

LENDING CLUB CASE STUDY

Analysis by:

- ❖ Mohammad Jawad Amina
- ❖ Karan Khosla

Purpose of this Case Study:

- This Lending club case study aims to give an idea of applying EDA in a real bank loan scenario. In this case study, we will develop a basic understanding of Charged Off Loan in banking services and understand how data is used to minimise the risk of losing money while lending to customers.
- Let's imagine when the bank receives a loan application, the bank have decided for loan approval based on the applicant's profile.
- Two types of risks are associated with the bank's decision:
 - Based on data and how it is related, we can deduce if the applicant is likely to repay the loan. If bank not approving the loan than the results is a loss of business to the company.
 - Suppose if the applicant is not willing to repay the loan, for example: he/she is likely to default. Then the approving the loan may lead to a financial loss for the bank.

→ Very first we have imported libraries which we are going to use in our entire analysis. The purpose to use these libraries are:

import pandas as pd → To work with dataset

import numpy as np → Math Library

import seaborn as sns → Graph Library that used matplotlib in background.

import matplotlib.pyplot as plt → To plot some parameters in seaborn.

→ Also, we have import warnings library, as it is triggering warning message on file size while uploading loan.csv.

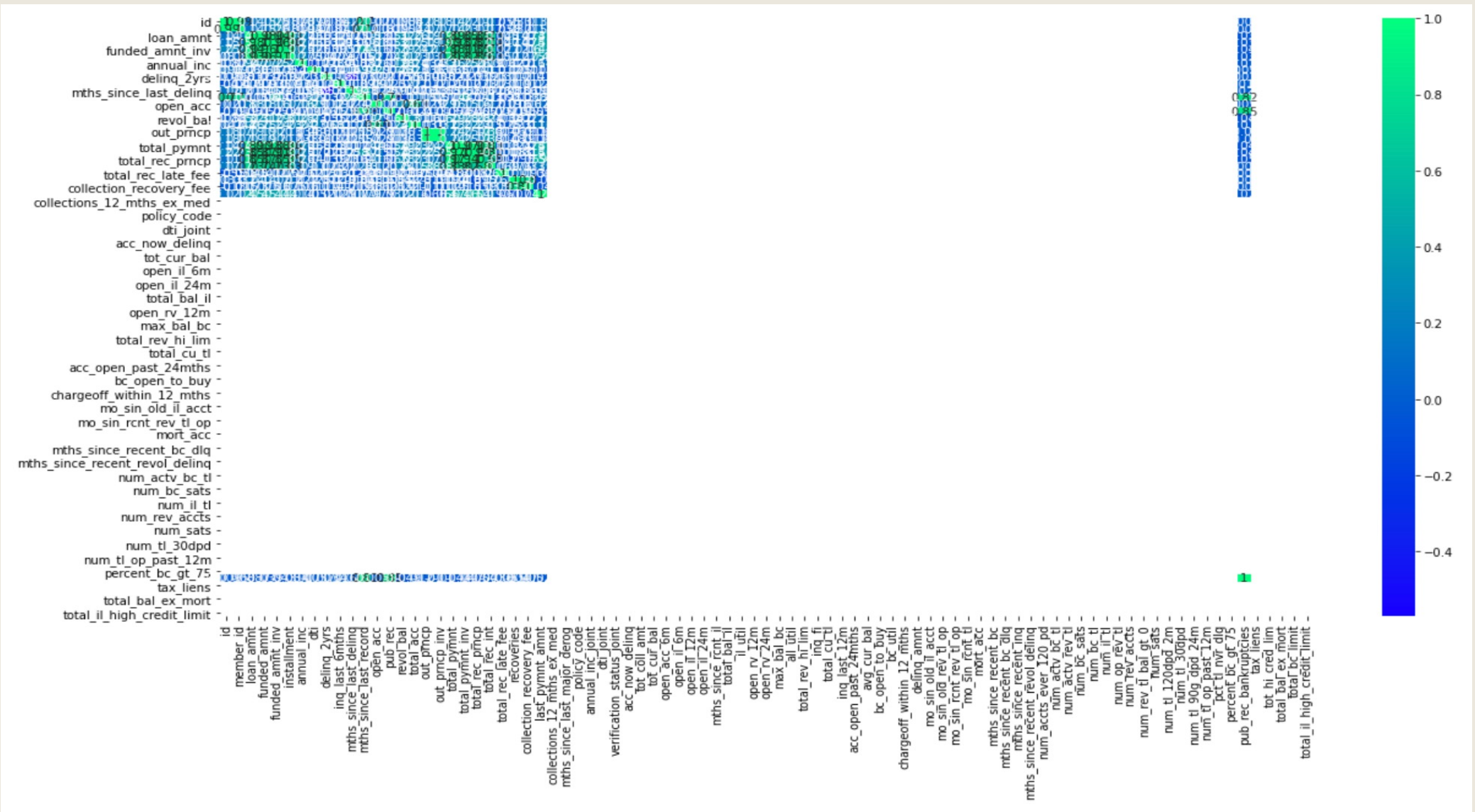
Import Python Libraries:

```
In [2]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Ignore Warnings:

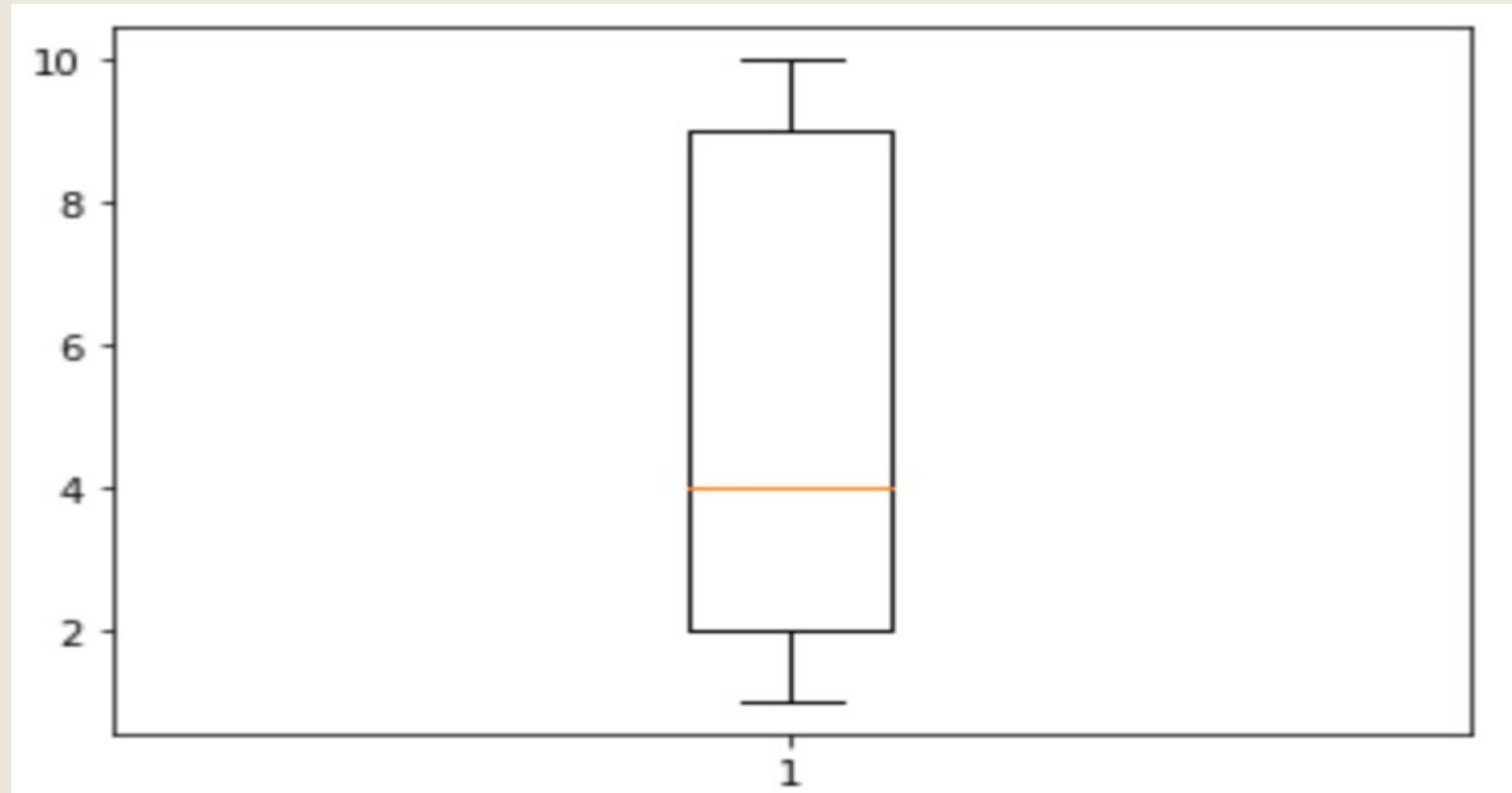
```
In [3]: import warnings
warnings.filterwarnings('ignore')
```

→ Loan data information through Heatmap, this is a RAW data information which is having entire column information.



→ While performing Data Cleaning process, we start working on “Interest Rate” (int_rate). Data type is in “Object” mode, where it’s had %. To utilise this column, very first we removed % and keep just numeric value. After that we have change data type to “float”.

→ Second data we start worked on “emp_length”, where we clear string value and convert data type from object to integer. While checking the “Null” value we have noticed that there are 1075 Null cell, we check the median value (4) and replace Null value with median value of “emp_length”. Below is the boxplot of emp_length where it showing median value.

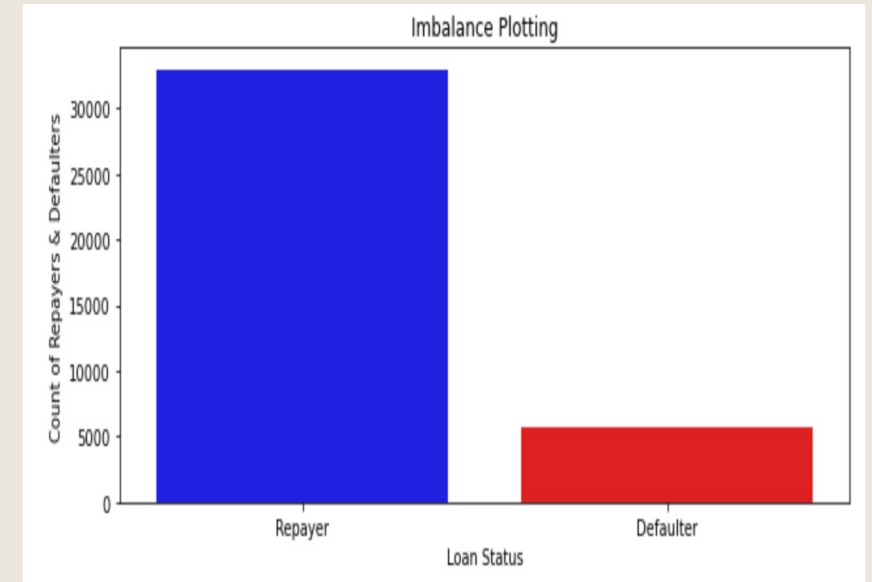
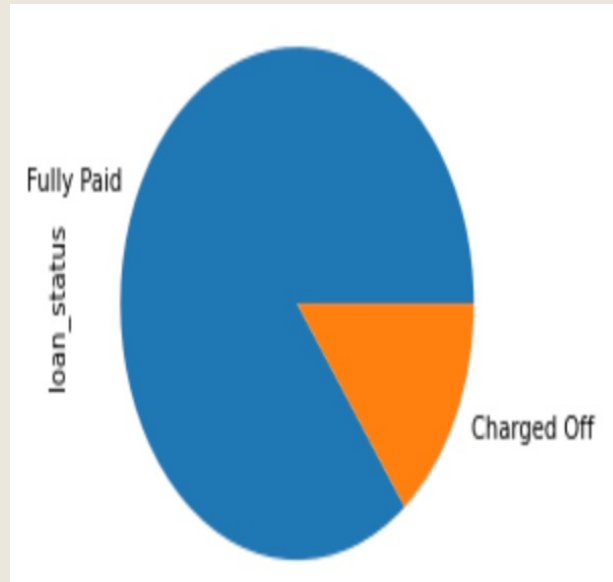
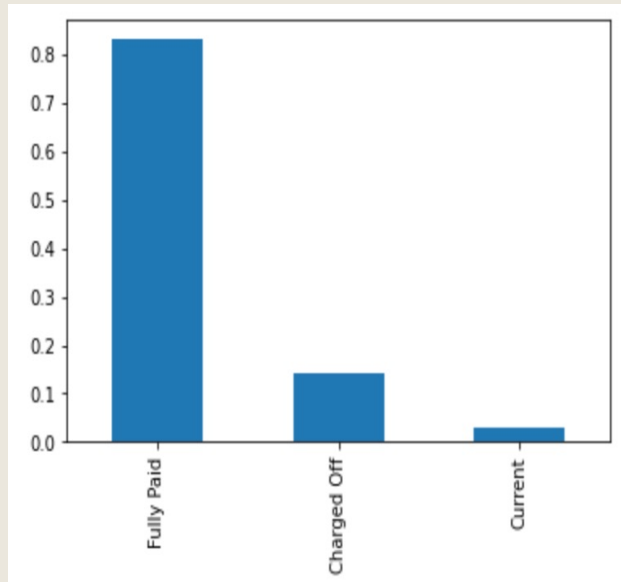


Univariate Analysis:

→ Loan status data (loan_status) which is our target data, in our analysis we have done comparison/correlation loan_status data with other data. Loan status data have three sub-categories (Fully Paid, Current & Charged Off). Please find the BAR chart on same.

→ We drop Current sub-category, as this information is for those customer which is currently running loan. We have shows in PIE chart which is having only two sub-categories (Fully Paid & Charged Off).

→ Also, you can see “Imbalance” analysis chart, which is representing the over all “Re-payer” and “Defaulter” customer.



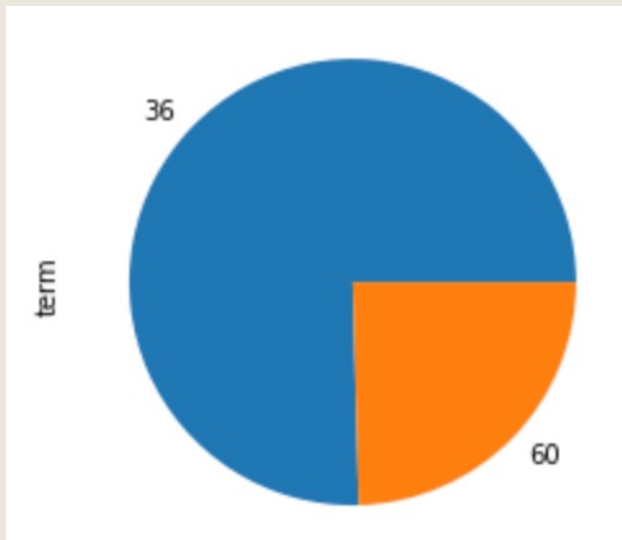
Bivariate Analysis:

→ For bivariate analysis, we did analysis target data “loan_status” with “term”, “int_rate”, “emp_length”, “dti” and “pub_rec_bankruptcies”. Let’s discuss one by one:

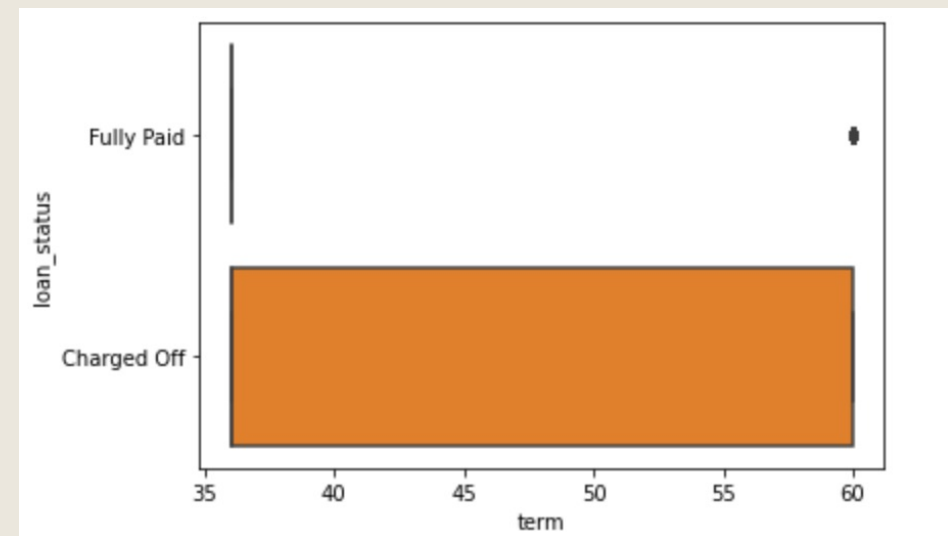
→ We did clean data of “term” as this is having string value and data type was “object”. Removed string value and keep it numeric, also we change data type to “float”.

→ Please find below ”Univariate Analysis” of “term” data and “Bivariate Analysis” using Boxplot for the “loan_status” & “term” data.

Univariate Analysis No. of Payments (term)



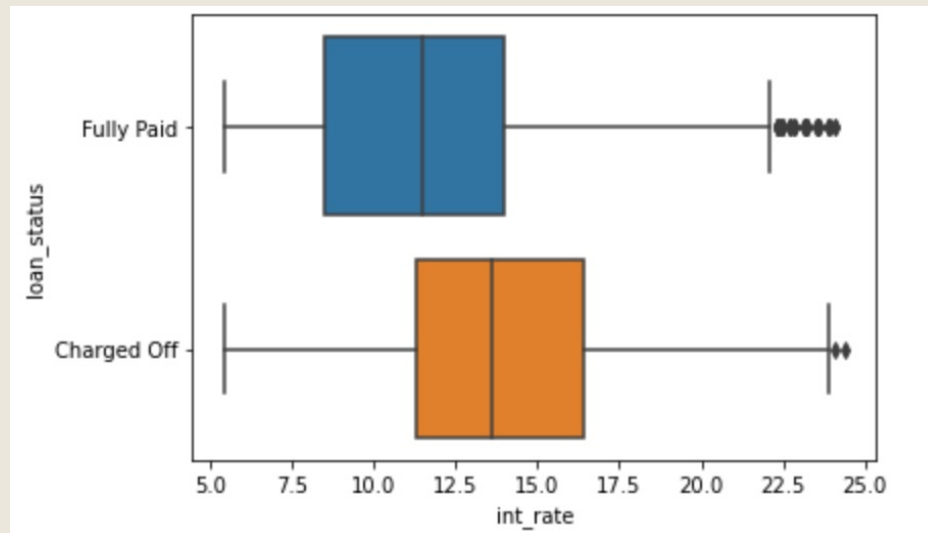
Bivariate Analysis Loan Status & Term



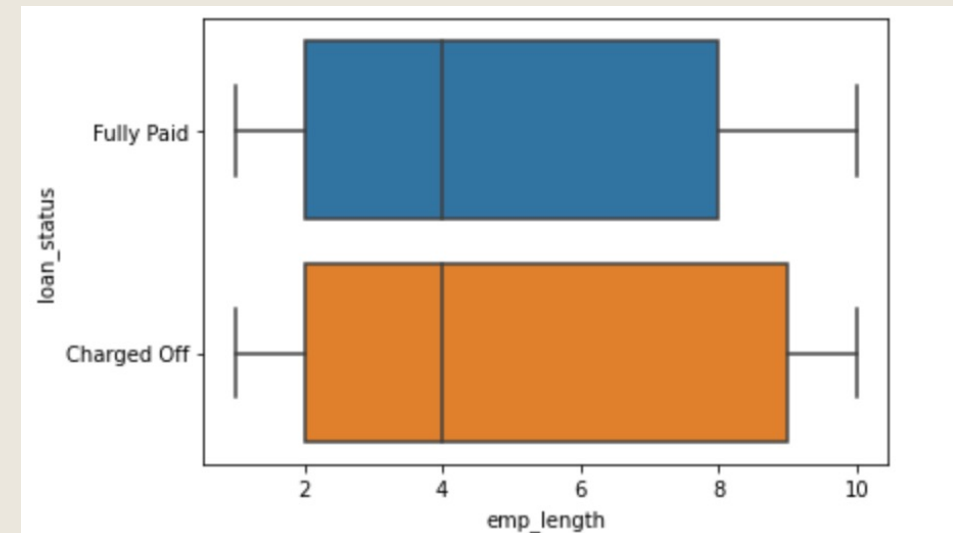
Bivariate Analysis:

- Another example of bivariate analysis, where we did analysis on target data “loan_status” with “int_rate” and “emp_length”.
- Analysis of interest rate and loan status. This will show the interest rates below 11% (approx) and interest rates have more fully paid status.
- Analysis between employee length and loan status. This will show the median is the same for both the status.
- It is very evident that more people in the higher employment length may not pay back the loan, the Q3 is higher for the charged off where employee length is between 8 to 10.

Loan Status & Interest Rate BOX Plot



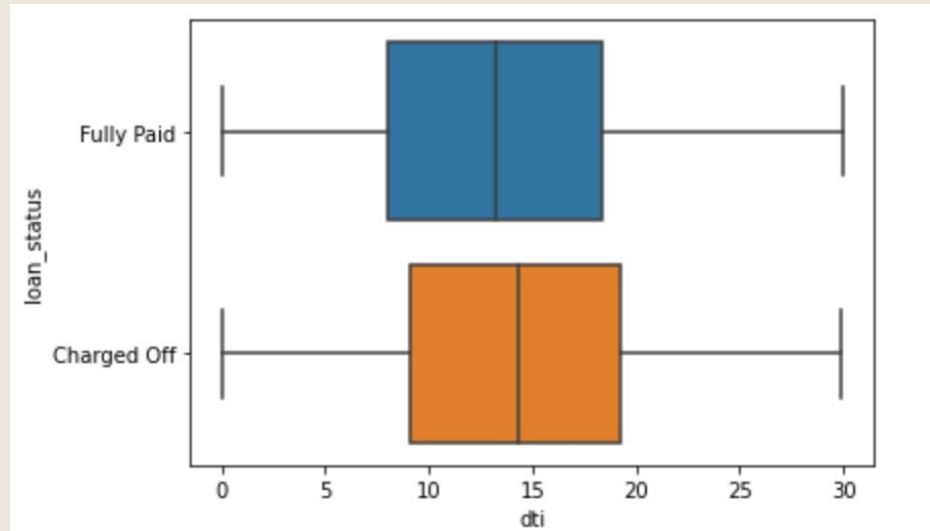
Loan Status & Employee Length BOX Plot



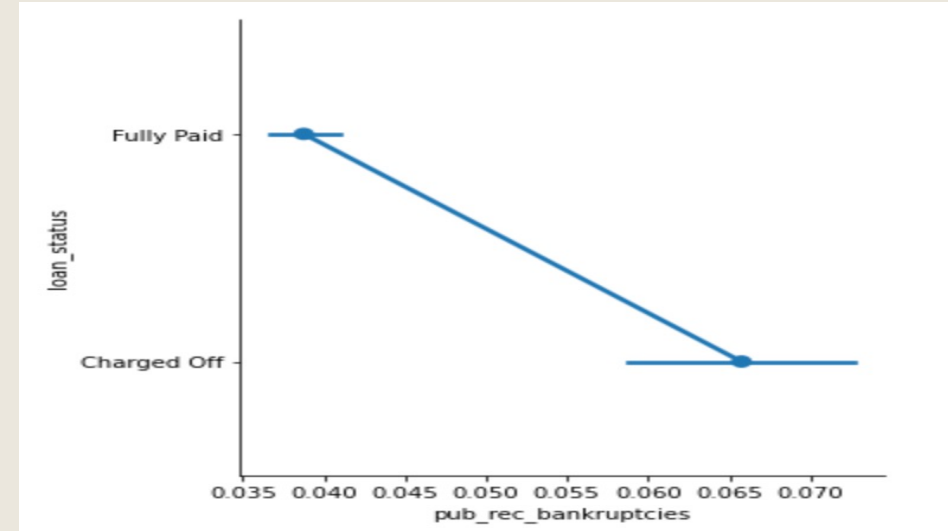
Bivariate Analysis:

- To further analysis, we did verify on target data “loan_status” with “dti” and “pub_rec_bankruptcies”.
- The values are higher for charged off where DTI is higher, this will prove that if the ratio between loan amount and monthly income is high then there is a slightly higher number (very little) of people who charged off the loan.
- Analysing the data between bankruptcies and loan status, this shows the loan status of fully paid is much higher when there are no bankruptcies.

Loan Status & DTI
BOX - Plot



Loan Status & Public Bankruptcies
CAT - Plot



Multi-variate Analysis:

→ As per below analysis, we can observe that the funded amount tends to increase the instalments which is rightly done.

→ We also observe that, as DTI does not have a correlation with factors such as instalments and “int_rate”. However, this would be a recommendation to increase the instalments. As the DTI decreases to improve the chances to get back the loan.

	installment	pub_rec	open_acc	dti	int_rate	funded_amnt	emp_length
installment	1.000000	-0.045706	0.172893	0.052038	0.277203	0.958035	0.122680
pub_rec	-0.045706	1.000000	0.000028	-0.004742	0.098635	-0.050576	0.060314
open_acc	0.172893	0.000028	1.000000	0.287849	0.006022	0.175682	0.096595
dti	0.052038	-0.004742	0.287849	1.000000	0.110913	0.062194	0.048340
int_rate	0.277203	0.098635	0.006022	0.110913	1.000000	0.304930	0.001998
funded_amnt	0.958035	-0.050576	0.175682	0.062194	0.304930	1.000000	0.148882
emp_length	0.122680	0.060314	0.096595	0.048340	0.001998	0.148882	1.000000

```

<bound method NDFrame._add_numeric_operations.<locals>.median of 0      B2
1      C4
2      C5
3      C1
5      A4
..
39712   A4
39713   C1
39714   A4
39715   A2
39716   E2
Name: sub_grade, Length: 38577, dtype: object> <bound method Series.mode of 0      B2
1      C4
2      C5
3      C1
5      A4
..
39712   A4
39713   C1
39714   A4
39715   A2
39716   E2
Name: sub_grade, Length: 38577, dtype: object>

```

→ Using Univariate method we find that the median and mode of the data is always B2, hence most of these loans in our data set are given to grade B and subgrade B2.

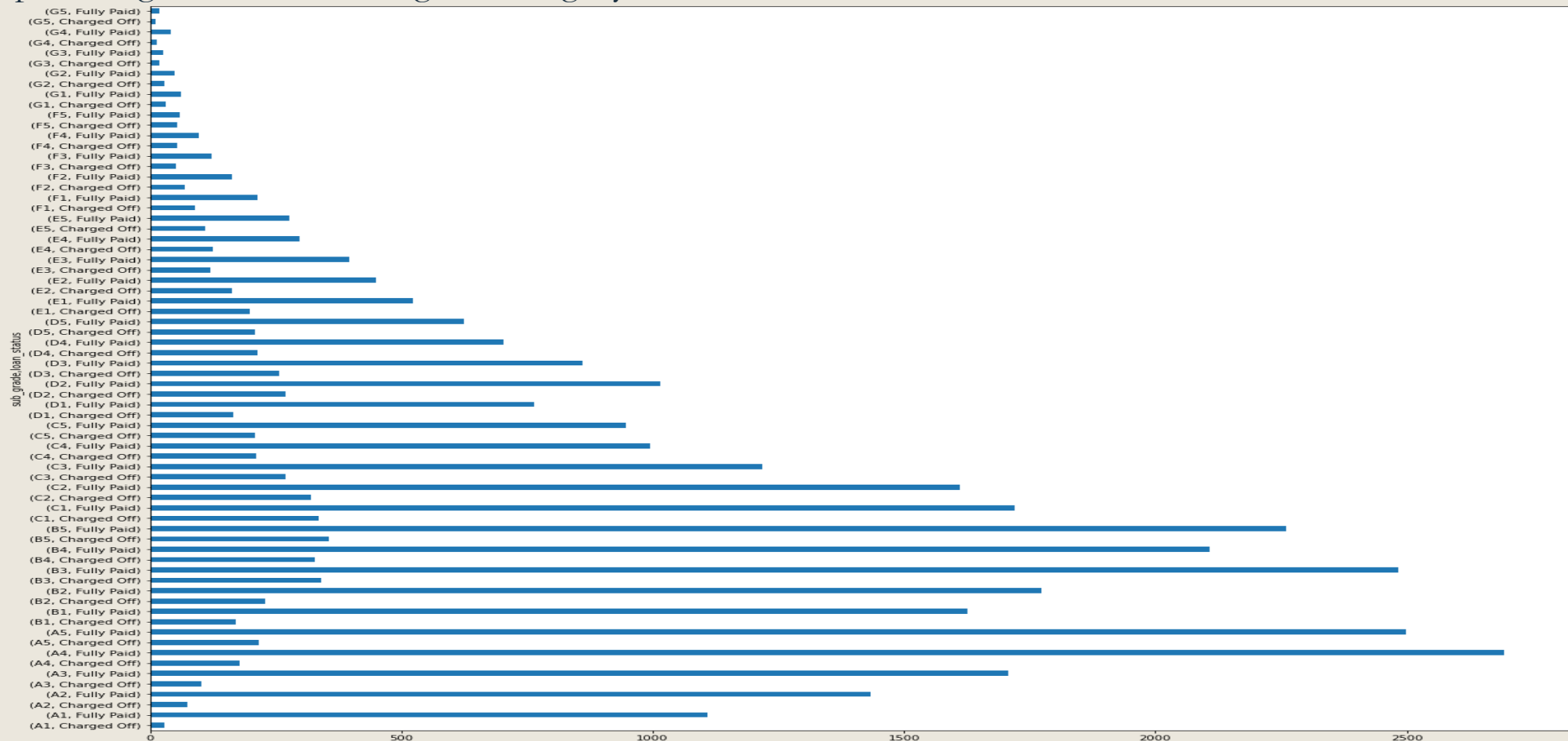
→ Using “sub_grade” we have found a count of individual subgrades of Fully Paid and Charged Off.

sub_grade	loan_status	
A1	Charged Off	30
	Fully Paid	1109
A2	Charged Off	74
	Fully Paid	1434
A3	Charged Off	103
	Fully Paid	1707
A4	Charged Off	178
	Fully Paid	2695
A5	Charged Off	217
	Fully Paid	2498
B1	Charged Off	171
	Fully Paid	1626
B2	Charged Off	228
	Fully Paid	1773
B3	Charged Off	341
	Fully Paid	2484
B4	Charged Off	329
	Fully Paid	2108
..		

Analysis based on Sub-Grades:

→ Hence, we can say the maximum loans that got fully paid are from the sub-category - A4 and the maximum that got charged off are from B5.

→ We also deduce that Sub-Grade B3 – B5, C1 – C2 and D1 – D2 might be calculated in a way that involves lot of risk. The reason to say this is because we find that the defaulters are increasing in this range and the fully paid loans are also dropping, we may need to make a preventive measure such as higher collateral while providing loans to this sub-grade category.



Recommendation:

- We observe that, as DTI does not have a correlation with factors such as instalments and “int_rate”. However, this would be a recommendation to increase the instalments where the DTI is low, as the DTI decreases can improve the chances to get back the loan without default.
- We also deduce that Sub-Grade B3 – B5, C1 – C2 and D1 – D2 might be calculated in a way that involves lot of risk. The reason to say this is because we find that the defaulters are increasing in this range and the fully paid loans are also dropping, we may need to make a preventive measure such as higher collateral while providing loans to this sub-grade category.
- Based on employee length we observe that the customer who have longer employee length does not able to paid back loan. As the recommendation we would suggest to give lower amount of loan to these type of customers.
- While analysis of interest rate and loan status. This will show the interest rates below 11%(approx) and interest rates have more fully paid status.
- In the “term” data we have noticed that the loans have 36 months as a term has more chances of getting fully paid.



Thank you