EXPLORATORY DATA ANALYSIS - EDA

Lending Club Case Study

Narayana Isanaka

OVERVIEW

Lending Club is 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 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

OBJECTIVE

Working as a Data scientist for Lending Club,

My goal is to analyse the data obtained from previous and existing clients to help the company to make informed data driven decisions.

This will help company operate efficiently and improve profitability.

As a Data Scientist,

I plan to perform the following steps to perform Exploratory Data Analysis on the provided data set.

1. Understanding the Data set
2. Cleaning and Manipulation of Data
3. Data Analysis
4. Recommendations and Presentation

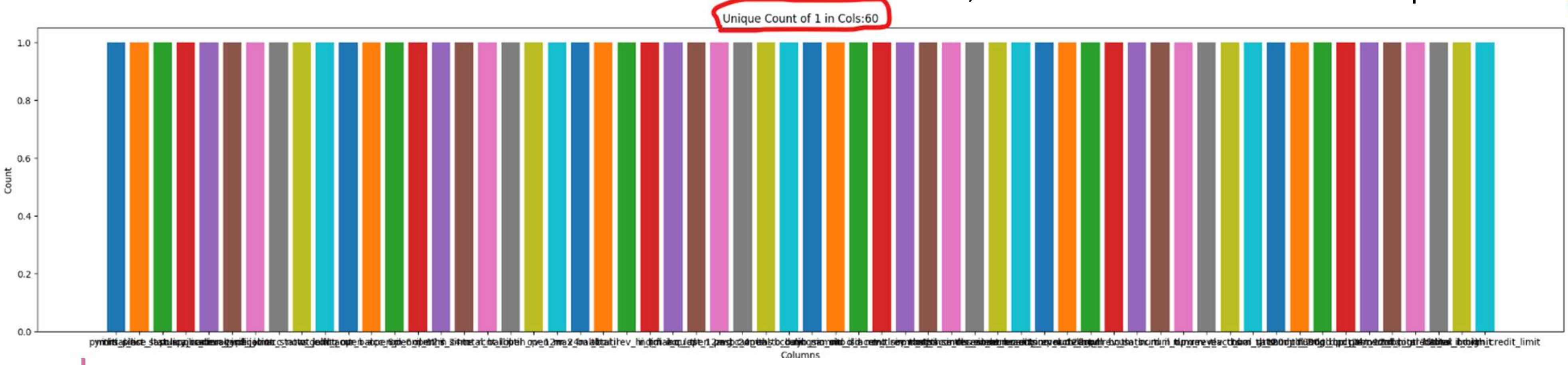
UNDERSTANDING THE DATA

Key Information about Data Set – Loan application data and current status of the loan payments.

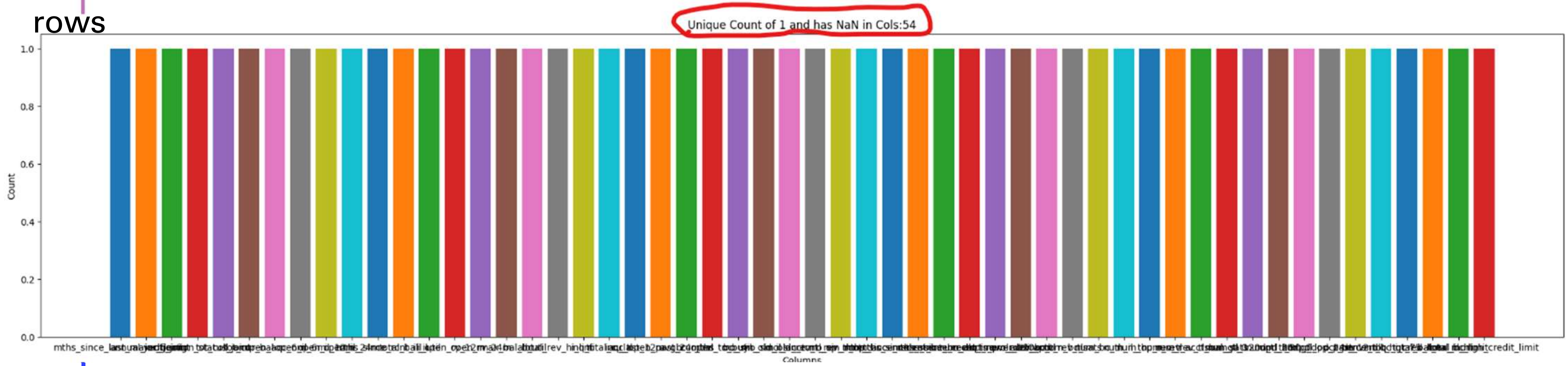
Each row is	A persons loan application data with the bank: Lending Club, with all the details.			
Number of rows:	39717	**Number of Columns:**	111	shape: (39717, 111)
Sampling Method:	All applicants who were granted loan between – **April 2008** and **Sep 2011**			
float64 - 74 **object** 24 **int64** 13	RangeIndex: 39717 entries, 0 to 39716 Columns: 111 entries, id to total_il_high_credit_limit dtypes: float64(74), int64(13), object(24)			
Key Categorical fields:	Grade, sub_grade, emp_title, home_ownership, verification_status, loan_status, purpose, title, addr_state,(Date cols:		2-10 unique values:	12
Key Fact fields:			11-150 Unique values:	12
Columns with NaN values only	54	**Columns with single value**	6	

Data understanding

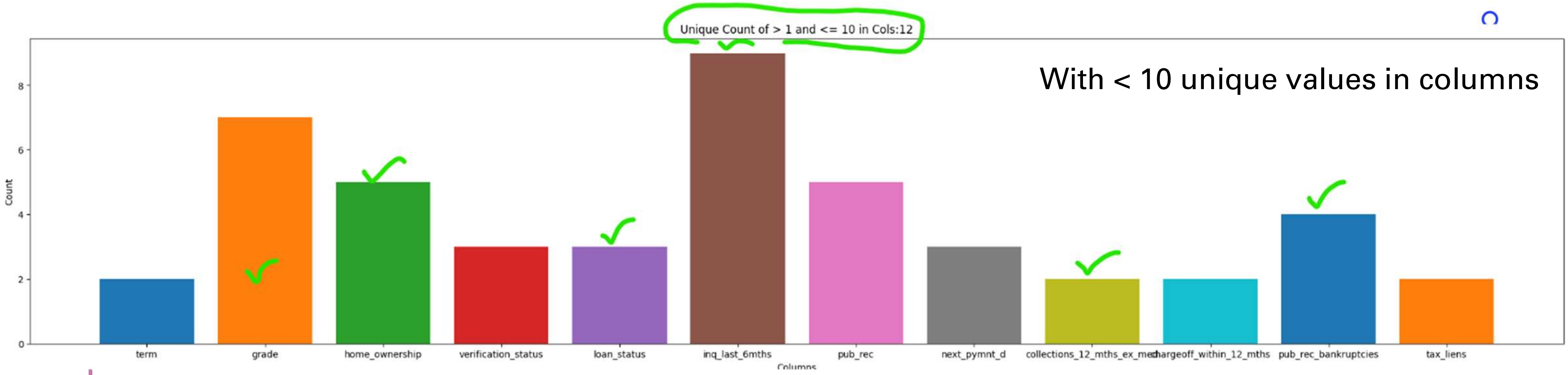
There are **60 columns** with same value for all the rows of **111 columns**, hence these columns are not required for Analysis



Of **60 columns** above **54 columns** has NaN value for all the rows

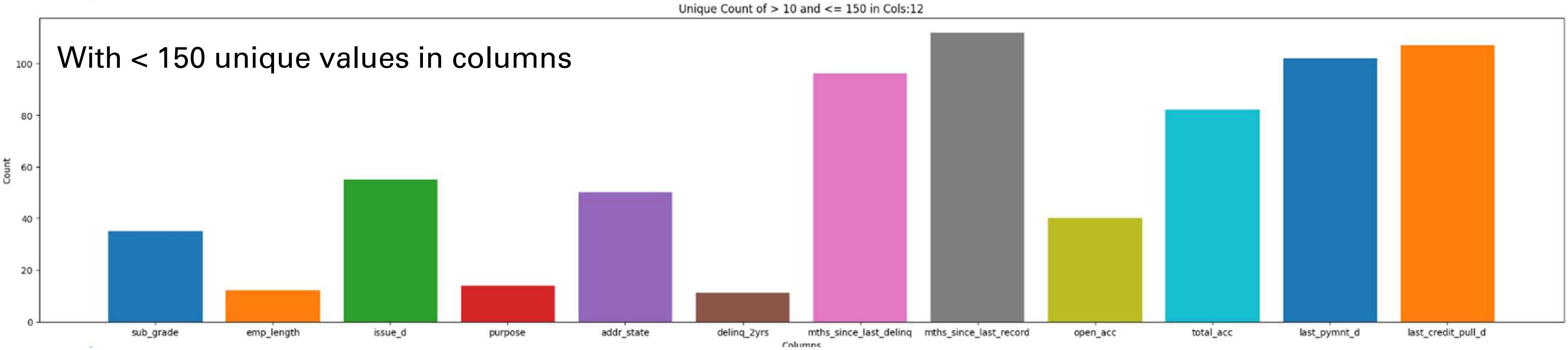


Data understanding



Unique Count of > 1 and <= 10 in Cols:12

With < 10 unique values in columns



Unique Count of > 10 and <= 150 in Cols:12

With < 150 unique values in columns

DATA UNDERSTANDING

Catogerical/Dimenssions	Dates	Facts/Numerical values
*term - 36/60 *grade *sub_grade *emp_title *emp_length *home_ownership *verification_status *loan_status *pymnt_plan - N *url *desc *purpose *title *addr_state *initial_list_status - f *application_type - INDIVIDUAL	issue_d earliest_cr_line last_pymnt_d next_pymnt_d last_credit_pull_d	*loan_amnt *funded_amnt ?funded_amnt_inv *int_rate *installment *annual_inc *dti *delinq_2yrs, *inq_last_6mths *mths_since_last_delinq *mths_since_last_record *open_acc , pub_rec, revol_bal, revol_util, total_acc out_prncp, out_prncp_inv total_pymnt, total_pymnt_inv , total_rec_prncp , total_rec_int, total_rec_late_fee Recoveries, collection_recovery_fee , *last_pymnt_amnt

DATA CLEANING AND STANDARDIZATION

As part of Data cleaning the following fixes were made:

Fixing Columns:

There are many columns with **ALL** NaN values & same value for all the records - this information is not helpful.

Fixing Missing Values & Standardizing values

"term" column is String, so removing "months" from feild value of "term" and rename it as "term_in_months“

# df.emp_title.unique().size is 23609 and will not get any data from this

Need to process text values - standardise

* lower

* remove multiple spaces

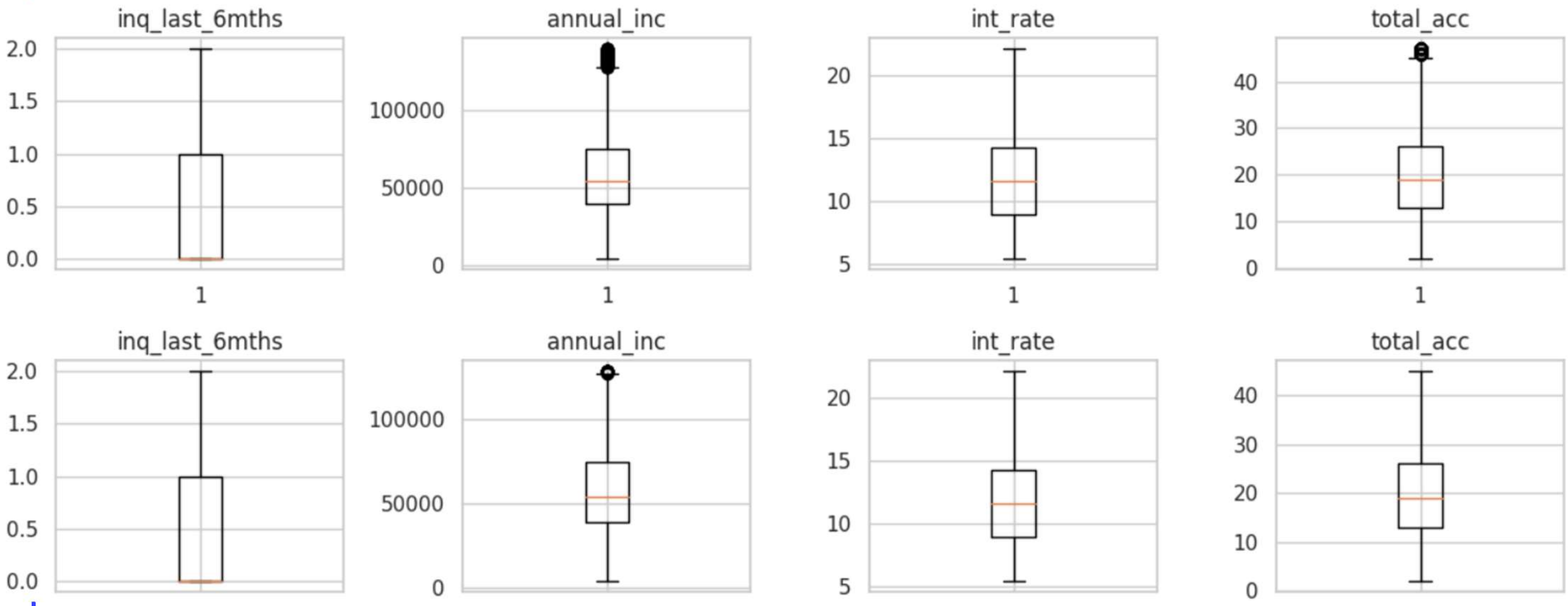
The following columns have same values and hence no specific/significant value to dataset.

1. droping.... policy_code [1]
2. droping.... acc_now_delinq [0]
3. droping.... delinq_amnt [0]

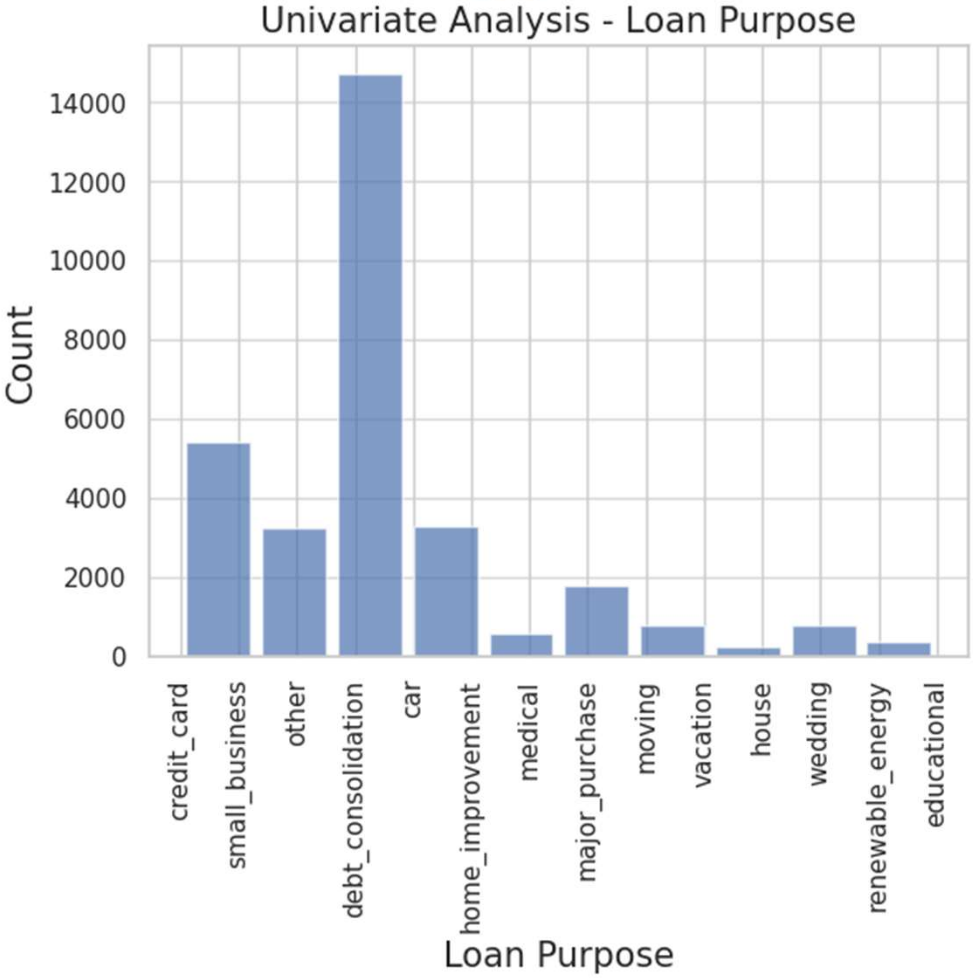
Delete outliers – rows for +/- 1.5 IQR

DATA CLEANING AND STANDARDIZATION

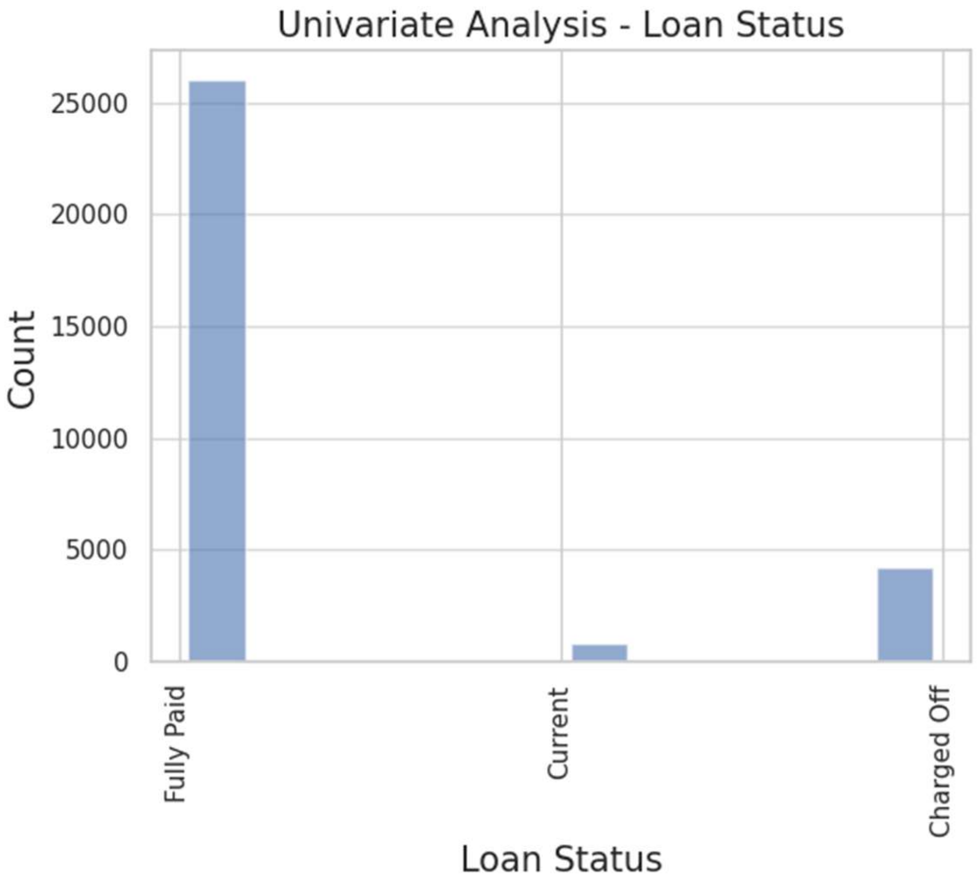
Cleaning Numerical values to remove outliers
❖ 1.5 X IQR above 75%Quartile
❖ 1.5 X IQR below 25%Quartile



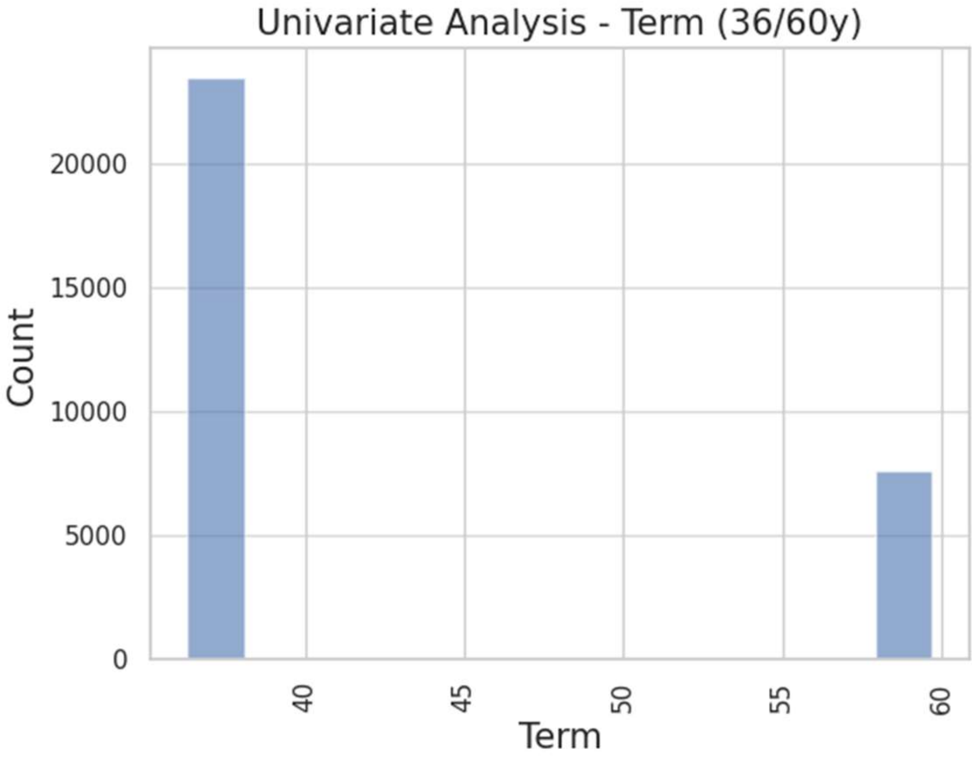
DATA ANALYSIS



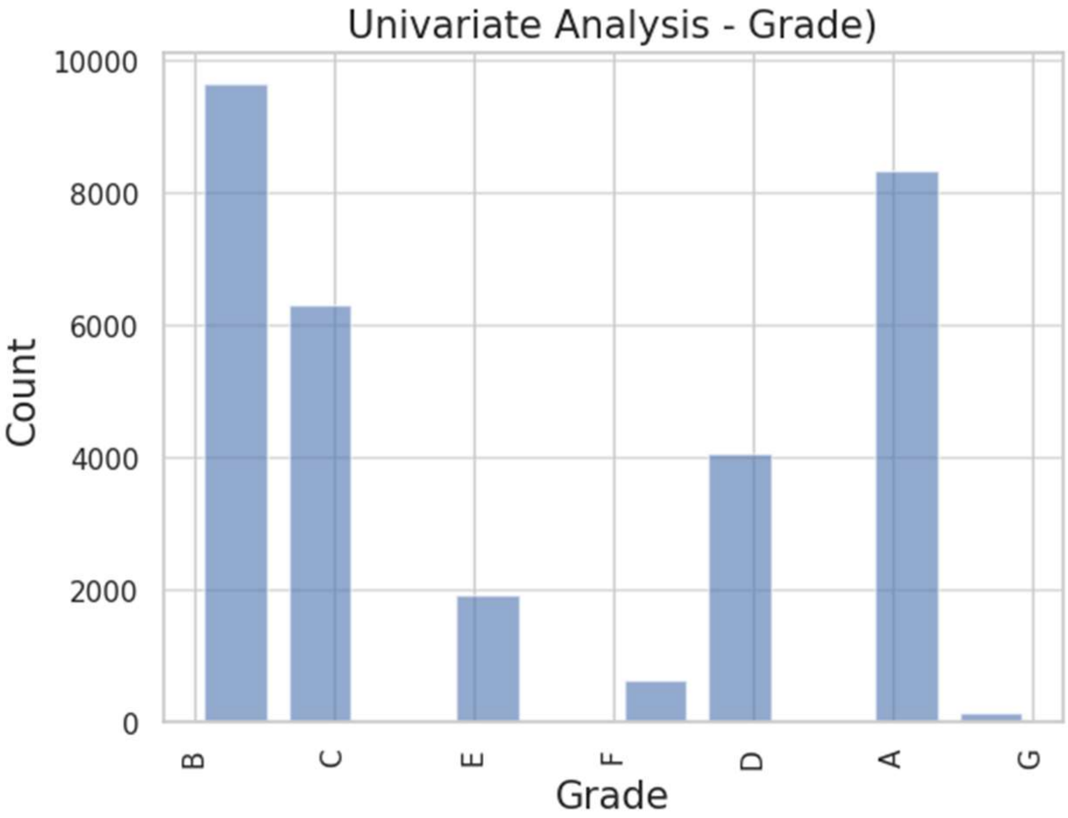
DATA ANALYSIS



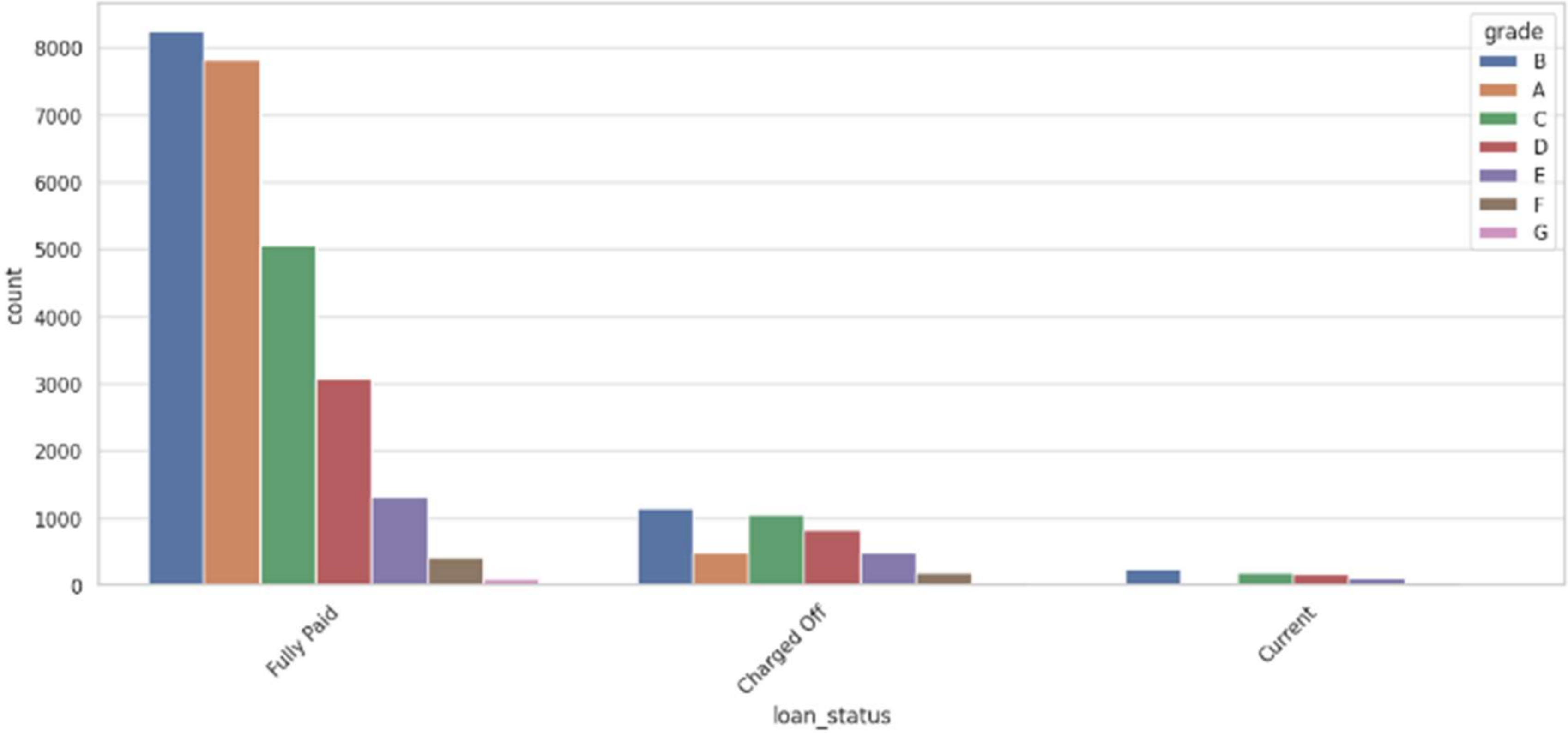
DATA ANALYSIS



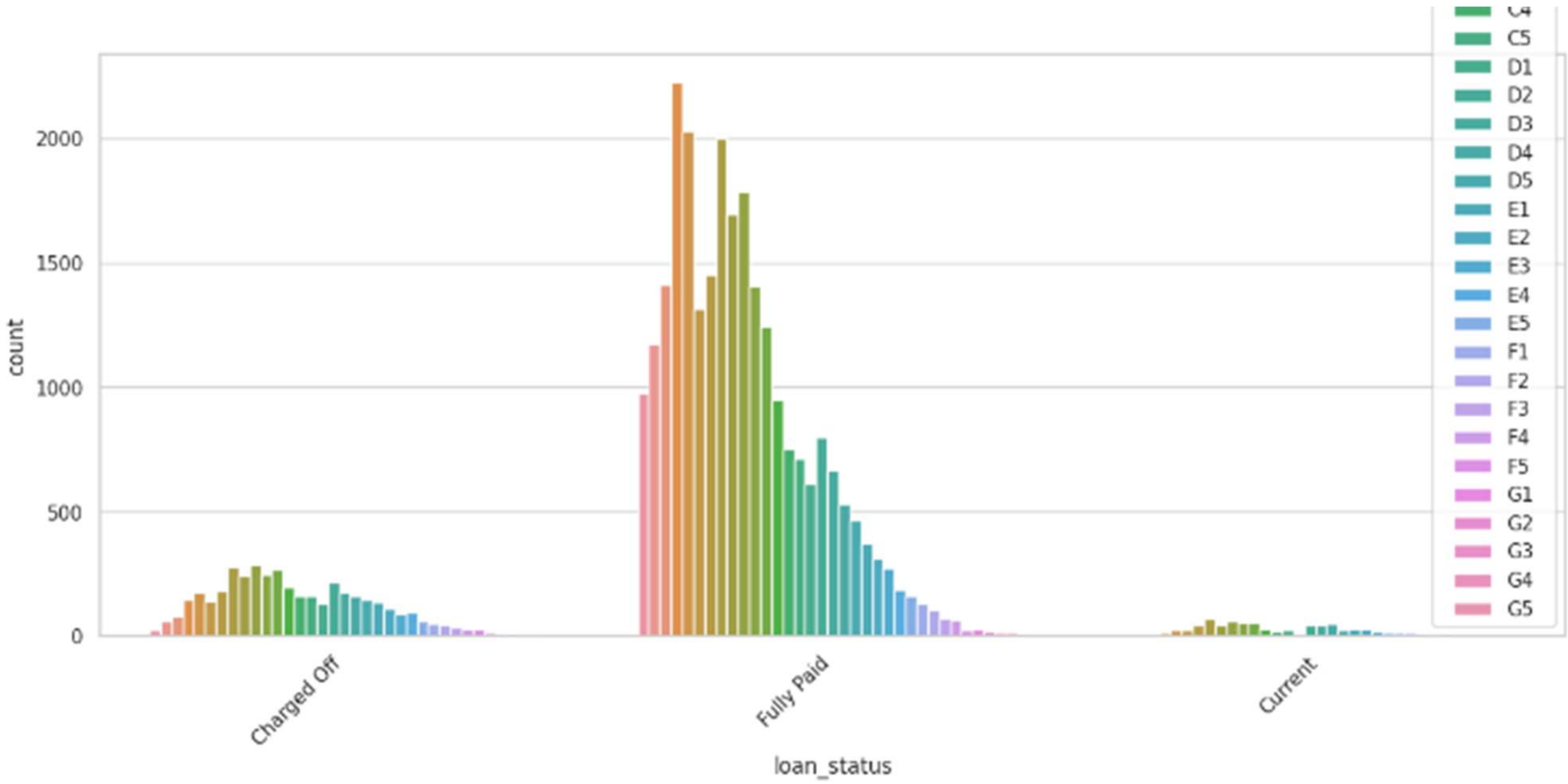
DATA ANALYSIS



DATA ANALYSIS



DATA ANALYSIS

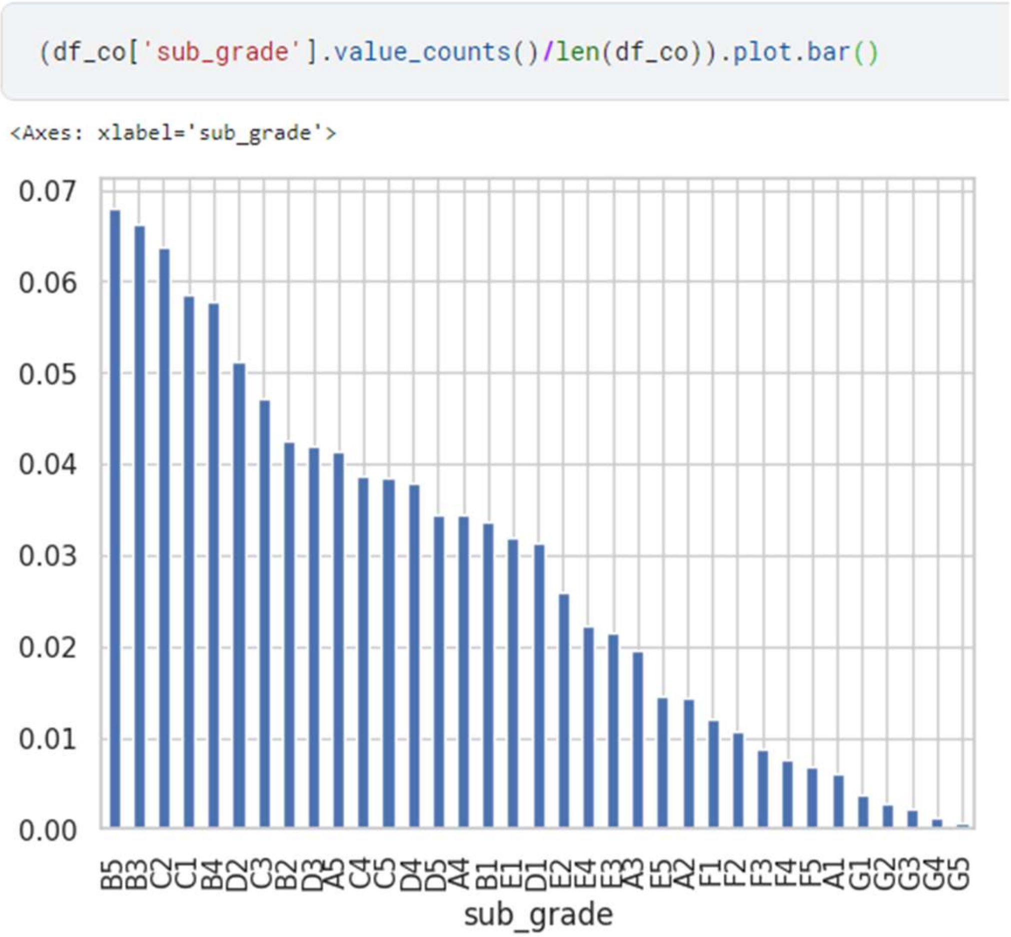


DATA ANALYSIS

DATA ANALYSIS

DATA ANALYSIS

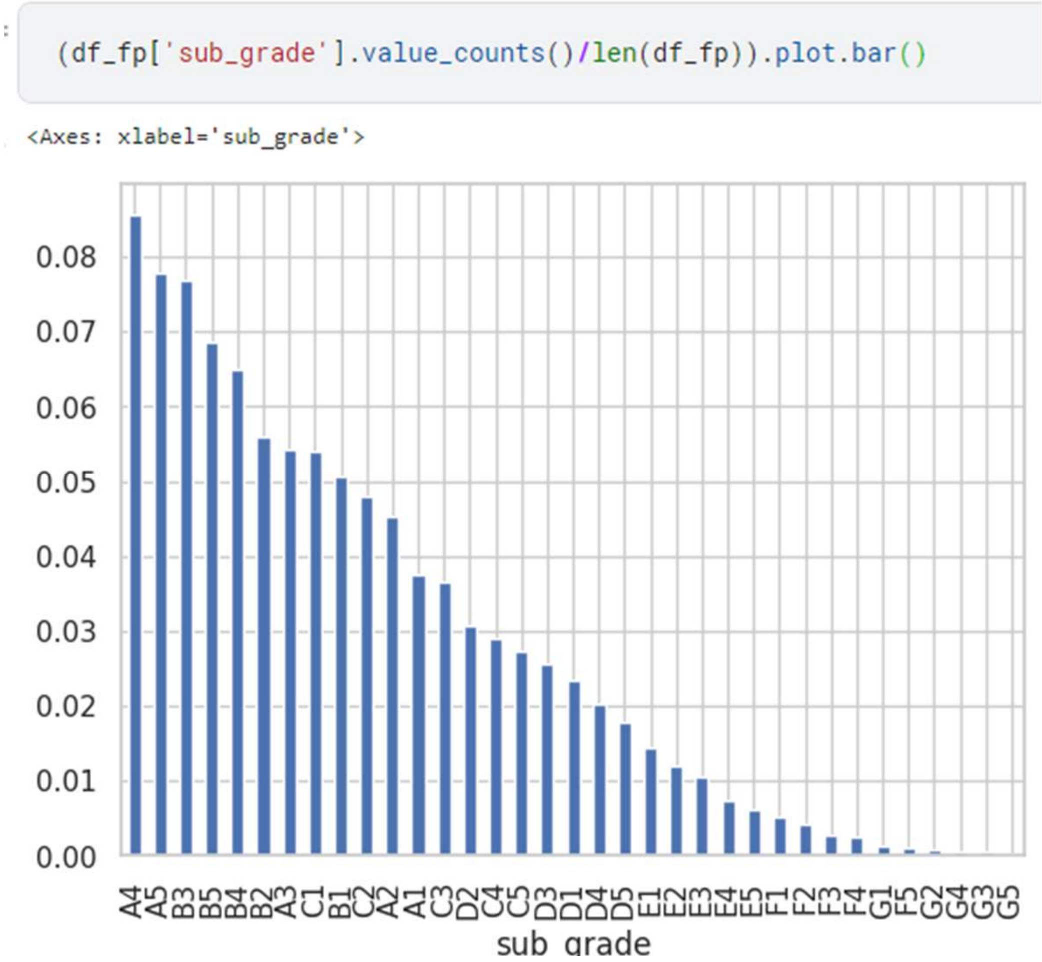
DATA ANALYSIS



```
(df_co['sub_grade'].value_counts()/len(df_co)).plot.bar()
```

```
<Axes: xlabel='sub_grade'>
```

DATA ANALYSIS



```
(df_fp['sub_grade'].value_counts()/len(df_fp)).plot.bar()
```

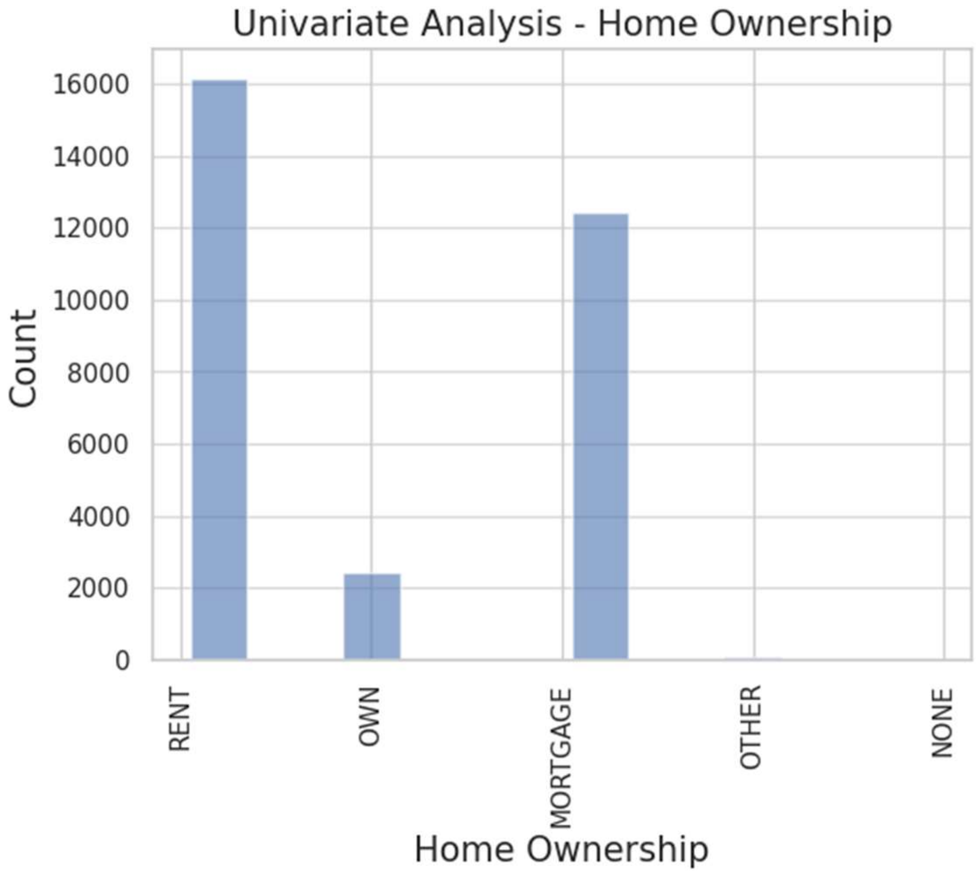
```
<Axes: xlabel='sub_grade'>
```

Insights

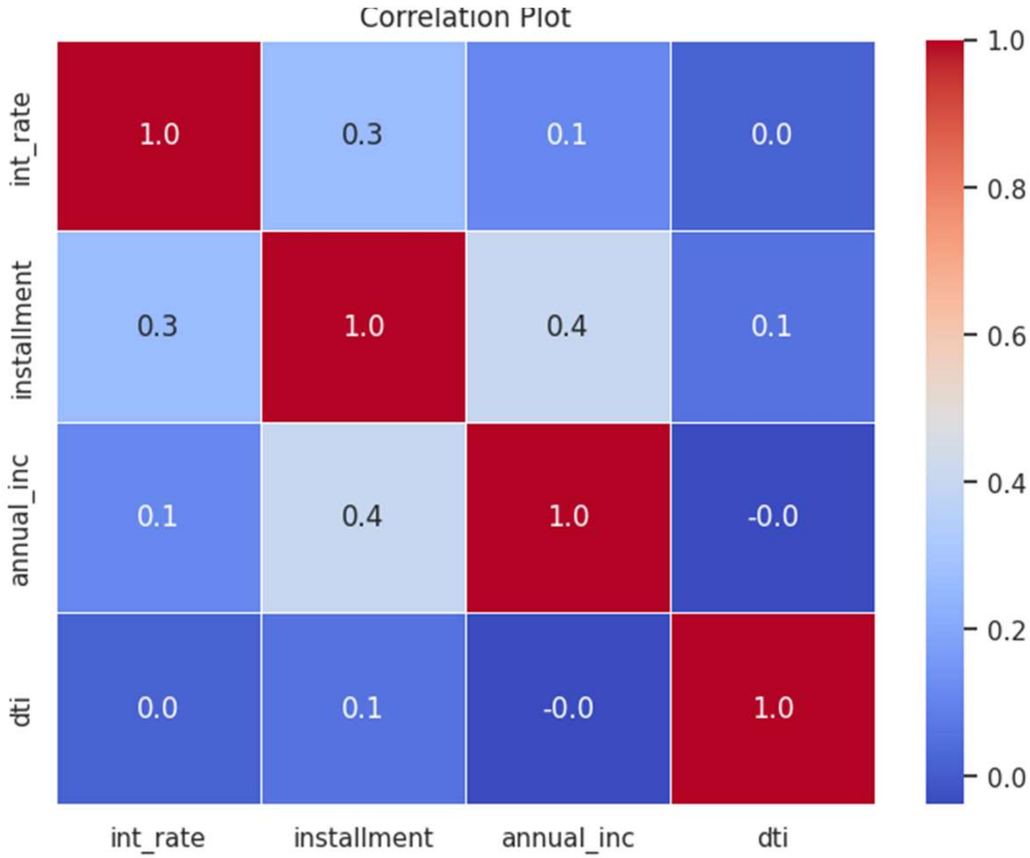
Segment Analysis can be done on "City Employes" to get default and income levels.

•most of the customers are paying the loans. 82.96% of customers are Fully Paid,14.16% are Charged Off and 2.87% are Current
•The loan businuss looks to be slowing, as we see only 2.87% are currently paying.
•top 2 resons loans are taken are due to - debt_consolidation & to pay of credit cards
•To drive the lending business, it is worth to try selling loans to people with high interest credit cards.
•**with a segmented and multivariate analysis:**
•it is found that Everyone applying with Purpose of "credit_card" are also "Charged Off" [TRUE]

DATA ANALYSIS



DATA ANALYSIS



THANK YOU

Narayana Isanaka