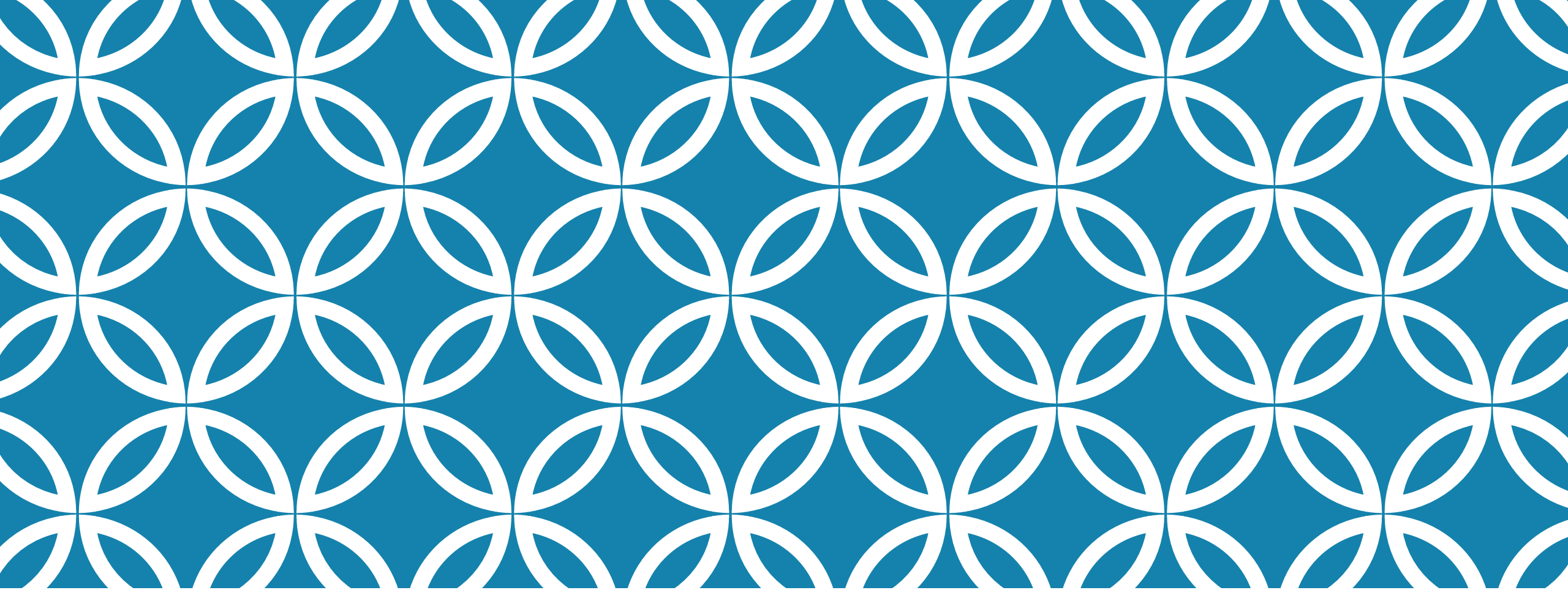


EXPLORATORY DATA ANALYSIS OF LOAN DATA

An overview of key patterns and trends

OBJECTIVE

To understand :

- Patterns
- Trends
- Insights

from loan data.

We have a dataset which contains historical data of loans provisioned.

It has 39717 rows and 111 columns with numerous details.

Our next step is to clean the data

METHODOLOGIES TO CLEAN DATA

1. Column cleaning

- Identified and dropped the columns which have NULL values more than 60% as it will not help in the analysis. We eliminated 51.35% of columns.

- Identified and dropped the columns with non unique values. We bring the columns down to 45 from 111.

- Identified and dropped further 25 columns which wouldn't help in section 1.3 of the notebook.

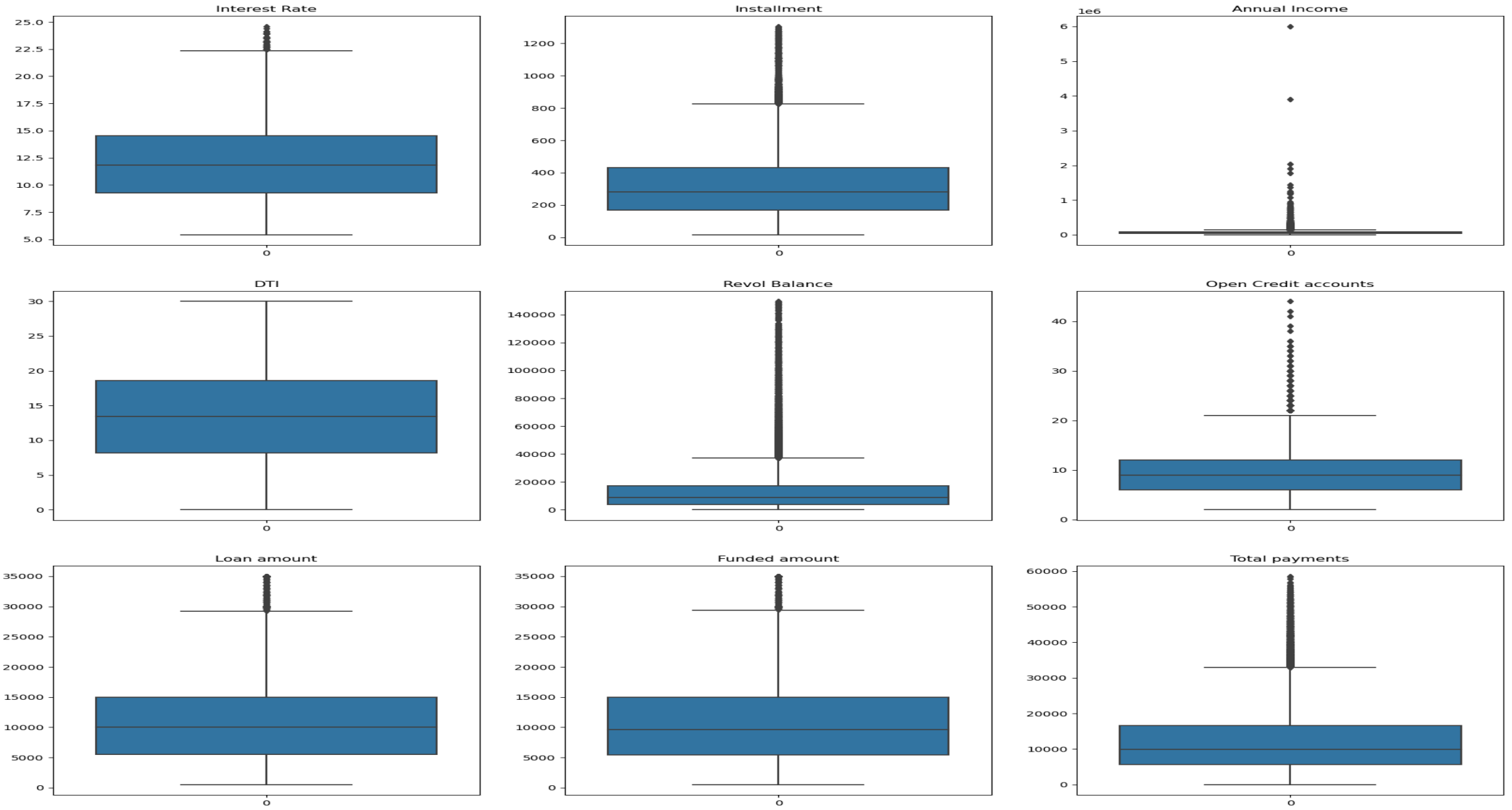
2. Row cleaning

- Identified the rows having NULL values (emp_length (1075) and revol_bal(50)).
- Imputed Employment length by figuring the mode of the annual income of those with emp_length as NULL ($36000.00). This helps as a benchmark to find the employment length.
- Found the mode of the employment length where the annual income equals the result we obtained ($36000.00) to get the best estimate – 3 years, and imputed the values.
- We had 50 NULL values of revol_bal. This could be eliminated instead of imputing.

3. Data transformation : Certain columns need to be correctly formatted and categorized.

- Employment length needs to be made numerical. Year, month and quarter should be extracted from the "Issue_date" field.
- Integer and float fields like loan amount, funded amount, installment, annual income, interest rate needs to be bucketed for effective categorical analysis.
- Outliers for these fields should be fixed. The next 3 slides shows outlier identification and treatment.

Annual Income and Revol Balance are fields having significant outliers



TREATING THE OUTLIERS

Make the IQR multiplier value 1.5. This means that the resultant should be within 1.5 times the IQR above or below the 25th percentile and 75th percentile. (IQR : Inter quartile range = 3rd quartile - 1st quartile).

Identify the 1st quartile, median and the 3rd quartile and the IQR.

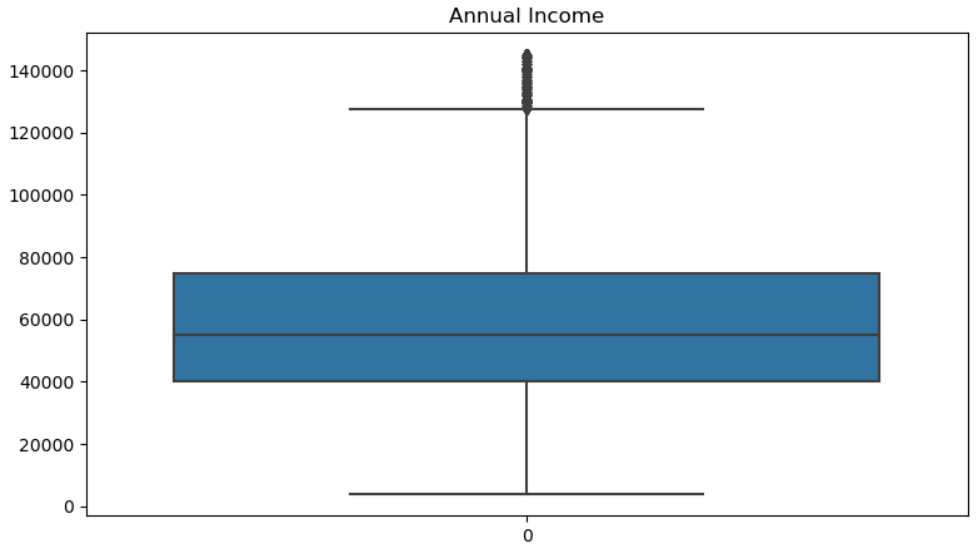
The upper and lower bounds are calculated by the formula :

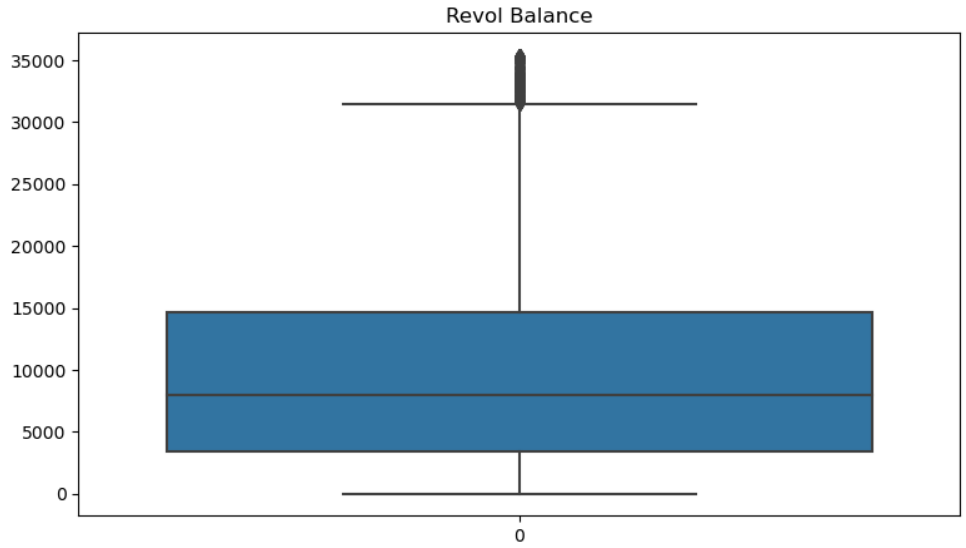
- Lower bound = (1st quartile – IQR * IQR multiplier)
- Upper bound = (3rd quartile + IQR * IQR multiplier)

Remove the rows lower than the lower bound and higher than the upper bound

Result in the next slide….

THE OUTLIERS FOR THE ANNUAL INCOME AND REVOL BALANCE FIELDS ARE NOW TREATED

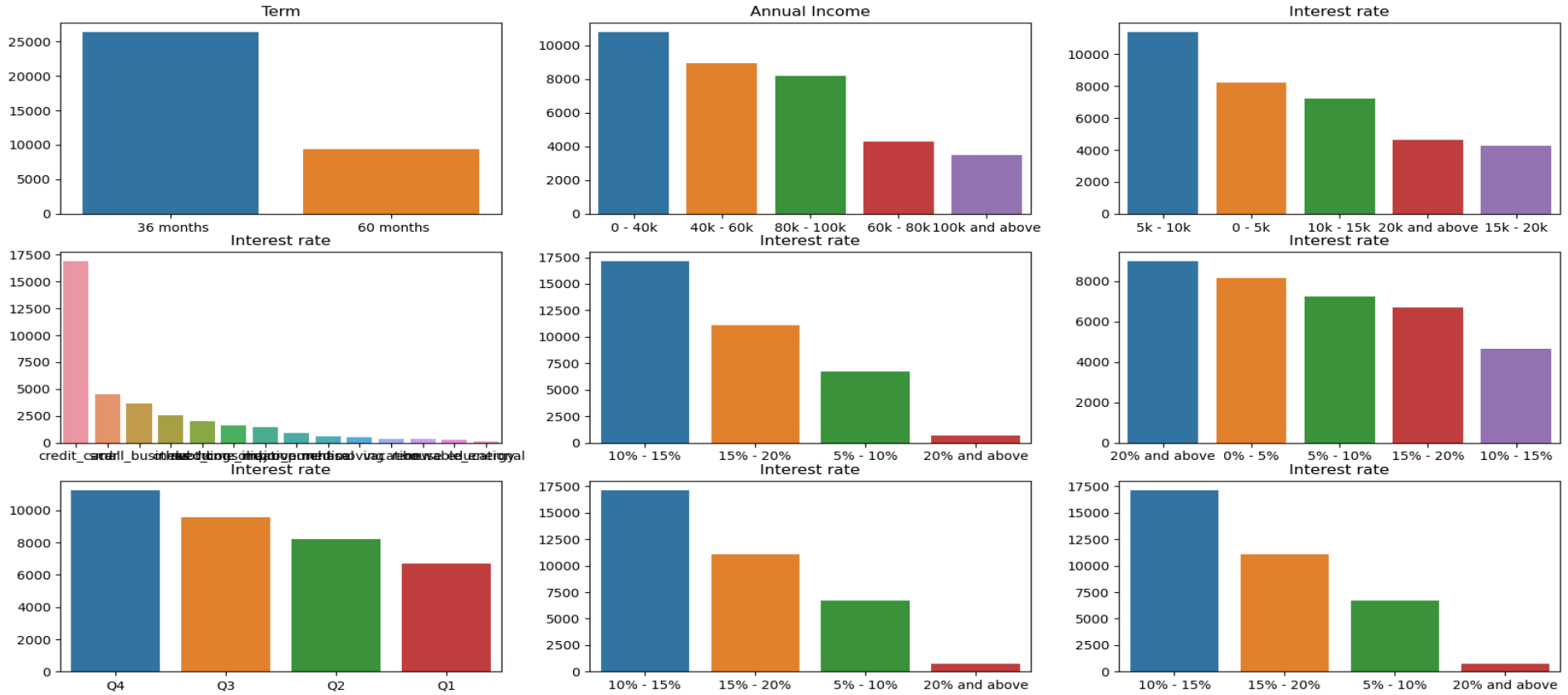




NUMERICAL VARIABLE SUMMARY STATISTICS

	Loan amnt	Funded amount	Interest rate	Instalments	Employment length	Annual Income	DTI	inq_last_6mths	Open accounts	Revol bal	Revol util	Total payment
count	35707	35707	35707	35707	35707	35707	35707	35707	35707	35707	35707	35707
mean	10603.5896	10369.05159	11.94994	307.485918	4.798667	59479.61379	13.277368	0.865349	9.04128	9955.16019	47.995156	11454.59353
std	6956.01551	6708.224957	3.692988	193.634714	3.491576	27146.59852	6.66388	1.063605	4.289218	8187.29348	28.216391	8367.371037
min	500	500	5.42	15.69	0	4000	0	0	2	0	0	0
25%	5000	5000	8.94	162.27	2	39996	8.16	0	6	3376	24.6	5427.287004
50%	9200	9000	11.83	268.83	4	55000	13.37	1	8	7986	48.1	9397.023804
75%	14787.5	14087.5	14.42	404.24	8	75000	18.55	1	11	14624.5	71.3	15508.78947
max	35000	35000	24.4	1288.1	10	145008	29.99	8	42	35430	99.9	58563.67993

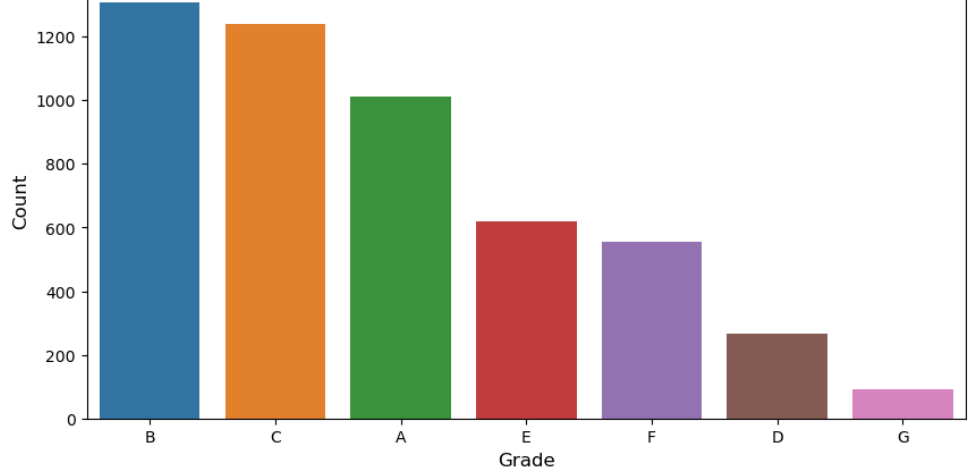
CATEGORICAL VARIABLE SUMMARY STATISTICS

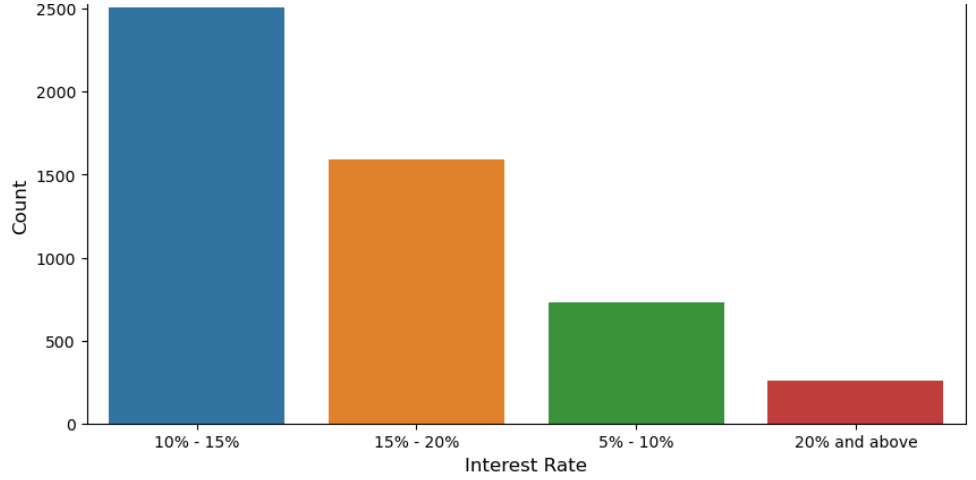


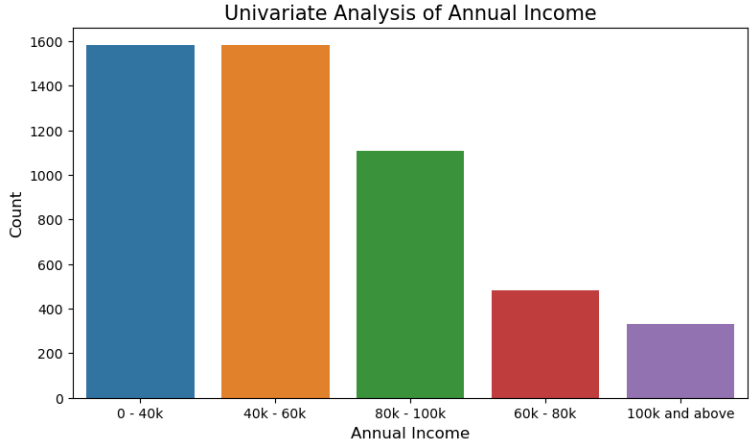
UNIVARIATE ANALYSIS

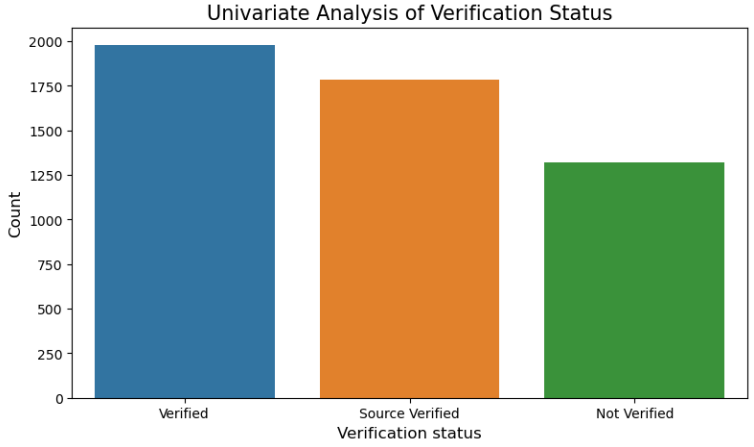
Univariate analysis was done on categorical variables using bar charts.

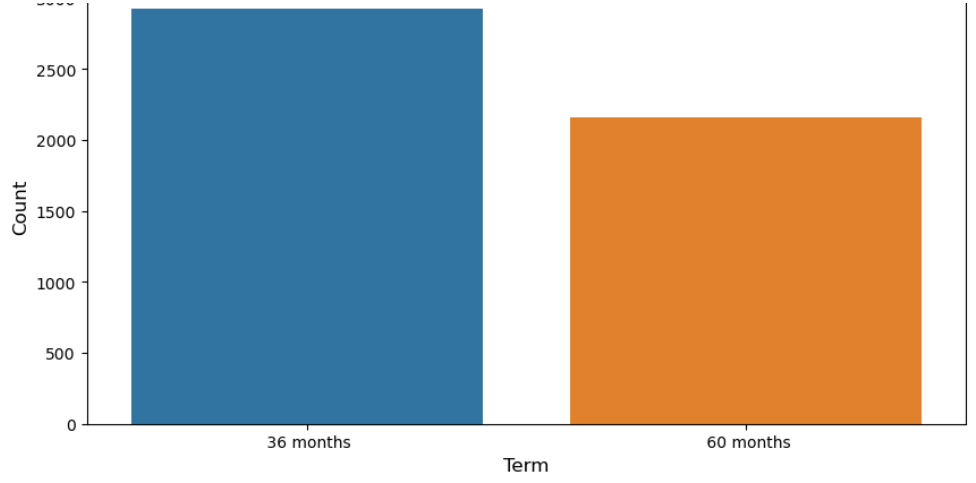
Primary approach was to see the counts for different categories for charged off loans only.











OBSERVATIONS USING UNIVARIATE ANALYSIS

We see that the analysis focuses on the counts.

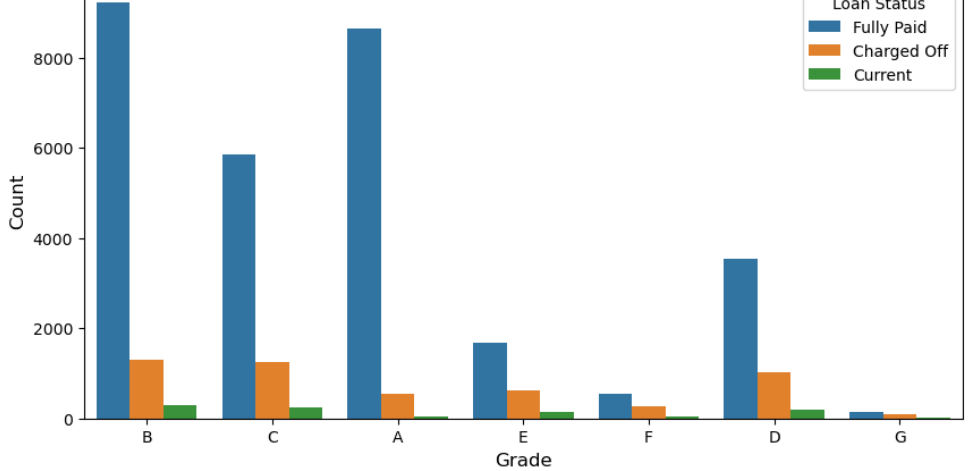
Like

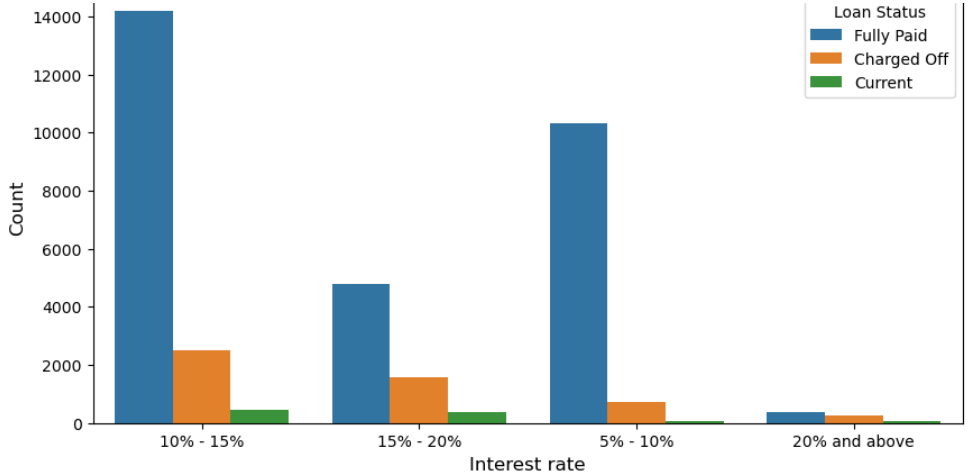
- Grade B loans default more
- 10 – 15% interest rate loans default more
- 36 month loans tend to default more
- People earning 0 – 60k per annum default more.
- Verified incomes default more

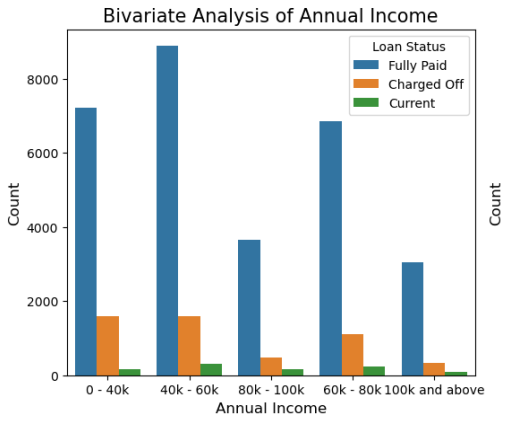
BIVARIATE ANALYSIS

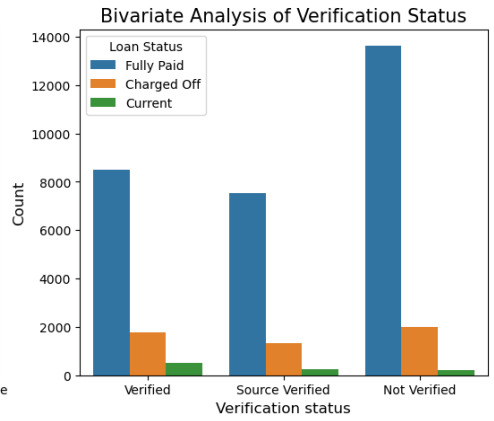
Bivariate analysis was also done on categorical variables using count plot charts.

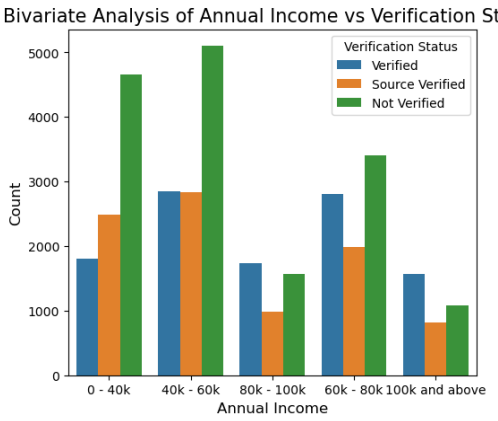
Primary approach was to see the counts for different categories across different loan statuses.

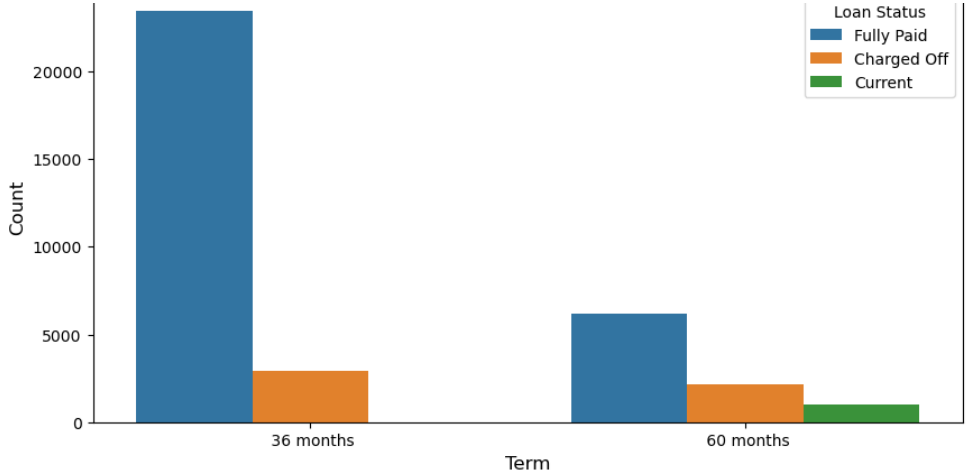












OBSERVATIONS USING BIVARIATE ANALYSIS

We see that the analysis is deeper.

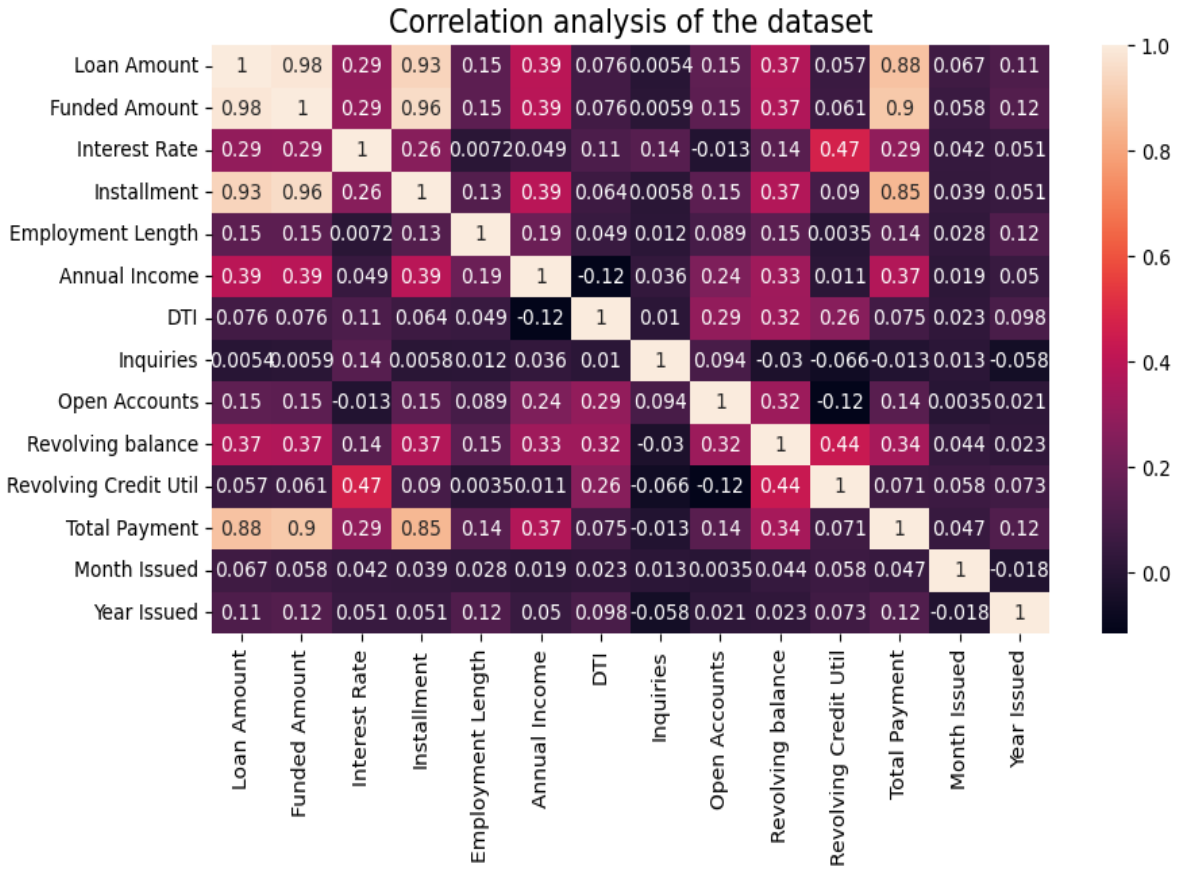
New points come forth lie

- Grade D loans default more in proportion than grade B as thought earlier.
- 15 – 20% loans default more in proportion as opposed to 10 – 15% interest rate loans.
- 60 month loans tend to default more in proportion.
- People earning 0 – 40k and 60 – 80k per annum default more.
- Verified incomes default more.

CORRELATION ANALYSIS

Conclusions on Correlation analysis:

- The DTI column has the lowest correlation with any column
- The DTI column has a negative correlation with Annual income
- Total amount, loan amount, installment and total payment, have very strong correlations with each other. Due to the strong correlation between these variables, we can analyse just the Loan amount.
- Month and year issued do not have any correlation with any variable, but it is not expected to have correlation as it is a measure of time. Any correlation which happens would be by chance.
- Inquiries column also does not have any strong correlation with any other variable.

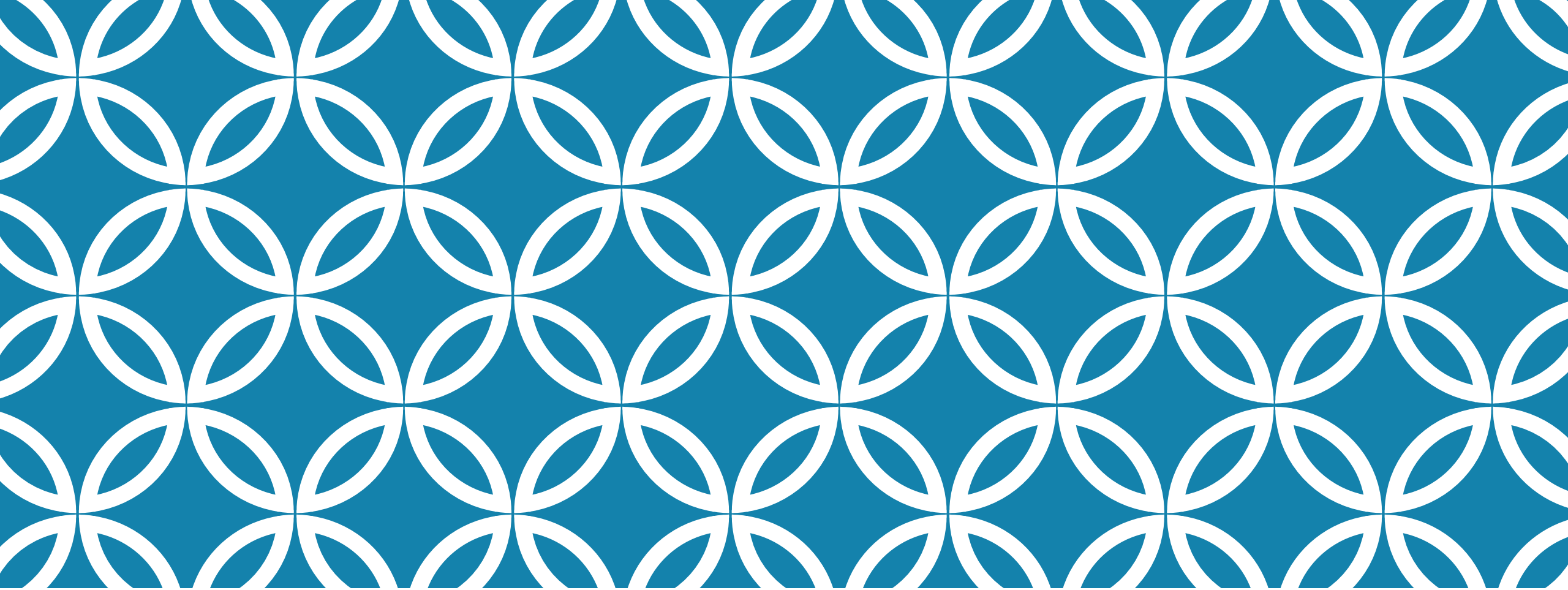


KEY FINDINGS

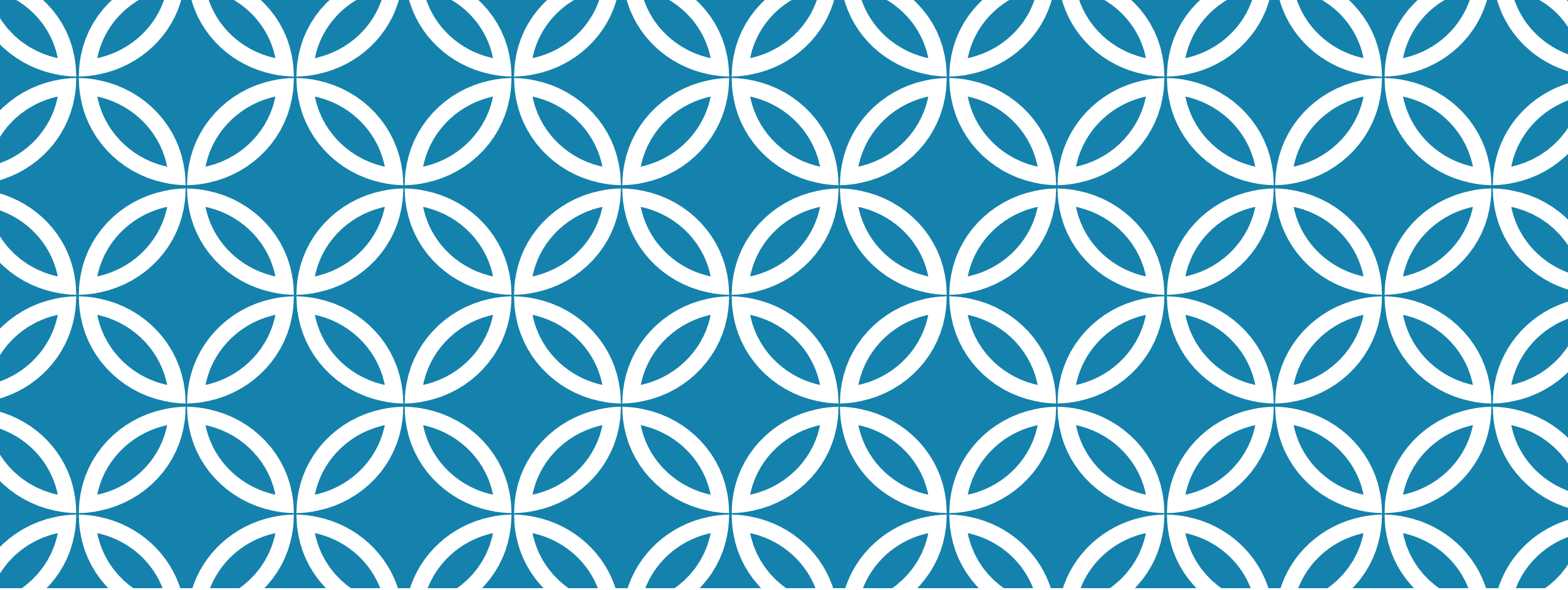
- The DTI column has the lowest correlation with any column
- The proportion of 60-month loans getting defaulted is higher.
- Proportion wise loans with grade D gets charged off the most with sub grade D3 and D4 which get charged off the most.
- Higher loan amounts are charged off more.
- People with 3 years of work experience and 10 years of work experience tend to default more
- People on Other Home ownership arrangements have a slightly higher percentage of loan defaulting, but they are extremely less in number. Then come people with mortgaged homes followed by people with rented accommodations.
- We see the least number of defaults in Q1 followed by a spike in Q2 and a bigger spike in Q4 after a slight decrease in Q3.
- The Q2 spike is contributed majorly by May. Analysis needed for the month of May

RECOMMENDATIONS AND CONCLUSION

- Higher loan amounts are charged off more. Hence proper vetting and background checks are necessary for higher loans.
- Stringent checks on business feasibility must be performed. Detailed RoI analysis should be performed as the percentage of defaulters that small business loans are more likely to default, followed by renewable energy projects.
- Perform an income and feasibility check for people with mortgaged homes and rental accommodations.
- To review loan policy changes post 2009 as we see that the number of customers have risen over the years exponentially.
- 2009 was the year when the defaults were t its lowest. But it increased ever since with increasing customers.



QUESTIONS?



THANK YOU