LENDING CLUB CASE STUDY

EXPLORATORY DATA ANALYSIS ON LENDING INSTITUTION DATA

1 EXECUTIVE SUMMARY

2 APPROACH

3 UNIVARIATE ANALYSIS
NUMBERICAL

4 UNIVARIATE ANALYSIS
CATEGORICAL

5 BIVARIATE ANALYSIS

6 MULTI VARIATE ANALYSIS

7 SUMMARY
CONCLUSION

SUBMITTED BY :
HARSHADA ROHAN KALSEKAR
IRSHAD AHAMED

EXECUTIVE SUMMARY

Introduction

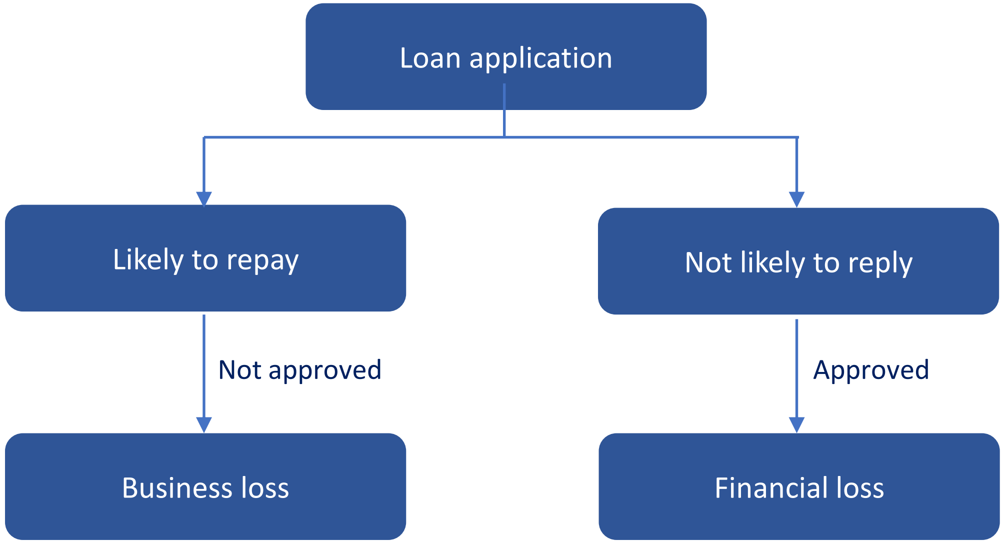
Company details :

Lending club is the consumer finance company which specializes in lending various types of loans to urban customers. LC is the largest online loan marketplace, facilitating Personal loans, Business loans, and Financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Business understanding:

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:

1. If the applicant is likely to repay the loan, then not approving the loan results in a **loss of business** to the company
2. 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



Loan application

Likely to repay

Not likely to reply

Not approved

Approved

Business loss

Financial loss

Objective

Customer focus – Identify risky applicants to reduce credit loss.

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Derive how consumer attributes and loan attributes influence the tendency of default.

Approach

Customer have provided a data set for all loans issued though the time period 2007 to 2011.
Selected approach is to use Exploratory Data Analysis (EDA) on the provided dataset , to understand the drivers of loan defaults to improve risk assessment and decision making.

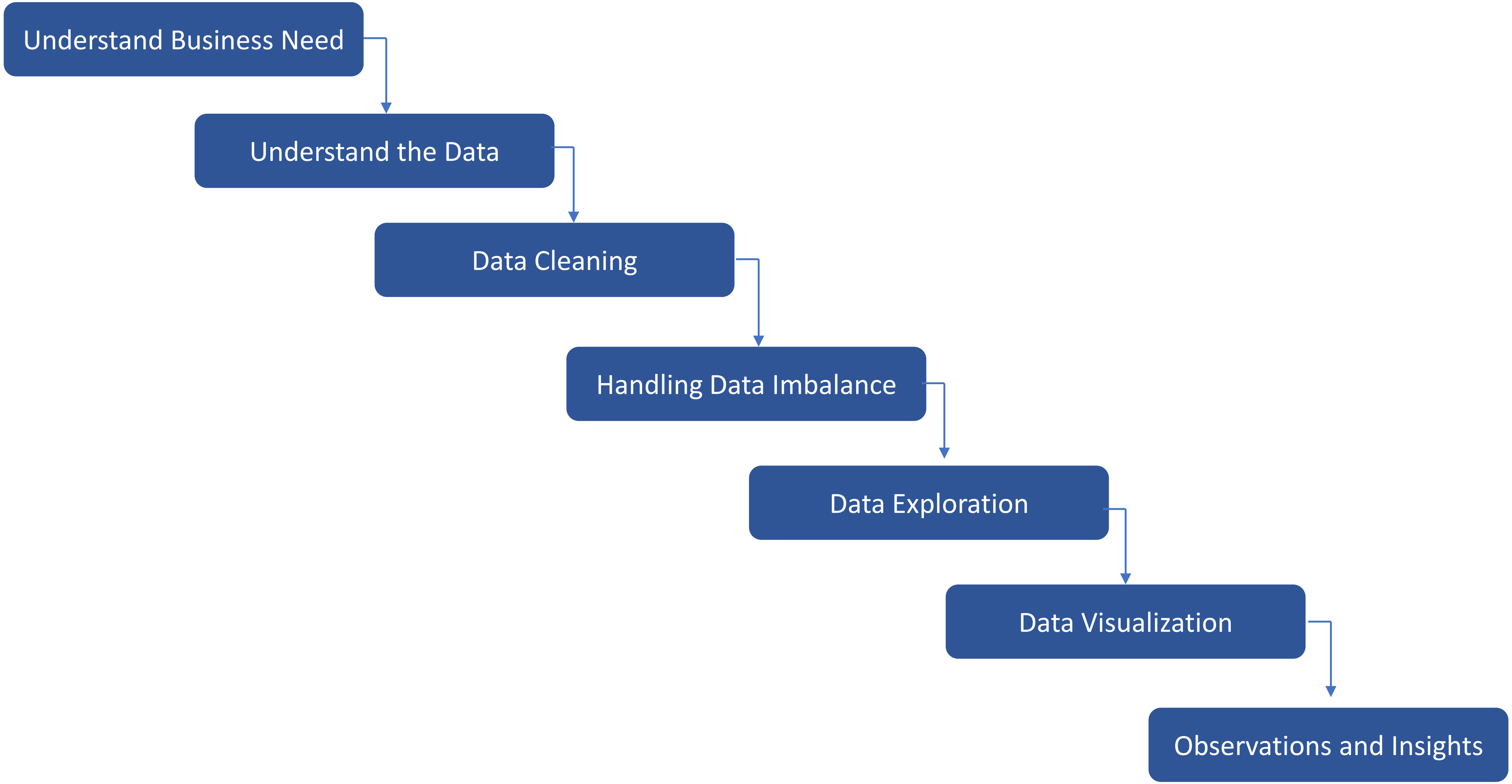
Tools/software used

Programming language for analysis – Python

Python libraries – Pandas, MatplotLib, Seaborn

UI interface for code development – Jupyter notebook

EDA APPROACH



Understand Business Need

Understand the Data

Data Cleaning

Handling Data Imbalance

Data Exploration

Data Visualization

Observations and Insights

DATA UNDERSTANDING

Dataset

The dataset contains the complete loan data for all loans issued through the duration 2007 to 2011.
The inputs contain:

- Dataset file – loan.csv
- Data dictionary - Data_Dictionary.xlsx

Data understanding

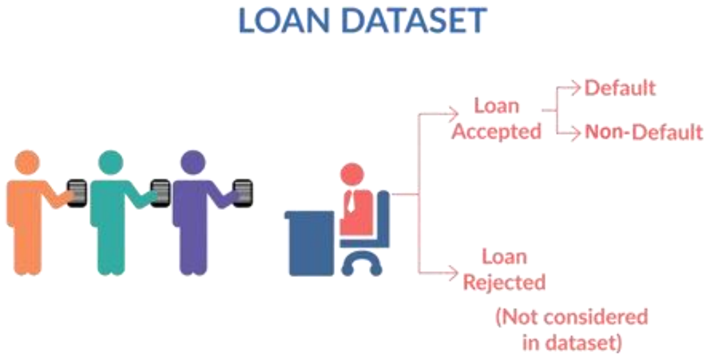
The step involves details manual inspection of dataset file.

- Go through the data dictionary document and understand the data content
- Check if all columns specified in data dictionary are available in data set (4 columns found missing. Same are not being taken into consideration)
- Brief through the dataset file to understand the data spread
- Identify the ‘target variable’

Target column

- The primary column based on which the results need to be predicted is identified as ‘loan_status’
- The variable in loan_status are :

  - ‘Fully paid’ - Applicant has fully paid the loan (the principal and the interest rate)
  - ‘Current’ - Applicant is in the process of paying the instalments
  - ‘Charged off’ - Applicant has not paid the instalments in due time for a long period of time



DATA CLEANING

Data cleaning process

The data cleaning step is important for identifying and rectifying :

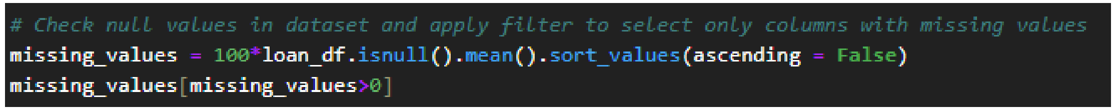
- ➢ Errors/Inconsistencies
- ➢ Missing values
- ➢ Outlier values

Steps

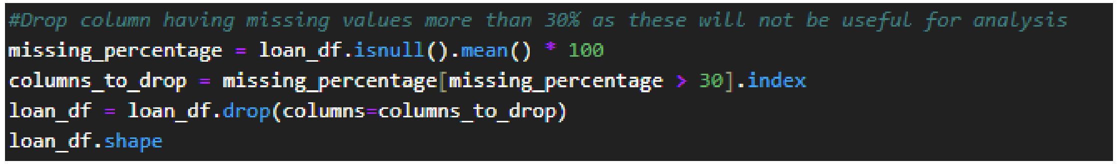
Following steps were followed to make sure dataset is cleaned before starting analysis :

- ➢ Manual inspection (Review dataset manually in detail)
- ➢ Identifying missing null values
- ➢ Fixing null values using suitable method (in this case median and mode)
- ➢ Fixing inconsistent data types

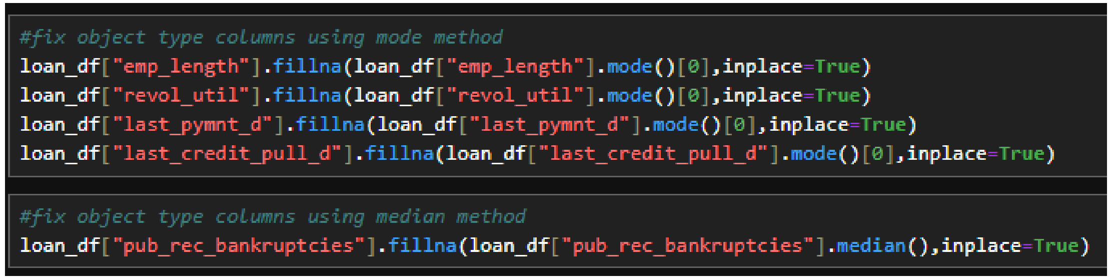
Null value fixing :



```
# Check null values in dataset and apply filter to select only columns with missing values
missing_values = 100*loan_df.isnull().mean().sort_values(ascending = False)
missing_values[missing_values>0]
```



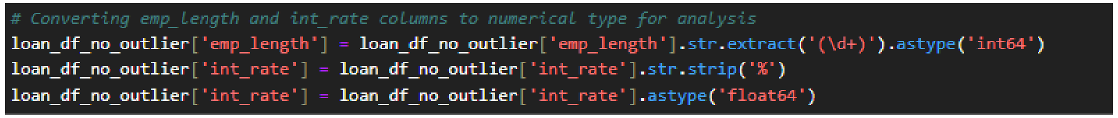
```
#Drop column having missing values more than 30% as these will not be useful for analysis
missing_percentage = loan_df.isnull().mean() * 100
columns_to_drop = missing_percentage[missing_percentage > 30].index
loan_df = loan_df.drop(columns=columns_to_drop)
loan_df.shape
```



```
#fix object type columns using mode method
loan_df["emp_length"].fillna(loan_df["emp_length"].mode()[0],inplace=True)
loan_df["revol_util"].fillna(loan_df["revol_util"].mode()[0],inplace=True)
loan_df["last_pymnt_d"].fillna(loan_df["last_pymnt_d"].mode()[0],inplace=True)
loan_df["last_credit_pull_d"].fillna(loan_df["last_credit_pull_d"].mode()[0],inplace=True)
```

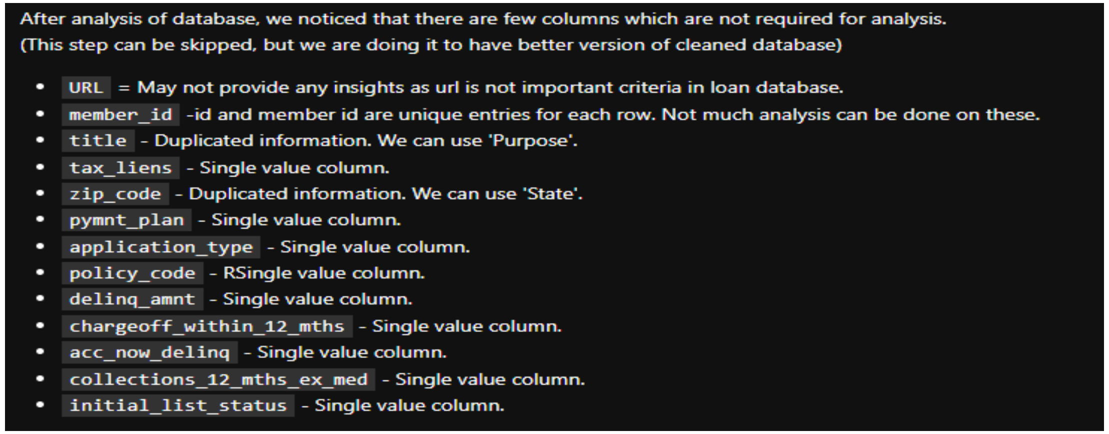
```
#fix object type columns using median method
loan_df["pub_rec_bankruptcies"].fillna(loan_df["pub_rec_bankruptcies"].median(),inplace=True)
```

Fix type of columns :



```
# Converting emp_length and int_rate columns to numerical type for analysis
loan_df_no_outlier['emp_length'] = loan_df_no_outlier['emp_length'].str.extract('(\d+)').astype('int64')
loan_df_no_outlier['int_rate'] = loan_df_no_outlier['int_rate'].str.strip('%')
loan_df_no_outlier['int_rate'] = loan_df_no_outlier['int_rate'].astype('float64')
```

Delete extra columns :



After analysis of database, we noticed that there are few columns which are not required for analysis.
(This step can be skipped, but we are doing it to have better version of cleaned database)

- `URL` = May not provide any insights as url is not important criteria in loan database.
- `member_id` -id and member id are unique entries for each row. Not much analysis can be done on these.
- `title` - Duplicated information. We can use 'Purpose'.
- `tax_liens` - Single value column.
- `zip_code` - Duplicated information. We can use 'State'.
- `pymnt_plan` - Single value column.
- `application_type` - Single value column.
- `policy_code` - RSingle value column.
- `delinq_amnt` - Single value column.
- `chargeoff_within_12_mths` - Single value column.
- `acc_now_delinq` - Single value column.
- `collections_12_mths_ex_med` - Single value column.
- `initial_list_status` - Single value column.

DATA IMBALANCE

Identifying outliers

As further step in data cleaning process is handling outliers. It is important step as it helps in :

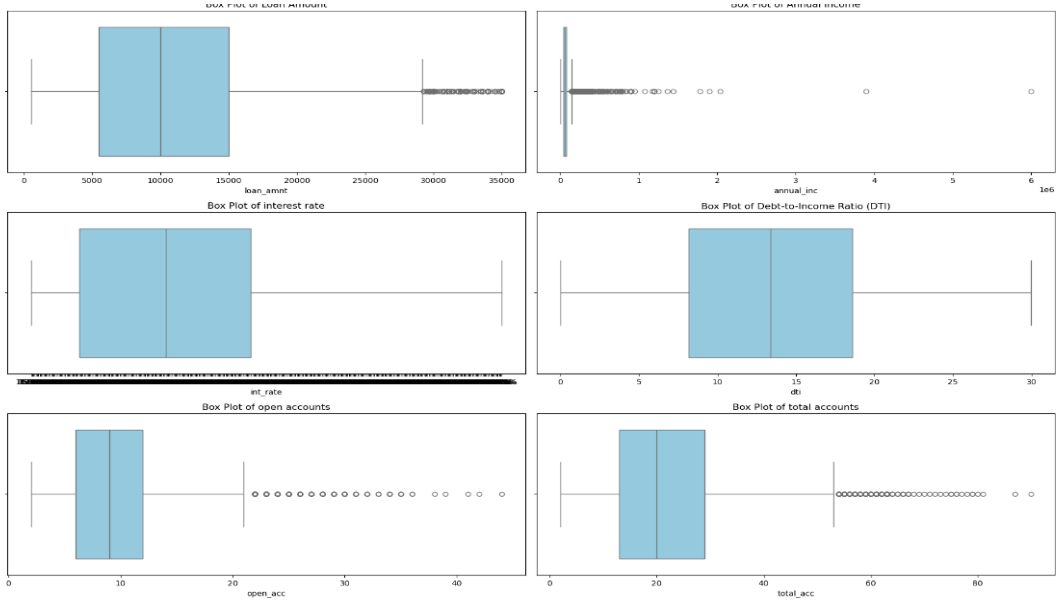
- Removing bias
- Better understanding of data range
- Improved accuracy (To remove incorrect entries)

Steps

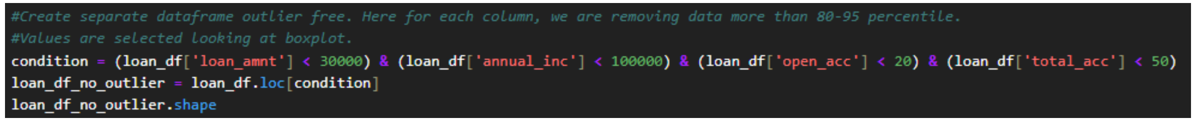
Following steps were followed to remove outliers:

- Select dependent variable columns (mostly numeric)
- Blot histogram or bar plot
- Understand the data distribution and central tendency (median)
- Use IQR or Z-Score method or decide based on boxplot range to remove outliers
- Create a separate dataframe which is outlier free

Boxplot :

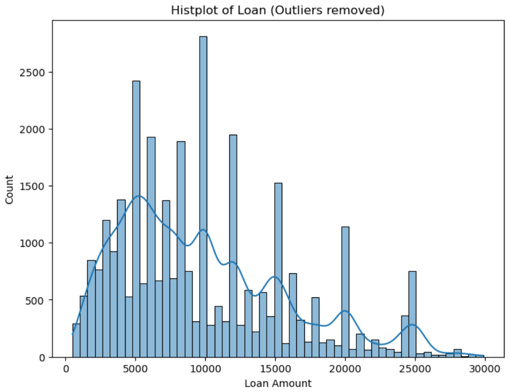


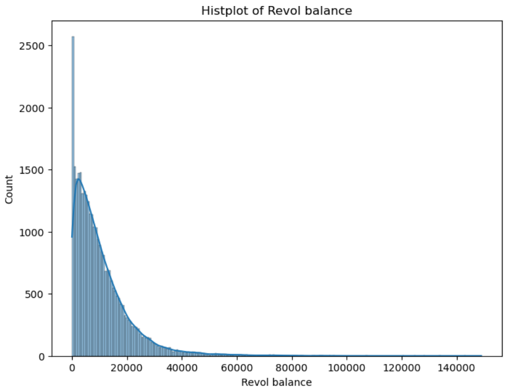
Remove outliers :



```
#Create separate dataframe outlier free. Here for each column, we are removing data more than 80-95 percentile.
#Values are selected looking at boxplot.
condition = (loan_df['loan_amnt'] < 30000) & (loan_df['annual_inc'] < 100000) & (loan_df['open_acc'] < 20) & (loan_df['total_acc'] < 50)
loan_df_no_outlier = loan_df.loc[condition]
loan_df_no_outlier.shape
```

Histogram to check distribution for outlier identification :





DATA Exploration

Univariate analysis

Univariate analysis carried out on Numerical and Categorical columns to understand :

- Distribution analysis
- Proportion analysis (counts)
- Find skewness
- Central tendency measures
- Outlier detection

Plots used

Following visualization methods were used for univariate analysis

- Countplot / Barplot
- Boxplot
- Histogram

Bivariate analysis

Bivariate analysis carried out to understand :

- Relation between two variables
- Find dependency on target variable
- Correlation

Plots used

Following visualization methods were used for univariate analysis

- Countplot / Barplot
- Scatterplot
- Jointplot

Multivariate analysis

Multivariate analysis carried out to understand :

- Correlation between more than one variable
- Relation dependent variables on target variable
- In-depth analysis of dataset

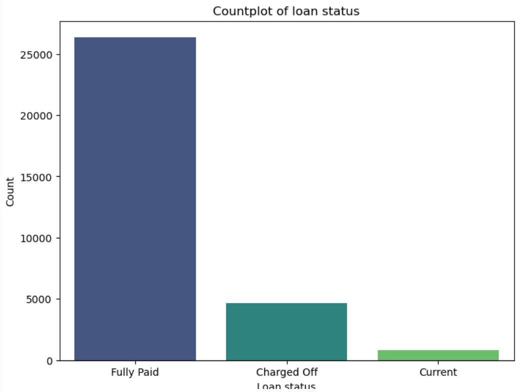
Plots used

Following visualization methods were used for univariate analysis

- Heatmap
- Cluster plot

Univariate analysis – Categorical columns

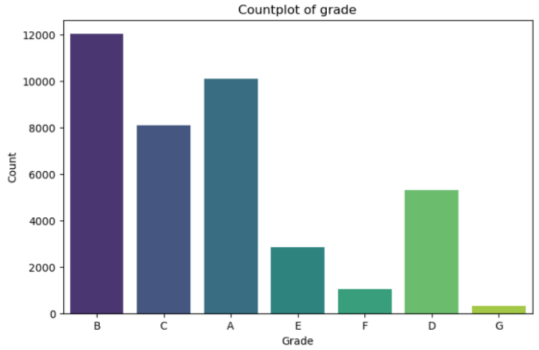
Loan status



Observations recorded

- Loan status has three parameters
- Fully paid, Charged off and Current
- The count-plot indicates, heighted number of loans are ‘Fully paid’
- The graph provide distribution of loan statuses in dataset.

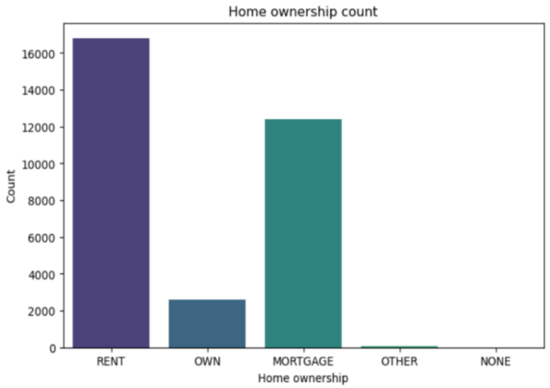
Grades



Observations recorded

- Grade category has seven parameters
- Grades in dataset include A,B,C,D,E,F,G
- Most of the loans given fall under grade B, followed by A,C,D,E,F,G respectively.
- A is Highest grade and G being lowest grade with respect to profile rating
- Less loans are granted in lower grade category (D, E,F,G) as compared to higher grade (A,B,C)

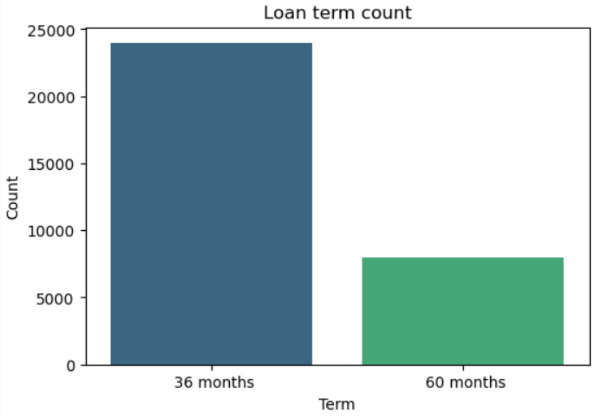
Home ownership



Observations recorded

- Home ownership categories are ‘Rent’, Own’, ‘Mortgage’, ‘Other’ and ‘None’
- ‘Rent’ is most dominant category in loan applicants, followed by ‘Mortgage’ and ‘Own’. However, the difference between ‘OWN’ and other categories is significantly high.
- The graph indicates that most of the loan applicant have debt (Rent or Mortgage).

Term

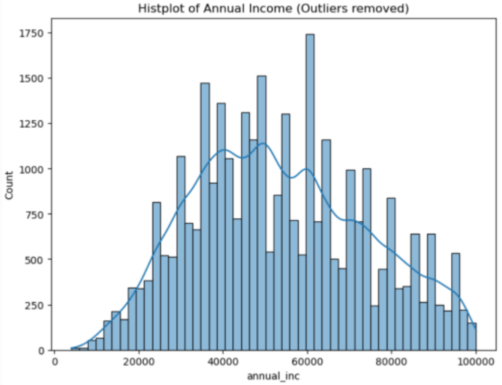


Observations recorded

- The bank gives loans for two category for ‘36 months’ and ‘60 months’.
- Most of the loan are approved for 36 months term as compared to 60 months term.

Univariate analysis – Numerical columns

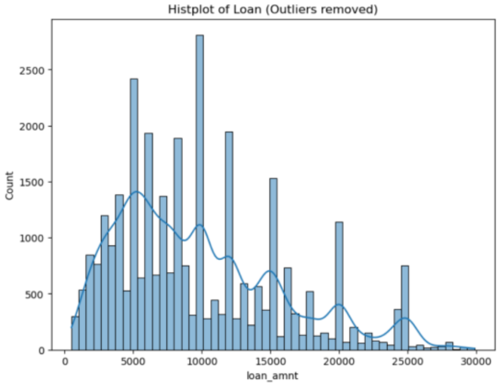
Annual income



Observations recorded

- ➢ The histogram shows distribution of annual incomer.
- ➢ The histogram is slightly right skewed, indicating there are few individual with comparatively higher income.
- ➢ The peak of the distribution is around 50000, suggesting it is the most common annual income.

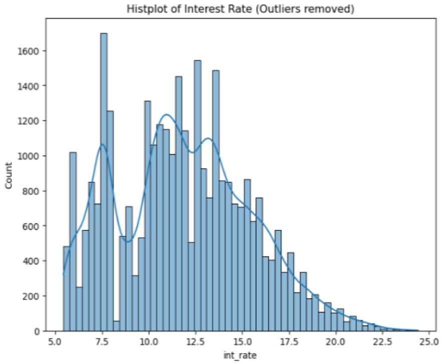
Loan amount



Observations recorded

- ➢ The graph shown distribution of loan amount.
- ➢ The histogram is skewed to the right, indicating there are few loans with very high amount.
- ➢ The peak of the graph is at 5000 indicating that, it is most common loan amount.

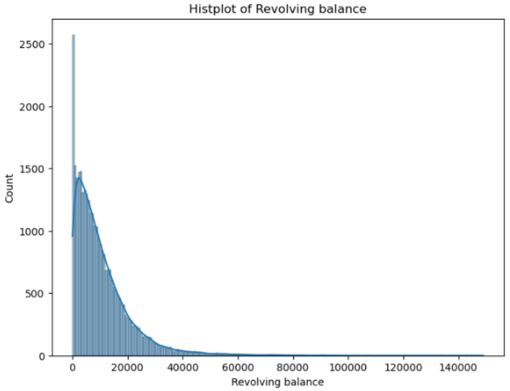
Interest rate



Observations recorded

- ➢ The graph shows the distribution of interest rates in a dataset.
- ➢ The histogram is skewed to the right, indicating that there are a few loans with very high interest rates.
- ➢ The peak of the distribution is around 10, suggesting that this is the most common interest rate.

Revolving balance

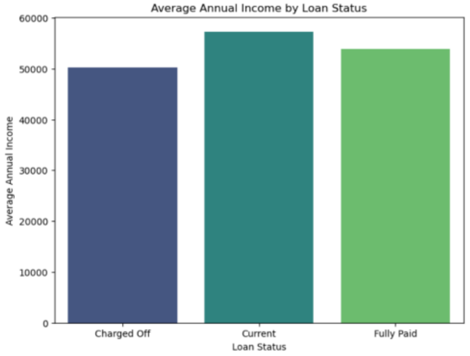


Observations recorded

- ➢ The graph shows the distribution of revolving balance.
- ➢ The histogram is heavily skewed to the right, indicating that there are a few individuals with very high revolving balances.
- ➢ The peak of the distribution is around 0, suggesting that large number of individuals have low or no revolving balance.

Bivariate analysis

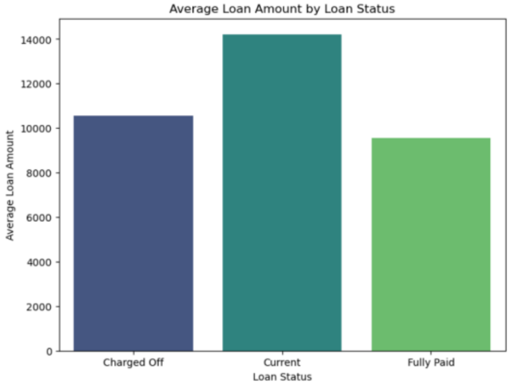
Annual income vs loan status



Observations recorded

- Individuals with loans that are "Charged Off" (meaning they defaulted) have the lowest average annual income compared to other loan status.
- However the difference is not major. This indicates that 'Annual income' criteria is not a deciding factor and hence can not be used to decide loan repayment possibility for a client.

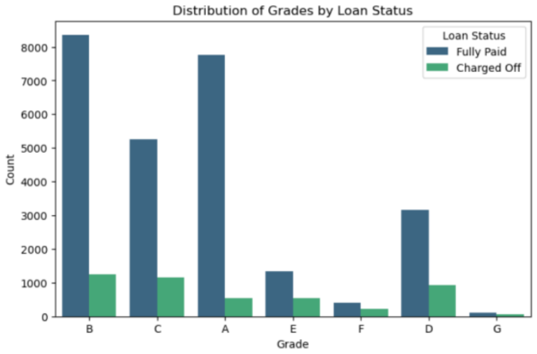
Loan amount vs loan status



Observations recorded

- The graph indicates that loan amount for 'Current' loan status is highest followed by 'Charged off' and 'Fully paid' respectively.
- Lenders might need to exercise more caution when granting larger loans, as they seem to be more likely to be defaulted on. However, it is not only the deciding criteria.

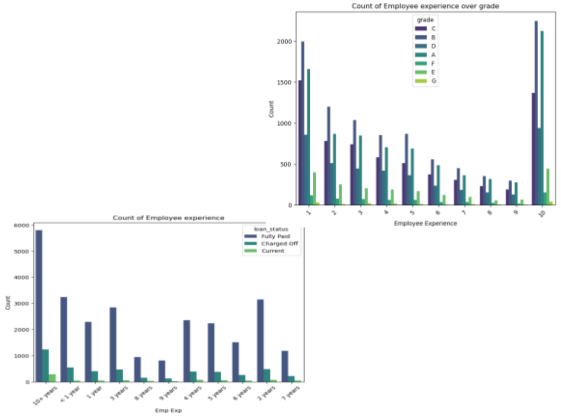
Grade vs loan status



Observations recorded

- By looking at difference between 'Charged Off' and 'Fully Paid' status for each grade, we can say that,Fully Paid loans are more likely to be in higher-grade categories (A, B, C), while Charged Off loans are more likely to be in lower-grade categories (D, E, F, G)
- Loan Risk: Loans with lower grades (D, E, F, G) are more likely to be charged off, indicating a higher risk of default.

Employee exp vs loan status and also grade

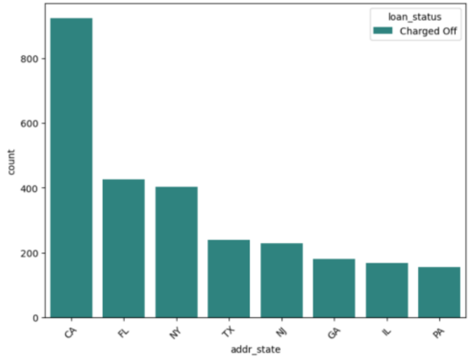


Observations recorded

- Most of loan applicants have either experience more than 10+ years followed by <=1 year.
- The highest concentration of high-risk grades is found among employees with less than or equal to 1 year of experience
- As experience increases, the proportion of high-risk grades generally decreases, indicating a possible correlation between tenure and lower-risk assessments.

Bivariate analysis

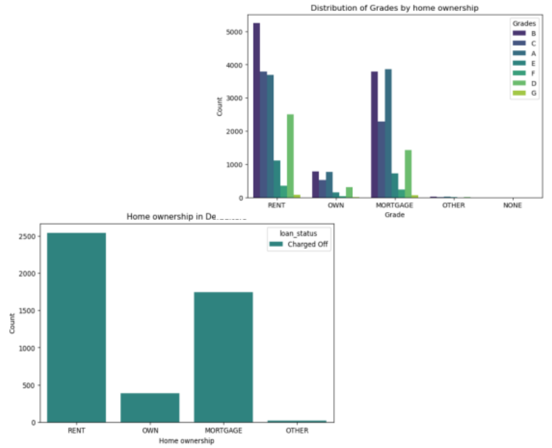
Address state vs loan status



Observations recorded

- ➢ Most of the loans which defaulted are applied from 'CA' (California) state almost more than 17% , followed by Florida, New york etc.

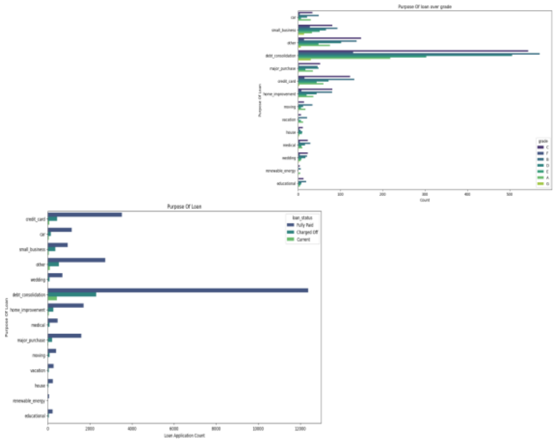
Home ownership vs loan status and grade



Observations recorded

- ➢ Loans for individuals who own their homes outright are more likely to be in higher-grade categories (A, B, C).
- ➢ For above count-plots we can derive that, when home ownership is of 'own' category, chances of defaulting loan are less.

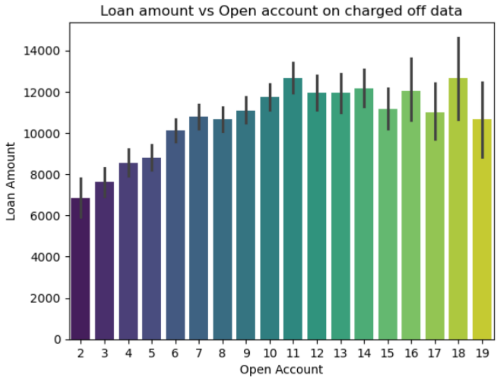
Purpose vs loan status and grade



Observations recorded

- ➢ Loans for debt consolidation have the highest concentration of high-risk grades (B, C, A, E), suggesting that individuals struggling to manage existing debt might be more likely to require higher-risk loans.
- ➢ Loans for credit card debt also show a significant number of high-risk grades, indicating a potential connection between credit card debt and financial difficulties.

Loan amount vs open accounts

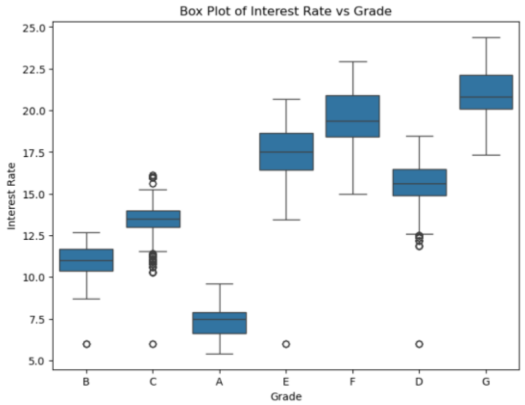


Observations recorded

- ➢ The increasing trend in loan amounts with more open accounts might suggest that individuals with a higher number of open accounts are more likely to default on their loans.

Bivariate analysis

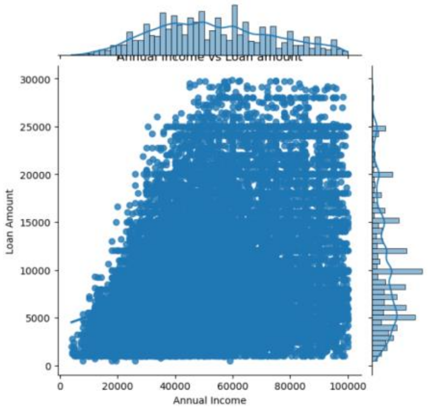
Grade vs Interest rate



Observations recorded

- ➢ As the loan grade decreases from A to G, the median interest rate generally increases.
- ➢ This suggests that borrowers with lower grades are offered higher interest rates.
- ➢ Customer with high risk profile tend to pay high interest rates.

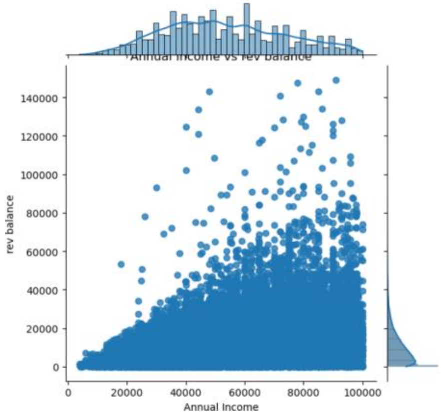
Annual income vs loan amount



Observations recorded

- ➢ The positive correlation between annual income and loan amount suggests that lenders may consider an individual's income when determining the loan amount they are eligible for.
- ➢ However, there is also a degree of variability. Some individuals with higher incomes receive lower loan amounts, while others with lower incomes receive higher loan amounts. This implies that the loan will be lended based on other parameters like DTI

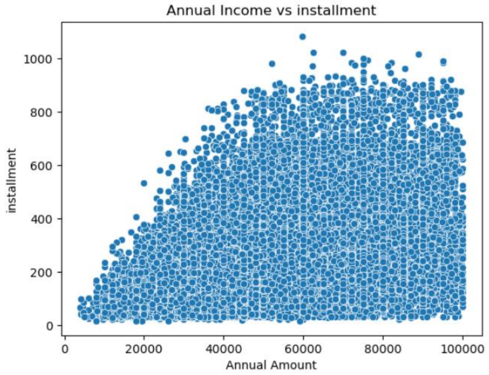
Annual income vs revolving balance



Observations recorded

- ➢ There seems to be a positive correlation between annual income and revolving balance.
- ➢ This suggests that, generally, as annual income increases, so does the revolving balance.

Annual income vs installment

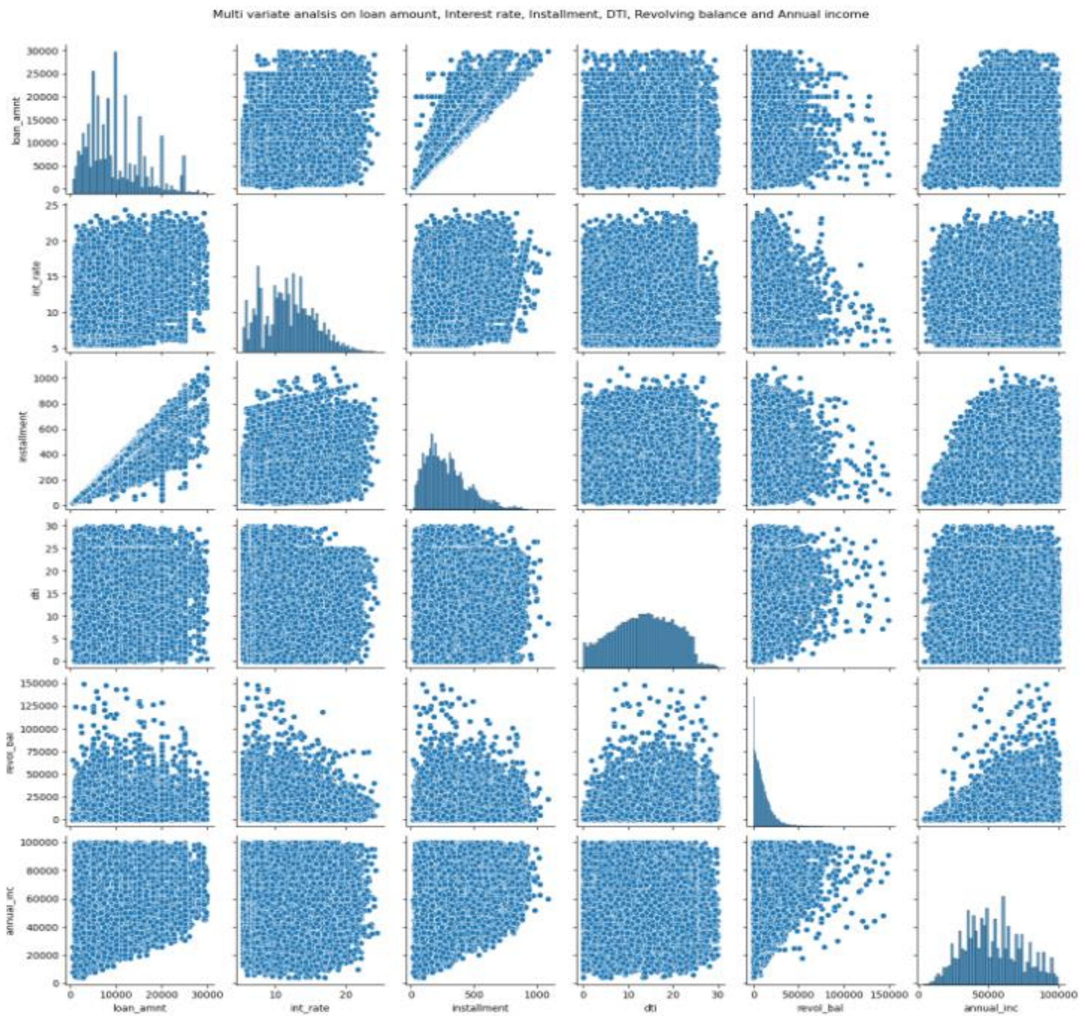


Observations recorded

- ➢ There appears to be a general positive correlation between annual income and installment amount.
- ➢ This means that as annual income increases, the installment amount tends to increase as well.

Bivariate analysis

Scatter plot for multiple numerical columns

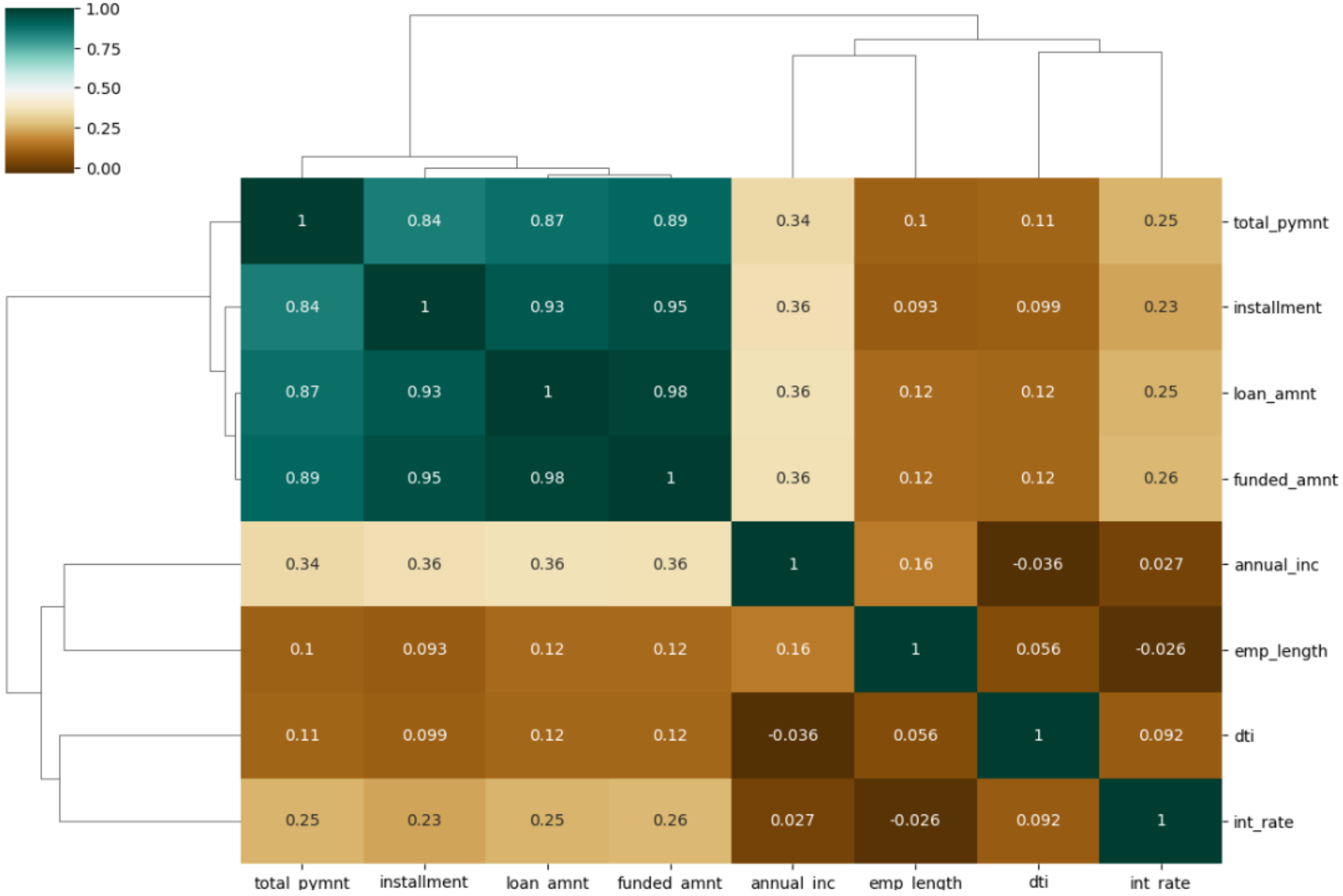


Observations recorded

- ➢ Loan amount vs. Annual income: There appears to be a weak positive correlation between loan amount and annual income. This suggests that individuals with higher annual incomes tend to be approved for larger loan amounts.
- ➢ dti vs. loan amount: There is no clear correlation between the dti and loan amount.
- ➢ Annual income vs. revol balance: There appears to be a weak postive correlation between annual income and revolving balance. This suggests that individuals with higher annual incomes tend to have higher revolving balances.
- ➢ Total Rec Principal vs. Total Int Rate: There appears to be a weak positive correlation between the total principal recovered and the total interest rate. This suggests that loans with higher interest rates might tend to have larger principal amounts.
- ➢ Total Rec Principal vs. Total Rec Late Fee: There seems to be a slight positive correlation between the total principal recovered and the total late fees. This could indicate that loans with higher principal amounts are more likely to incur late fees.
- ➢ Total Int Rate vs. Total Rec Late Fee: There is no clear correlation between the total interest rate and the total late fees.
- ➢ Total Rec Late Fee vs. Recoveries: There seems to be a weak negative correlation between the total late fees and recoveries. This might suggest that loans with higher late fees are less likely to have recoveries.
- ➢ Total Int Rate vs. Last Pmt Amt: There is no clear correlation between the total interest rate and the last payment amount.
- ➢ Total Rec Late Fee vs. Last Pmt Amt: There is no clear correlation between the total late fees and the last payment amount.
- ➢ Recoveries vs. Last Pmt Amt: There is no clear correlation between recoveries and the last payment amount.Total Rec
- ➢ Principal vs. Last Pmt Amt: There is a weak positive correlation between the total principal recovered and the last payment amount. This suggests that loans with larger last payments might have higher total principal recoveries.
- ➢ Total Int Rate vs. Last Pmt Amt: There is no clear correlation between the total interest rate and the last payment amount.

Multivariate analysis

Cluster plot for multiple numerical columns and its correlation



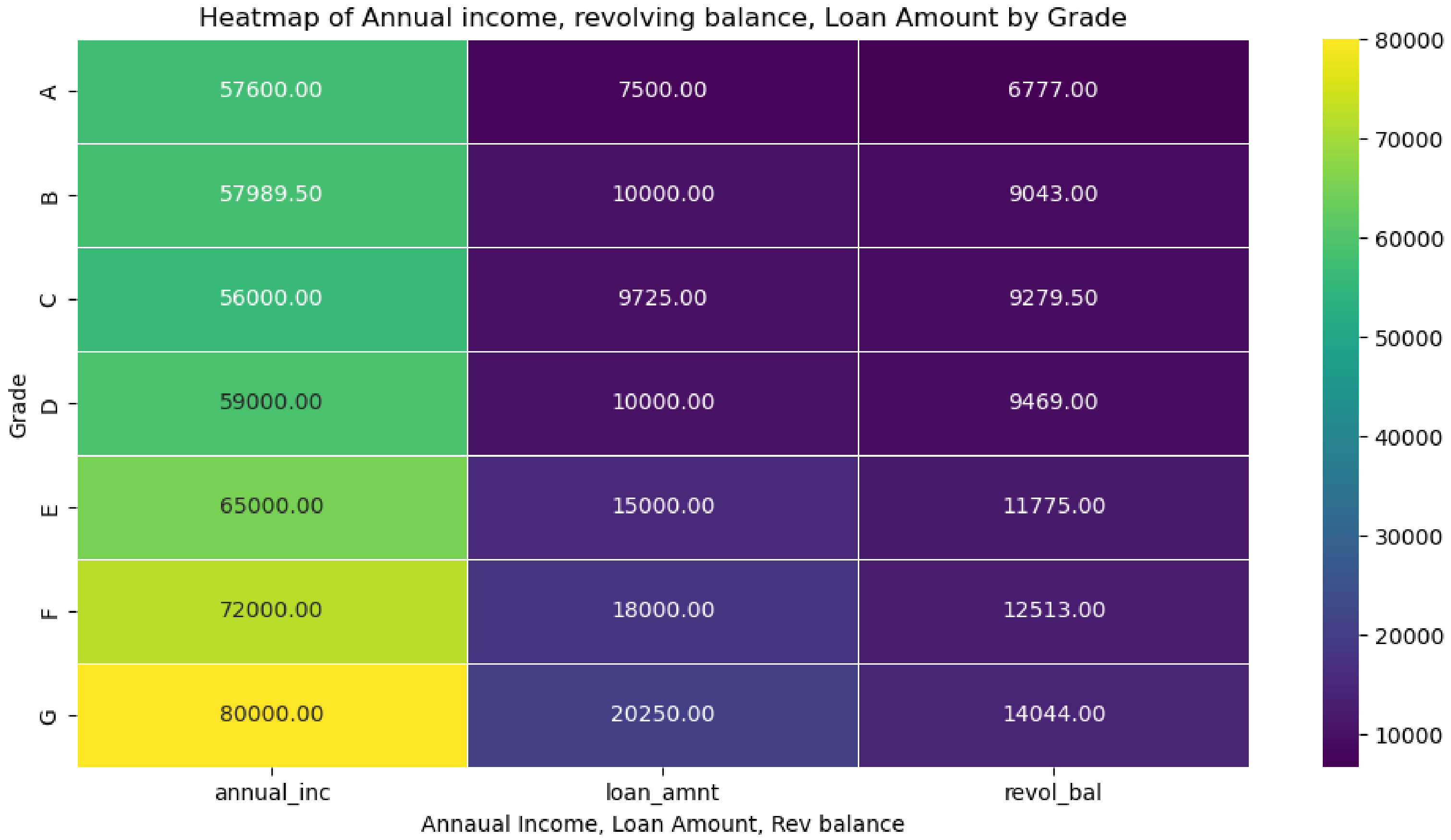
	total pymnt	installment	loan amnt	funded amnt	annual inc	emp length	dti	int rate
total_pymnt	1	0.84	0.87	0.89	0.34	0.1	0.11	0.25
installment	0.84	1	0.93	0.95	0.36	0.093	0.099	0.23
loan_amnt	0.87	0.93	1	0.98	0.36	0.12	0.12	0.25
funded_amnt	0.89	0.95	0.98	1	0.36	0.12	0.12	0.26
annual_inc	0.34	0.36	0.36	0.36	1	0.16	-0.036	0.027
emp_length	0.1	0.093	0.12	0.12	0.16	1	0.056	-0.026
dti	0.11	0.099	0.12	0.12	-0.036	0.056	1	0.092
int_rate	0.25	0.23	0.25	0.26	0.027	-0.026	0.092	1

Multivariate analysis

Heat map for Annual income, Loan amount, revolving balance over grade

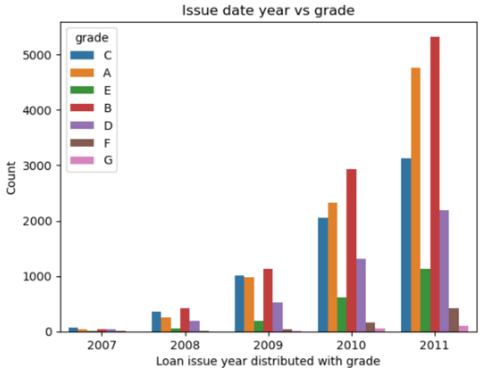
Observations recorded

- As the grade increases, the average annual income, loan amount, and revolving balance also increase.
- The average loan amount increases more rapidly with grade than the average annual income or revolving balance.
- The average revolving balance increases with grade.



Derived metrics

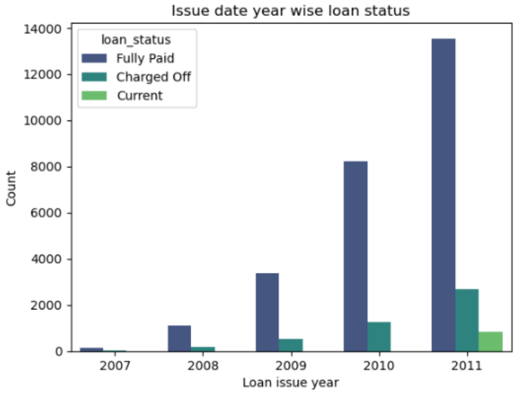
Loan issue date (year) vs grades



Observations recorded

- The number of loans issued increased significantly from 2007 to 2011, with a spike in 2011.
- In earlier years (2007-2009), most loans were issued with grades A, B, and C.
- In 2010 and 2011, there was a noticeable increase in the number of loans issued with lower grades (D, E, F, and G).

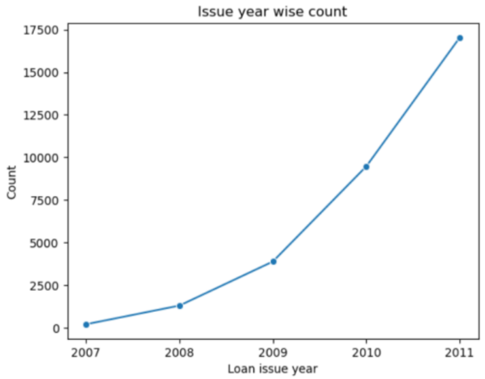
Loan issue date (year) vs loan status



Observations recorded

- The number of loans issued increased each year from 2007 to 2011.
- In 2011, the number of loans issued was significantly higher than in previous years.
- The majority of loans issued in each year were Fully Paid.

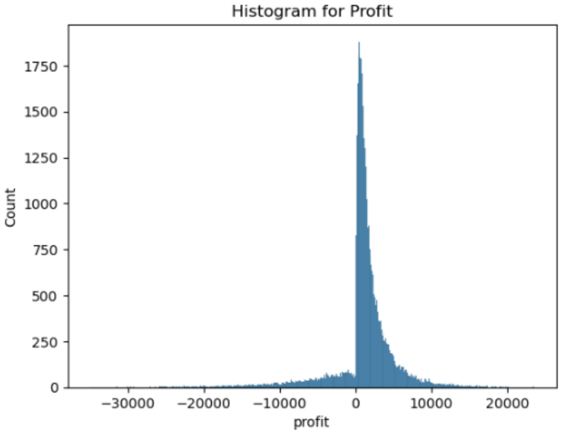
Year-wise loan issue counts



Observations recorded

- The loans are increasing at massive scale year on year achieving 100% increase year on year.
- 2011 marked as most customer had applied for the loan

Profit analysis



Observations recorded

- The majority of the data points cluster around a profit value close to 0. This suggests that a large number of transactions or entities have profits near zero.
- The presence of negative profits indicates potential risks associated with the business

Conclusion

Summary

- Annual income of loan applicants are falling between 30,000 and 70,000 indicating that mid-range salary drawer may tend to buy more loans.
- Applicants having 10+ years of experience and less than 1 year may tend to buy more loans.
- Chances are, applicants may default due to other monthly debts, in other words increase it debt-to-income ratio. It also includes that their home ownership is rented or mortgage.
- Applicants with grade E, F and G may tend to pay more interest as they fall under high-risk profiles.
- State CA, NY and FL records the most loan buyers.
- Applicant profile may get riskier when the number of inquires increases.
- Major purpose for the loan is debt consolidation, credit card and others.
- More the annual income, more the revolving balance and more the loan amount.
- Annual income source verified is lower charged off applicant when compared with verified.
- As the grade increases, the average annual income, loan amount, and revolving balance also increase.
- More the loan amount, loan enquiries and open accounts, the chances of loan defaulting increases.

Conclusion

To minimize the credit loss, Lending club can monitor below parameters:

- Higher loan amounts,
- More inquiries,
- Open accounts
- High-risk profiles (grades E, F, G)
- High debt-to-income ratios.
- Geographical patterns

By understanding these key trends, we can make smarter lending decisions that not only reduce risk but also empower borrowers to achieve financial success.