LENDING CLUB CASE STUDY

Presented By-
Kritik Mohanty
Srishti Garg

PROBLEM STATEMENT:

- We are provided with the data of a consumer finance company which specializes 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 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.
- The company can take the decision to accept the loan proposal or reject the loan proposal. In the event the company accepts the loan proposal the loan can progress into three states: Fully Paid, Current and Charged Off.
- The company incurs loss if the loan is approved but the borrowers default while paying it back. The company also incurs a loss if they don't approve the loan and thus losing the opportunity to gain capital.
- ***Objective:*** Use Exploratory Data Analysis (EDA) to understand how customer attributes and loan attributes influence the tendency to default.

Exploratory Data Analysis:

- Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data. EDA is an important first step in any data analysis.
- In this dataset we will perform EDA.
- EDA consists of 4-5 parts or steps namely:
  - Understanding Data
  - Data Cleaning
  - Data Shaping and deriving.
- Data Analysis: Univariate Analysis ,Segmented Univariate Analysis , Correlation Analysis and Bivariate Analysis

Understanding Data:

This is the first step in the analysis. In this step we observe the skeletal structure or the shape of the dataset by obtaining the rows count and columns count of the dataset as well as we take a look at the number of null values , duplicate values and unique values in the dataset.

```
[6]: ## import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
[2]: #reading the data
loan_data = pd.read_csv("loan.csv")
```

```
[3]: loan_data.head()
```

[3]:

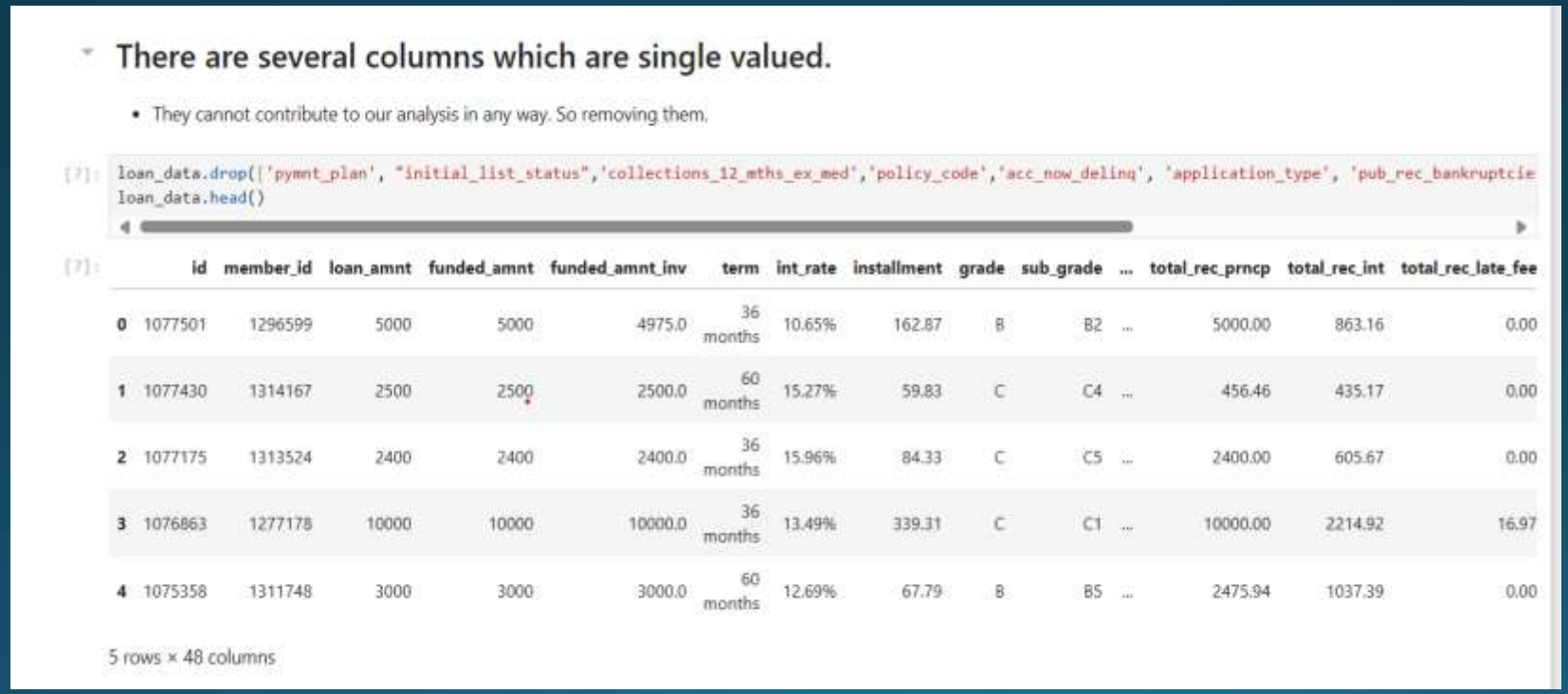
	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_past_12m	pc
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	NaN	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	NaN	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	NaN	
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	NaN	NaN	
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	NaN	NaN	

5 rows × 111 columns

Data Cleaning:

Data Cleaning is the process of getting rid of impurities and redundancies present in the data , which might cause errors in analysis process in later stages. In data cleaning we get rid of any null values , duplicate values and any variables which might not contribute to the data analysis.

In the picture below we can see the removal of the unnecessary columns from the dataset:



There are several columns which are single valued.

- They cannot contribute to our analysis in any way. So removing them.

```
[7]: loan_data.drop(['pymnt_plan', "initial_list_status",'collections_12_mths_ex_med','policy_code','acc_now_delinq', 'application_type', 'pub_rec_bankruptcie
     loan_data.head()
```

[7]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	total_rec_prncp	total_rec_int	total_rec_late_fee
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	5000.00	863.16	0.00
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	456.46	435.17	0.00
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	2400.00	605.67	0.00
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	10000.00	2214.92	16.97
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	2475.94	1037.39	0.00

5 rows × 48 columns

Standardizing the data

Standardizing the data is a very crucial part of Data Analytics . We have standardize the data like below

Standardizing the data

- "revol_util" column although described as an object column, it has continous values.
- So we need to standardize the data in this column
- "int_rate" is one such column.
- "emp_length" --> [(< 1 year) is assumed as 0 and 10+ years is assumed as 10]
- Although the datatype of "term" is arguable to be an integer, there are only two values in the whole column and it might as well be declared a categorical variable.

[17]:
```
loan_data.revol_util = pd.to_numeric(loan_data.revol_util.apply(lambda x : x.split('%')[0]))
```

[18]:
```
loan_data.int_rate = pd.to_numeric(loan_data.int_rate.apply(lambda x : x.split('%')[0]))
```

[19]:
```
loan_data.emp_length = pd.to_numeric(loan_data.emp_length.apply(lambda x: 0 if "<" in x else (x.split('+')[0] if "+" in x else x.split()[0])))
```

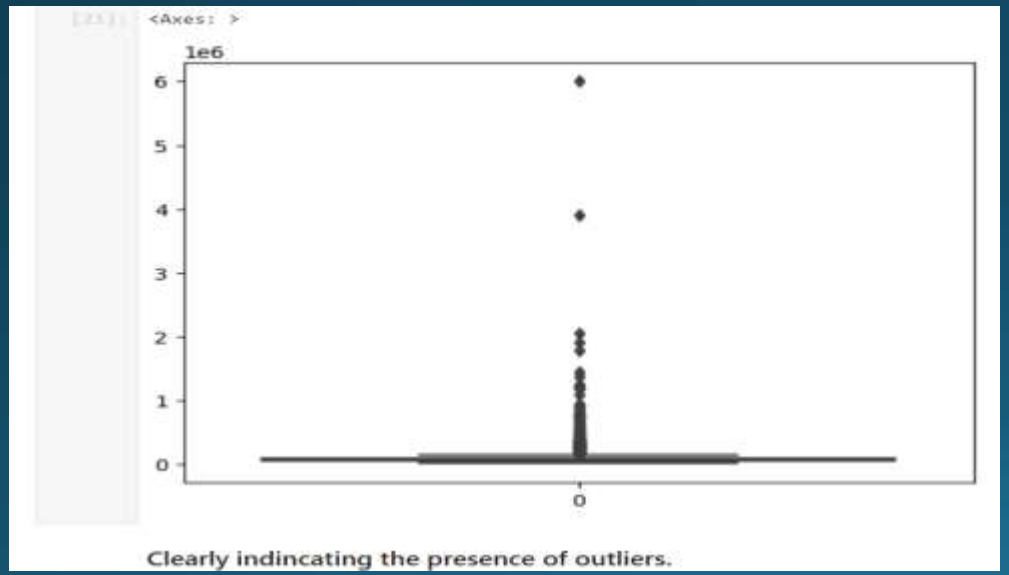
[20]:
```
loan_data.head()
```

[20]:

	loan_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_length	home_ownership	annual_inc	...	issue_d	loan_status	purpose	d
0	5000	4975.0	36 months	10.65	162.87	B	B2	10	RENT	24000.0	...	Dec-11	Fully Paid	credit_card	27.6
1	2500	2500.0	60 months	15.27	59.83	C	C4	0	RENT	30000.0	...	Dec-11	Charged Off	car	1.0
2	2400	2400.0	36 months	15.96	84.33	C	C5	10	RENT	12252.0	...	Dec-11	Fully Paid	small_business	8.7
3	10000	10000.0	36 months	13.49	339.31	C	C1	10	RENT	49200.0	...	Dec-11	Fully Paid	other	20.0
5	5000	5000.0	36 months	7.90	156.46	A	A4	3	RENT	36000.0	...	Dec-11	Fully Paid	wedding	11.2

Data Shaping(Removing Outliers):

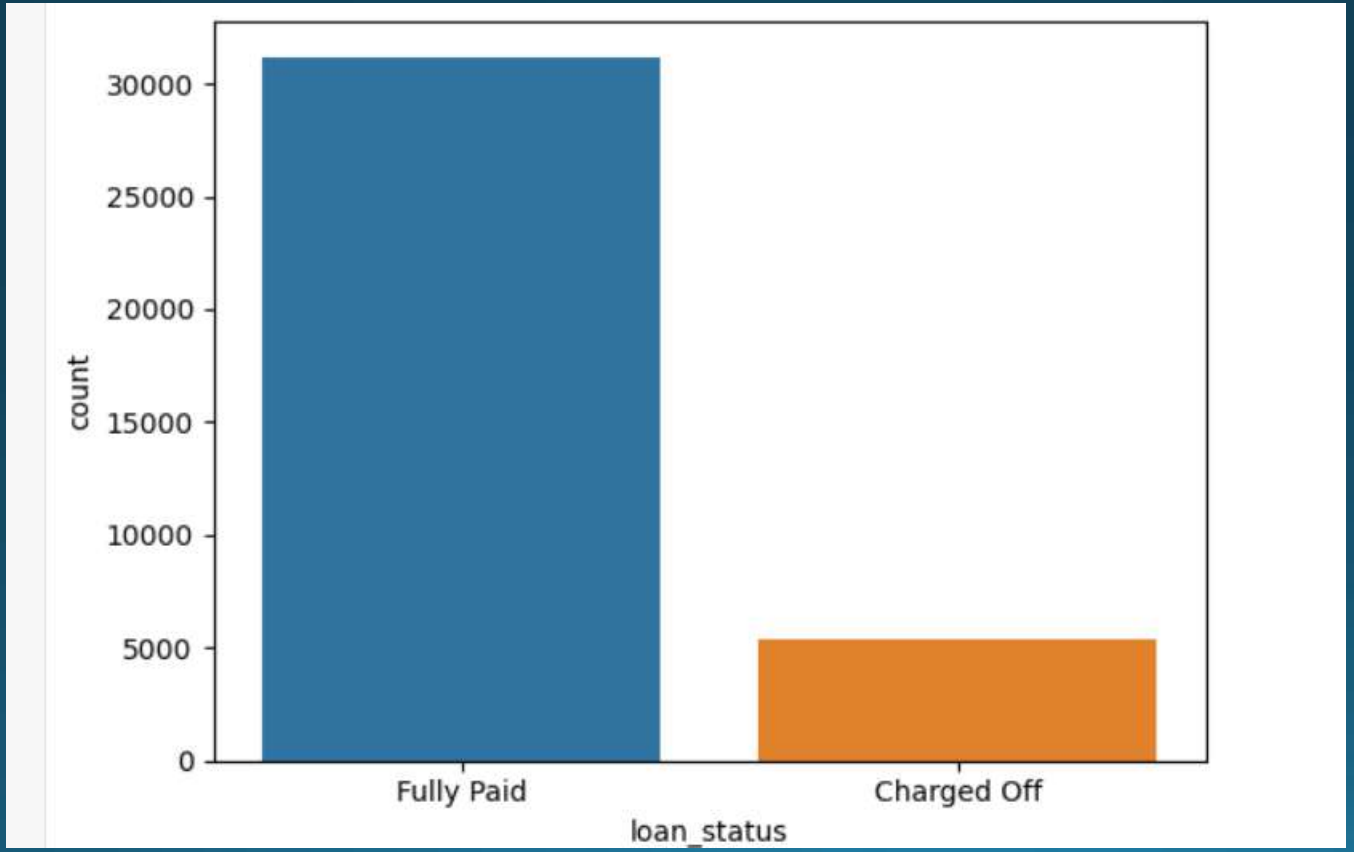
Before heading into analysis , its also imperative that we remove all the outliers present in the dataset(if any). Thus we charted the boxplots of all the numerical columns to check for outliers as we can see below. From the boxplots we inferred that there were outliers present for the columns :



Clearly indincating the presence of outliers.

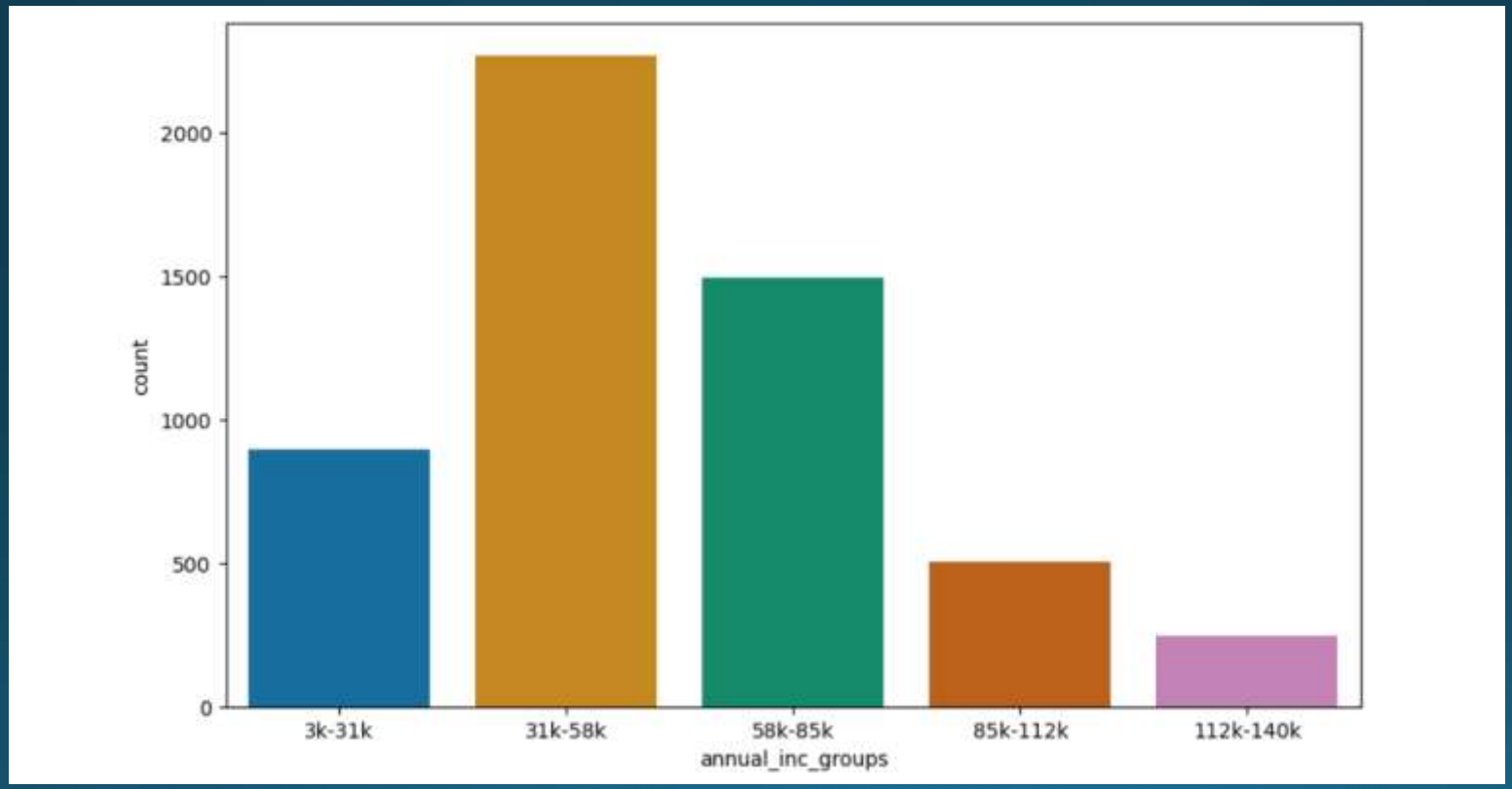
Univariate Analysis:

Analysis of Loan Status: The analysis of loan status showed that the number of people who have defaulted is way less compared to the people who have fully paid their loan back as we can see below:



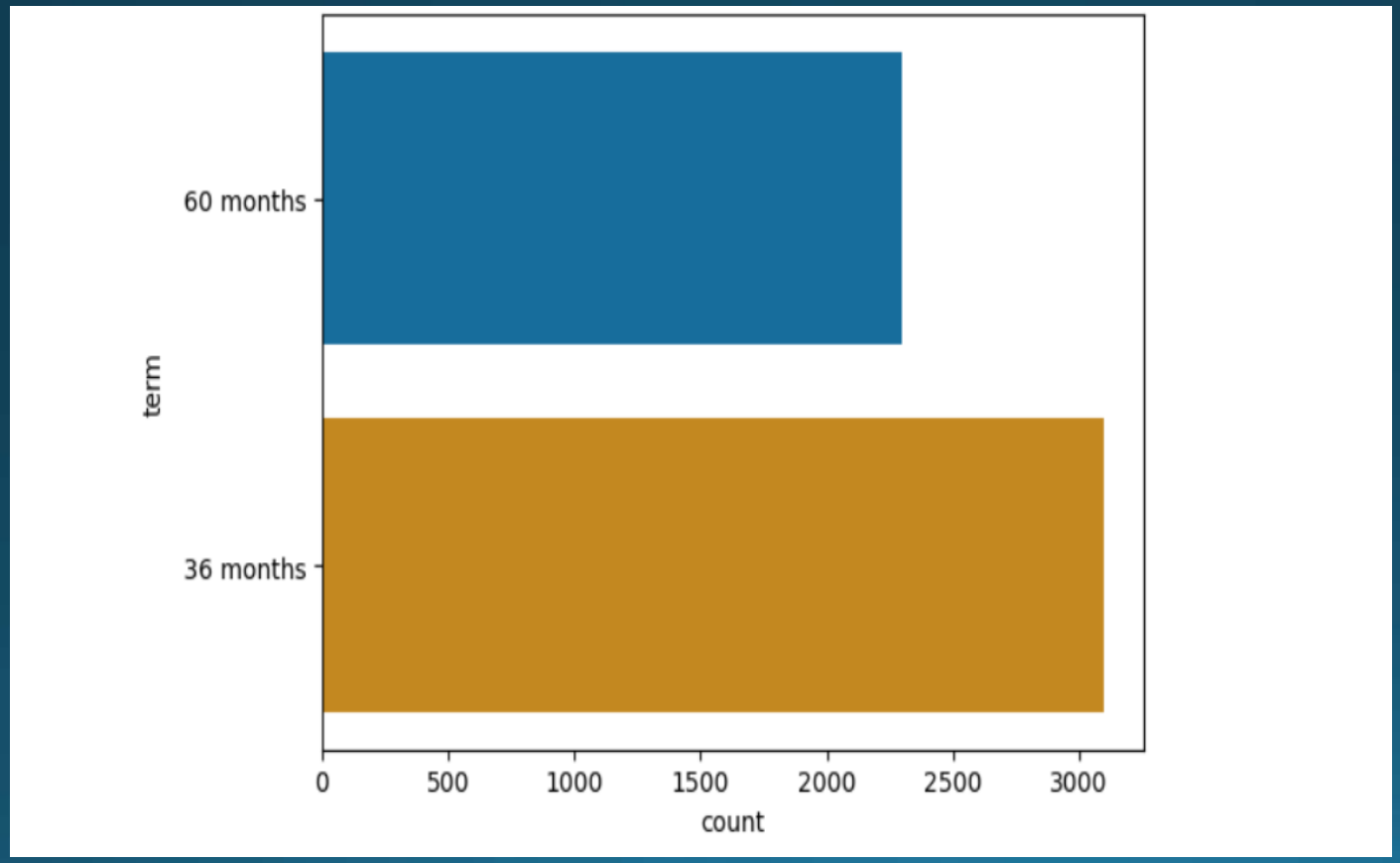
Analysis of Annual Income:

The analysis of annual income showed that most of the borrowers have an annual income within the range of 31000-58000 .This suggests that most people who have high annual income have a low tendency to borrow money as shown below:



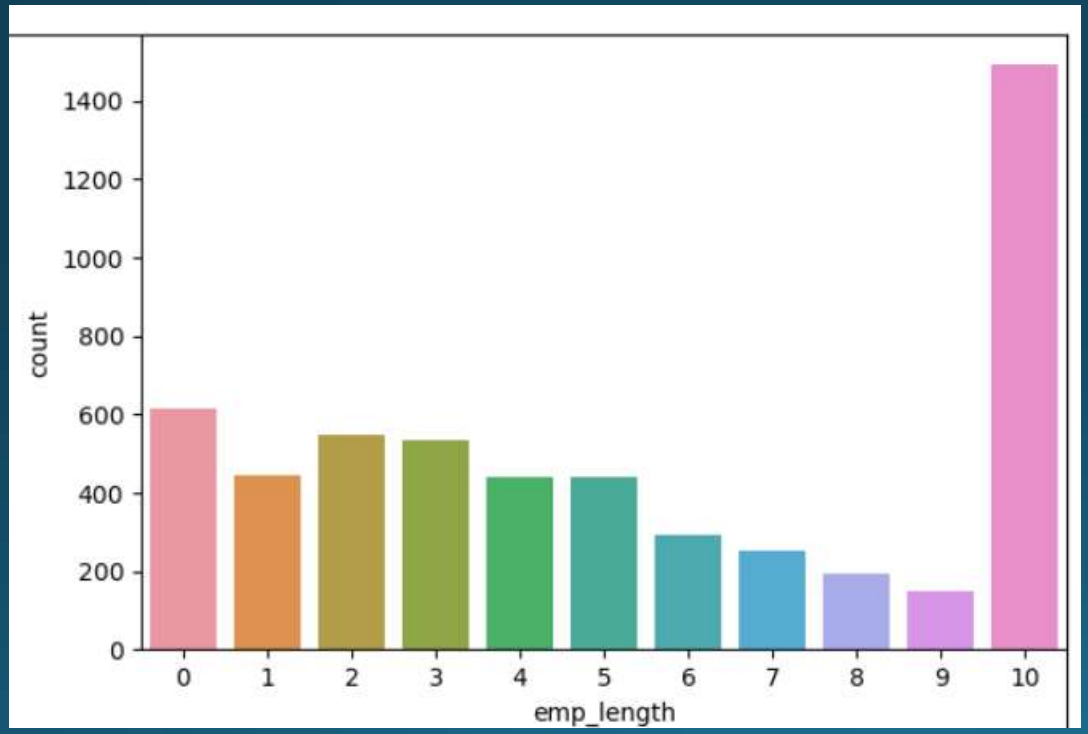
Analysis of loan term:

The analysis of loan term showed that majority of loans have a 36 month term. This suggests that most loans which were borrowed, the borrowers opted for the 36 month plan as shown below:



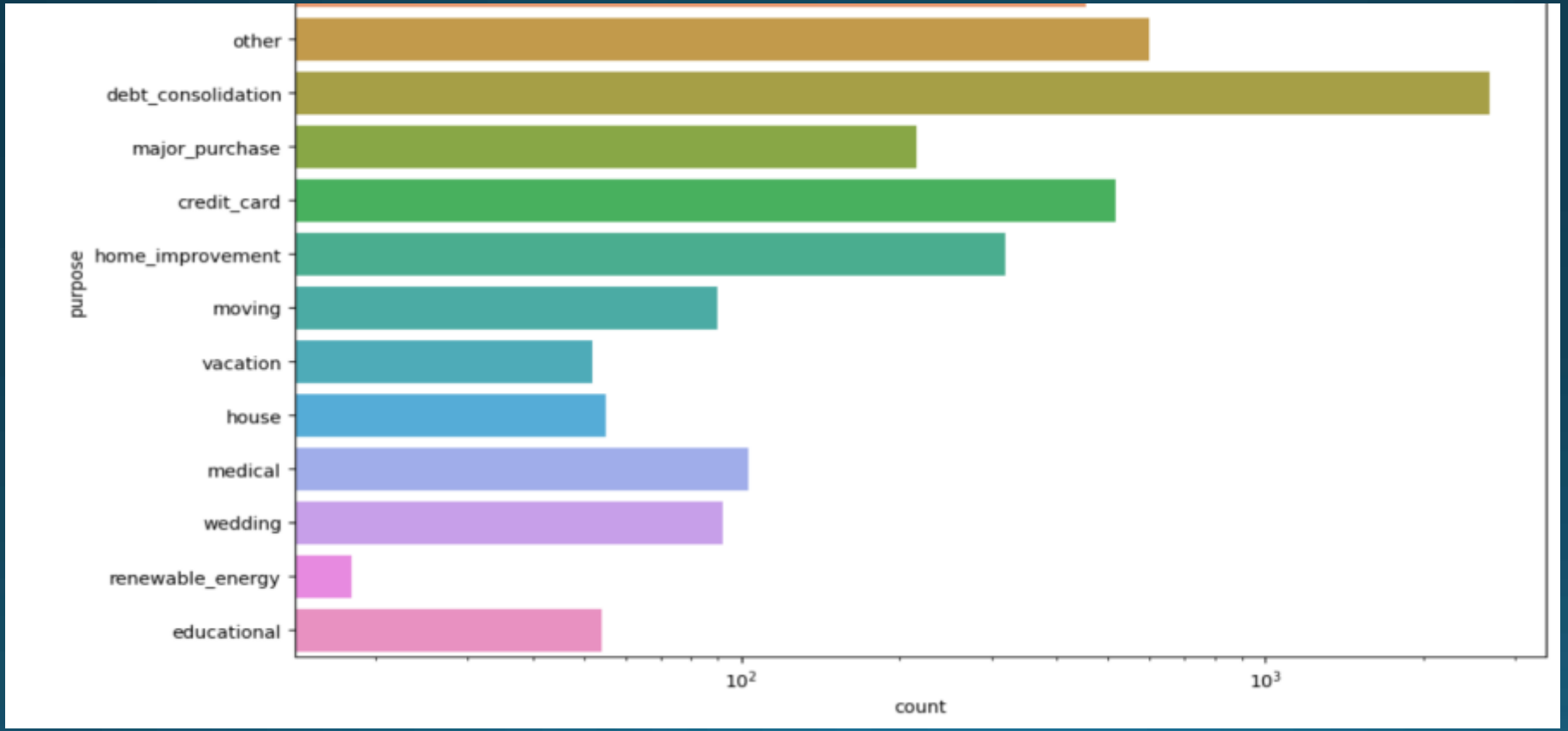
Analysis of Employment Length:

The analysis of employment length showed that majority of the borrowers have a work experience of over 10 years as shown below:



Analysis of Purpose of Loan:

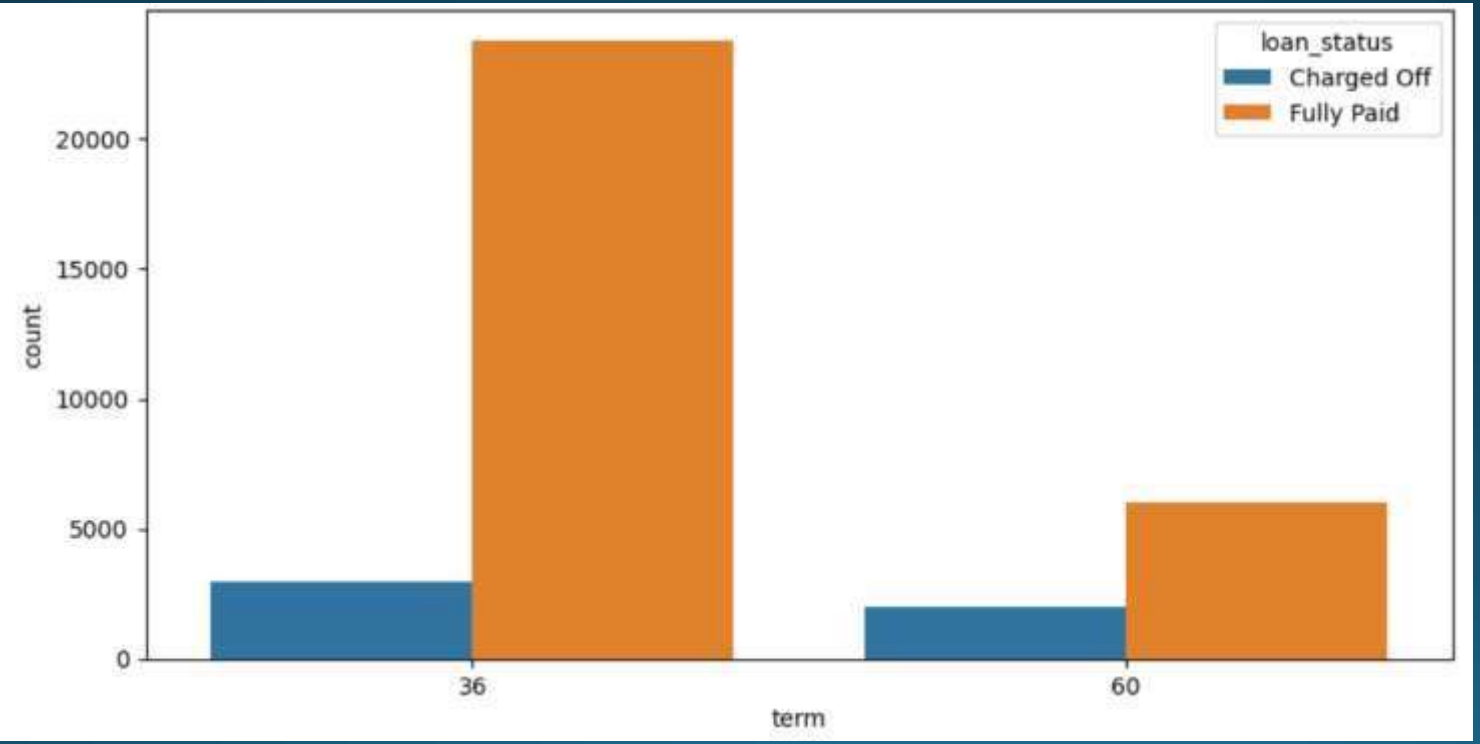
The analysis of purpose showed that the majority of the loans are being borrowed for the purpose of debt consolidation followed by credit card debt as shown below:



Analysis of Loan Term based on Loan Status:

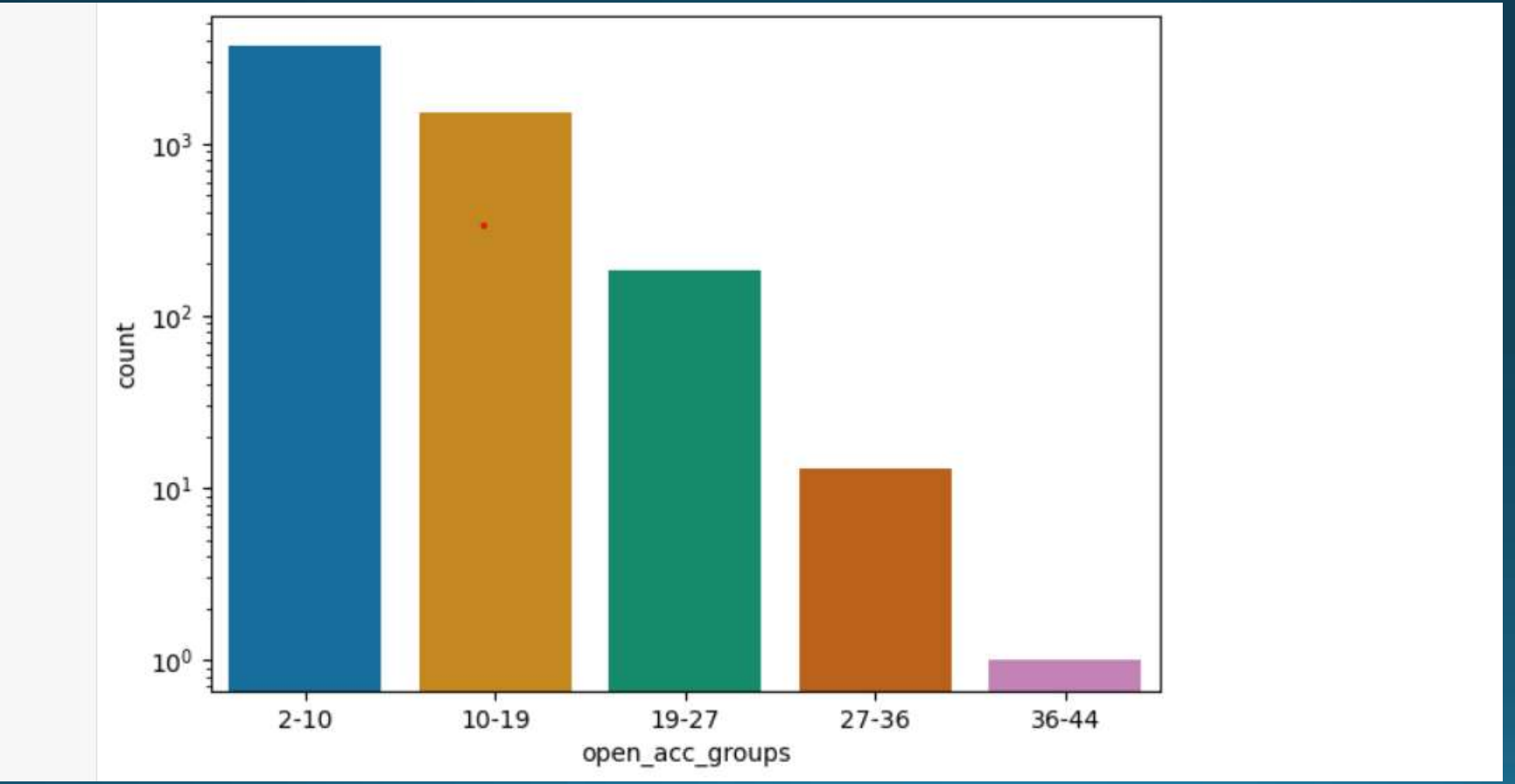
The analysis of loan term based on loan status showed that for loans of term period 36 months, the likelihood of getting paid in full is higher than the loans of term period 60 months.

This suggests that the loans which have a term period of 60 months have a high chance of getting defaulted as shown below:



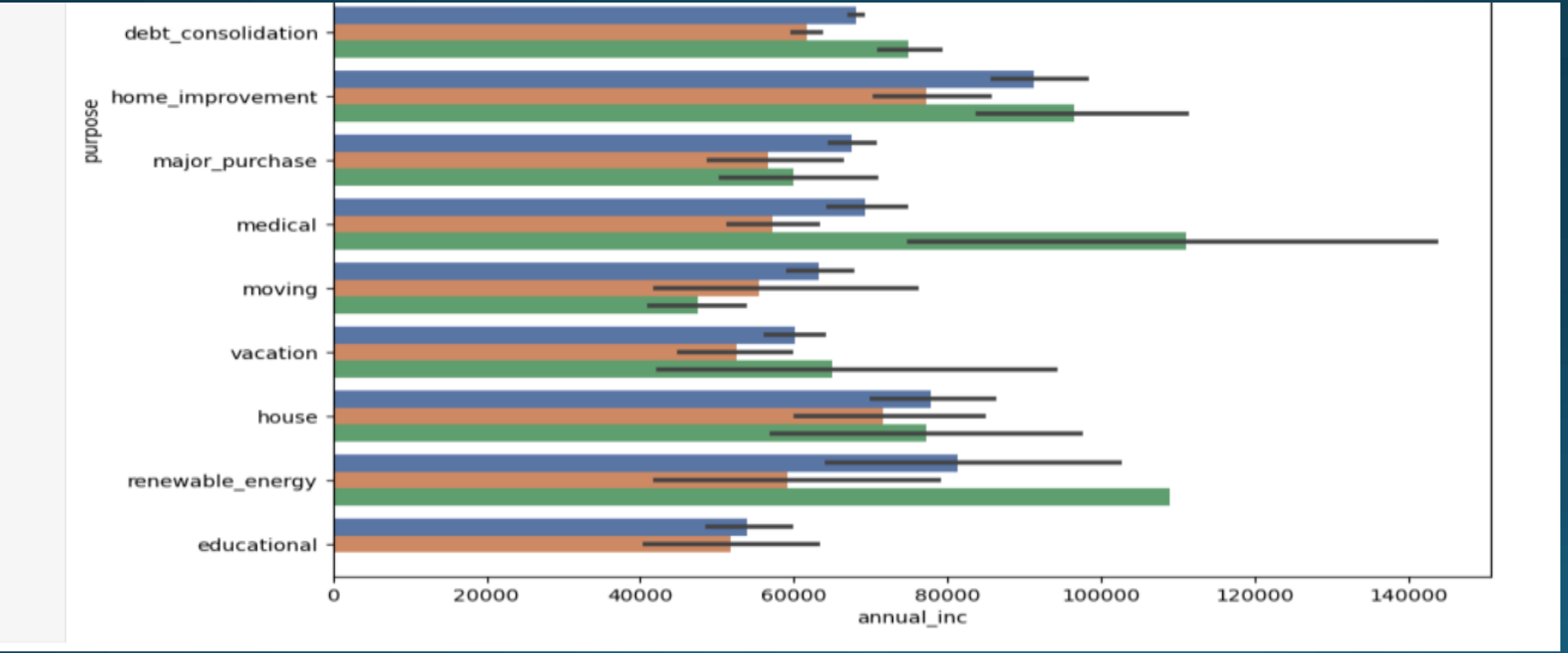
Analysis of Open account Groups based on Loan Status:

This suggests that the borrowers with open account groups within 2-10 have a higher chance of defaulting and paying in
full as shown below:



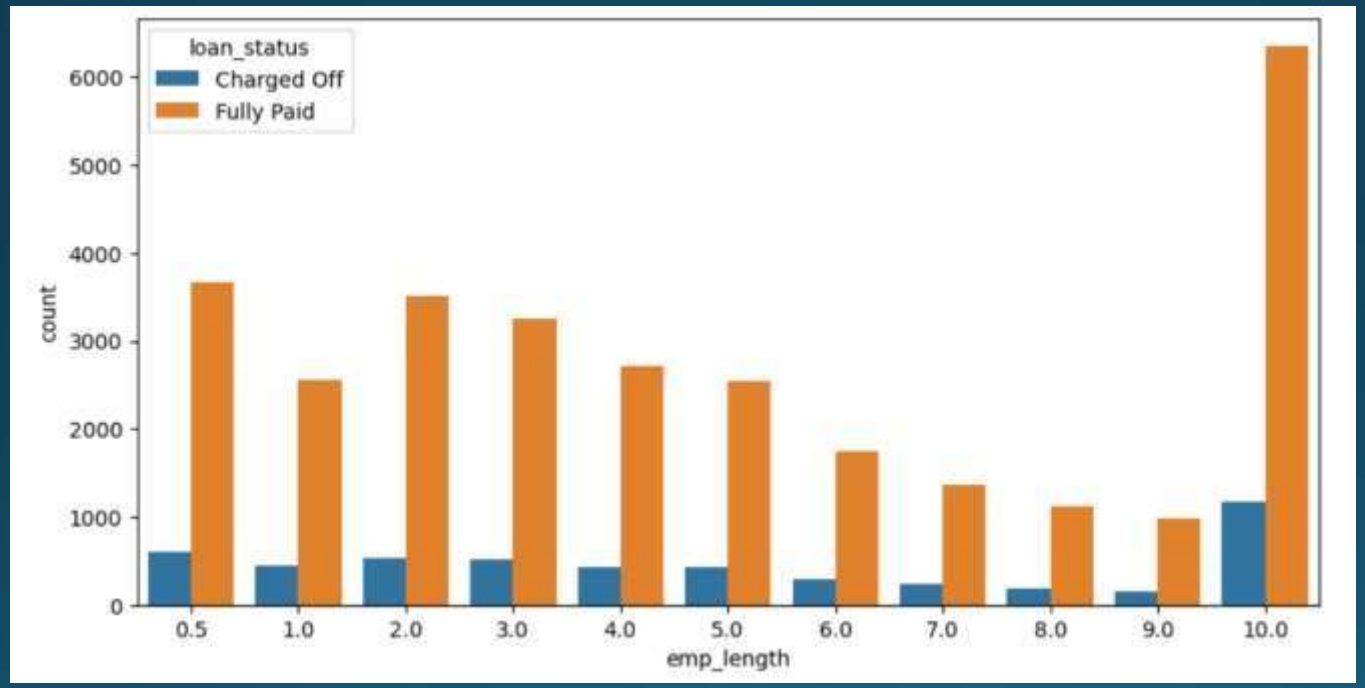
Analysis of Annual Income based on Purpose:

The analysis of annual income based on Purpose showed that the borrowers who earn 140000 annually have a high purpose of mediccal



Analysis of Employment Length based on Loan Status:

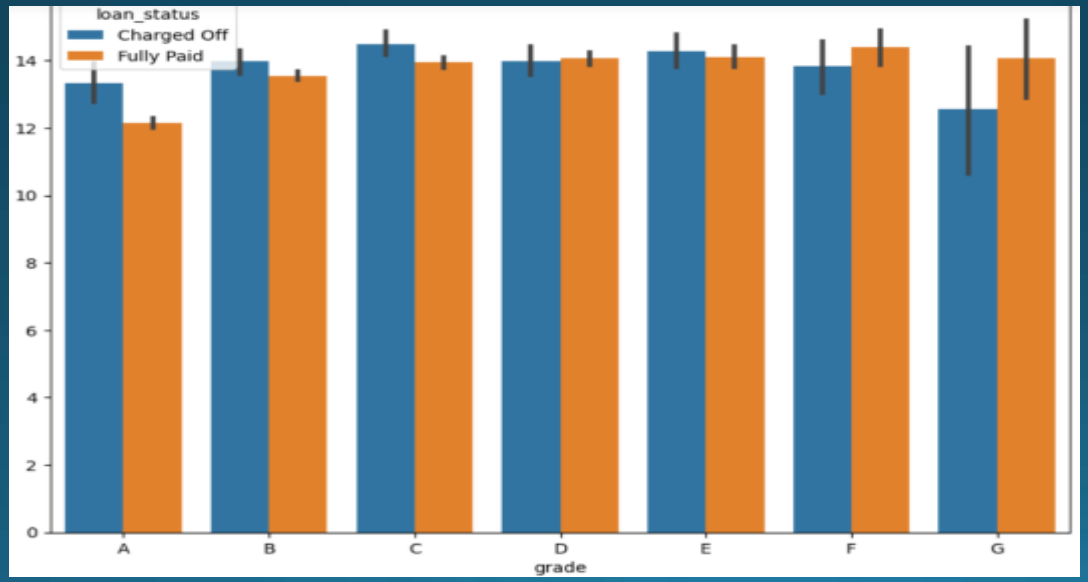
The analysis of employment length based on loan status showed that the borrowers who have more than 10 years of working experience have a high likelihood to fully pay as well as default on their loans as shown below:



Bivariate Analysis:

Analysis of DTI over Grade for Loan Status: The analysis of DTI over grade in terms of loan status showed that for loans of grade A-C the higher the DTI the higher the chance of the loan being defaulted, where as for higher grade loans the value of DTI and fully paid and charged off loans hold a similar ratio and even at times the fully paid loans for higher grades have even DTI.

This suggests that for low grade loans with borrowers having high DTI in comparison are more likely to default on their loans ,as shown below:



Observation:

- Applicants taking loan for 'home improvement' and have income of 60k -70k
- Applicants whose home ownership is 'MORTGAGE and have income of 60-70k
- Applicants who receive interest at the rate of 21-24% and have an income of 70k-80k
- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %
- Applicants who have taken a loan for small business and the loan amount is greater than 14k
- Applicants whose home ownership is 'MORTGAGE and have loan of 14-16k
- When grade is F and loan amount is between 15k-20k
- When employment length is 10yrs and loan amount is 12k-14k
- When the loan is verified and loan amount is above 16k
- For grade G and interest rate above 20%

Recommendations:

Major driving factors which can be used to predict loan defaulters are:

1) DTI
2) Loan Term Period
3) Loan Amount
4) Annual Income
5) Grade
6) Home Ownership
7) Interest Rate

We can refer to the comparative analysis of all these variables to figure out and predict loan defaulters and then come up with strategy to minimize defaulting by addressing all these factors.