Lending Club Case Study

Table of Contents

Problem Statement and Business Case Scenario

- We received loan dataset and dictionary for the columns containing the information about past loan applicants and whether they ‘defaulted’ or not. The 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.
- Perform an analysis to get some insights for the driving factors which leads to default Loans.

Exploratory Data Analysis

- Data Preview.

  We have received data which has 39717 rows and 111 columns.

Dataset Preview

In [46]:
```
df.head()
```

Out[46]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_past_12
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	Na
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	Na
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	Na
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	NaN	Na
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	NaN	Na

5 rows × 111 columns

Data cleanup: Filter all unwanted columns , which have no effect in analysis such as zip_code, address , url etc.

In [47]:
```
df.shape
```

Out[47]:
```
(39717, 111)
```

Exploratory Data Analysis

- Data Cleansing .

  We cleansed data on following strategy.

- Clean up unwanted columns, having no impact on analysis such as id, member_id, url ,etc

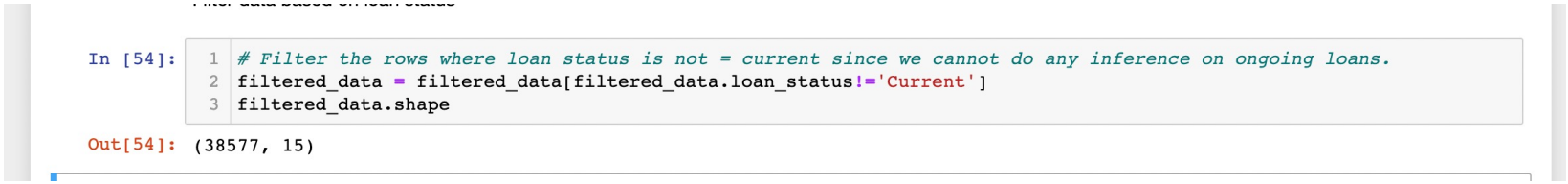
In [48]:
```
# Filter all unwanted columns , which have no effect in analysis such as zip_code, address , url etc.
unwanted_columns = ['url', 'zip_code', 'delinq_2yrs','emp_title', 'desc', 'title', 'addr_state', 'earliest_cr_line'
'mths_since_last_record', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'initial_list_status', 'ou
'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d',
'id', 'member_id']
filtered_data = df.drop(unwanted_columns, axis=1)
filtered_data.shape
```

Out[48]: (39717, 69)

Exploratory Data Analysis

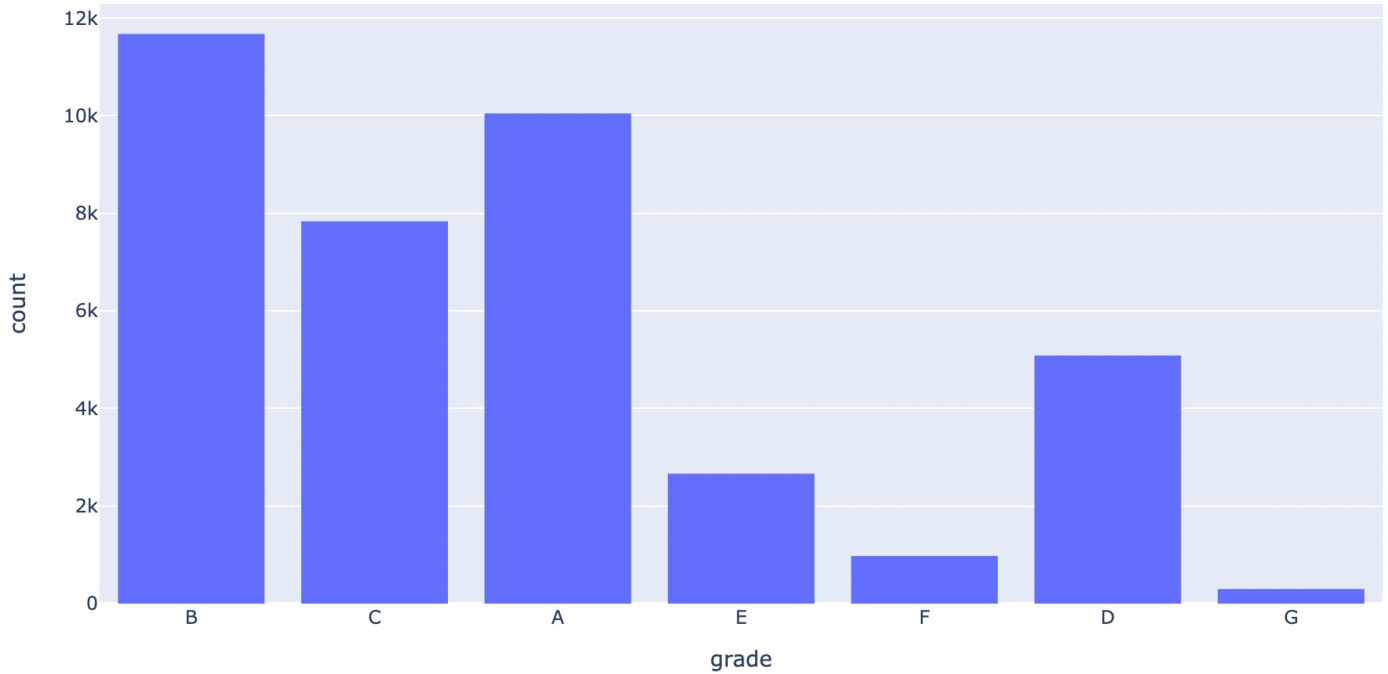
- Data Cleansing .
  - Drop all duplicates
  - Filter the column with only null values and remove it
  - Filter rows which only contain null values
  - Filter the rows where loan status is **Current** since we cannot do any inference on ongoing loans.

  Below is the final dataset used for analysis



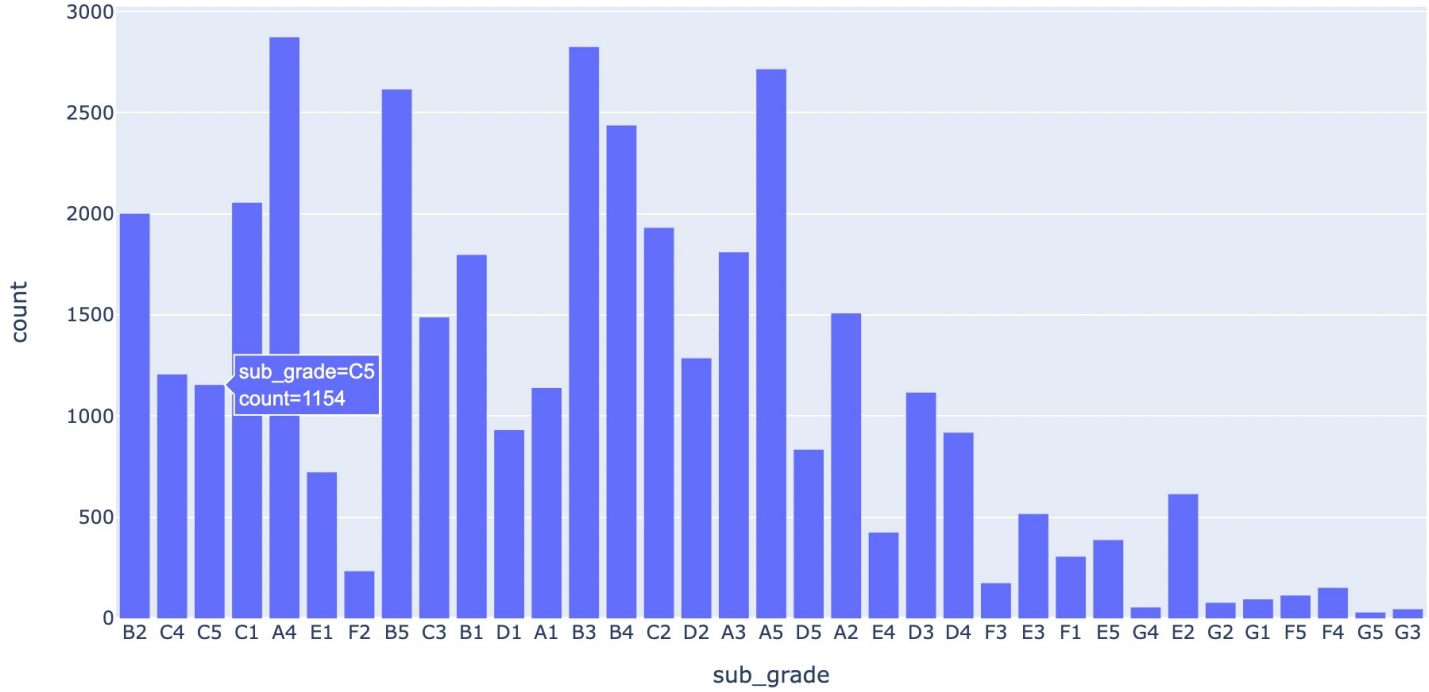
Exploratory Data Analysis

- Univariate Analysis.
    1. Majority of the loans in the dataset belong to Grade B



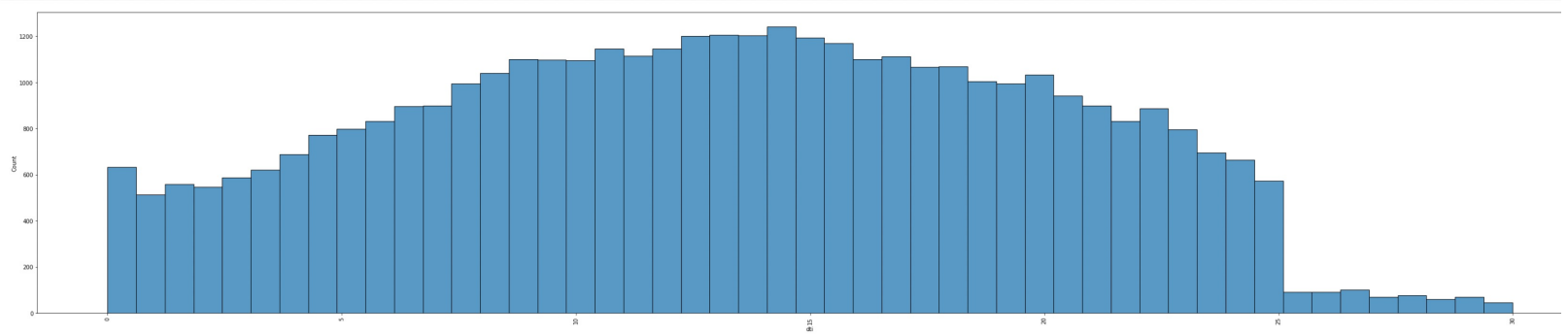
Exploratory Data Analysis

- Univariate Analysis.
  2. Majority of the loans in the dataset belong to Grade A4, Good Loan Category, highest risky loan grade G5 is the least.



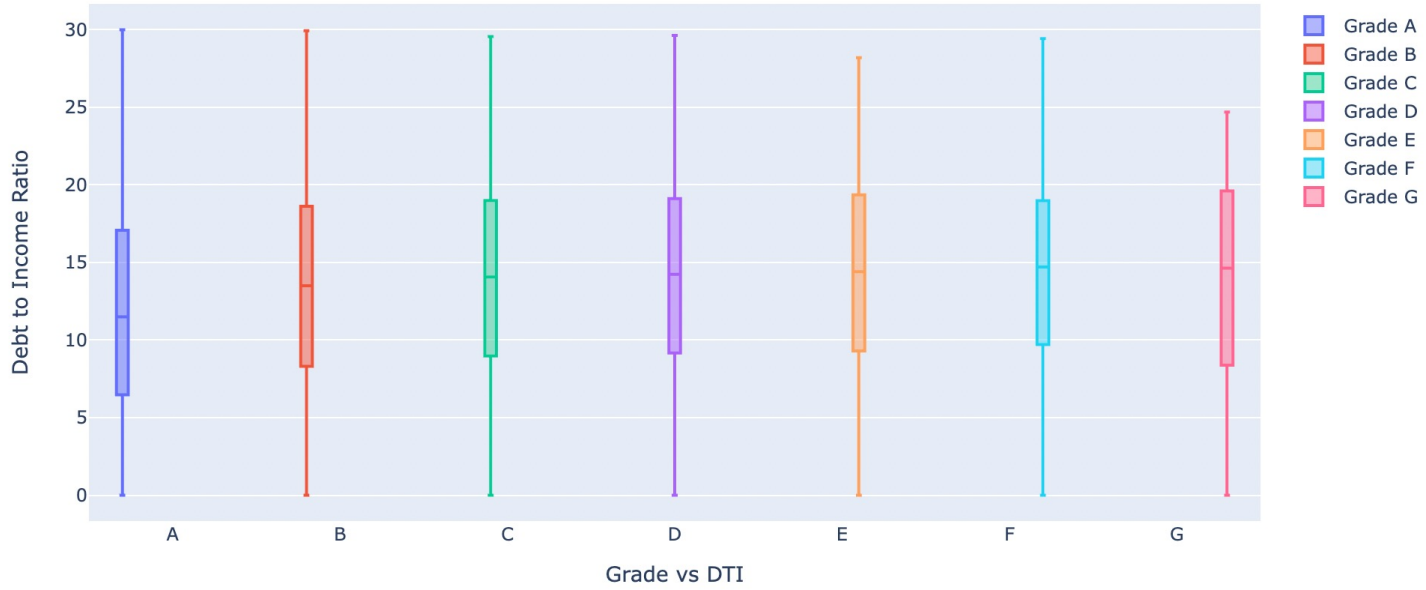
Exploratory Data Analysis

- Univariate Analysis.
  3. Majority Borrowers in the dataset spread across **dti** value between 10 to 20, which means no risky loans.



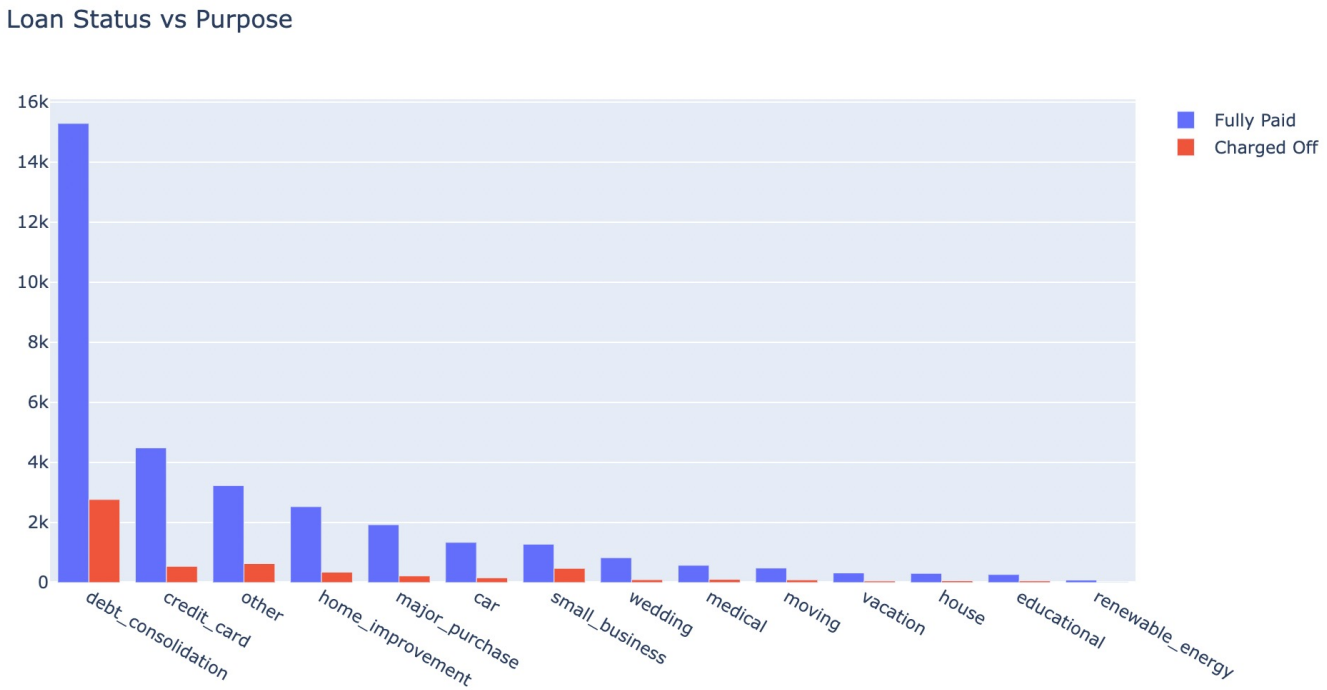
Exploratory Data Analysis

- Bivariate Analysis.
  1. We see inversely correlation between loan **grade** and **Debt to Income(dti)** Value.We see the median is increasing with risky loan grade. Lending to members with high **dti** will be slightly riskier.



Exploratory Data Analysis

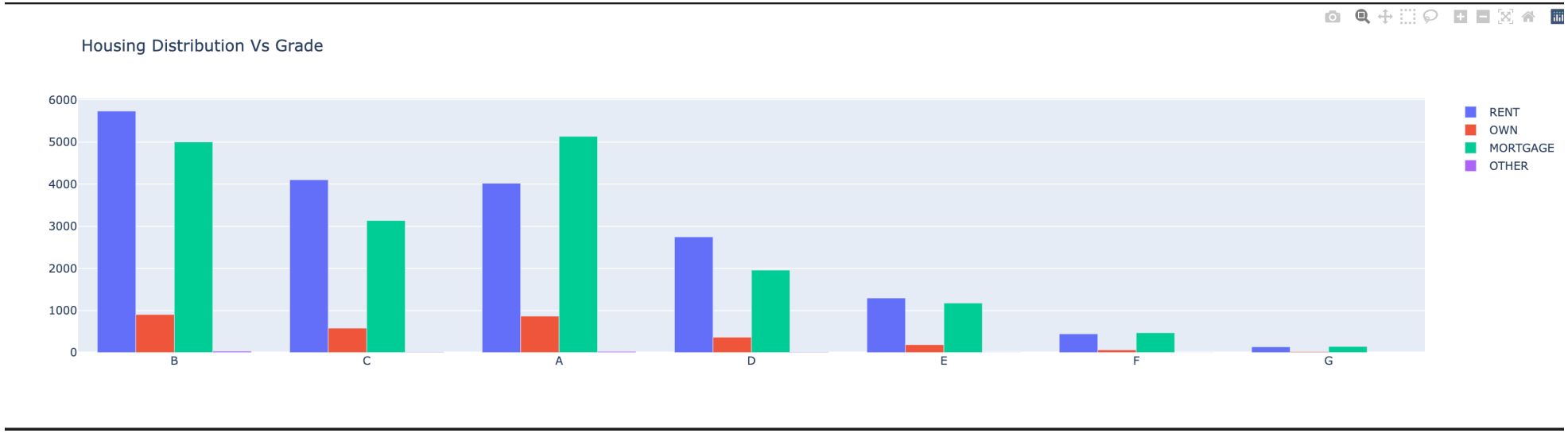
- Bivariate Analysis.
  2. Loans provided for the purpose of car, major_purchase,wedding have high percentage of fully paid to charged off ratio and are risky.



Exploratory Data Analysis

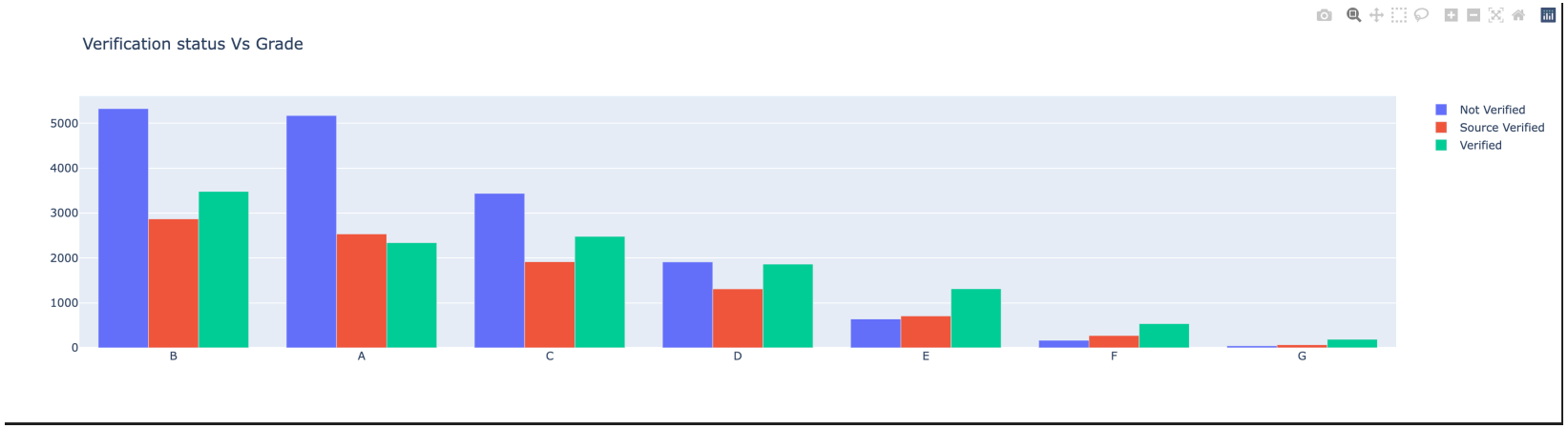
- Bivariate Analysis.

  3. Number of people having mortgage have good grade loans when compared to rented and own house. Hence we can conclude that lending to people having **Mortgage** is likely to have least risk .



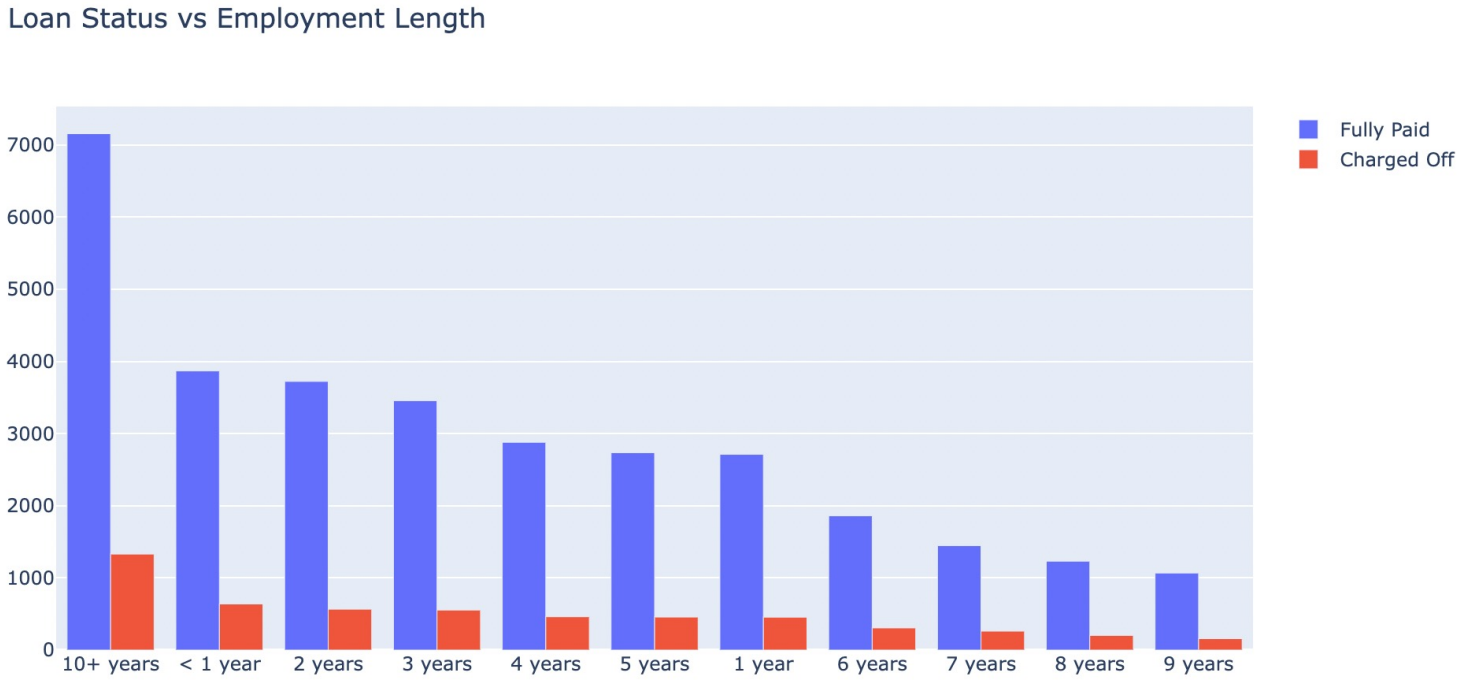
Exploratory Data Analysis

- Bivariate Analysis.
    4. Higher risk in lending loan when income is not **Source verfied**.



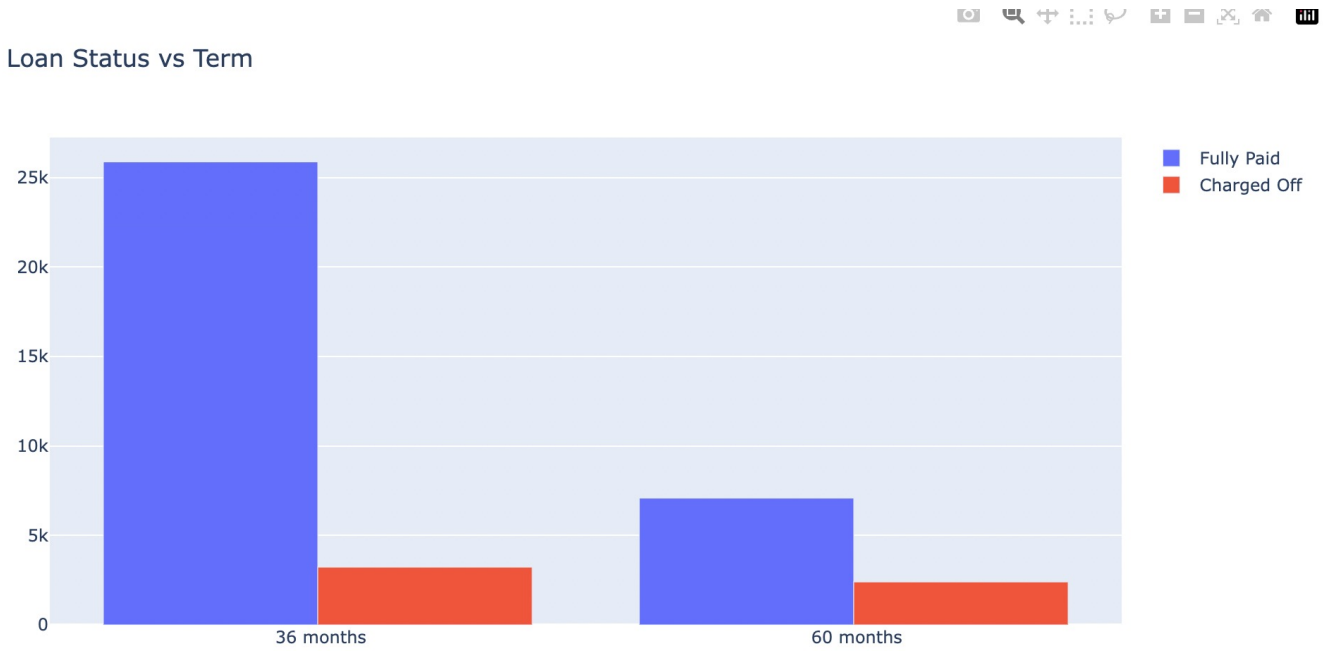
Exploratory Data Analysis

- Bivariate Analysis.
  5. We see the ratio between Fully Paid and Charged Off is least in 10+years, compared to others. Hence it is risky to loan to 10 +years.



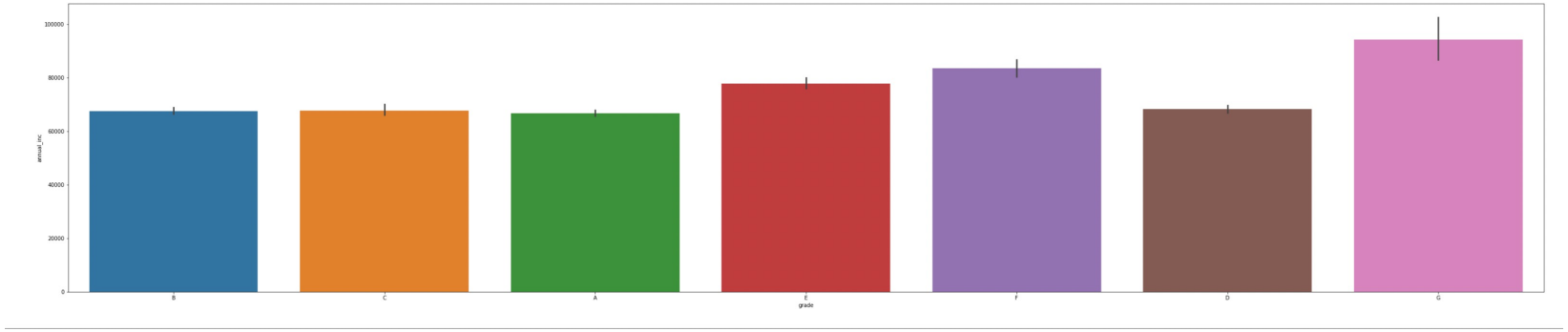
Exploratory Data Analysis

- Bivariate Analysis.
  6. Loan lent for 36 months has fully paid to charged off ratio is high and hence less risky compared to 60 months. Hence Loan lent for higher term is risky.



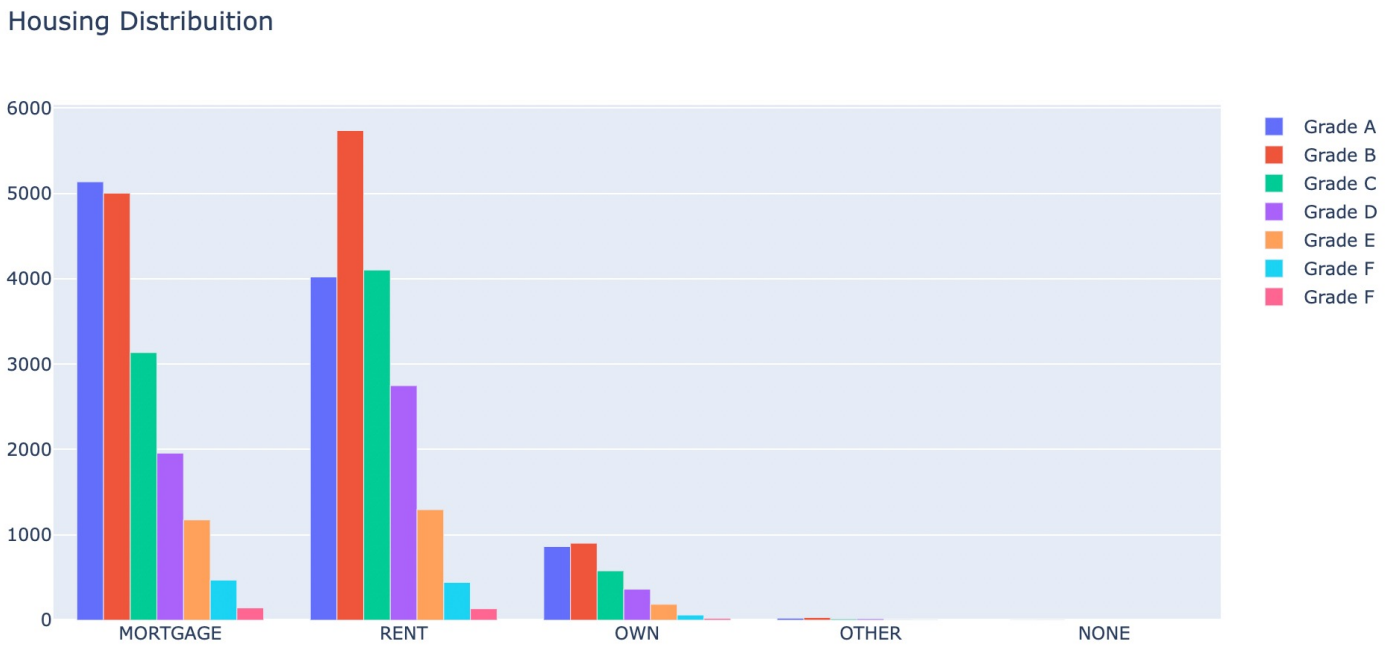
Exploratory Data Analysis

- Bivariate Analysis.
  7. People having income below 70000 belong to less risk category(A,B)



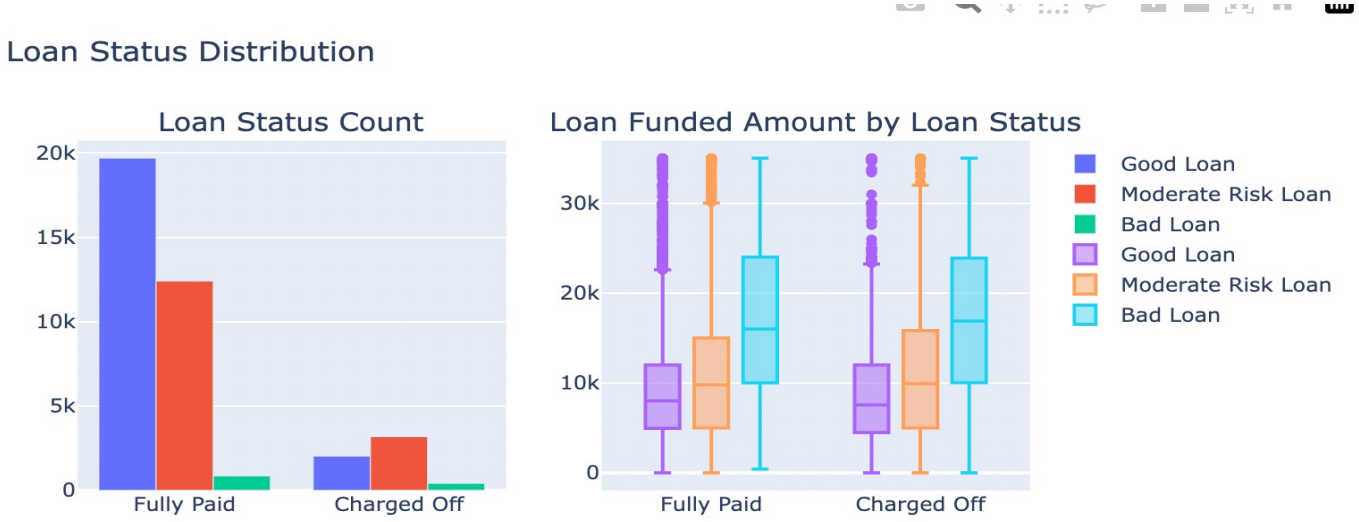
Exploratory Data Analysis

- Bivariate Analysis.
    9. Loan lent to people having Mortgage and own house are less risky.



Exploratory Data Analysis

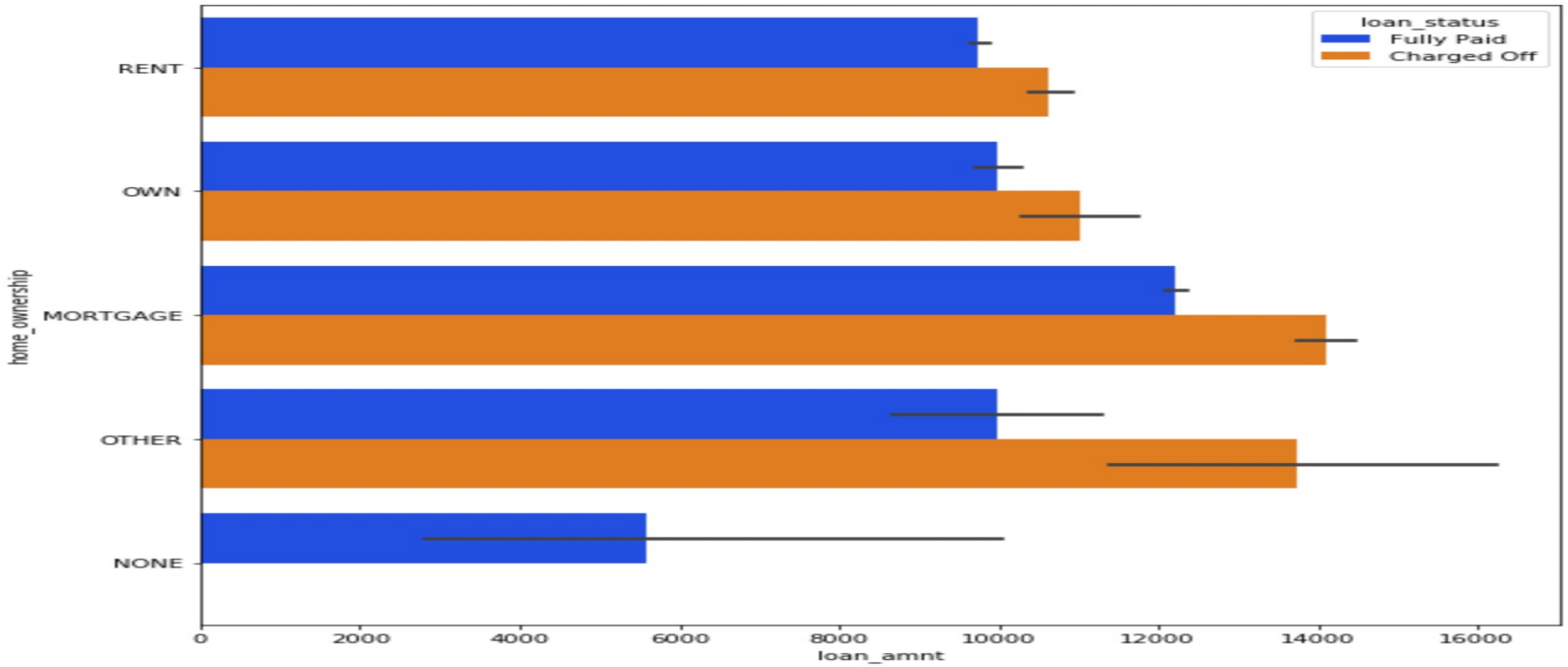
- Multivariate Analysis.
  1. In Loan Status Count Good Loans are fully Paid.
  2. From Loan Funded Amount by Loan Status we conclude that the Bad Loans have higher median value for loan amount and loan risk increases as funded amount increases



Exploratory Data Analysis

- Multivariate Analysis.

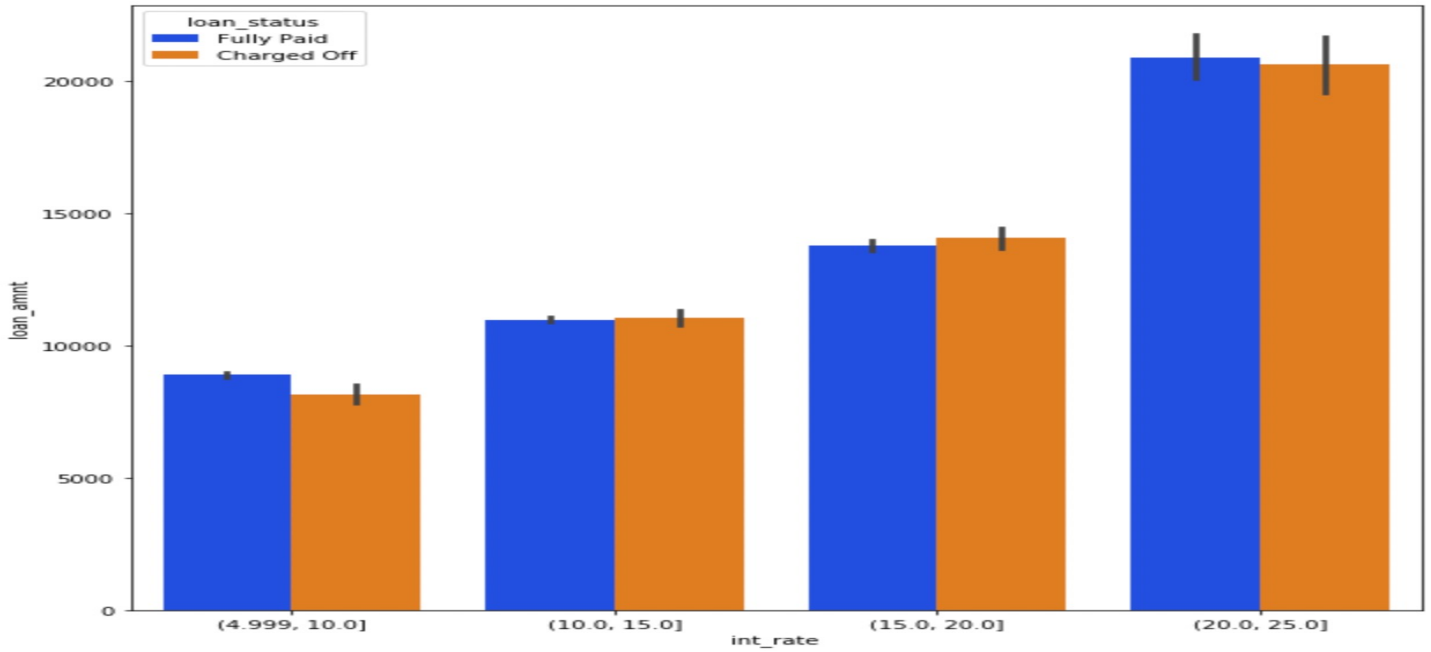
  2.Charged_off loans are higher when loan amount is higher and lent to **Other** categories



Exploratory Data Analysis

- Multivariate Analysis.

  3. Loans are defaulted when the have high interest rates and loan amount is higher, Charged off loans are high when loan amount is 10-15k at the interest rate of 15-20%



Observations and Conclusions

There is a more probability of defaulting when :

- Applicants taking loan for other purposes , car, major_purchase,wedding.
- Applicants having income of above 70k are risky.
- Applicants who take loan amount in the range 10-15k at the interest rate of 15-20%.
- When employment length is 10yrs charged off percentage is higher
- When the income source is not verified.
- When loan amount is higher and lent to OTHER categories.

Thank You

- Shruthi Annappa
- Sivakanth Jayaram