Lending Club Case Study

TEAM – PIYUSH CHOUDHARI

ABHISHEK SHARAN

Problem Statement

Lending club is currently facing challenges with their loan approval/declines process. This case study involves performing exploratory data analysis (EDA) to identify patterns in consumer attributes and loan characteristics that might be strong indicators of loan default. Based on the loan default patterns, we can take business decisions minimizing the risk associated with the loan.

There are majorly two types of risks that can be minimized:

1. When the customer will not default on the loan but the loan is not approved - risk losing revenue

2. WHen the customer will default on the loan but the loan is approved - risk of losing capital

Business Objective and data understanding

Loan default is the largest financial loss to a financial lending company. Our objective is to analyze customer and loan attributes and identify patterns that indicate a strong possibility of default.

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

In the loan.csv file, each row corresponds to the different loan and customer attributes during a loan application. On initial look, customer attributes such as annual income and loan attributes such as interest rate, term seem to inituitively govern if the customer will default on the loan.

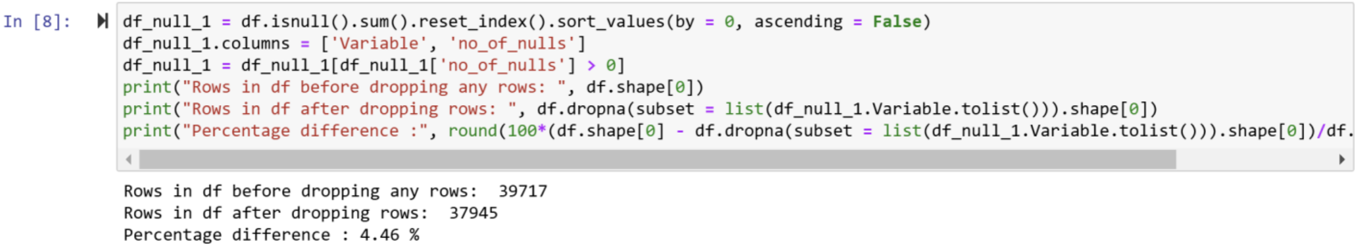
Data Cleaning and preparation

The following steps were taken in the data cleaning and preparation stage:

1. On checking null values, many columns with null values > 60% were found. All these columns were directly dropped from the data.

2. From the remaining data, some redundant columns were dropped and rows with null values were dropped retaining ~96% of the data.

3. Further, some more of the redundant and unecessary columns such as member_id, desc etc were dropped to clean the dataset for understanding.



```
df_null_1 = df.isnull().sum().reset_index().sort_values(by = 0, ascending = False)
df_null_1.columns = ['Variable', 'no_of_nulls']
df_null_1 = df_null_1[df_null_1['no_of_nulls'] > 0]
print("Rows in df before dropping any rows: ", df.shape[0])
print("Rows in df after dropping rows: ", df.dropna(subset = list(df_null_1.Variable.tolist())).shape[0])
print("Percentage difference :", round(100*(df.shape[0] - df.dropna(subset = list(df_null_1.Variable.tolist())).shape[0])/df.
```

```
Rows in df before dropping any rows:  39717
Rows in df after dropping rows:  37945
Percentage difference : 4.46 %
```

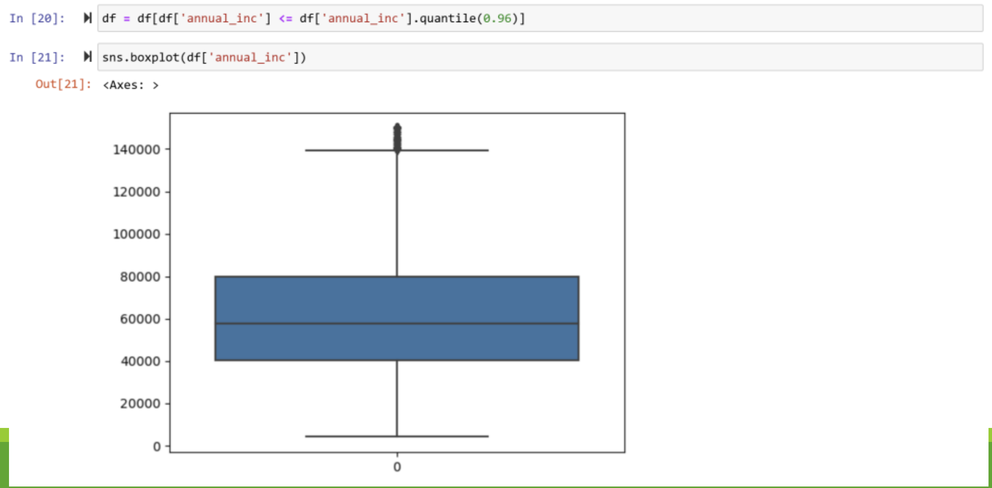
Feature Extraction, Data Conversion and outlier treatment

1. The columns “term” and “int_rate” were cleaned by performing the necessary string operations on the columns

2. The issue date column was converted to the pandas datetime format for smoother analysis.

3. The target variable, loan status, was converted to a 1 or 0 flag using the logic – 1 if loan status is “Charged Off” else 0

4. There were outliers found in the annual income column. To deal with this, all records with annual_income above 96th percentile of annual income were removed:

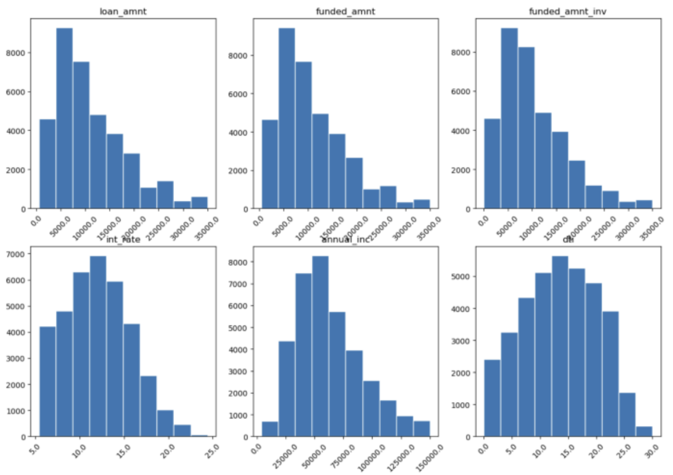


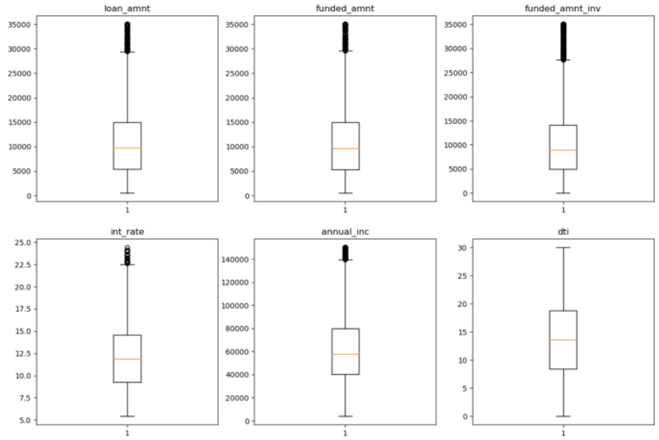
```
In [20]:  df = df[df['annual_inc'] <= df['annual_inc'].quantile(0.96)]

In [21]:  sns.boxplot(df['annual_inc'])
Out[21]: <Axes: >
```

Univariate Analysis – Box Plots and histogram,

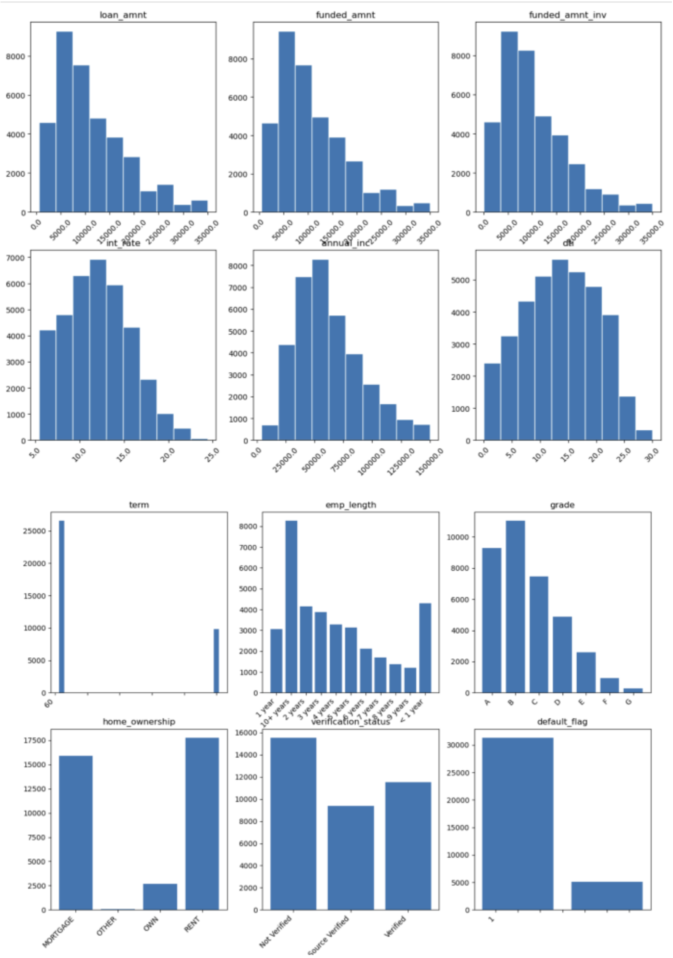
- The median loan amount is around 9500 indicating that the data has few rows with high loan amounts.
- The median interest rate seems to be between 11 and 12.5
- The curves above almost look like a bell curve with a long tails towards the right.
- Most customers have theur incomes between 25K to 60K. With fewer customers with high income (100K+)
- Most customers have received an interest rate between 11 - 12.5%
- Majority of the customers have taken a loan amount between 5K - 11K





Univariate Analysis

- The curves above almost look like a bell curve with a long tails towards the right.
- Most customers have theur incomes between 25K to 60K. With fewer customers with high income (100K+)
- Most customers have received an interest rate between 11 - 12.5%
- Majority of the customers have taken a loan amount between 5K - 11K
- Majority loans are in the data are of grade B
- Most loan applicatants have 10+ years of employment. With < 1 year being the second highest number.
- Very few loan applicants own a home. Majority are on Rent or are paying mortgage for their homes.
- **Most loans are taken for debt_consolidation**. I.E to combine all loans into a single loan with objective of having a lower interest rate.
- In each year, the highest number of loans were issued in **Q4**



Bivariate Analysis - Quantitative

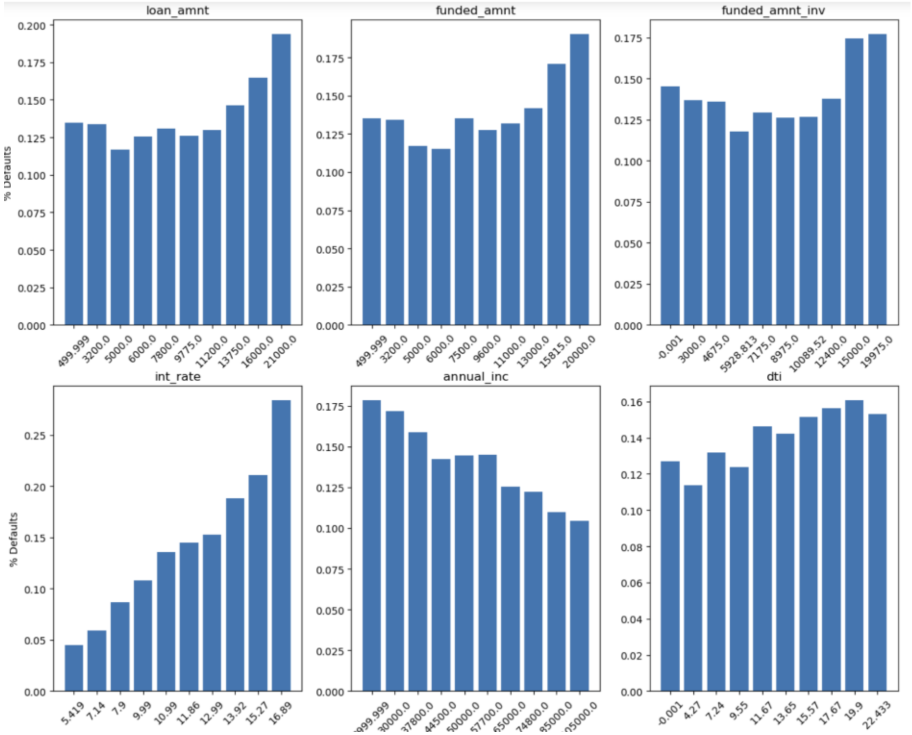
On checking the loan Default percentage (Total defaults/Total Application)*100 in a bin, some really interesting trends were spotted:

- Defaults are higher in customers with low annual income VS customers with higher annual income.

- Defaults are high when the intest rate is high.

- The debt to income ratio seems to be a good predictor of defaults

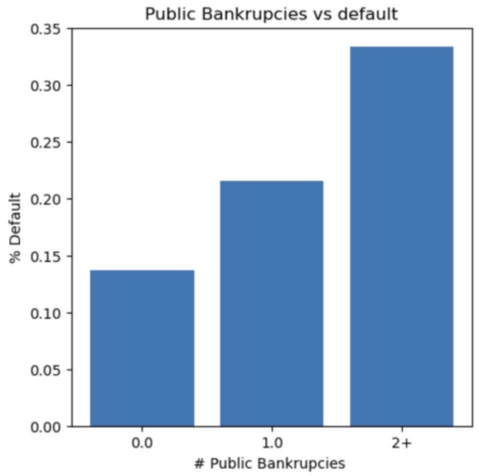
- Intuitively, the higher the loan amounts, higher are the defaults.

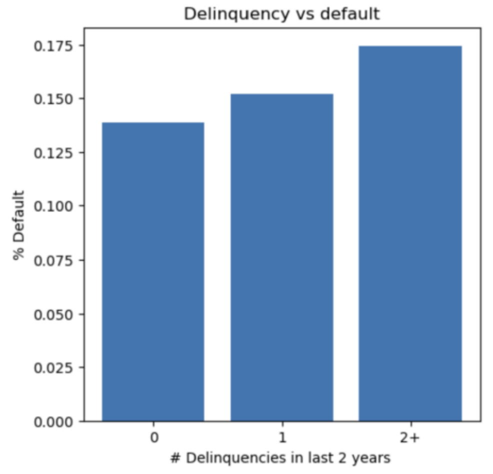


Bivariate Analysis – Public Bankrupcies and Delinquency

- On creating variables for Public Bankrupcies, we found the customers with 2+ instances of bankruptcy, have ~35% chance of default as compared to just ~15% with 0 instances of bankruptcies.

- Similarly, Customers who were delinquent of their other loans in the past two years 2+ times have a higher probability of default





Summary of all the findings

From the **univariate analysis**, we saw that:

- - Most loan Amounts are between 5K to 10K
- - Most of the customers have an annual income of 50K - 60K
- - Customers usually get an interest rate of 11% - 13% on their loans
- - The most common reason why the loan is taken is Debt Consolidation
- - Majority of our customers are from California, and on the second place we have New York, Indicating that our customers are highly concentrated in states with big cities.
- - Most of the loans are taken in Q4 of the year

Findings from the **Bivariate analysis**:

- - There is a high Default percentage when the loan amount is high.
- - Similarly, the default percentage goes up when the interest rate increases.
- - As expected, the default % decreased with increase in income. Meaning, low income customers have a higher probability of defaulting on our loans.
- - Also, Customers with high Debt-to-income ratio have a higher probability of default.
- - Loans taken for Small Business have the highest probability of default.
- - Looking at the location, customers from nevada have a higher probability of default.
- - We also saw that as the public bankrupcies are more than 2, there is a 34% chance that the customer will default on their loan
- - Similarly, customers who were delinquent on their payments 2+ times in the last two years have a higher probability of loan defaults.

Conclusion

In this analysis we have used Loan and customer attributes to profile a customer and identify variables that impact the chances of default.

- Some of the customer attributes we found have an imnpact were - annual income, dti, bankrupcies, delinquncies on other loans location etc.
- Some of the Loan attributes which have a high chance of default are interest rates, term, Loan amount and the purpose of the loan.

Based on these we can profile prospective customers and take the following decisions:

- Reducing the loan amount
- Increasing/Decreasing the interest rate
- Declining the loan