**EDA:**

Univariate Analysis:

Univariate analysis includes exploration of data wrt a single variable.

This includes histograms, pie charts, bar charts, box plots.

Histograms: continuous data displayed as frequency charts based on intervals

Barcharts: Categorical data within a variable can be displayed in multiple columns

Pie charts: frequency distribution of multiple categories as a circle. (Use only when less categories are present)

Boxplot: Representation displaying the medians, quartiles, and upper/lower bounds for outliers.

[When Should You Delete Outliers from a Data Set? - Atlan | Humans of Data](https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/)

Bivariate analysis:

Correlation tests with Covariance.

Scatter plot analysis. (Has to be checked. There might be multiple segments of data each providing a different analysis). If there is enough support for the outliers/multiple segments, each of them might require a different analysis.

**EDA insights**:

**min\_cibil\_score**: from the data 7.3 % of the population have min cibil score -1.

This cibil score is given to users who don't have any credit history at all. By trying to fit people with non existent credit history in the same model might create errors. Thus a new subset for this has to be created for people whose credit score is -1 and whose isnt. This population of 7.3% may influence the balance of the model. The min\_cibil\_score follows a bell curve(normal distribution) for the values that are not equal to -1. Hence this distribution can have a model of it’s own.

A screenshot of a cell phone

Description automatically generated

On checking the distribution of data by splitting on the basis of cibil score equal to -1 or not, We find no influence/correlation with other variable. Hence min\_cibil\_score for applications=-1 can also be resolved by replacing it with median or statistically proven quantile for financial risk analysis.

**d\_date\_of\_birth**:

The distribution is skewed showing a majority of users born between 1978 and 1995 using the service indicating users between the age of 23 and 40 i.e gen y and very few gen x.

However this graph also indicates the working population distribution by age thus not providing a lot of insights.

**A close up of a logo

Description automatically generated**

**tnc\_tenure\_first**: The distribution of tenures:

A screenshot of a social media post

Description automatically generated

Thus with such an imbalance of tenures, this can be another attribute to be taken to create different models for each individual tenures of 24 and 12 months and different for another.

**Correlation factors:**

The amount and interest rates have negative correlation. This implies the higher the loan amount the lower the interest rate applicable on it.

The tenure and amount have positive correlation. This trend implies that higher the tenure the loan amount would also be higher.

Correlation is also seen between 3m\_avg\_bal\_new and amount. Implies higher loan amounts are disbursed only to people with higher average recently recorded balance.

On comparing based on loan types.

The interest rates for lvl based loans are higher than ecom.

Yellow is loan\_type=‘lvl\_loan’ and blue is loan\_type=‘ecom’

A close up of a logo

Description automatically generated

**Boxplot analysis:**

Plots show insignificant insights regarding whether a variable is an outlier or not. Since these are correlated dependent variables replacing the particular field with mean/median is not enough. The correlated variables also have to be changed. Thus in this case the record has to be dropped. But since the data is required/mandatory to prevent overfitting of data. Specific case data cannot be excluded. If there were more data points for these outliers to have their own model, data can be segmented for certain intervals