EXPLORATORY DATA ANALYSIS (EDA)

LENDING CLUB CASE STUDY PRESENTATION

NARAYANA ISANAKA

LENDING CLUB

- IS A **CONSUMER FINANCE COMPANY**
- SPECIALIZES IN LENDING VARIOUS TYPES OF LOANS TO URBAN CUSTOMERS.
- THE COMPANY IS LOOKING FOR HELP TO MAKE DATA DRIVEN DECISIONS ON LOAN APPLICATIONS RECEIVED.

RISKS:

- IF THE APPLICANT IS **LIKELY TO REPAY THE LOAN**,
  THEN NOT APPROVING THE LOAN RESULTS IN A **LOSS OF BUSINESS**
- IF THE APPLICANT IS **NOT LIKELY TO REPAY THE LOAN** (DEFAULT ON LOAN),
  THEN APPROVING THE LOAN MAY LEAD TO A **FINANCIAL LOSS.**

OBJECTIVE

WORKING AS A DATA SCIENTIST FOR LENDING CLUB, GOAL IS TO ANALYZE THE LOAN APPLICATION DATA FROM PREVIOUS AND EXISTING CUSTOMERS, TO HELP COMPANY MAKE DATA DRIVEN DECISIONS.

Exploratory Data Analysis

1. Data Cleaning
2. Data Analysis
3. Recommendations
4. Understanding the Data set

Understand the Data set

1. Nature of data, related data sets, domain, timeframe and size of the data set.
2. Metadata

Data Cleaning

1. Fix Rows and columns
2. Fix missing values
3. Standardise values
4. Fix invalid values
5. Filter data

Data Analysis

1. Perform Univariate Analysis
2. Segmented Univariate Analysis
3. Bivariate analysis
4. Derived metrics

Insights & Recommendations

1. Extract insights and provide recommendations

UNDERSTANDING THE DATA SET.

KEY FINDINGS

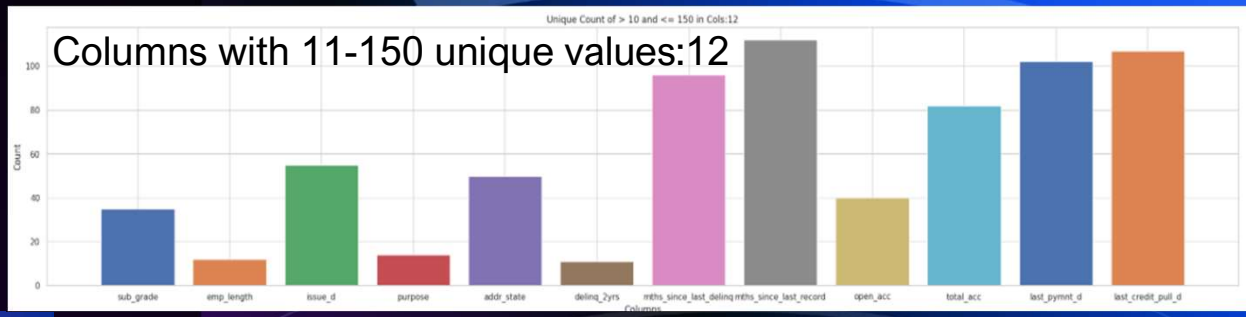
1. Each row has person's loan application details from Apr 2008 to Sep 2011.
2. Shape of the data set (Rows: 39717, columns:111)
3. **Key columns missing from data set are(in data dictionary):** fico_range_high, fico_range_low, **last_fico_range_high and last_fico_range_low**
4. There are: float64 – 74 , object – 24 and int64 – 13 as column data types.

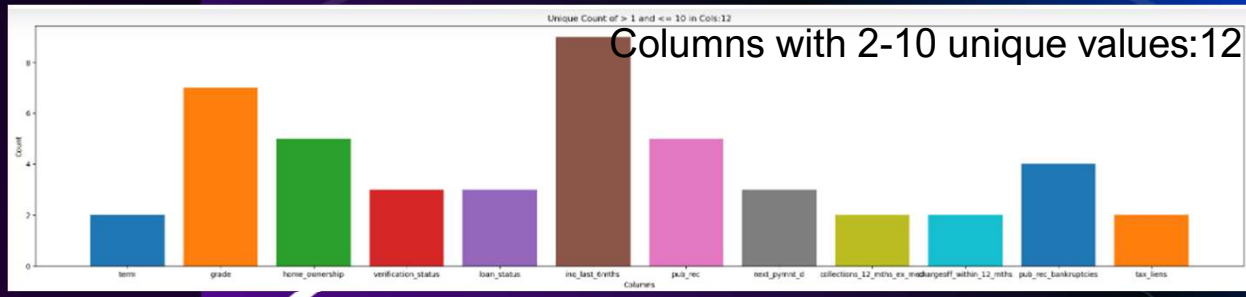
Column	Type	Description
loan status	Categorical/ Dimension	Current status of the loan
Grade/sub_Grade	Categorical/ Dimension	LC assigned loan grade
home_ownership	Categorical/ Dimension	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
purpose	Categorical/ Dimension	A category provided by the borrower for the loan request.
addr_state	Categorical/ Dimension	The state provided by the borrower in the loan application
loan_amnt	Numeric/ Fact	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
Int_rate	Numeric/ Fact	Interest Rate on the loan
annual_inc	Numeric/ Fact	The self-reported annual income provided by the borrower during registration.
dti	Numeric/ Fact	A ratio of borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
recoveries	Numeric/ Fact	post charge off gross recovery
emp_length	Numeric/ Fact	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
installment	Numeric/ Fact	The monthly payment owed by the borrower if the loan originates.

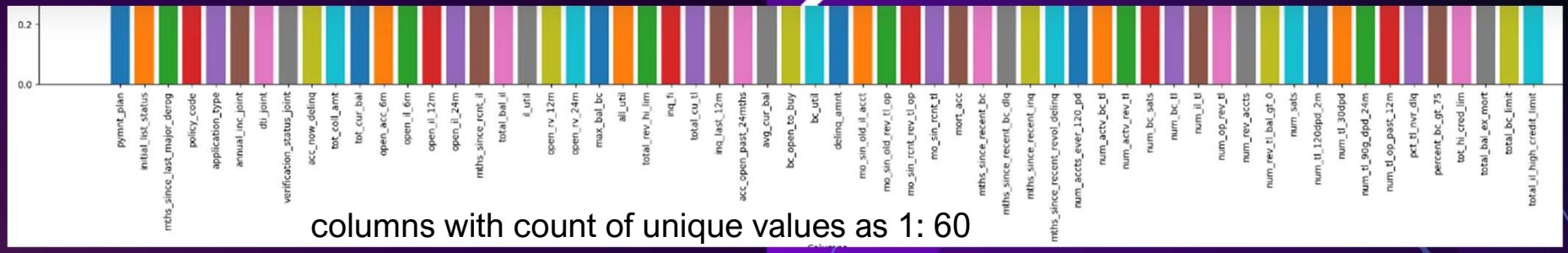
UNDERSTANDING THE DATA SET.

KEY FINDINGS

1. columns with count of unique values as 1 is 60
2. Columns with 2-10 unique values -12
3. Columns with 11-150 unique values -12
4. Columns with only Null values - 54
5. Columns with only 1 value – 6







UNDERSTANDING THE DATA SET.

KEY FINDINGS

Summary

Looking at the dataframe.describe() we can see that the overall statistical averages.

1. get a feel of how the data is spread for each of the columns
2. Identify any critical numerical fields that are not in the "describe()" report, and note the Data cleaning requirements.

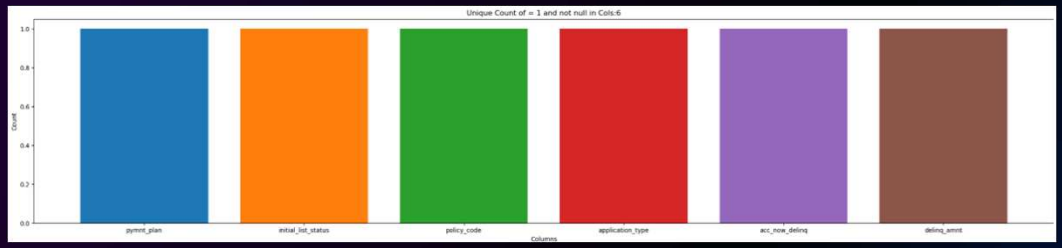
Column	mean	Min	50%	max
loan_amnt	11219	500	9600	35000
installment	324.56	15.69	280.22	1305
annual_inc	The self-reported annual income provided by the borrower during registration.			
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.			
recoveries	post charge off gross recovery			
Int_rate – is a object data type and is not part of this report. Data cleaning is required for this Column to remove % and change the data type.				
emp_length - is a object data type and is not part of this report. Data cleaning is required for this Column to remove year/years/> characters and change the data type.				
term - is a object data type and is not part of this report. Data cleaning is required for this Column to remove 'months' characters and change the data type.				

DATA CLEANING.

KEY FINDINGS

- Cleaning based on Null values

1. The shape of the data set is : shape: (39717, 111)
2. There are 54 columns with **ALL** Null values value for all the records
3. There are 6 columns with 1 unique value and is not null. These columns are not useful and hence dropping from the data frame.
4. The shape after removing columns is : (39717, 51)



The colum is :**pymnt_plan** and the only value in the column is:('n')

The colum is :**initial_list_status** and the only value in the column is:('f')

The colum is :**policy_code** and the only value in the column is:(1)

The colum is :**application_type** and the only value in the column is:('INDIVIDUAL')

The colum is :**acc_now_delinq** and the only value in the column is:(0)

The colum is :**delinq_amnt** and the only value in the column is:(0)

Though there are accounts with status "Charged Off", the delinq_amnt is 0 – this is a data quality issue.

The columns with the same value does not provide any specific insights for each application/row.

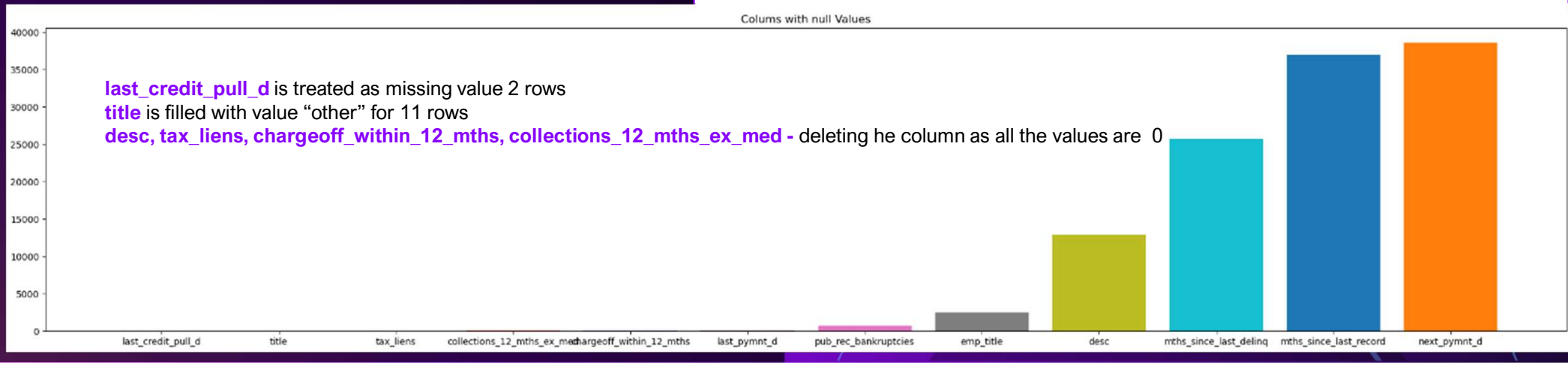
DATA CLEANING.

KEY FINDINGS

- Fixing missing/invalid values.
- Columns > 25000 null values can be dropped as 62% are null values.

Description of Data cleaning by categories

Column: **last_credit_pull_d** has 2 null values
Column: **title** has 11 null values
Column: **tax_liens** has 39 null values
Column: **collections_12_mths_ex_med** has 56 null values
Column: **chargeoff_within_12_mths** has 56 null values
Column: **last_pymnt_d** has 71 null values
Column: pub_rec_bankruptcies has 697 null values
Column: emp_title has 2459 null values
Column: desc has 12942 null values
Column: **mths_since_last_delinq** has 25682 null values
Column: mths_since_last_record has 36931 null values
Column: next_pymnt_d has 38577 null values



last_credit_pull_d is treated as missing value 2 rows
title is filled with value “other” for 11 rows
desc, tax_liens, chargeoff_within_12_mths, collections_12_mths_ex_med - deleting he column as all the values are 0

DATA CLEANING.

KEY FINDINGS

- Standardizing values.

1. emp_title has multiple spaces, leading spaces , trailing spaces, caps, mixed cases etc.
2. To fix this we change it to lowercase and remove all spaces and replace multiple spaces with single space.

Description of Data cleaning by categories

Column: **last_pymnt_d** has 71 null values
Column: pub_rec_bankruptcies has 697 null values
Column: emp_title has 2459 null values

emp_title

Col: emp_title has 2459 null values

Analyse data for last_pymnt_d fit the median date of
Feb -12

last_credit_pull_d

It has 2 missing values, took meadian of month and year and it came out to be Oct-12 as the median value for the column.

UNIVARIATE ANALYSIS: Field : recoveries

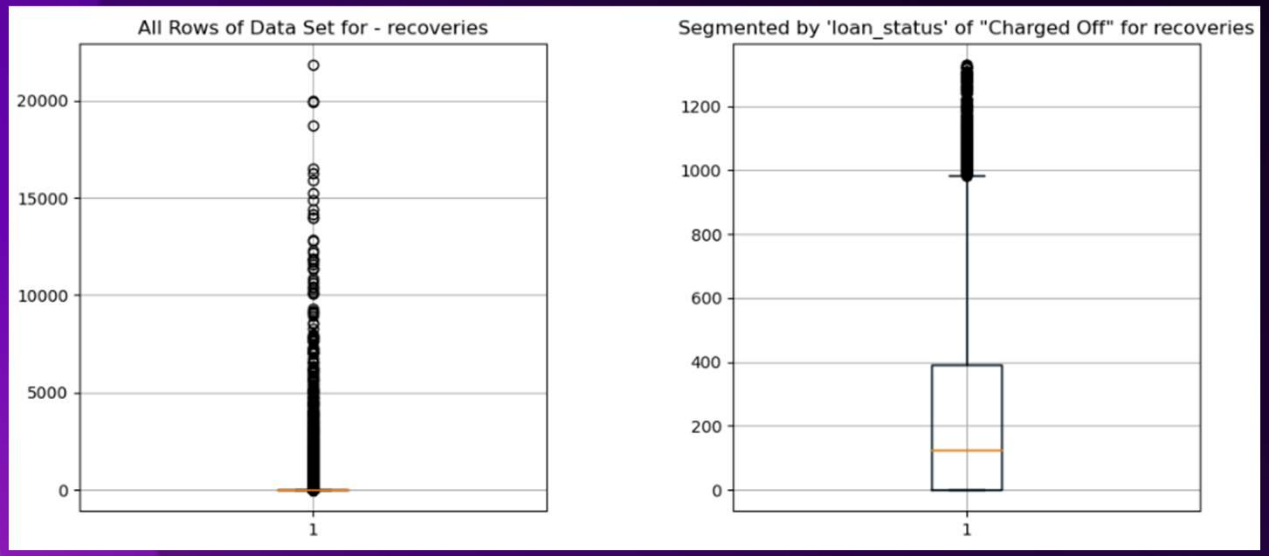
#RECOVERIES:

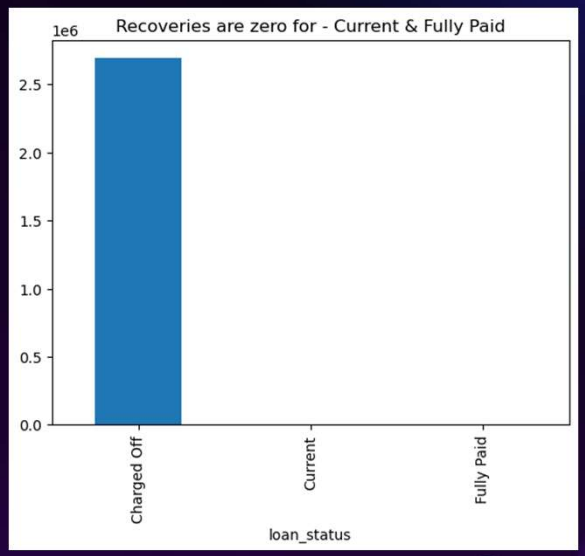
SUMMARY METRICS:

This is a special case field that requires segmented analysis to perform the distribution analysis.

This is because the value for this field only exists when the loan_status is "Charged_Off".

Hence considering all fields will distort summary metrics.





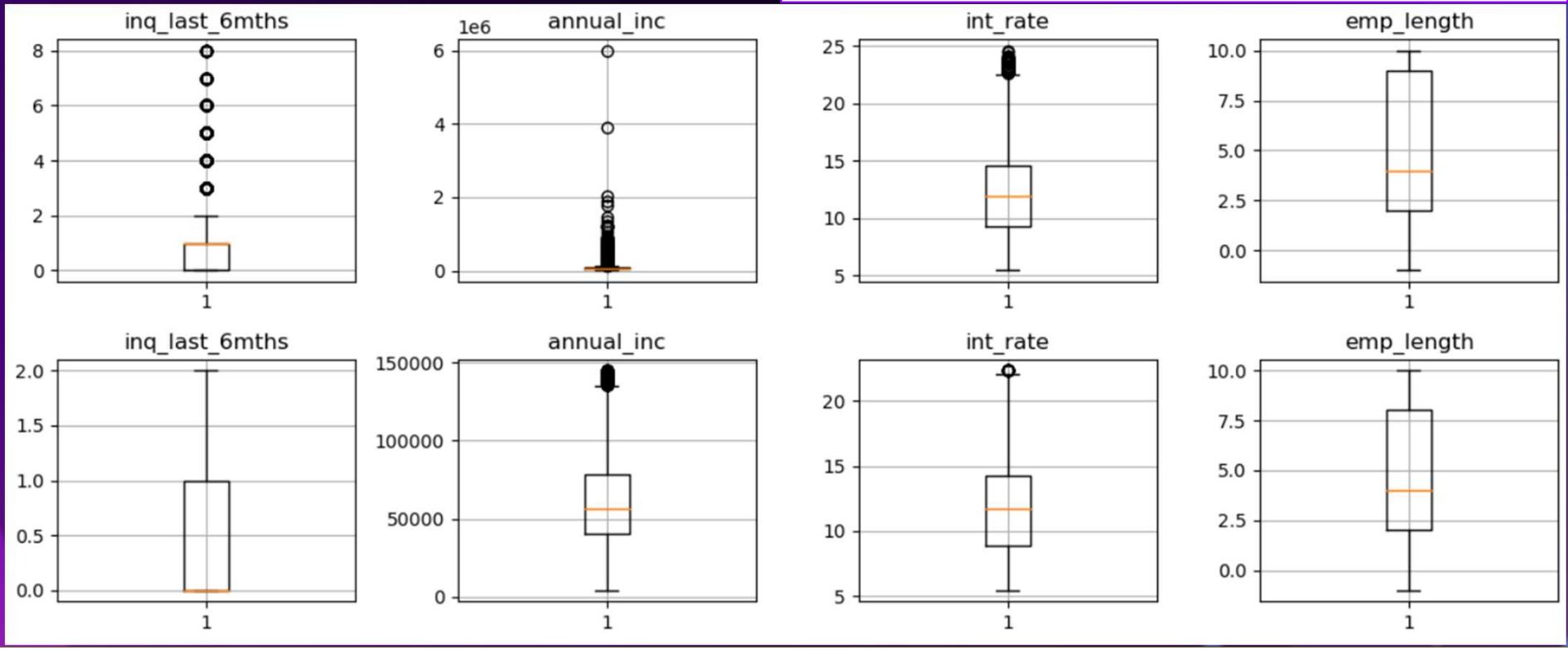
- Calculating "**mean**" covering all rows of "recoveries" columns will result in "**wrong analysis**".

UNIVARIATE ANALYSIS: Fields: Inq_last_6mths, annual_inc, int_rate, emp_length

OBSERVATION:

- Outliers were removed from the dataset by excluding data points below Q25-1.5*IQR and above Q75+1.5*IQR, enabling focused analysis on the core dataset.
- we can observe the mean for columns moved as noted in the table.

column	Mean before IQR adjustment	Mean after IQR adjustment
inq_last_6mths	0.86	0.62
annual_inc	$68,968	$ 60,635
Int_rate	12.02%	11.76%
emp_length	4.81	4.76

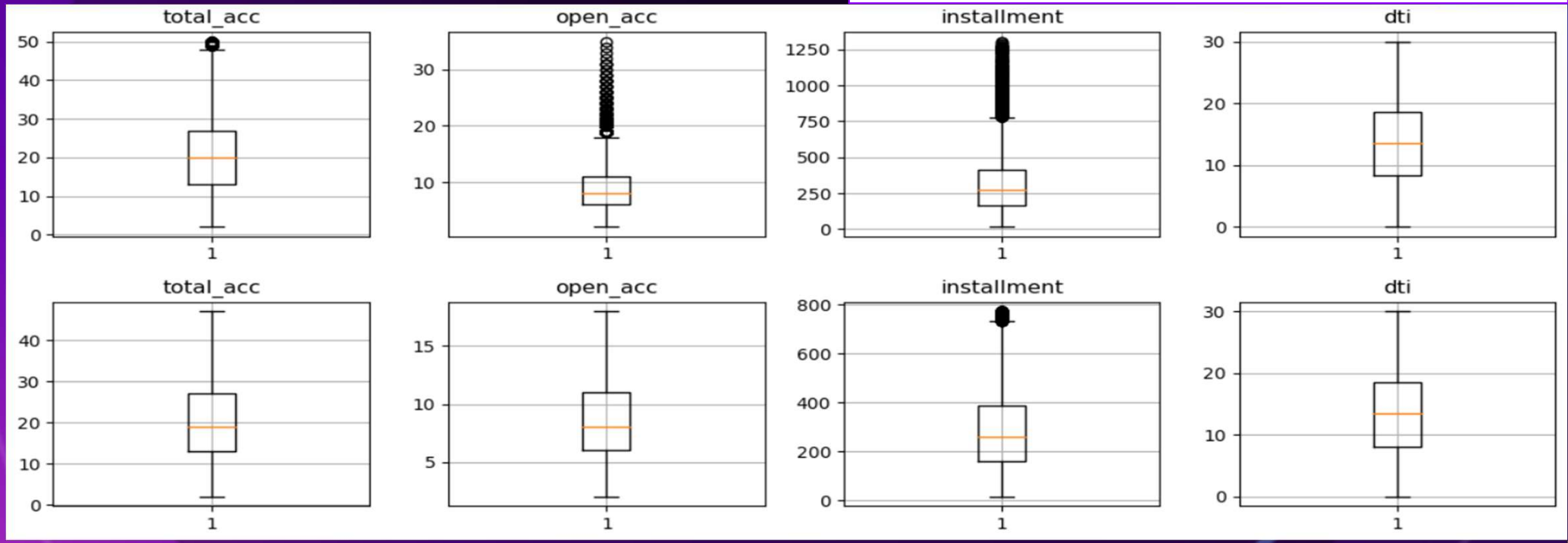


UNIVARIATE ANALYSIS: Fields: total_acc, open_acc, installment, dti

OBSERVATION:

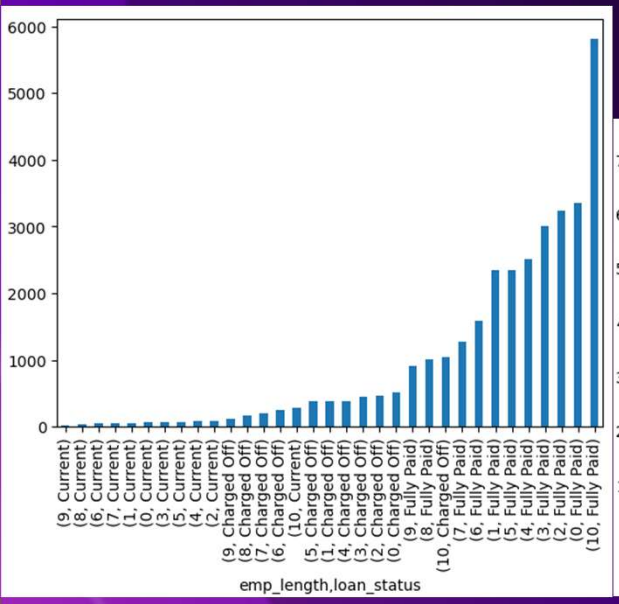
- Outliers were removed from the dataset by excluding data points below Q25-1.5*IQR and above Q75+1.5*IQR, enabling focused analysis on the core dataset.
- we can observe the mean for columns moved as noted in the table.

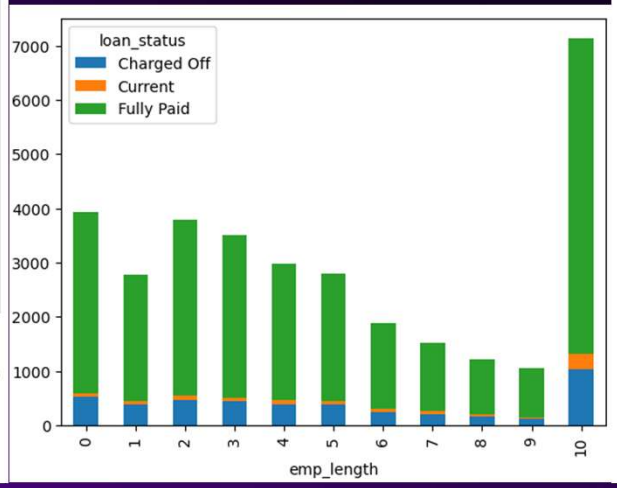
column	Mean before	Mean after
total_acc	22.08	20.02
Dti	13.32	13.29
open_acc	9.29 to 8.58	
Installment	324.56	$ 289.22

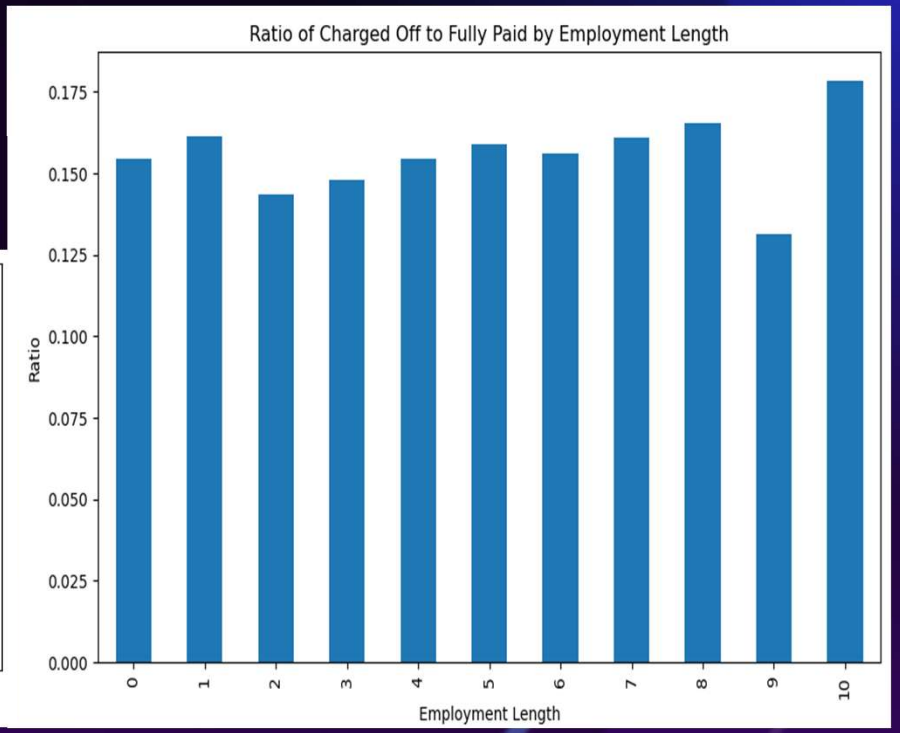


SEGMENTED-UNIVARIATE ANALYSIS

- Segmented by (Charged Off/Current/Fully Paid) loan_status vs emp_length
- We can see that the risk of default is less during the year 2,3,4 and 9.
- The risk of “Charge Off” slightly higher in the band 1,5,6,7,8 and 10
- 9th year employees are at less risk of charge off, compare to all others







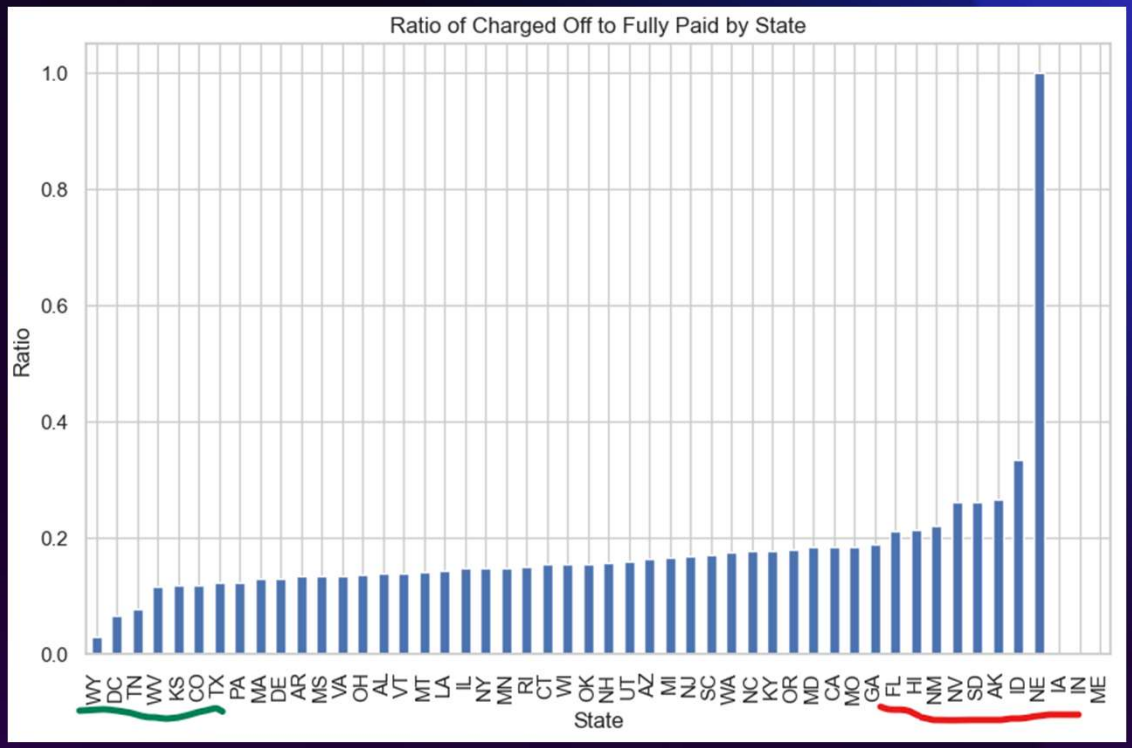
SEGMENTED-UNIVARIATE ANALYSIS

By looking at the ratio of default to paid off:

Low Risk States:
WY, DC, TN, WV, KS, TX,PA and MA

High Risk States:
FL, HI, NM,NV,SD,AK,ID,NE

We do not have business in :
IA, IN ME



This **Analysis** could be **wrong** as we are considering all states including where the business volume is low.
On next slide – lets examine states with high business volume and calculate ratios.

SEGMENTED-UNIVARIATE ANALYSIS

The volume of the business is more in CA, NY, FL, TX, NJ, PA, VA…

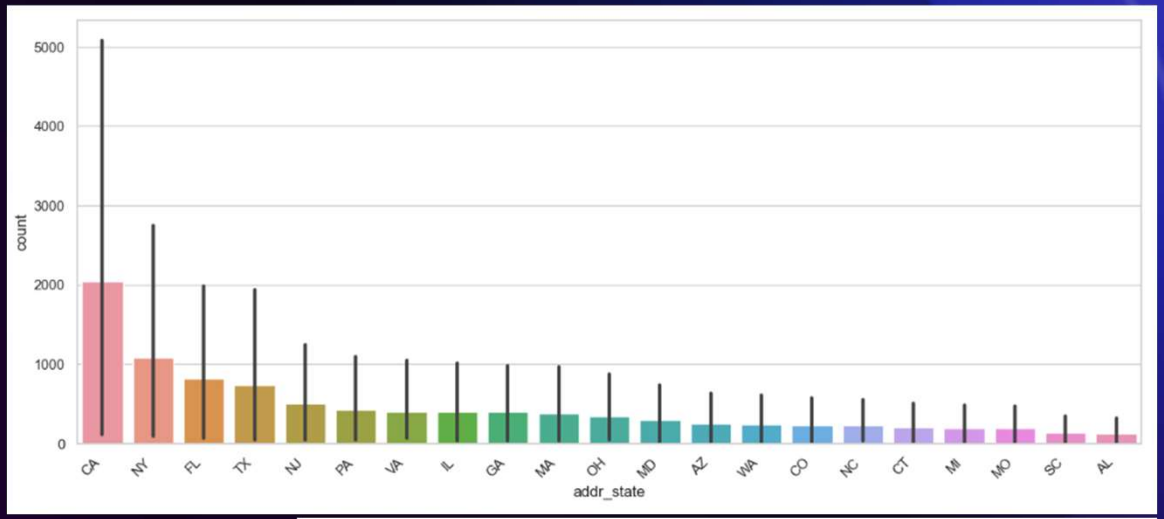
Lets now examine the ratios of these states and find comparatively risky states.

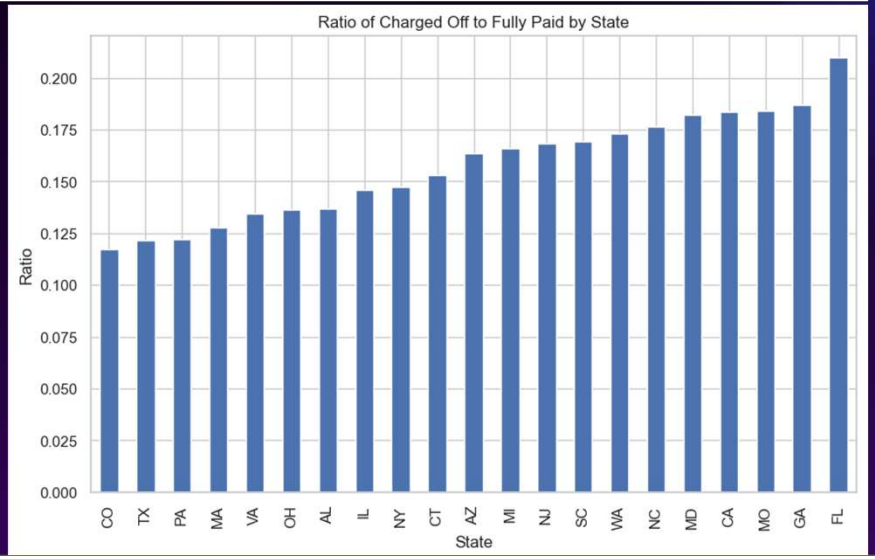
From the ratio bar-chat: we can observe

Low Risk states:
CO,TX,PA,MA,VA

High Risk States:
FL, GA, MO, CA, MD and NC.

This **Analysis is more meaningful** as we eliminated low volume business states.

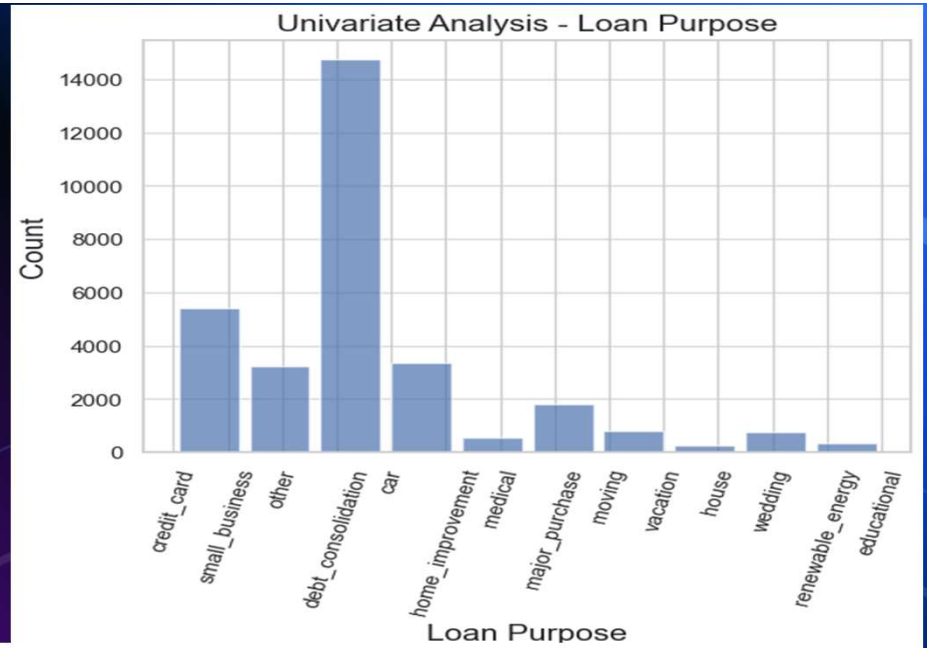


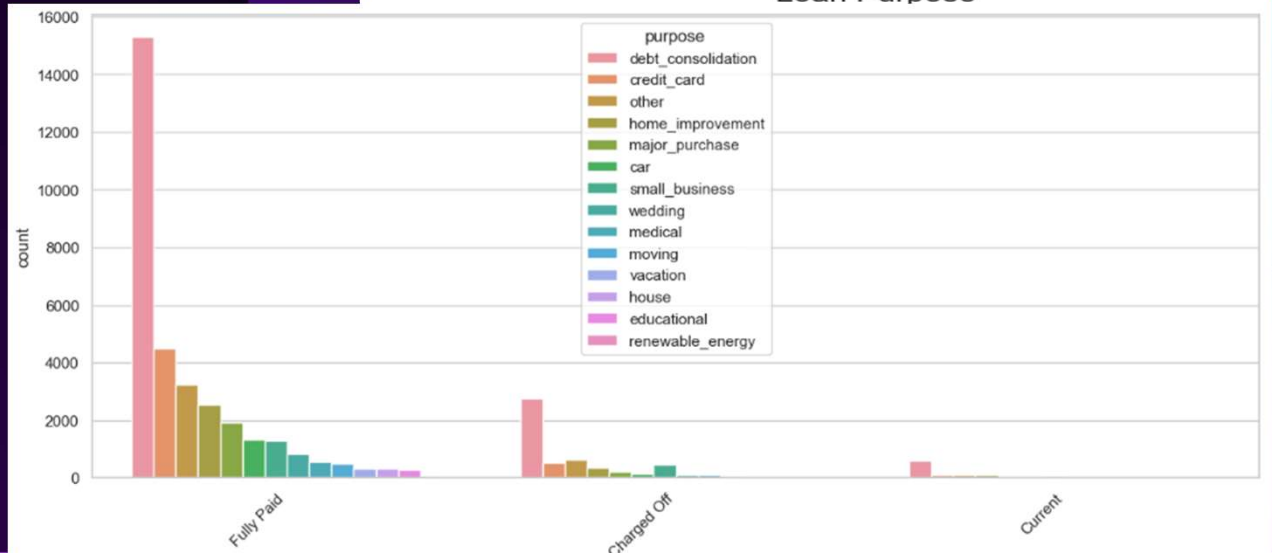


SEGMENTED - UNIVARIATE ANALYSIS.

LOAN PURPOSE

- Most of the loans are taken for debt consolidation and pay off credit card debt.
- It is worth to market loans to customers with LOW credit card debt with high interests and income

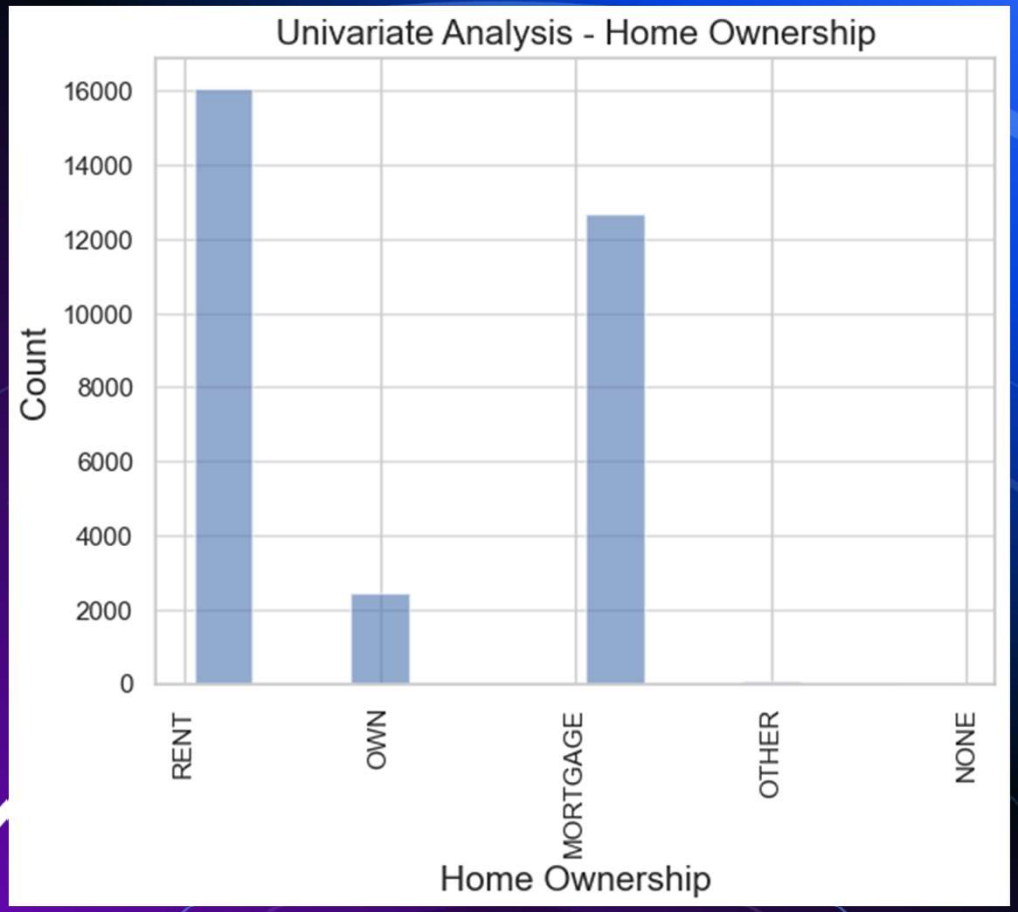




SEGMENTED - UNIVARIATE ANALYSIS.

HOME OWNERSHIP

- People on rent are taking more loans, followed by Mortgage and Own.
- It is worth to **market** loans to **Own** segment as they are fully paid off on Mortgage and it is likely that they will “Fully Pai the loan.



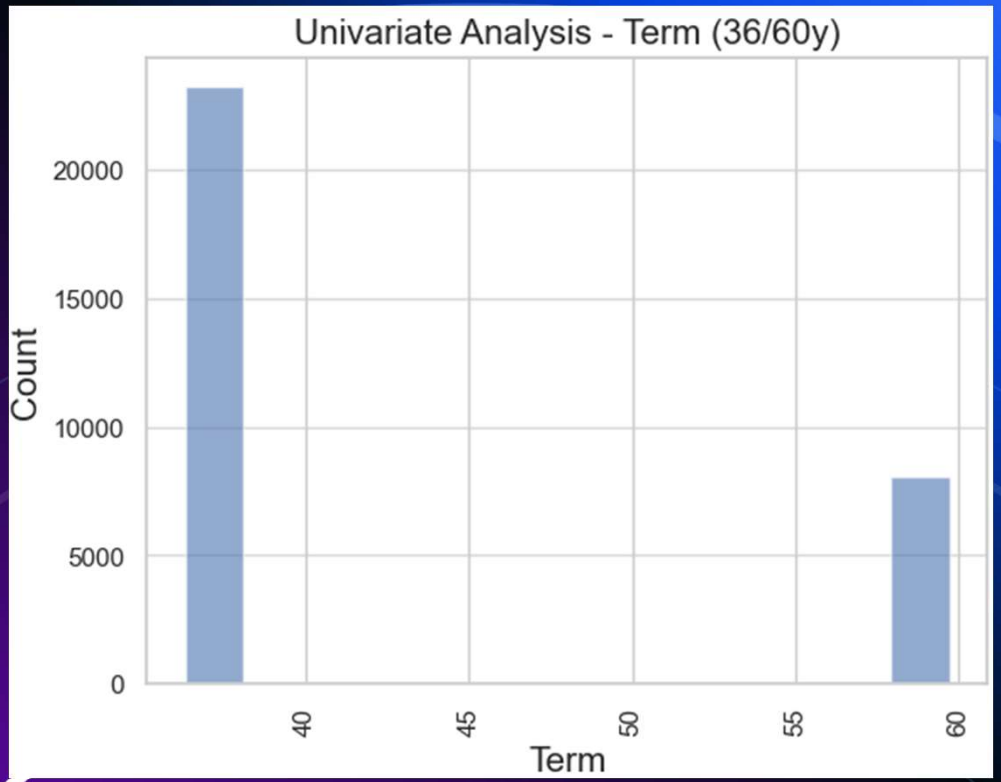
SEGMENTED - UNIVARIATE ANALYSIS.

LOAN TERM VS INT%

- Surprisingly the 60year mean interest rate is grater than 30year loan. "**Long term loans are risky**" 14.8% vs 11% for 30y.

	term	int_rate
0	36	11.004656
1	60	14.805912

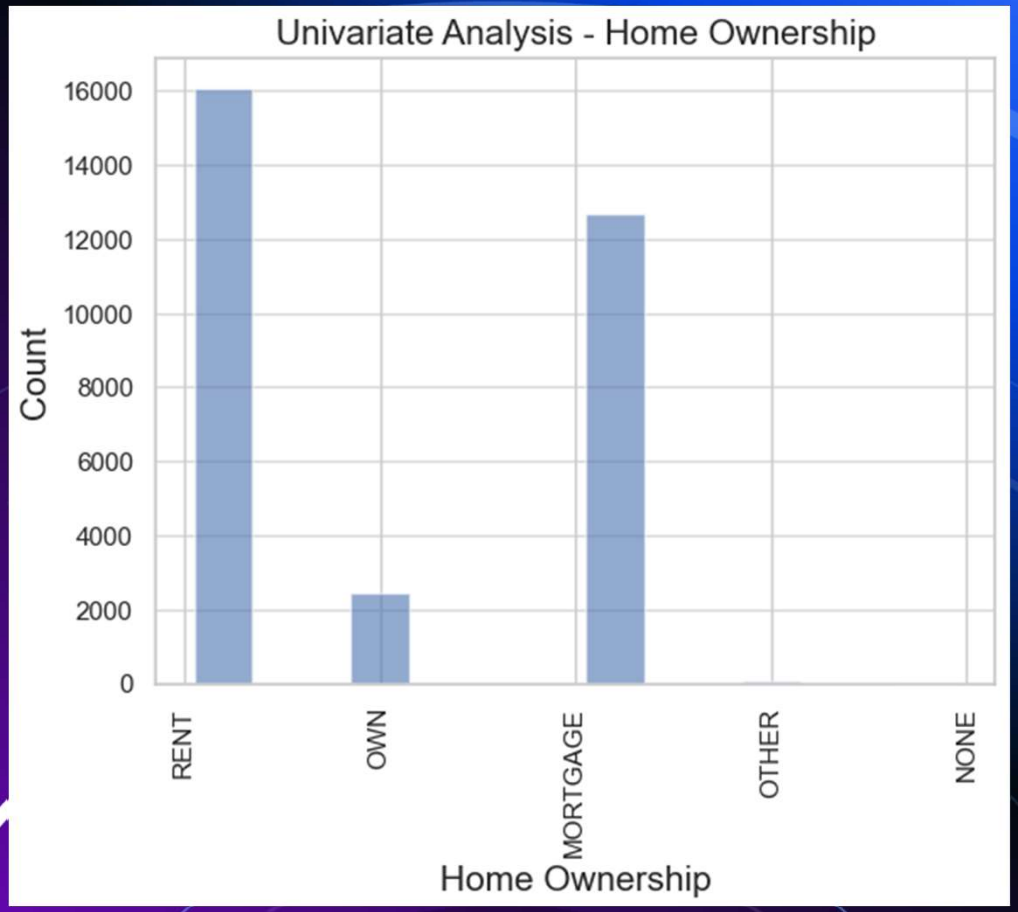
- Most customers go for 36y term loan.



SEGMENTED - UNIVARIATE ANALYSIS.

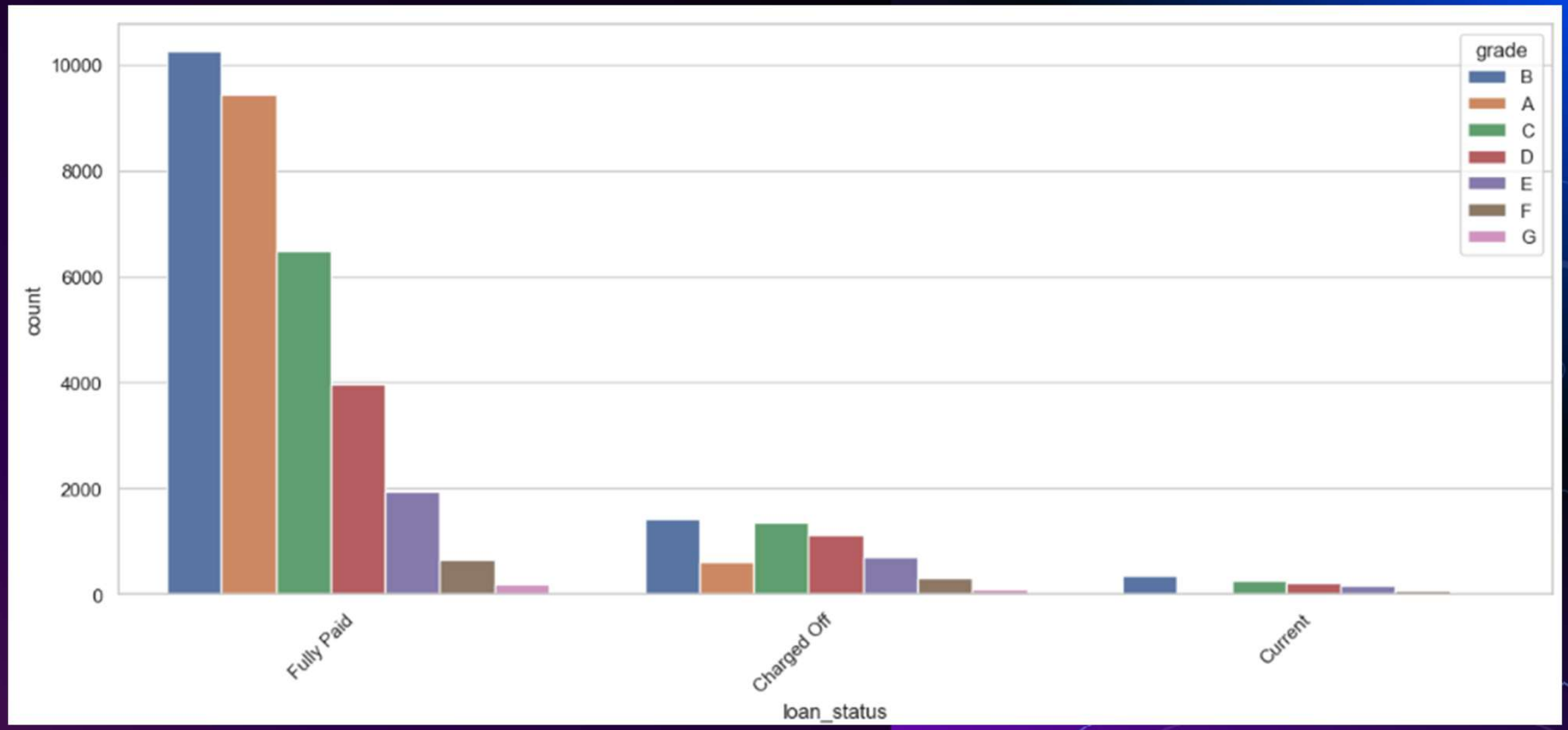
HOME OWNERSHIP

- People on rent are taking more loans, followed by Mortgage and Own.
- It is worth to **market** loans to **Own** segment as they are fully paid off on Mortgage and it is likely that they will “Fully Pai the loan.



SEGMENTED -UNIVARIATE ANALYSIS.

LOAN STATUS BY GRADE

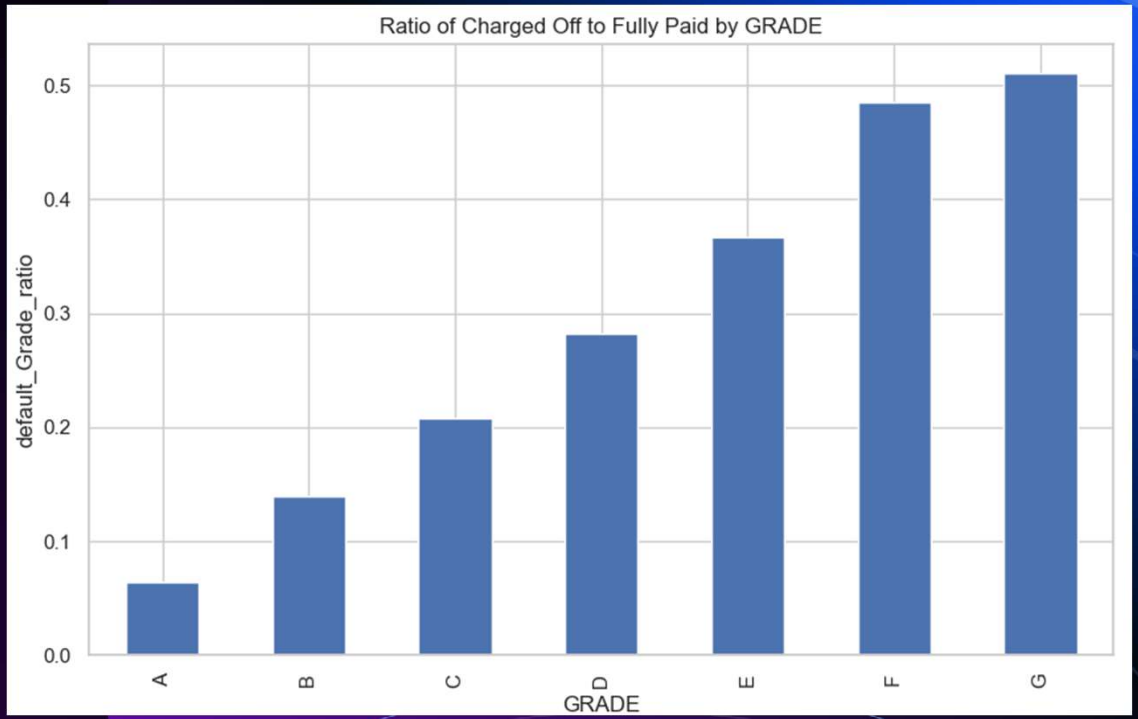


- We can see that grade “A” has maximum ration for Fully Paid /Charged Off

SEGMENTED - UNIVARIATE ANALYSIS.

RATIO OF CHARGED OFF TO FULLY PAID BY GRADE

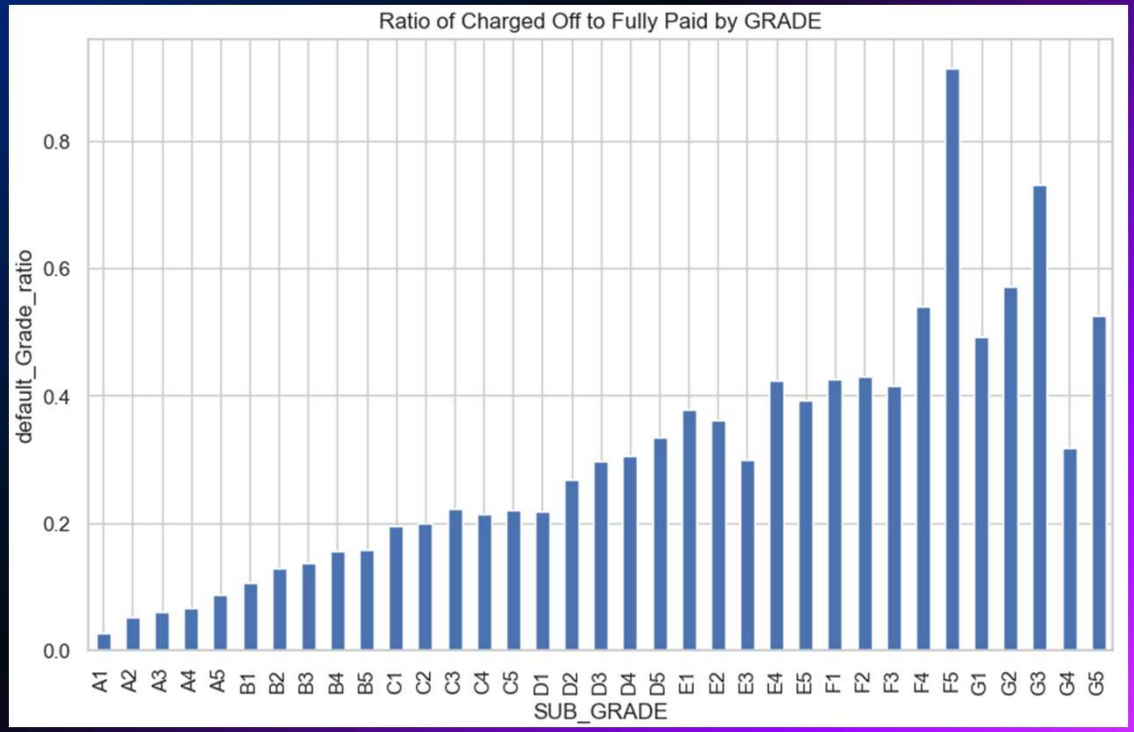
- Selling Loans by grade is the best way to be profitable
- **Grade A** has least default compared to other grades



SEGMENTED - UNIVARIATE ANALYSIS.

RATIO OF CHARGED OFF TO FULLY PAID BY SUB_GRADE

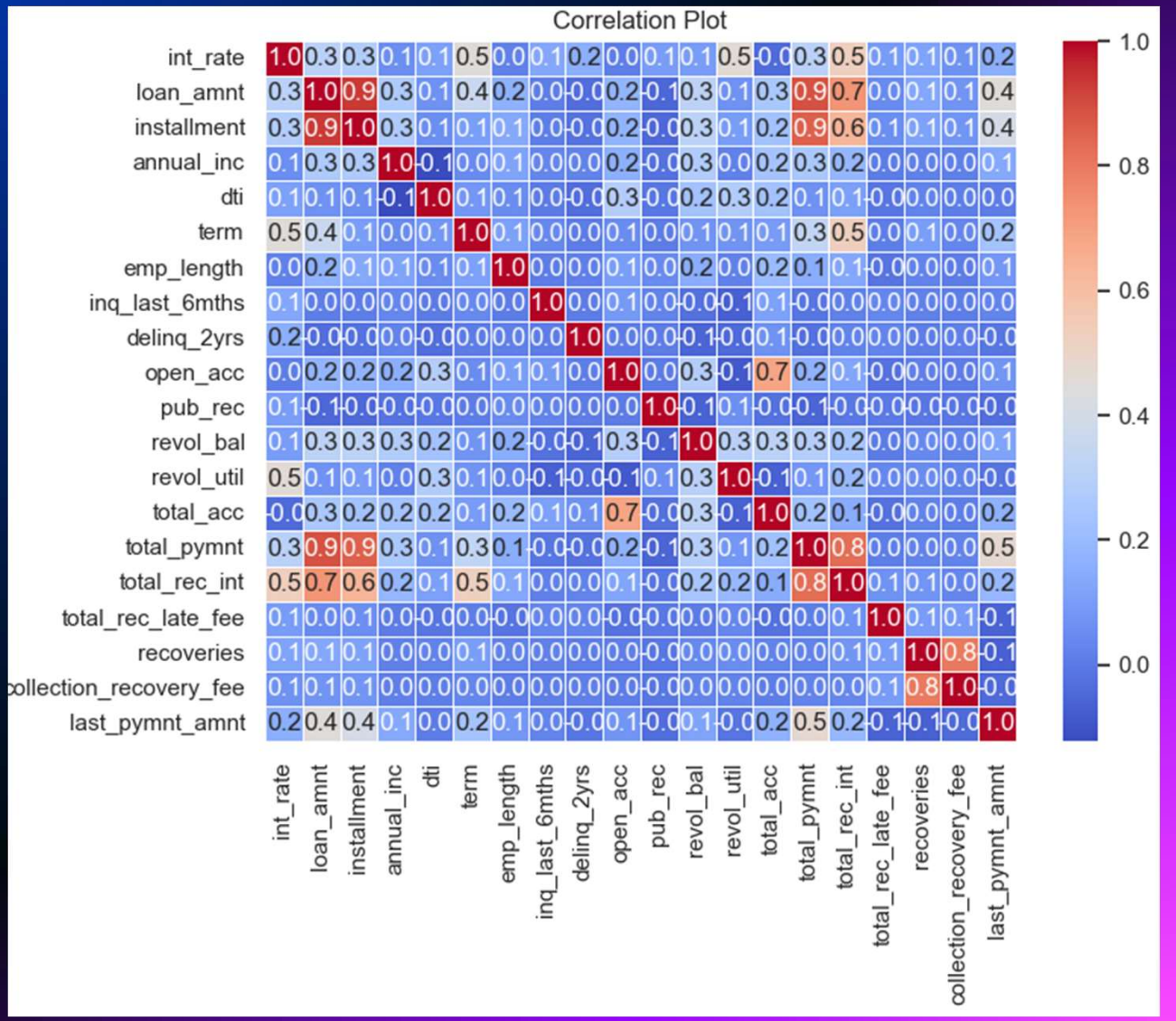
- Selling Loans by sub_grade is the best way to be profitable
- **However SubGrades F,G series are risky.**
- **F5 is the most risky sub_group.**



CORRELATION PLOT FOR NUMERIC FIELDS.

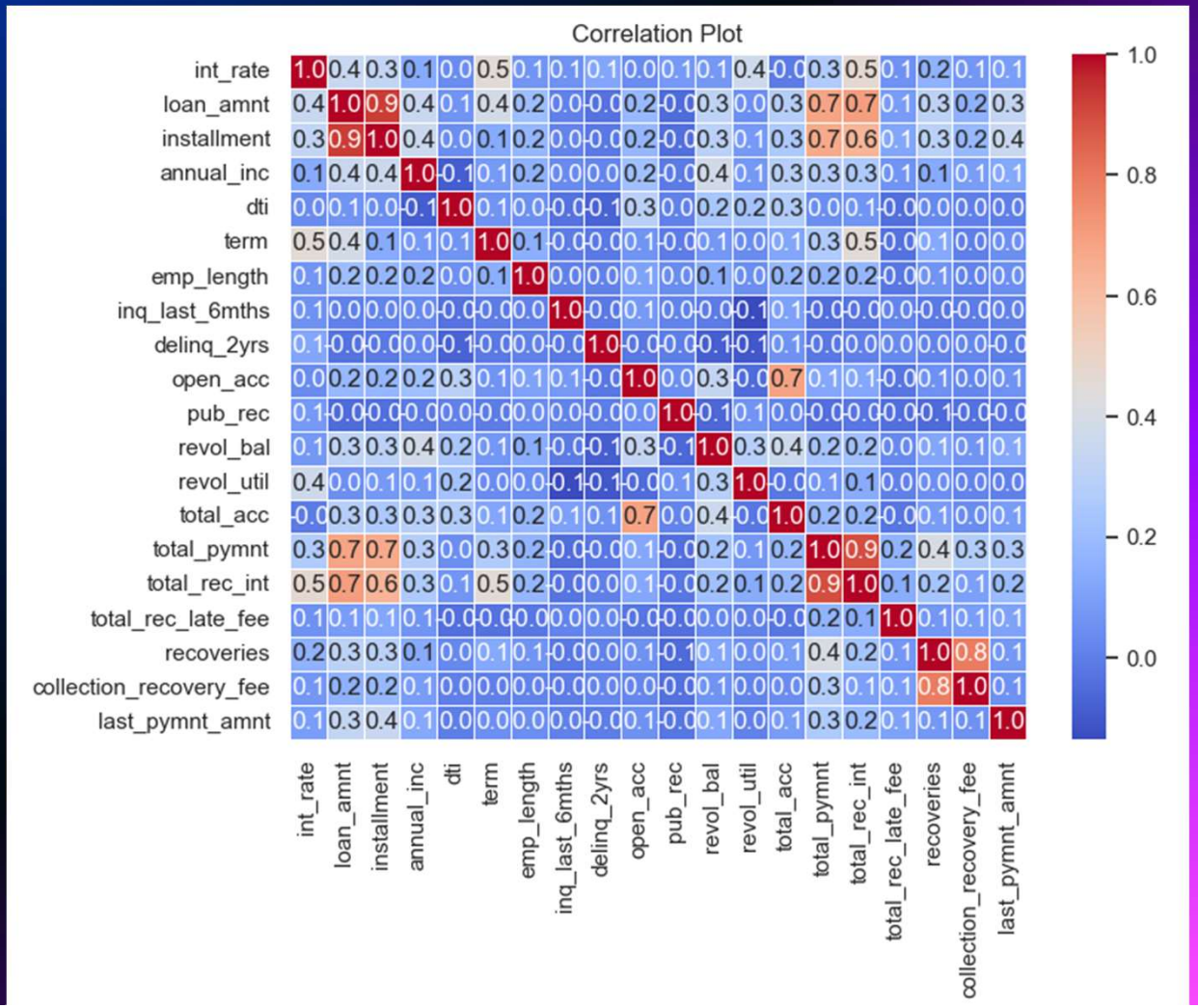
- We can notice the dynamics of the data and feel the correlation shows appropriate moves in related fields.
  Example: loan_amnt vs installment.

- This shows that the data quality of the dataset is as expected.



CORRELATION PLOT FOR NUMERIC FIELDS. SUBSET: "CHARGED OFF"

- We can notice that correlation of some fields increases with the subset of Charged Off"



DATA INSIGHTS & RECOMMENDATIONS

- Selling Loans by grade is the best way to be profitable
  - **Grade A** has least default compared to other grades
  - Sub_Grade **F5 most risky category of the sub grades.**
- Most of the customers are paying the loans.

  82.96% - Fully Paid, 14.16% Charged Off and 2.87% are Current
- The Loan business looks to be slowing, as we see only 2.87% are currently paying.
- Top 2 reasons loans are taken are due to debt consolidation and to pay of credit cards.

DATA INSIGHTS & RECOMMENDATIONS

- Loans originating from states: CO,TX,PA,MA,VA,OH are comparatively at Low Risk.
- Loans originating from states: FL, GA, MO, CA, MD and NC are comparatively at High Risk.
- The risk of “Charge Off” slightly higher for emp_length in the band 1,5,6,7,8 and 10
- It is worth to **market** loans to **Own** segment as they are fully paid off on Mortgage and it is likely that they will “Fully Pai the loan.
- Surprisingly the 60year mean interest rate is grater than 30year loan. “**Long term loans are risky**” 14.8% vs 11% for 30y.

THANK YOU!!

Narayana Isanaka