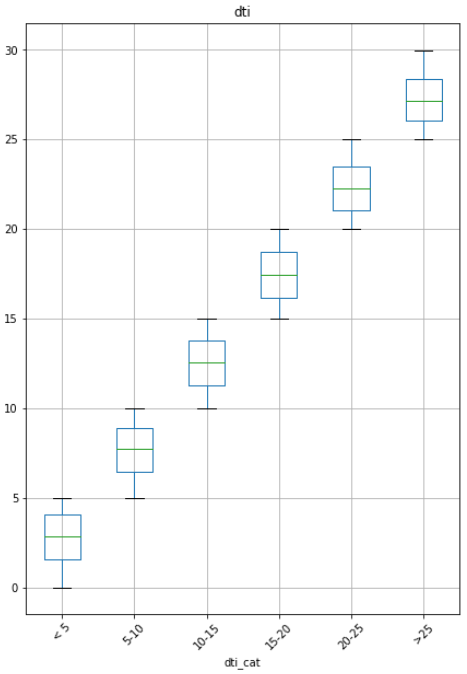
- Creating new columns (categorical varibales)
  - installment -> installment_cat
  - annual_inc -> annual_inc_cat
  - int_rate -> int_rate_cat
  - funded_amnt_inv -> funded_amnt_inv_cat
  - dti -> dti_cat











Univariate Analysis

Analysis of Variable Types

Unordered Categorical

Ordered Categorical

Quantitative

Unordered Categorical Variables

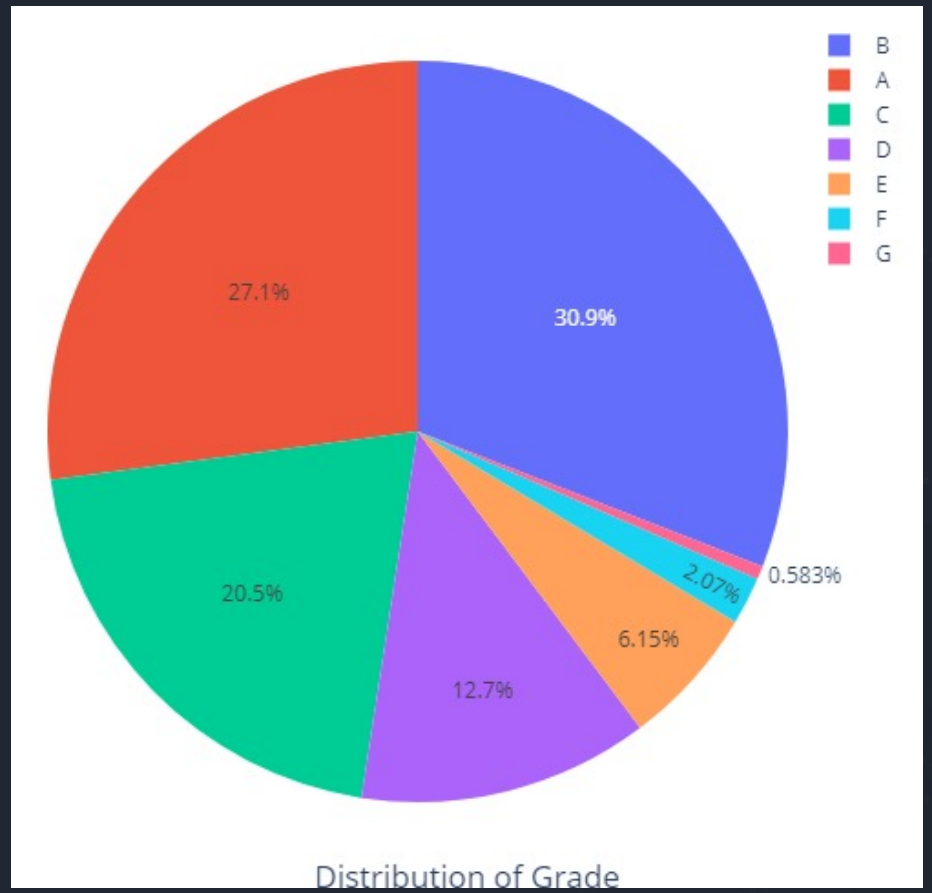
We brought all the columns falling under Unordered Categorical Variables as mentioned below and performed Univariate Analysis.

Below slides have a few pictorial representation of the Univariate Analysis

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

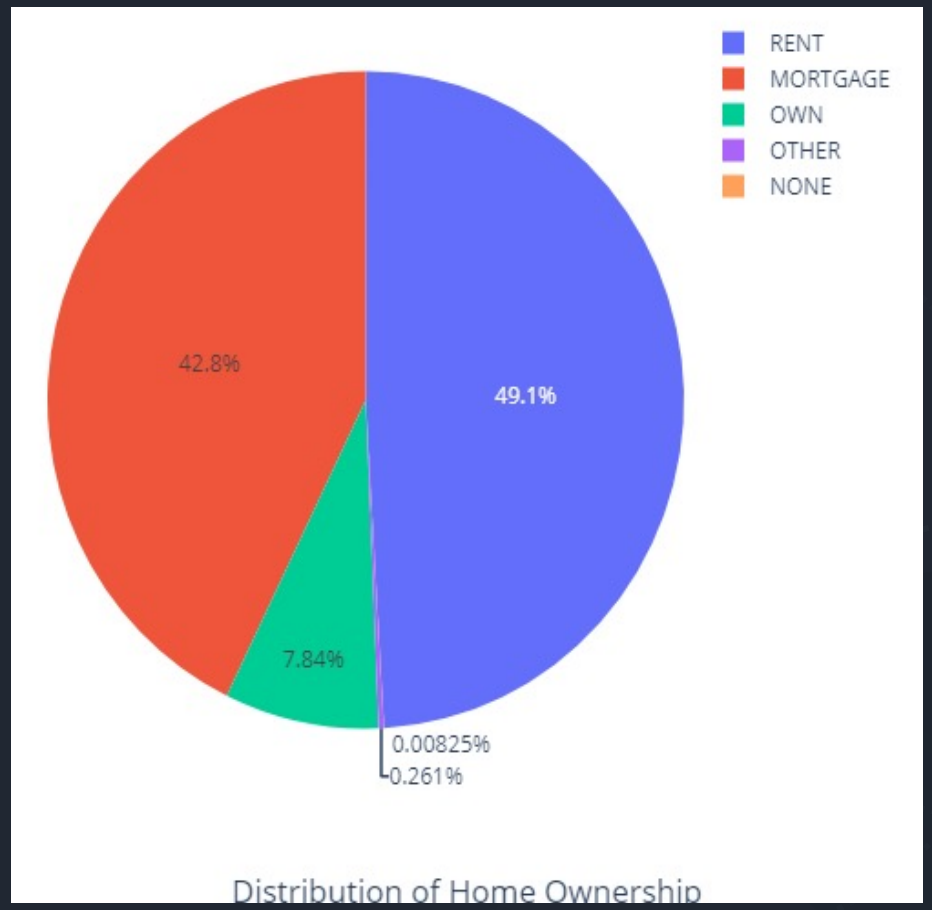
Univariate Analysis: *Grade*

- Takeaway
- Gdf
- Gdfgdfgdf
- Gfdf
- g



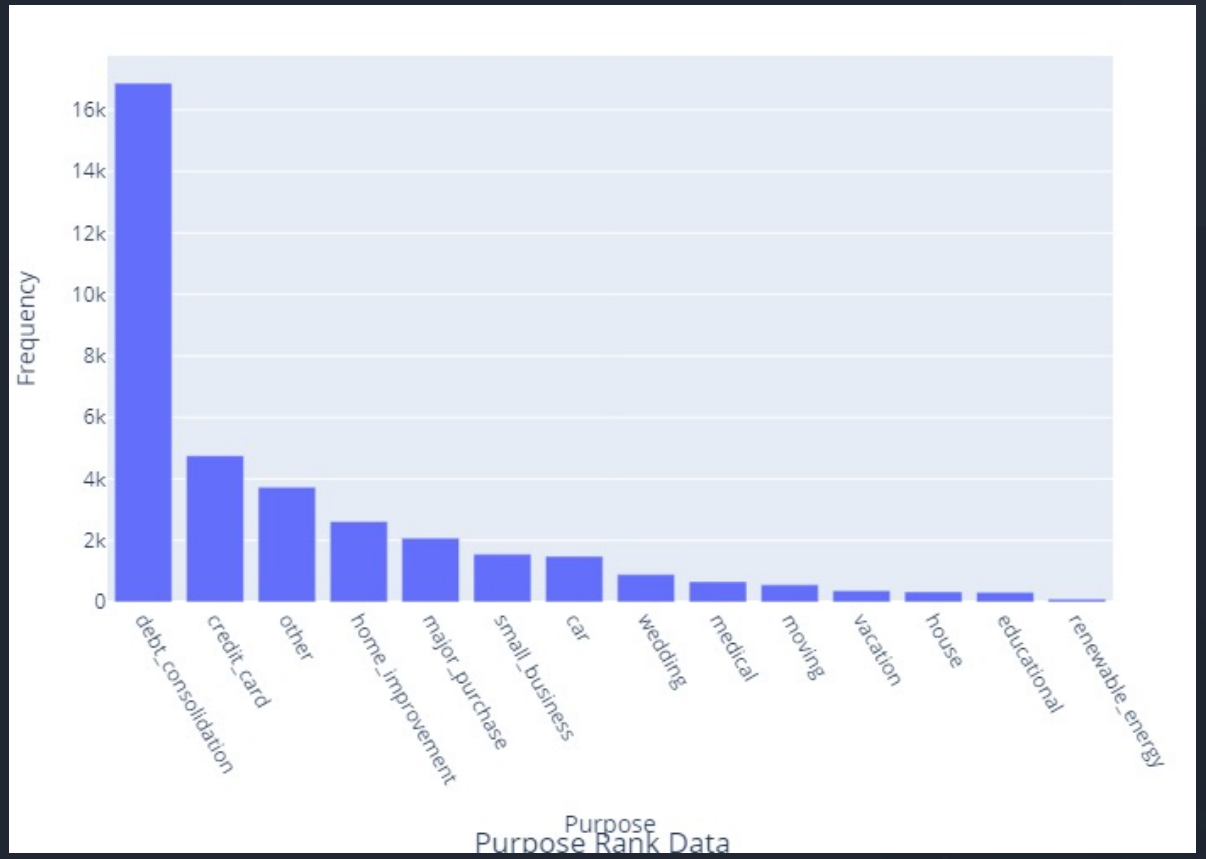
Univariate Analysis: Home Ownership

- Takeaway
- Rented - 49.1%
- Mortgage – 42.8%
- Gfdf
- g



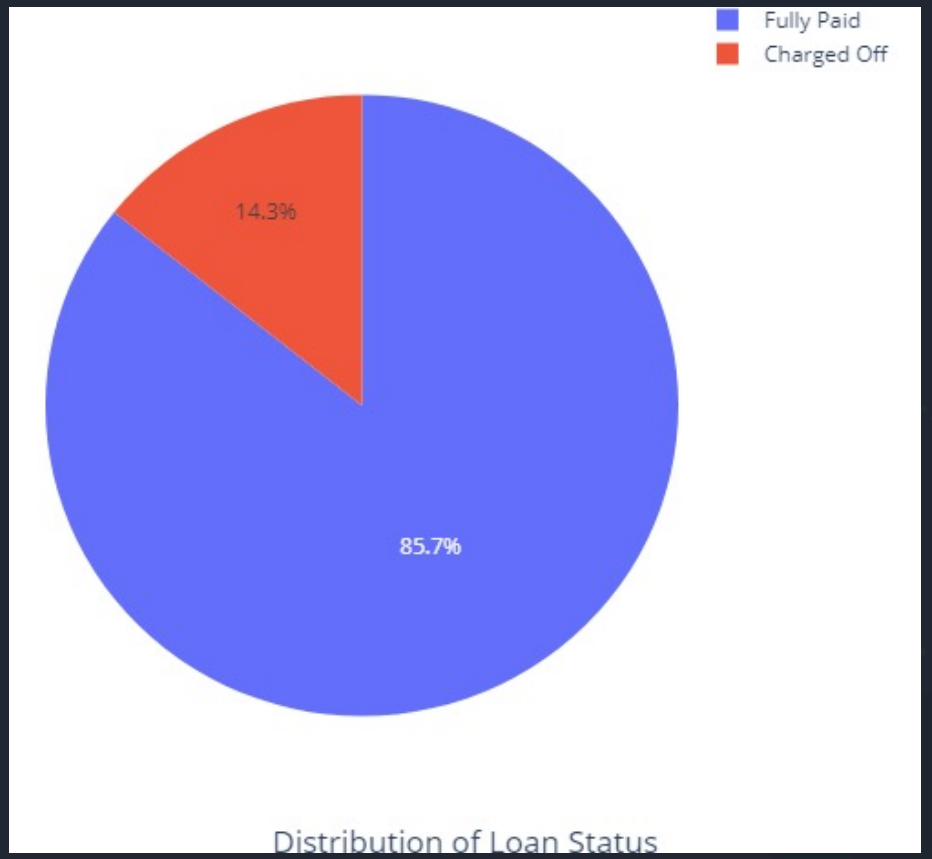
Univariate Analysis: *Loan Purpose*

- Takeaway



Univariate Analysis: *Loan Status*

- Takeaway
- Gdf
- Gdfgdfgdf
- Gfdf
- g



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

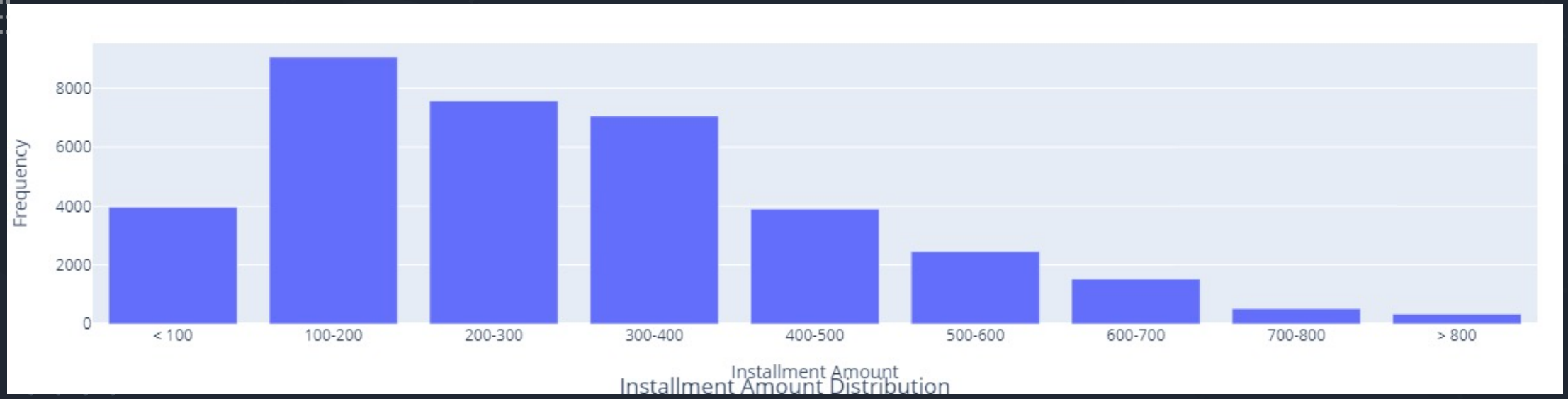
- Loan Term
- Employment Length
- Issue Month
- Issue Year
- Installment
- Annual Income
- Interest Rate
- Funded Amount Invested
- Debt-to-Income Ratio

Univariate Analysis: *Employment Length*

- Takeaway

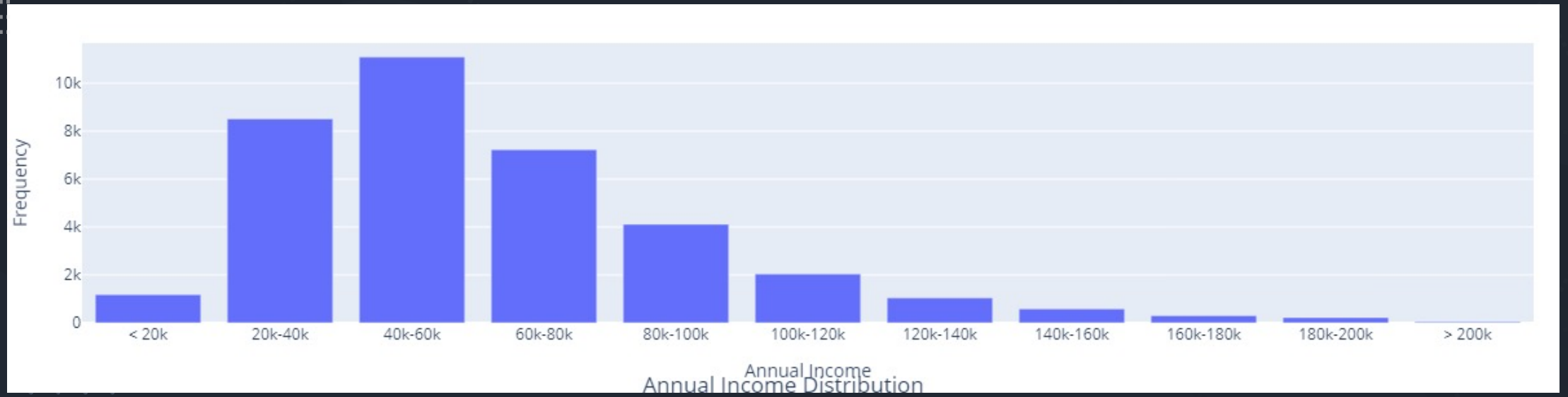
Univariate Analysis: *Installment Amount*

- Takeaway



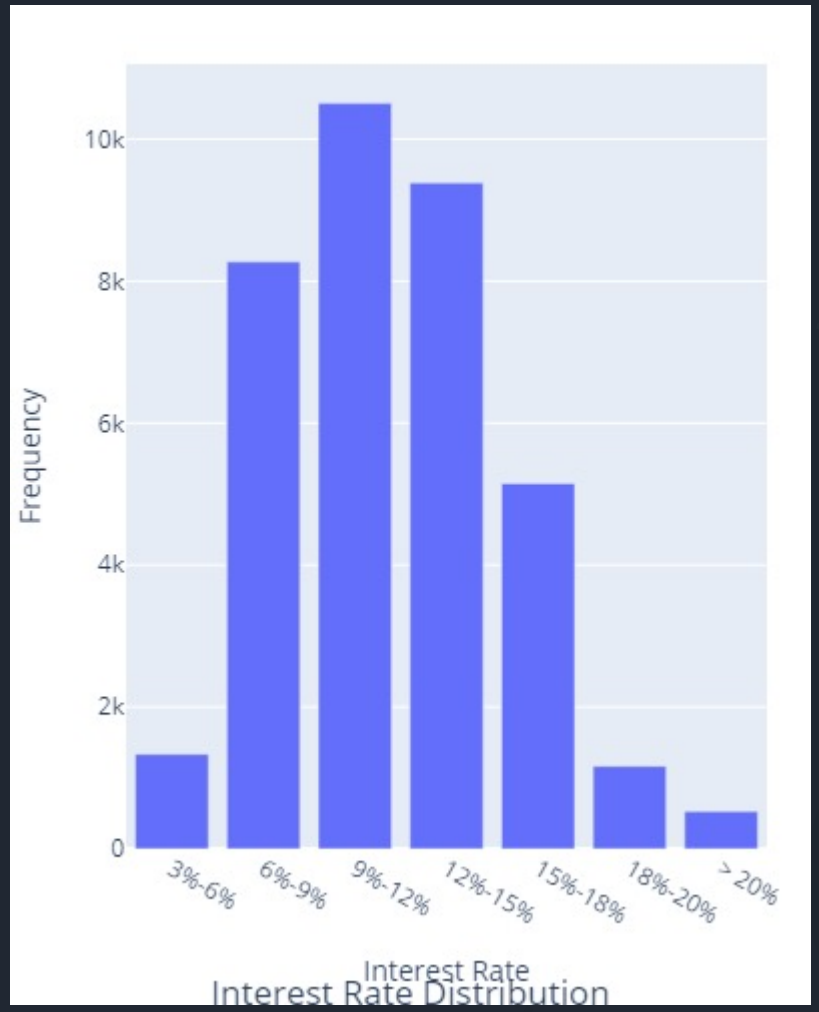
Univariate Analysis: *Annual income*

- Takeaway



Univariate Analysis: *Interest Rate*

- Takeaway
- Gdf
- Gdfgdfgdf
- Gfdf
- g



Quantitative Variables

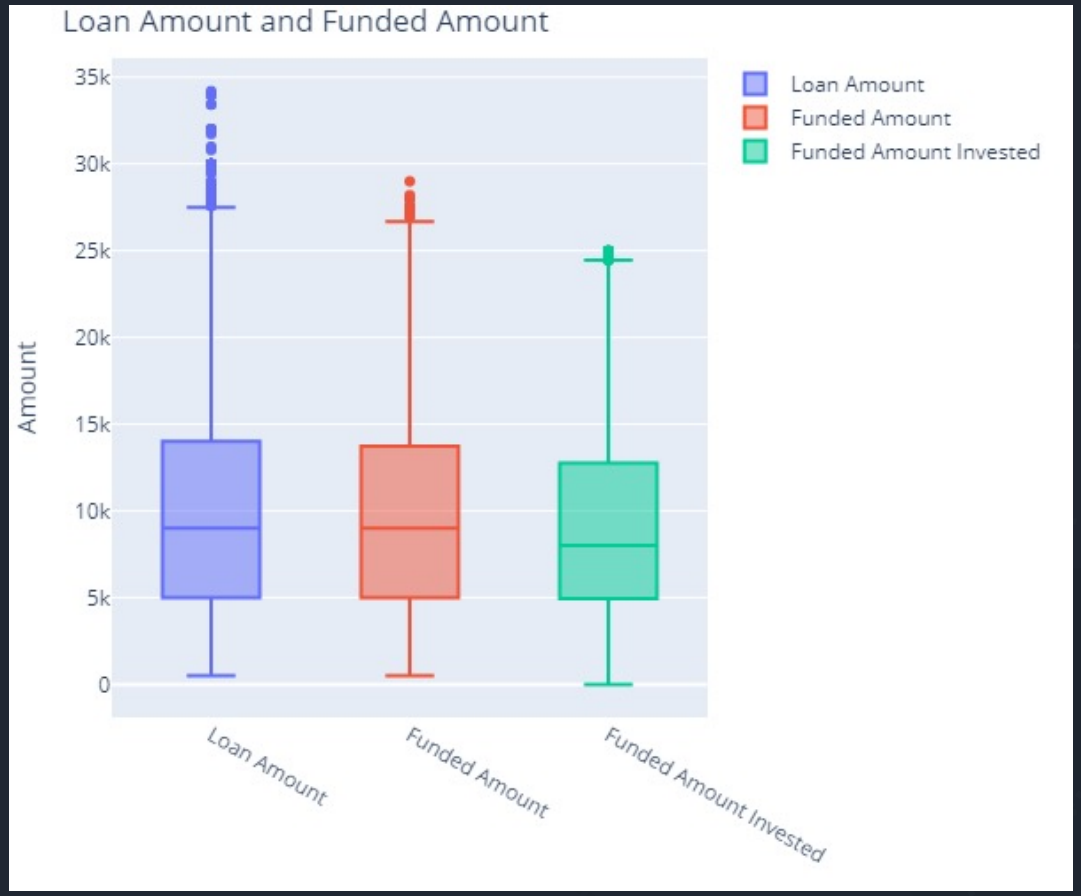
We brought all the columns falling under Quantitative Variables as mentioned below and performed Univariate Analysis.

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

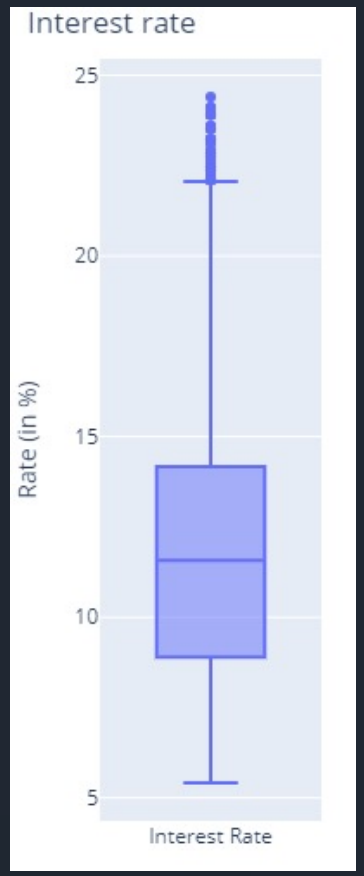
Univariate Analysis: *Loan Amounts*

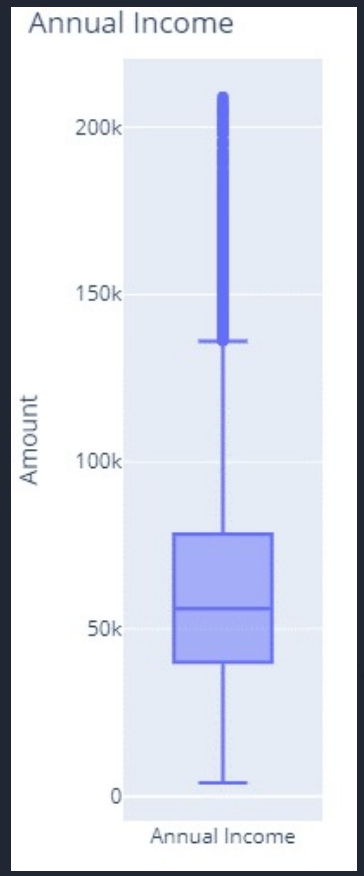
- Takeaway
- Gdf
- Gdfgdfgdf
- Gfdf
- g

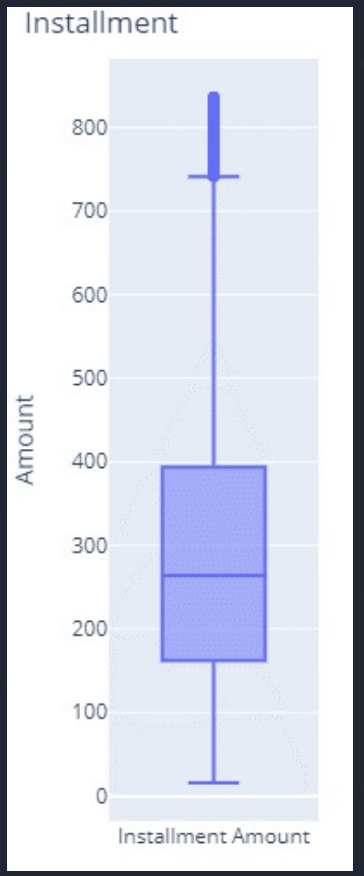


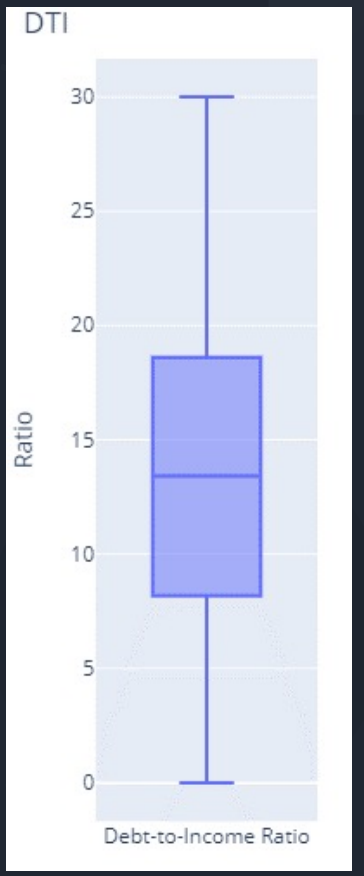
Univariate Analysis: *Others*

- Takeaway
- Gdf
- Gdfgdfgdf
- Gfdf
- g









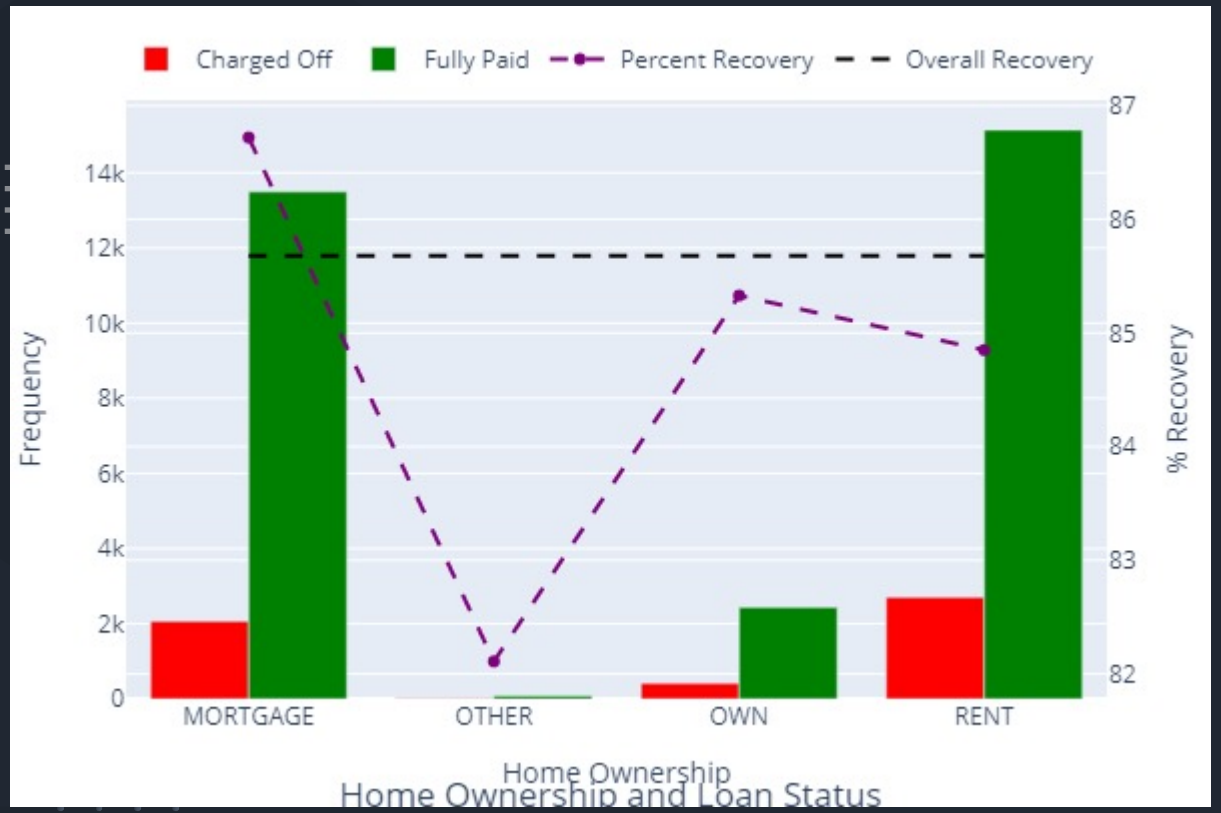
Segmented Univariate Analysis

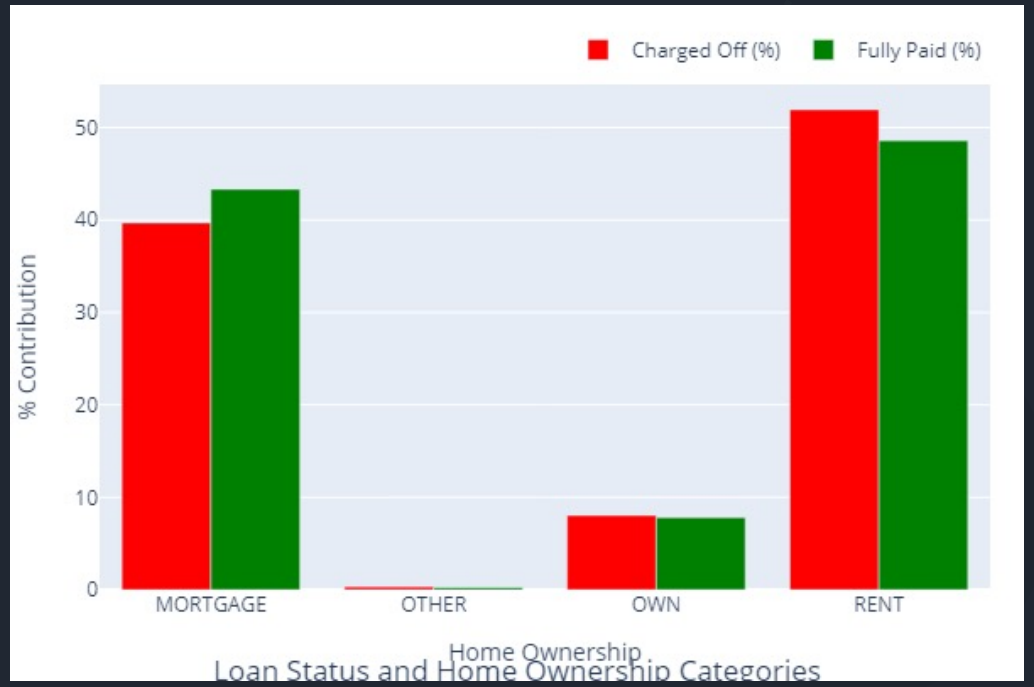
Categorical Variables for Analysis

- 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*

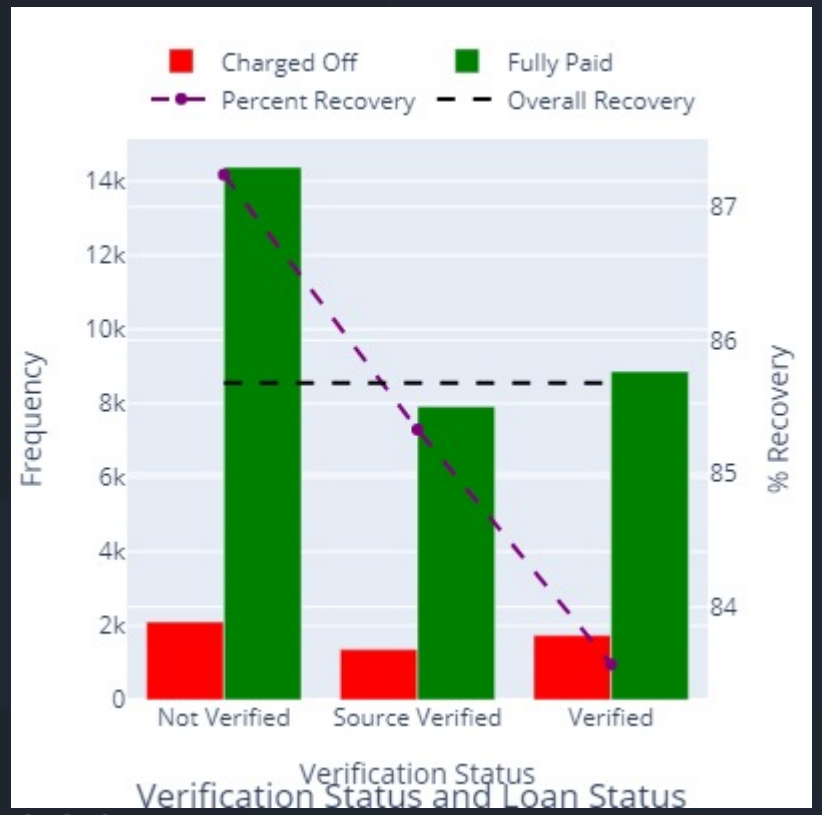
- Charged Off
  - Maximum contribution: RENT (51.94%)
  - Minimum contribution: OTHER (0.33%)
- Fully paid
  - Maximum contribution: RENT (48.61%)
  - Minimum contribution: OTHER (0.25%)

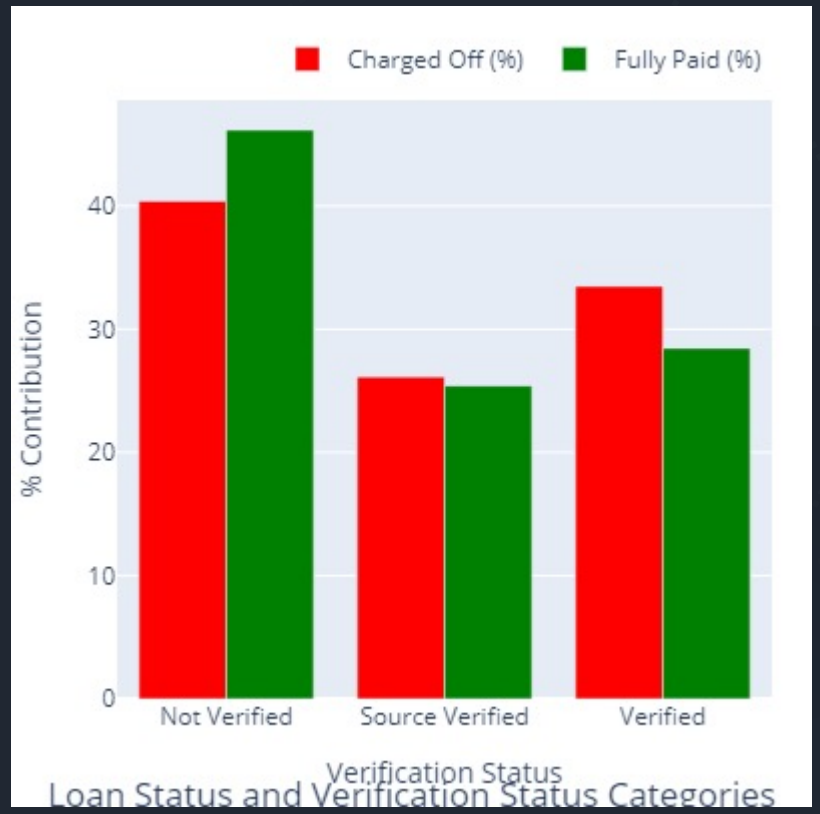




Segmented Univariate Analysis: *Verification*

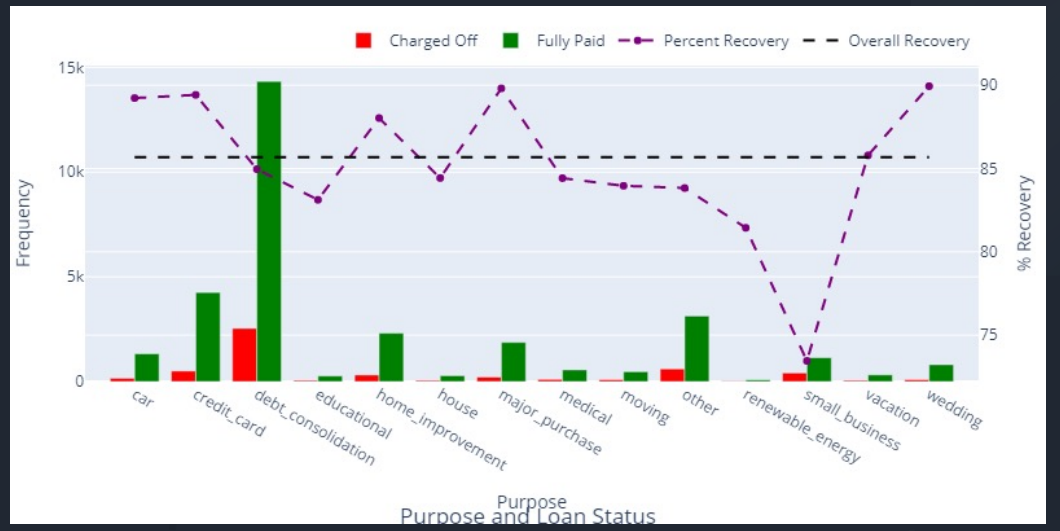
- Charged Off
  - Maximum contribution: Not Verified (40.39%)
  - Minimum contribution: Source Verified (26.13%)
- Fully paid
  - Maximum contribution: Not Verified (46.15%)
  - Minimum contribution: Source Verified (25.4%)

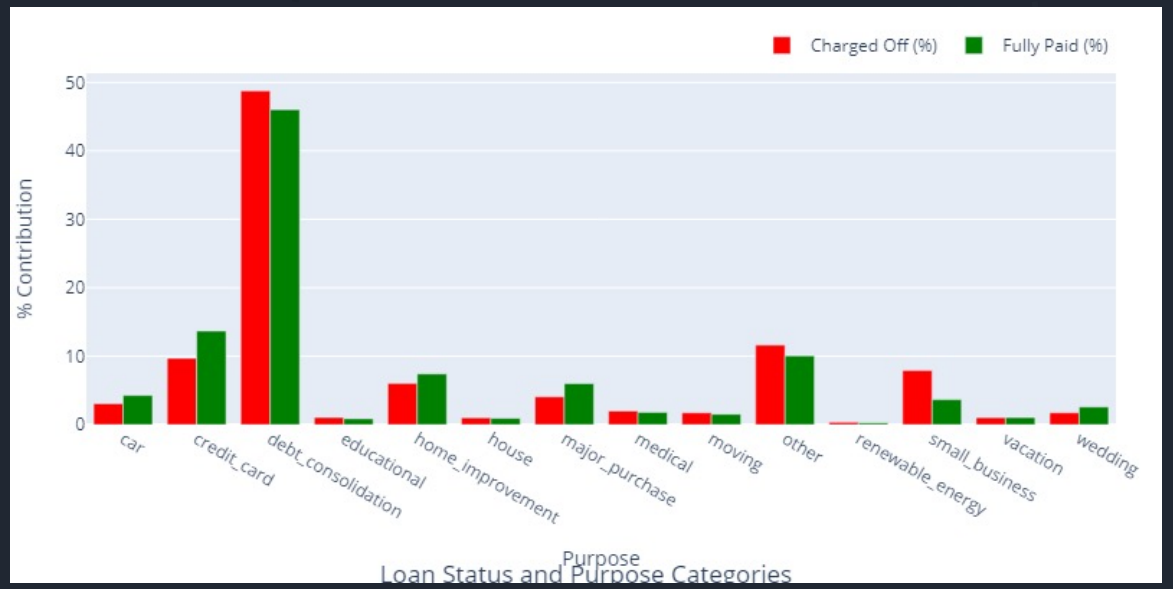




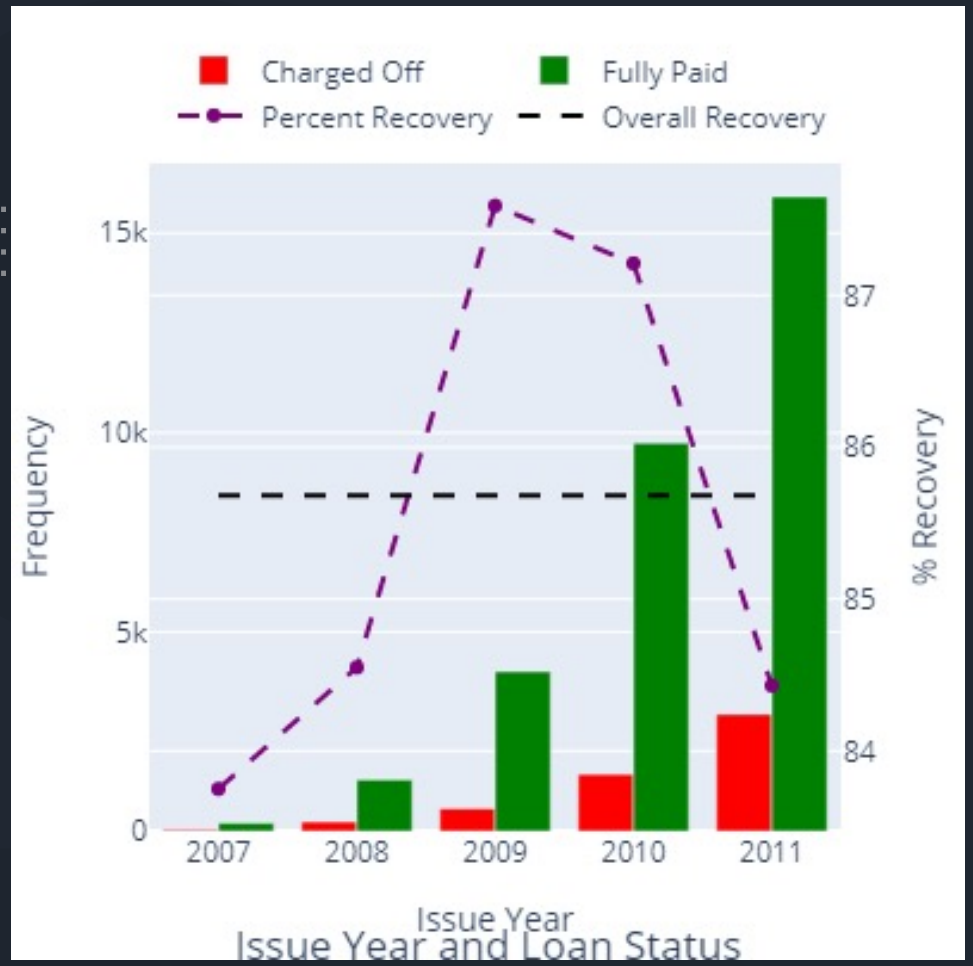
Segmented Univariate Analysis: *Loan Purpose*

- 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%)

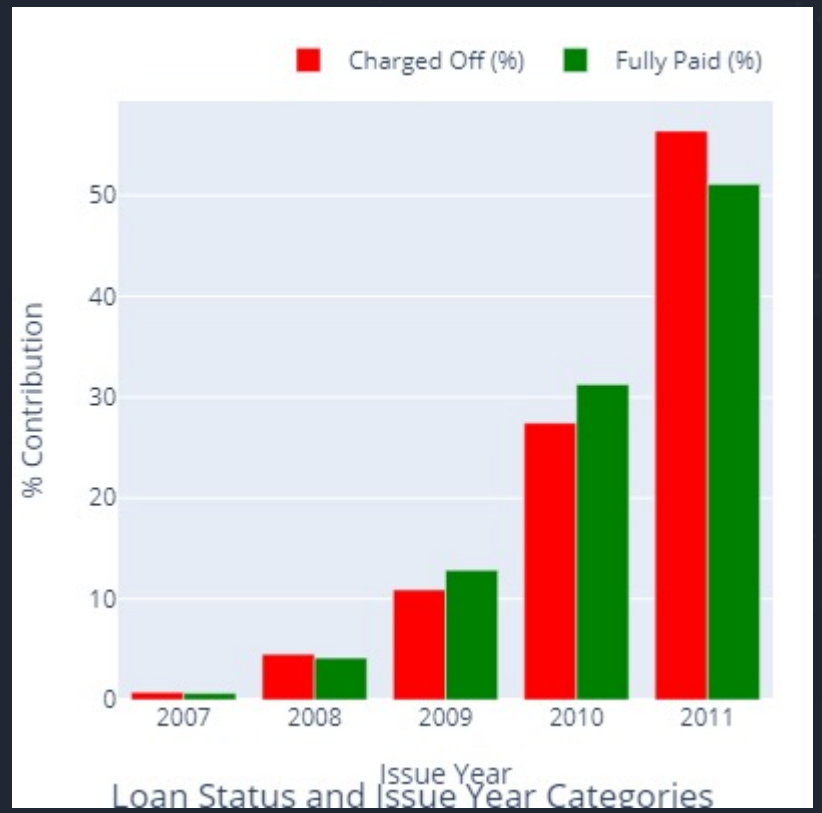




Segmented Univariate Analysis: *Issue Year*

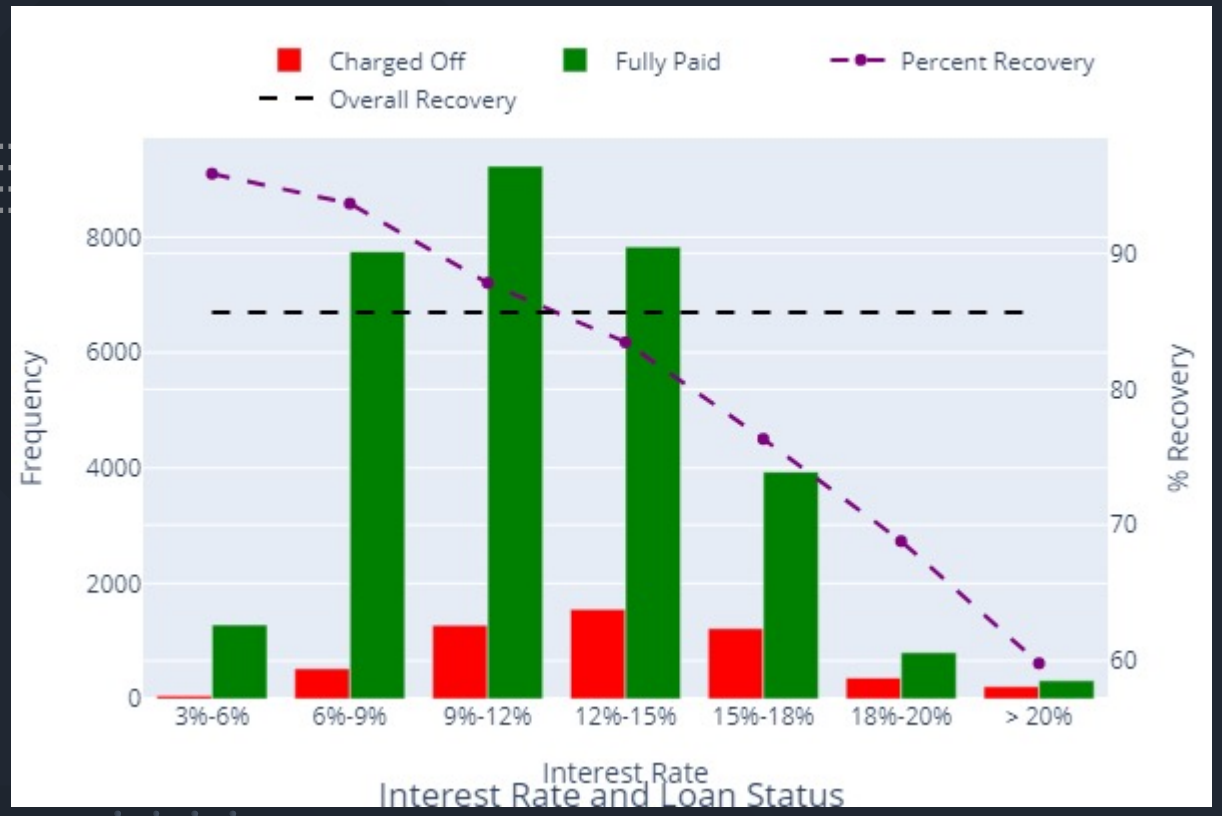


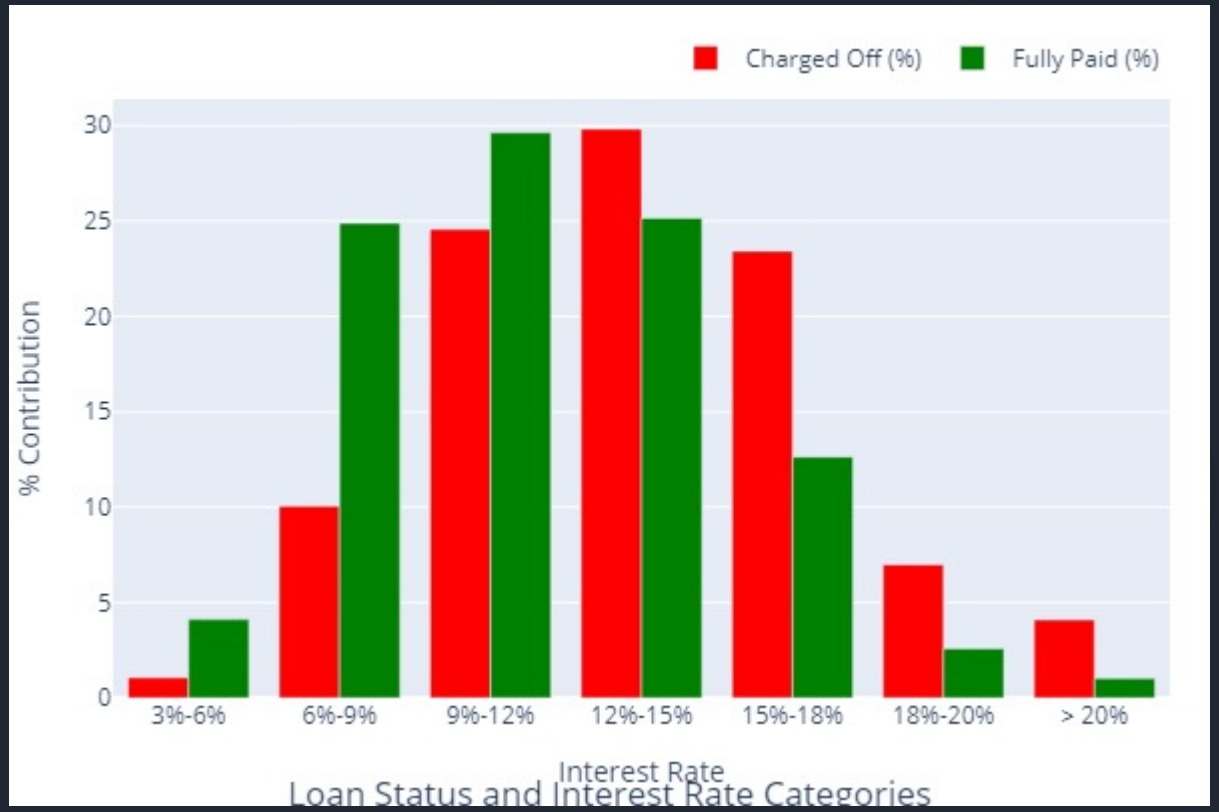
- Charged Off
  - Maximum contribution: 2011 (56.38%)
  - Minimum contribution: 2007 (0.75%)
- Fully paid
  - Maximum contribution: 2011 (51.1%)
  - Minimum contribution: 2007 (0.65%)



Segmented Univariate Analysis: *Interest Rate*

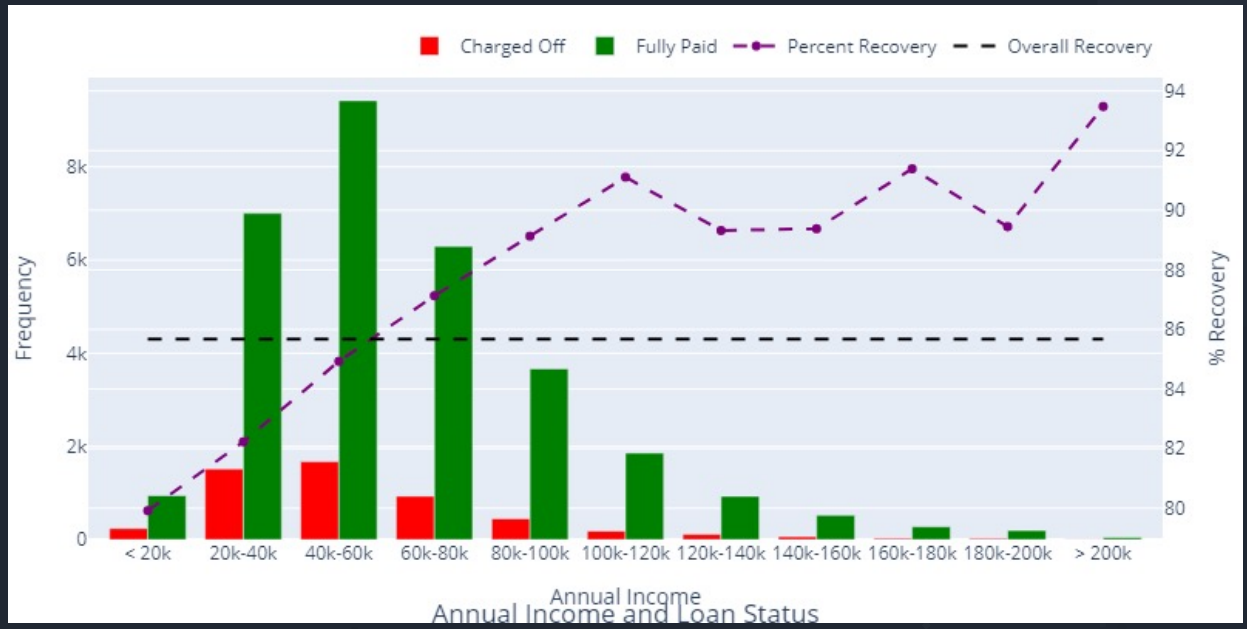
- Charged Off
  - Maximum contribution: 12%-15% (29.82%)
  - Minimum contribution: 3%-6% (1.06%)
- Fully paid
  - Maximum contribution: 9%-12% (29.63%)
  - Minimum contribution: > 20% (1.02%)

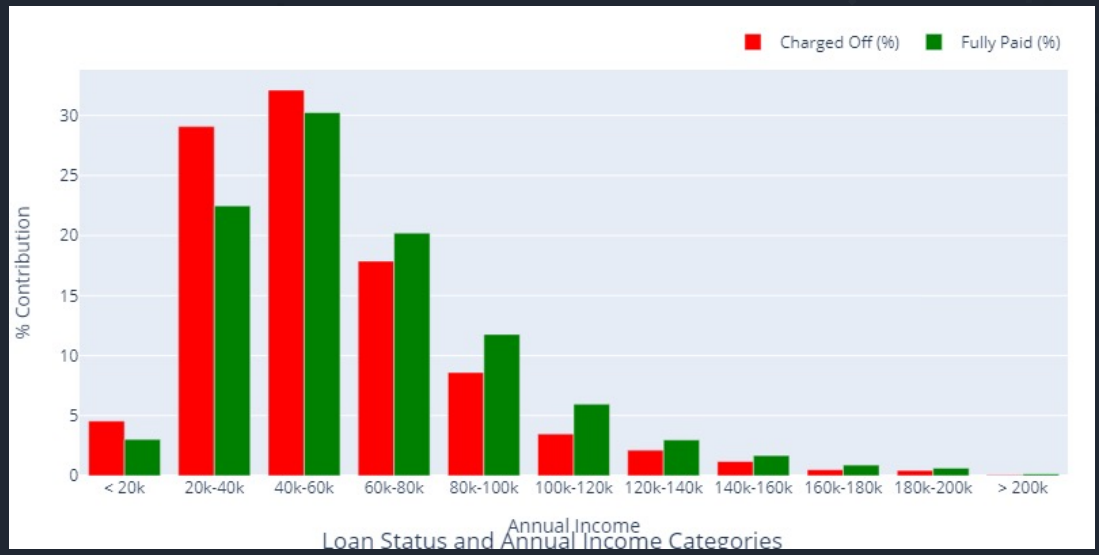




Segmented Univariate Analysis: *Annual Income*

- 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

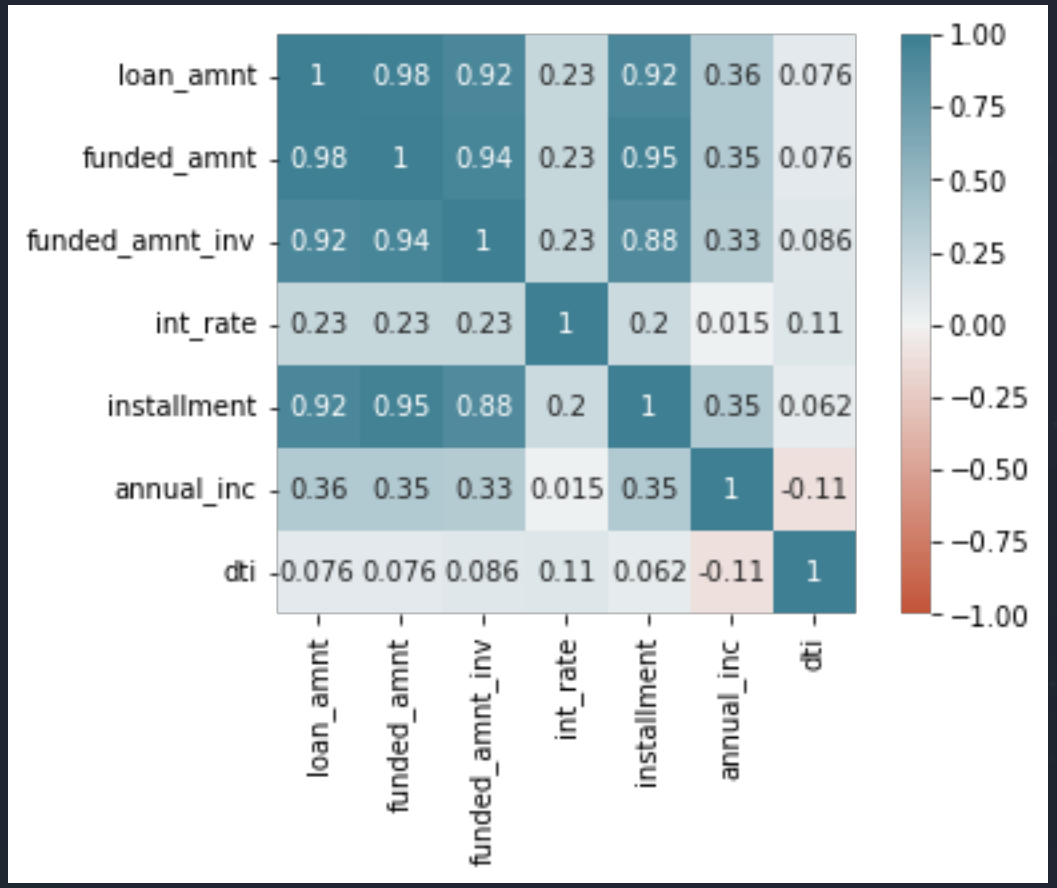
Analysis of Variable Types

Continuous

Categorical

Bivariate Analysis: *Continuous Variables – Correlation Matrix*

- Takeaway
- Gdf
- Gdfgdfgdf
- Gfdf
- g



Recommendations