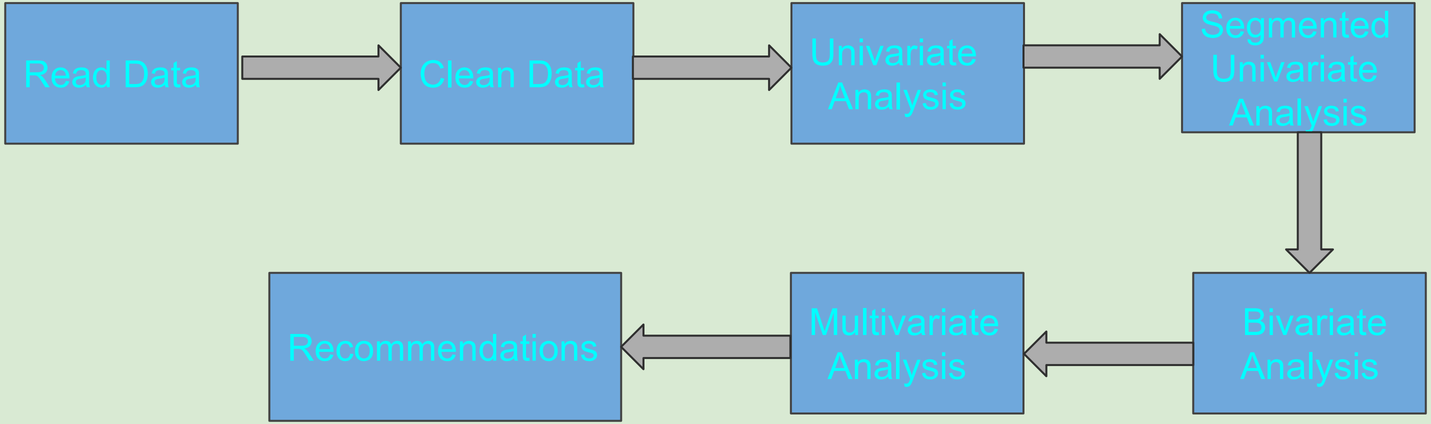
Lending club case study

Group members :

1. Raj Huligal
2. Rakesh Bhadra

Methodology



Problem Statement

The company wants to understand the driving factors and the variables which are strong indicators of loan default, so that the company can utilise this knowledge for its portfolio and risk assessment.

The data that has been provided are :

1) Loan Data
2) Loan Data Dictionary

Read Data & Clean Data

1. Read data from the loan data file in a dataframe.
2. Check for na values in each column.
3. Drop all columns where there are 40 percent of na values.
4. Check for the columns whose values needs special treatment like removing certain symbols or a specific text and clean such columns.
5. Check if all columns have proper data type, and if required convert the columns needed to the required data type.
6. Fill “na” values with any specific string if required for special columns. eg: Desc column with “Unknown” for “na” values.
7. Now for all the columns which still have “na” values , replace them with median for int and float data type columns, and mode for string and object data type columns.

Read Data & Clean Data (Continued..)

8. Drop if any duplicate rows from the dataframe.
9. Check for “na” values and make sure there are no such values.
10. Remove if any duplicate rows present.
11. Remove rows for columns which have outliers. Since it is financial and loan data, we have removed outliers only in specific columns like loan amount , annual income.
12. Now the data is ready for analysis and generate any required dataframes using various conditions.

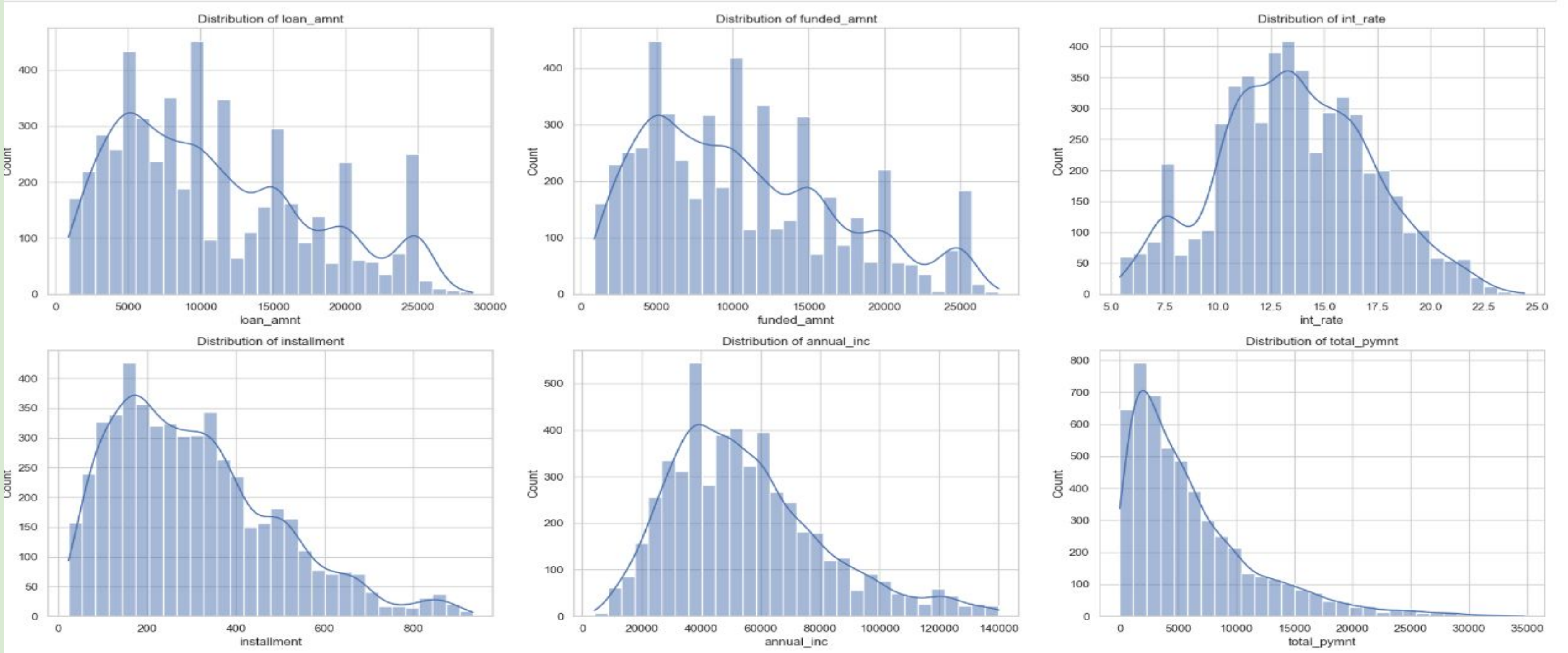
Univariate Analysis

1. Make a list of numerical columns and generate a hist plot for them which is based on counts
2. Make a list of categorical columns and generate a count plot for them which is based on counts.

Univariate Analysis for Numerical Columns

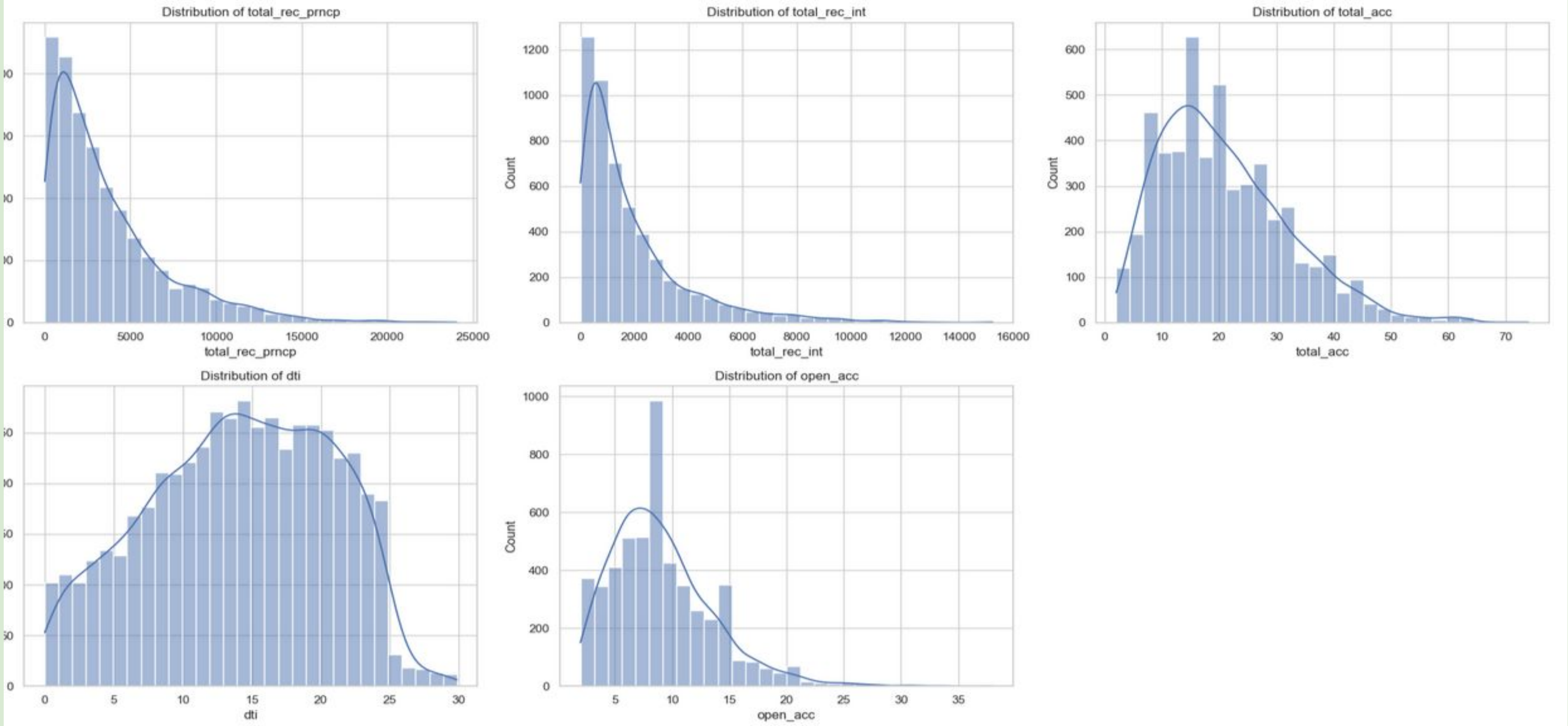
1. It seems lower amount loans are defaulting more for charged-off loans.

2. It seems that annual income less than 80000 are defaulting more.



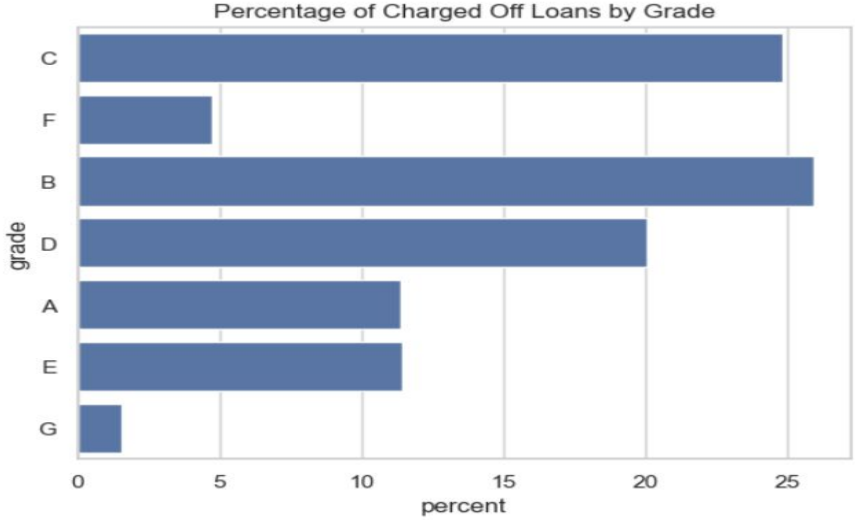
Univariate Analysis for Numerical Columns (Continued..)

1. It seems the dti is in the range of 7 to 24 for defaulted loans.

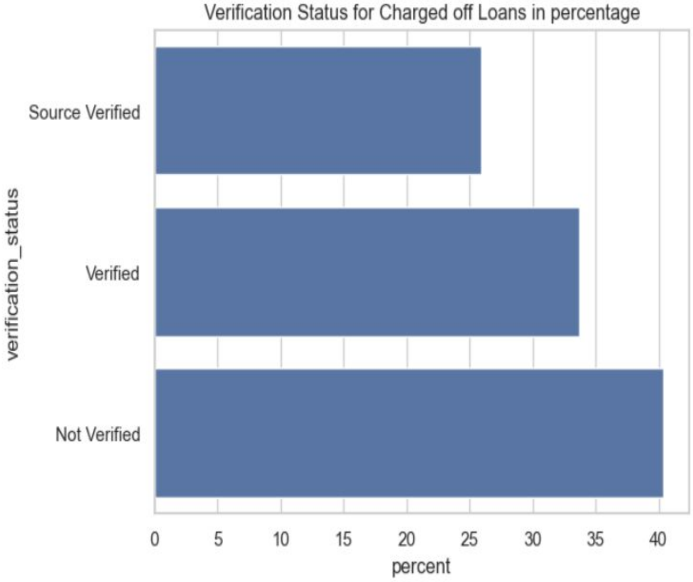


Univariate Analysis for Categorical Columns

1. Grade - B (25.9%), C (24.85%), D (20.04%) loans contain about 70% of the defaulted loans.
2. About 40% of the defaulted loans are not verified.



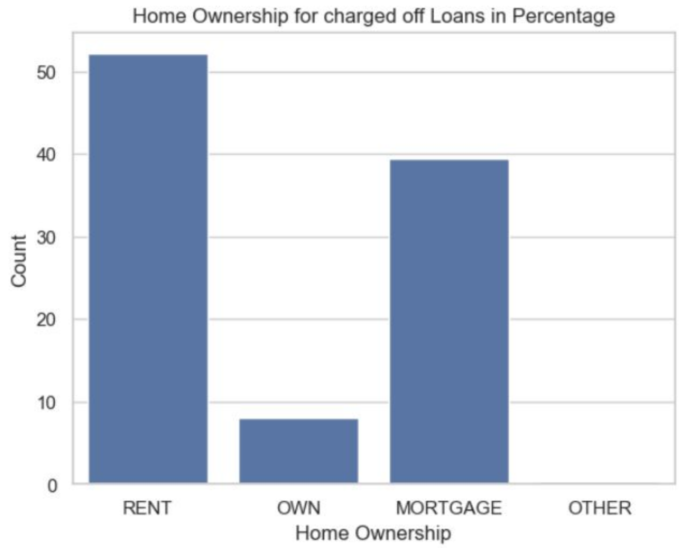
```
Percentage of Charged Off Loans Count by Grade
grade
B     25.930928
C     24.850473
D     20.046305
E     11.460544
A     11.383369
F      4.765580
G      1.562801
```



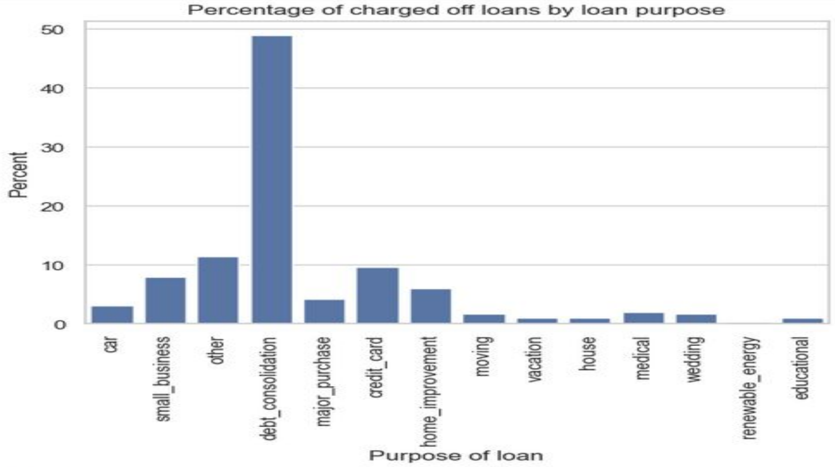
```
Verification Status for Charged off loans in Percentage
verification_status
Not Verified       40.401312
Verified           33.687054
Source Verified    25.911634
```

Univariate Analysis

1. Most of the defaulted loans have home ownership as Rent - 52.13 percent followed by Mortgage - 39.47 percent.
2. The loan purpose of most of the defaulted loans is debt consolidation (48.8 percent) followed by other which is (11.46 percent)



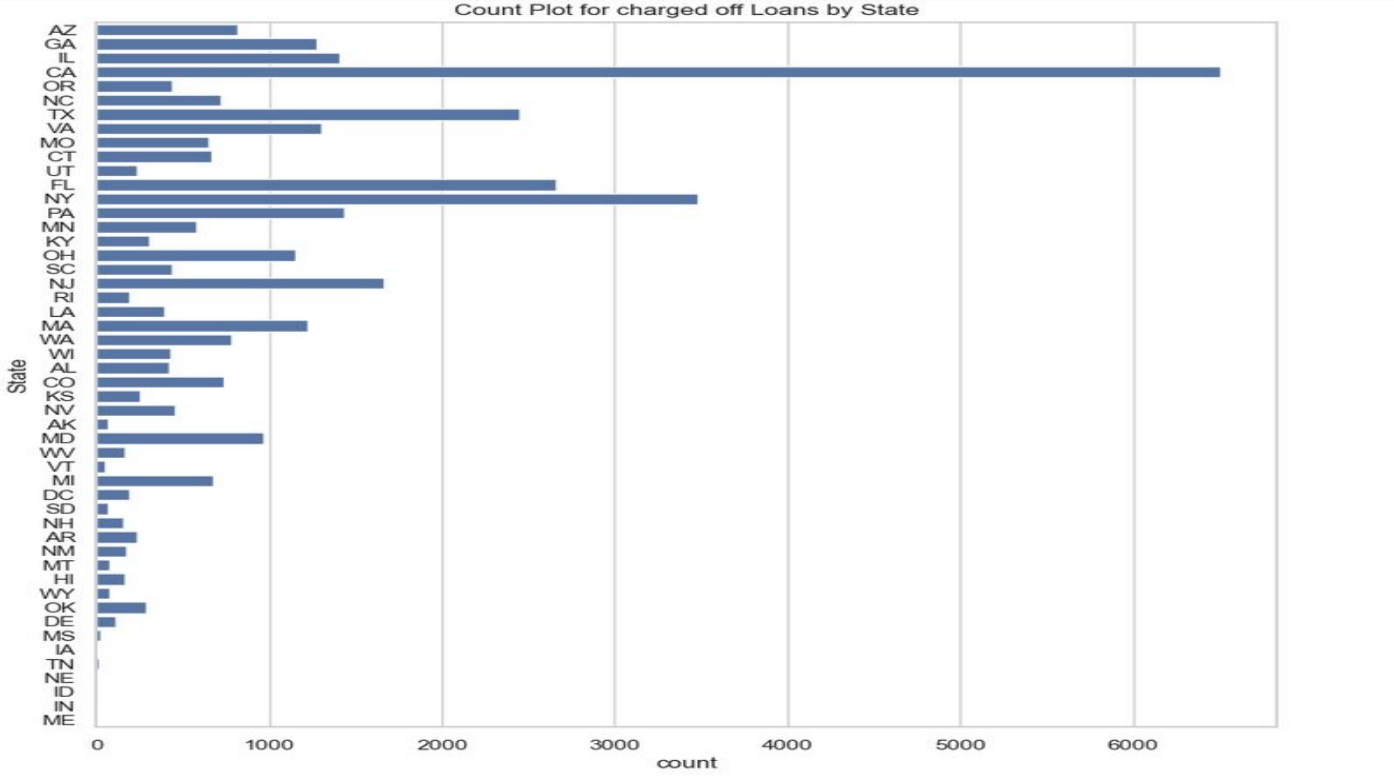
```
home_ownership
RENT          52.131970
MORTGAGE      39.475207
OWN            8.064827
OTHER          0.327995
```



```
purpose
debt_consolidation    48.890604
other                 11.460544
credit_card            9.666216
small_business         7.987652
home_improvement       5.981092
major_purchase         4.148177
car                    3.067721
medical                1.948678
moving                 1.736446
wedding                1.736446
educational            1.041868
vacation               1.003280
house                  0.983986
renewable_energy       0.347289
```

Univariate Analysis

1. Most of the defaulted loans have originated in CA state, followed by NY, FL and TX.



Univariate Analysis

Univariate Analysis

1. Make a list of numerical columns and generate a hist plot for them which is based on counts
2. Make a list of categorical columns and generate a count plot for them which is based on counts.

Univariate Analysis

1. Make a list of numerical columns and generate a hist plot for them which is based on counts
2. Make a list of categorical columns and generate a count plot for them which is based on counts.

Univariate Analysis

1. Make a list of numerical columns and generate a hist plot for them which is based on counts
2. Make a list of categorical columns and generate a count plot for them which is based on counts.

Univariate Analysis

1. Make a list of numerical columns and generate a hist plot for them which is based on counts
2. Make a list of categorical columns and generate a count plot for them which is based on counts.

Final Conclusion

1.