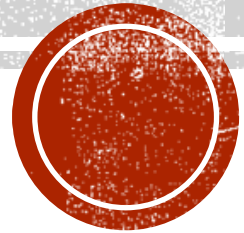
LENDING CLUB CASE STUDY

Exploratory Data Analysis – Assignment

Executive PG Programme in Machine Learning



OUR UNDERSTANDING OF REQUIREMENTS

A consumer finance company which specializes in lending various types of loans wants to perform analysis on the existing data to learn about the risks and arrive to a solution which will enable them to lend the loans in a way which will reduce the risks.

For this activity we have used the concepts of exploratory data analysis and analyzed the data using python

This presentation describes about the data analysis done on the provided data based loan.csv on the company's loan applicant's history.

APPROACH

Data Cleaning

- Removing Duplicates
- Removing null Columns
- Removing special Characters from Number columns

Univariate Analysis

- Analyze the data for loan amount, Interest Rate, Total Payment.
- Discard outliers
- Derive and analyze the correlation matrix for loan.

Bivariate Analysis

- Perform analysis on loan amount and purpose of loan
- Check the no. of loans provided vs status of loans
- Analyze the data based on annual income vs charged off

Conclusion

- Based on the analysis provide observations

DATA CLEANSING

- Based on the data provided it is observed that there are total 111 columns and 39717 rows in the loan.csv dataset.
- 54 columns from these are identified as null columns and hence discarded.
- Further unwanted columns like are removed and the dataset is finally left with 39717 rows and 51 columns which can be used for analysis.
- Further it was observed that columns like interest rate have special characters hence the data is cleaned to remove the special characters.
- All the columns related to amount are being converted into numeric values.

OBSERVATIONS (AFTER DATA CLEANING)

- Initial Observations based on loan status and total loans issued are shown in table.
- Based on the data 82% of loans issues are fully paid and around 14% got charged off.

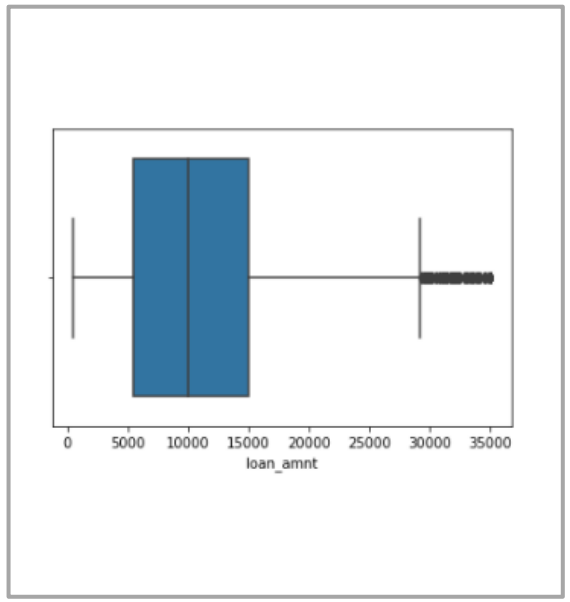
Fully Paid	82.96%
Charged Off	14.17%
Current	2.87%

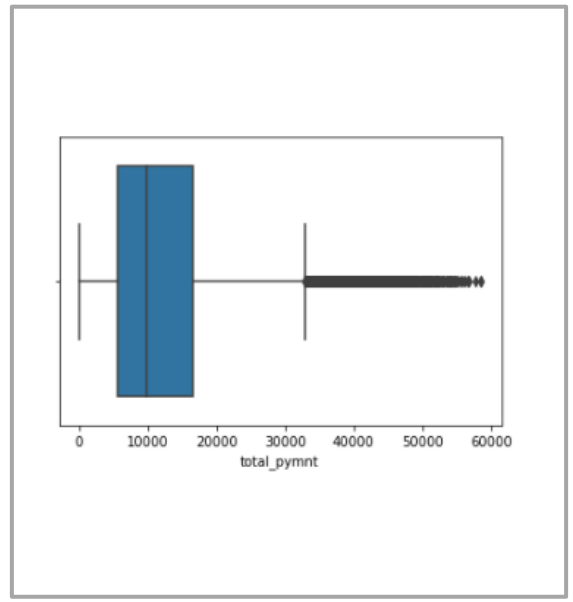
- Initial Observations based on purpose of loan are shown in adjacent table.
- It is observed the major chunk of loans were taken for debt_consolidation and credit card is also the trailing reason after debt consolidation.

debt_consolidation	46.93%
credit_card	12.92%
other	10.05%
home_improvement	7.49%
major_purchase	5.51%
small_business	4.60%
car	3.90%
wedding	2.38%
medical	1.74%
moving	1.47%
vacation	0.96%
house	0.96%
educational	0.82%
renewable_energy	0.26%

UNIVARIATE ANALYSIS

- Basic Statistics are derived for quantitative variables.
- Variables used are:
  - Loan_amount
  - Annual_Inc
  - Total_payment
- Observations:
  - While performing basic analysis on Quantitative variables it is observed that there are outliers Annual Income column. Hence the outliers are removed for further analysis.
  - The data related to loan amount and total payment seems to be fine for further analysis.



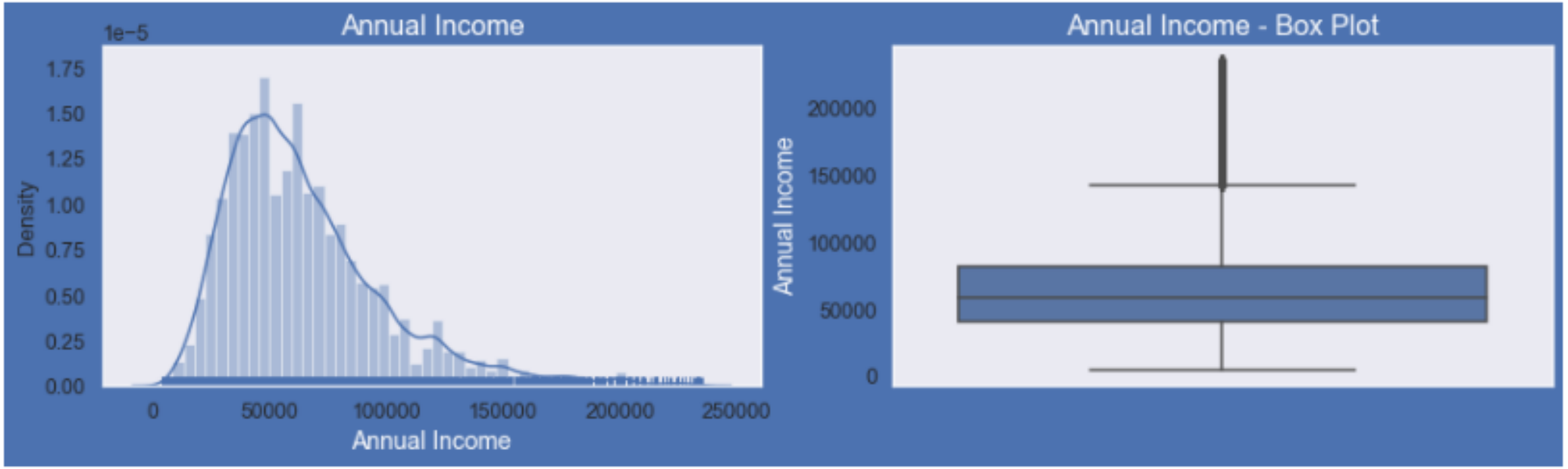


Before Removing Outliers in Annual Income :

```
count     39717.00
mean      68968.93
std       63793.77
min        4000.00
25%       40404.00
50%       59000.00
75%       82300.00
max     6000000.00
Name: annual_inc, dtype: float64
```

After Removing Outliers from annual Income :

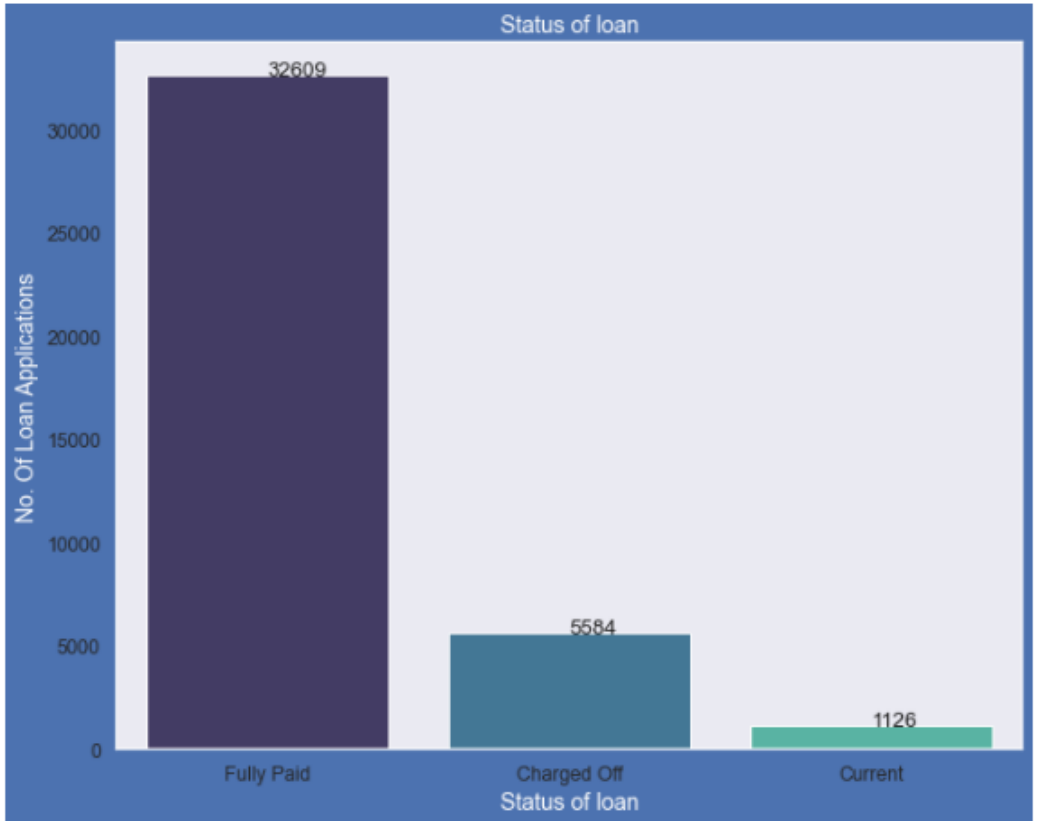
```
count     39319.00
mean      65524.22
std       35215.89
min        4000.00
25%       40000.00
50%       58000.00
75%       81000.00
max      234996.00
Name: annual_inc, dtype: float64
```



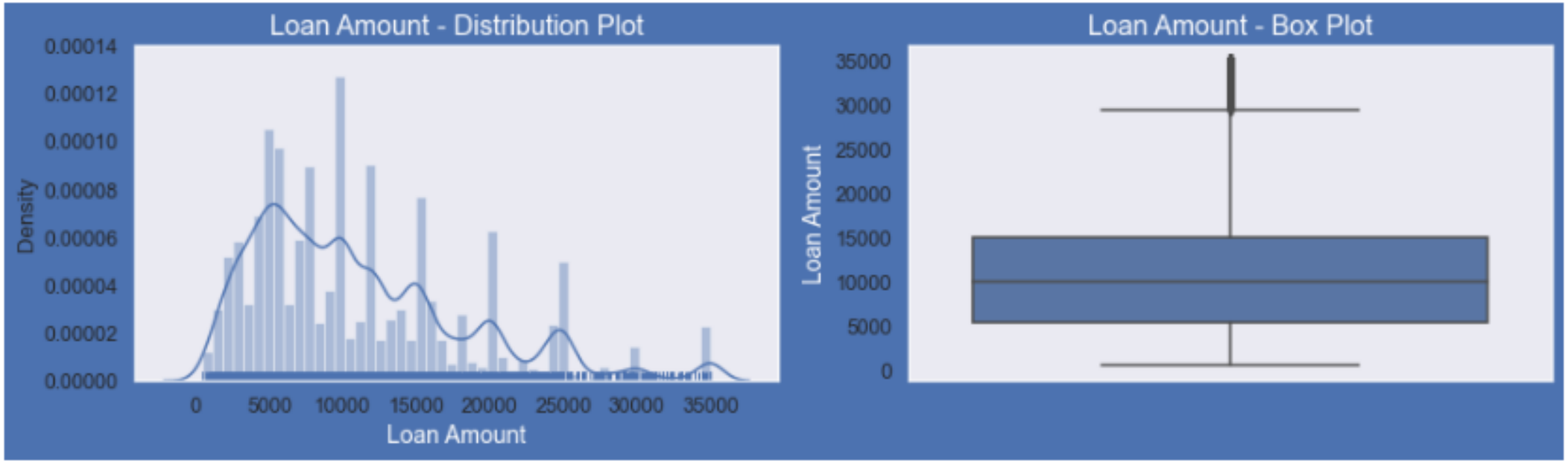
UNIVARIATE ANALYSIS . . .

- Quantitative Variables [Annual Income]
- The plot indicates that the most of the borrower's income is in the range of 40000 to 80000

UNIVARIATE ANALYSIS



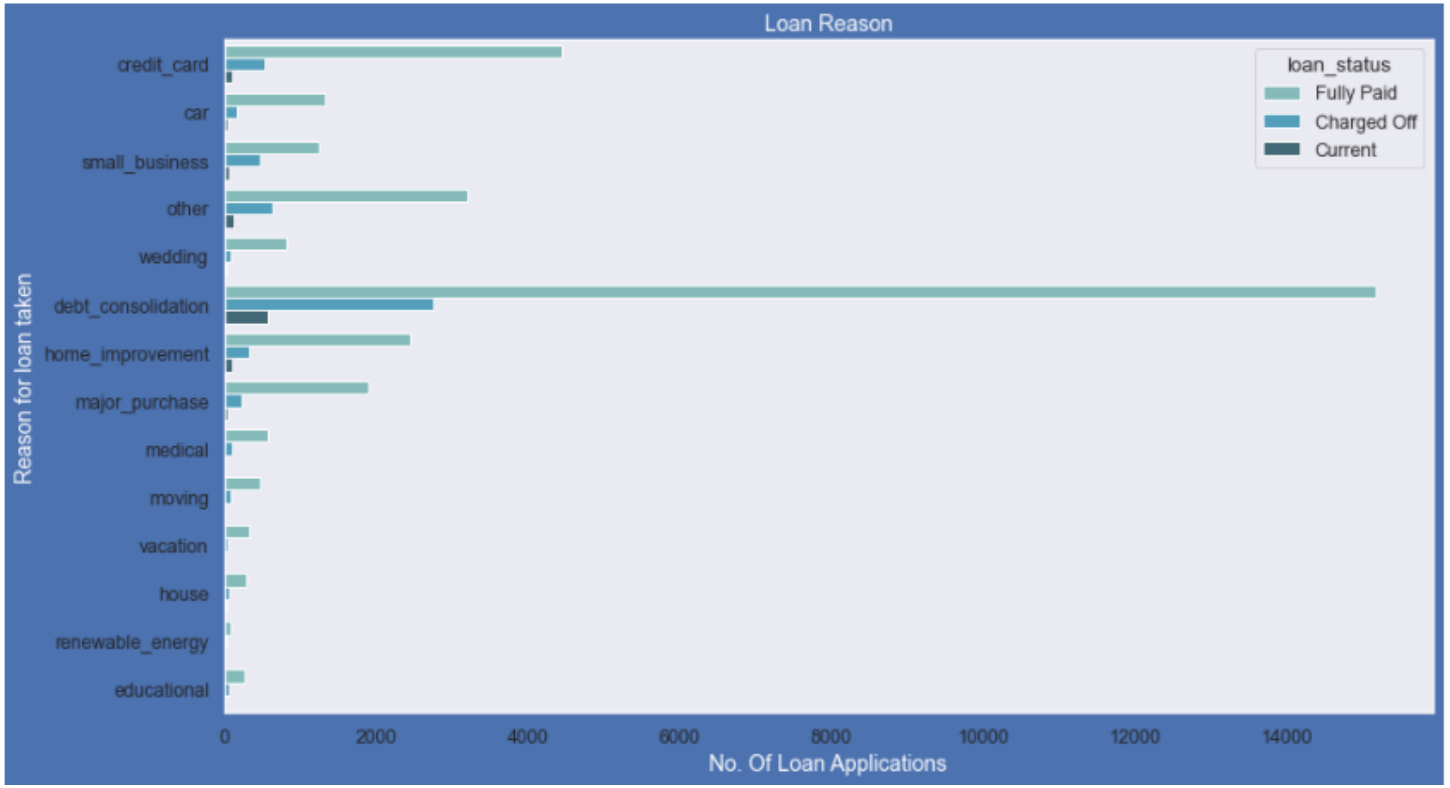
- Categorical Variables – Un Ordered [Loan Status]
- The bar plot indicates that around 14% of total loans were charged off out of total loans issued.



UNIVARIATE ANALYSIS

The box plot for loan amount shows that most of the loan amounts range from 5000 to 15000.

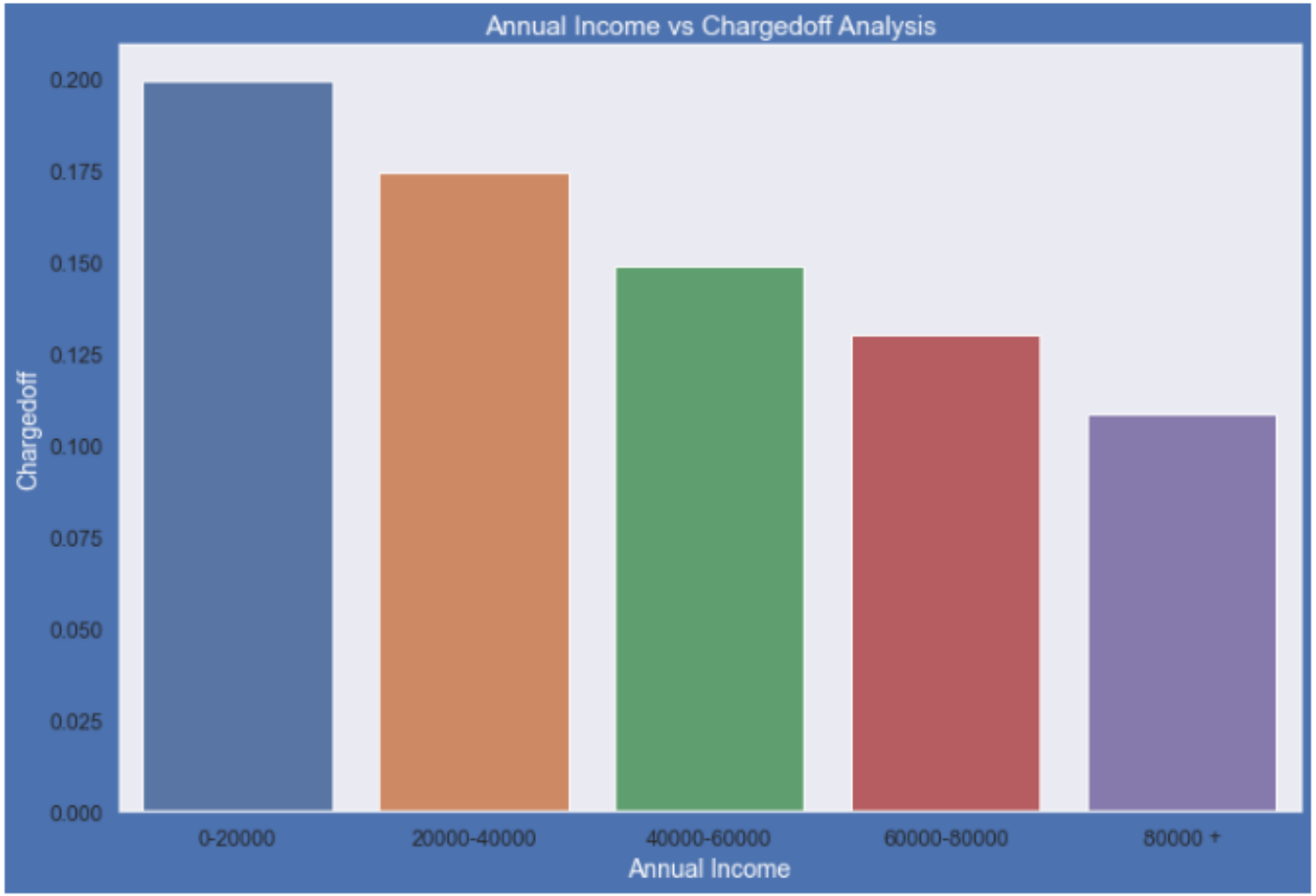
BIVARIATE ANALYSIS



- The plot indicates that the No. of loans taken for debt consolidation and credit card is high
- This also shows that in these categories the charged of percentage is more too.

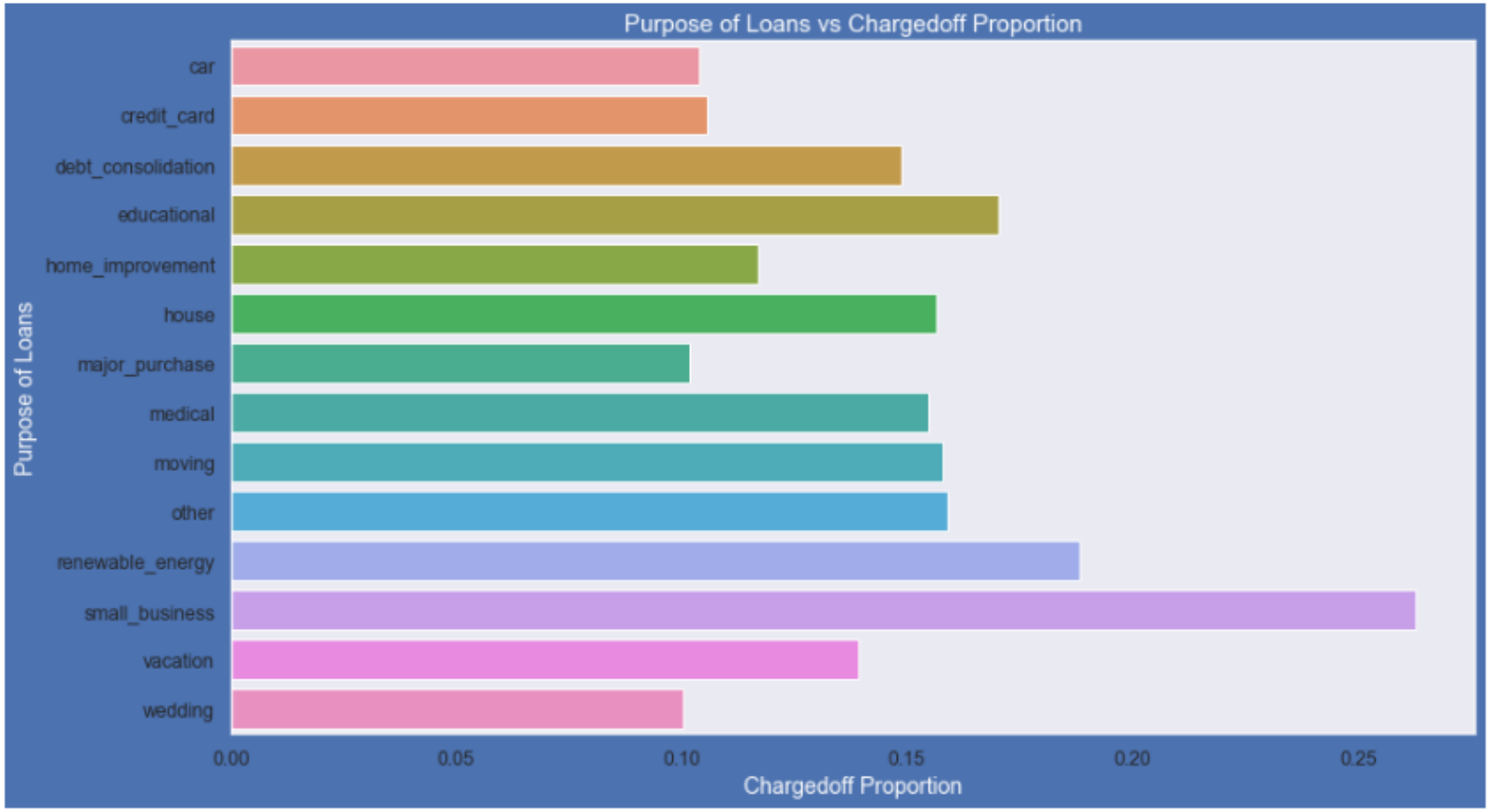
BIVARIATE ANALYSIS

- The plot shows that charged off percentage is more for the people who opted for 60-month loan tenure.



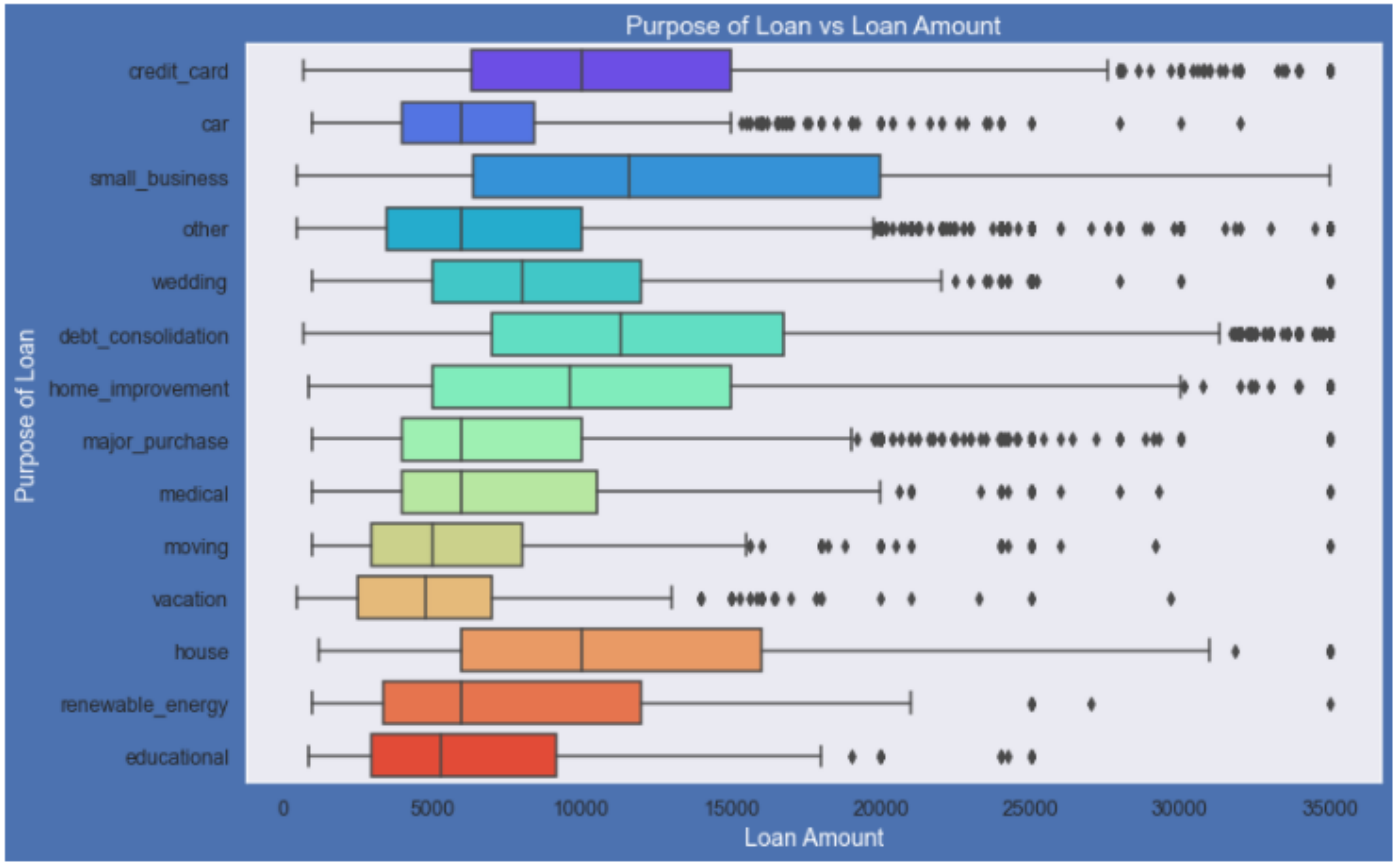
BIVARIATE ANALYSIS

- Below are the observations on loan charge off in respect to the annual income of the borrower.
  - Borrowers having income 0-20000 are the highest to rate of charge off
  - Charge off lones are minimum in case of borrowers who have income 80000+



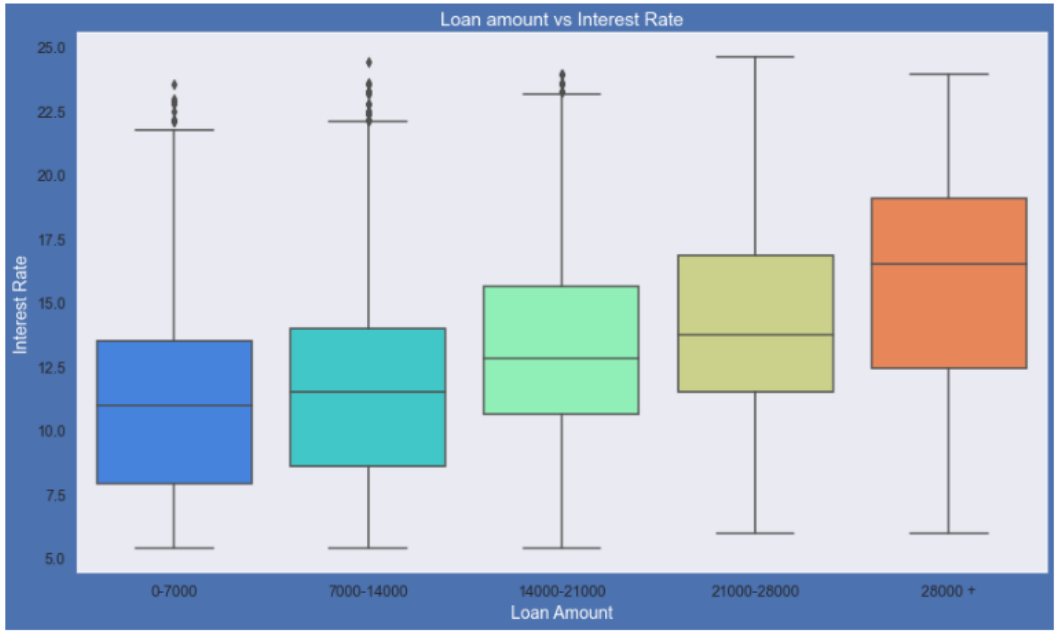
BIVARIATE ANALYSIS LOAN REASON VS CHARGED OFF PROPORTION

- Small business applicants have high chances of getting charged off whereas charged off proportion is better for renewable energy than other categories.



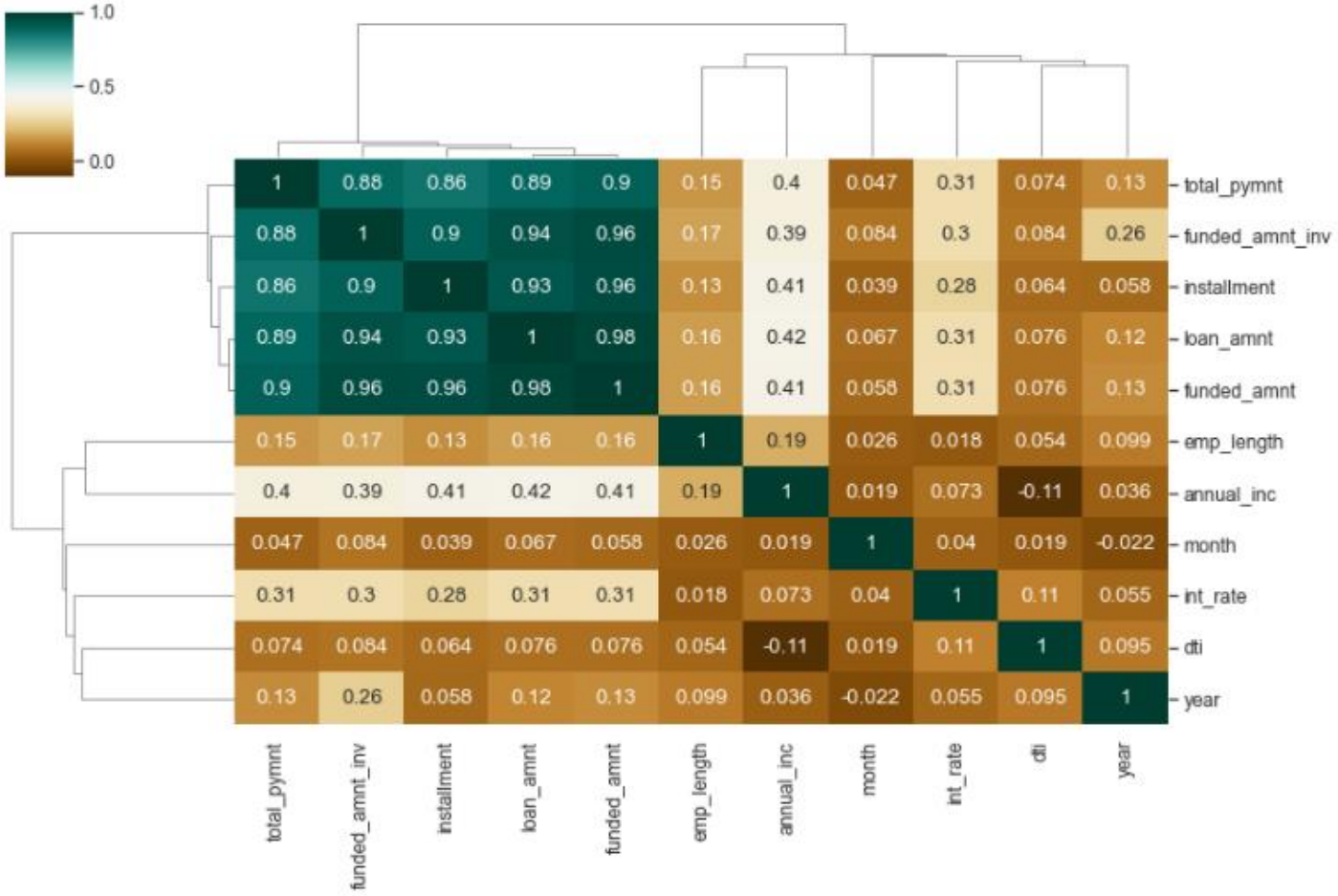
PURPOSE OF LOAN VS LOAN AMOUNT BIVARIATE ANALYSIS

- Median,95th percentile,75th percentile of loan amount is highest for loan taken for small business purpose among all purposes.
- Debt consolidation is second and Credit card comes 3rd.



LOAN AMOUNT VS INTEREST RATE BIVARIATE ANALYSIS

- This indicates that as the loan amount increases the interest rate also increases.



BIVARIATE ANALYSIS – CORRELATION MATRIX

- Funded amount and loan amount are strongly correlated.
- Employee length and annual income is also strongly correlated with each other.