Group Name: Tech Geeks
Name: Dhanashri Jagadale
Email: dhanashri.jagadale1998@gmail.com
College: Munster Technological University
Specialization: Data Science

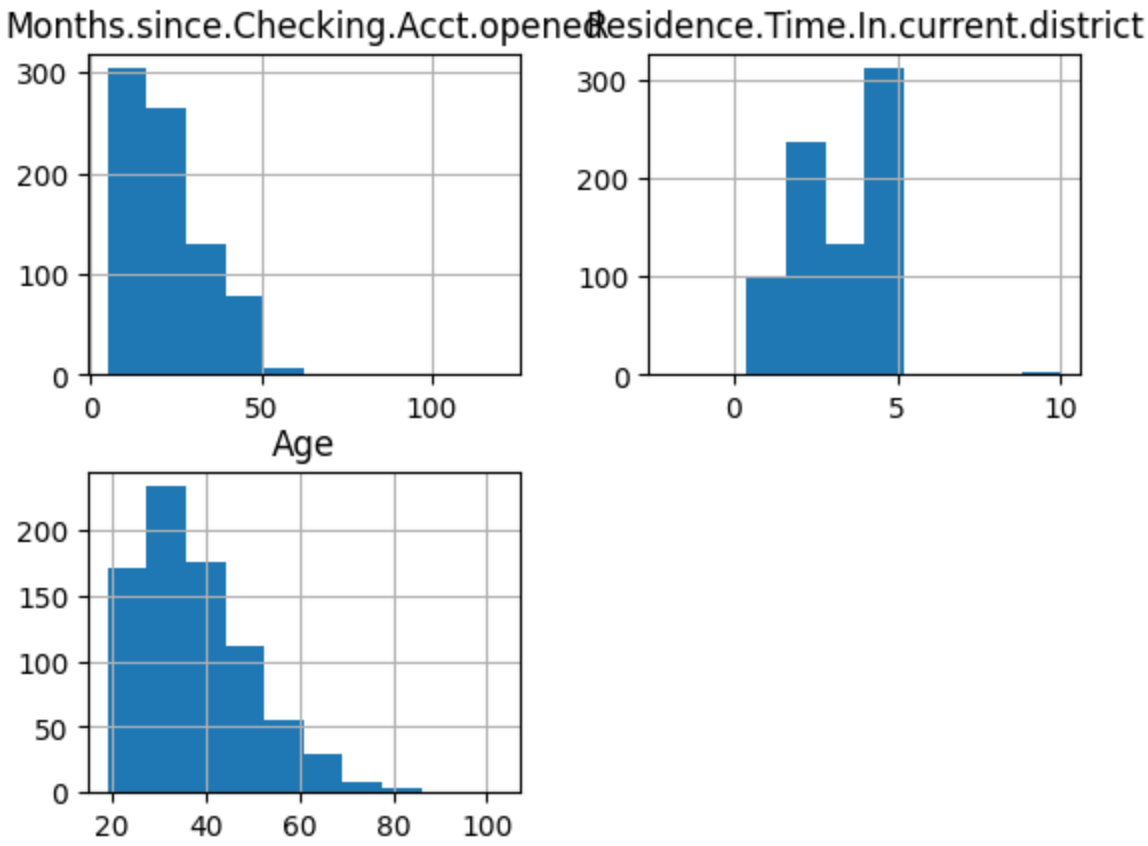
Problem Description: The problem is to develop a predictive model that can assess the credit worthiness of potential future customers of a financial institution. The available data set consists of 807 past loan customer cases, each with 14 attributes including financial standing, reason for the loan, employment, demographic information, foreign national status, years of residence in the district, and the outcome/label variable Credit Standing, which classifies each case as either a good loan or bad loan. The objective is to build a model that accurately predicts the credit standing of new loan applications, using the available data as the training set. The model should be able to identify the key factors that determine creditworthiness and provide insights to help the financial institution make better lending decisions.

Data understanding: The dataset contains information about various individuals who have applied for loans. The data includes information such as the individual's checking account status, credit history, loan reason, savings account status, employment status, personal status, housing status, job type, foreign national status, months since checking account opened, residence time in the current district, age, and credit standing. There are 807 rows with 14 attributes.

As the data contains missing values, we need to handle them as well. We can either remove the missing values or fill them with some values like the mean or median of the feature. The number of missing values is very small and there is no meaningful pattern to their occurrence therefore, Deleting missing values can be a reasonable approach. I have used dropna() function to remove any missing values before calculating the skewness. This is important because missing values can cause errors in the skewness calculation.

There are three numerical columns in the dataset. 'Months.since.Checking.Acct.opened', 'Residence.Time.In.current.district', 'Age'. Plotted Histograms and boxplots for these three columns to find out the outliers and skewness in the data.

Months.since.Checking.Acct.opened - 1.333082832090534
Residence.Time.In.current.district - 0.2396128529730903
Age - 0.9944826745950124



A skewness value of 0 indicates a perfectly symmetrical distribution. A positive skewness value indicates that the distribution has a longer right tail, meaning that there are more extreme values on the right side of the distribution. Conversely, a negative skewness value indicates that the distribution has a longer left tail, meaning that there are more extreme values on the left side of the distribution.

In my case, the "Months.since.Checking.Acct.opened" column has a skewness value of 1.333082832090534, which is positive, indicating that there are more extreme values on the right side of the distribution. The "Residence.Time.In.current.district" column has a skewness value of 0.2396128529730903, which is also positive but much closer to zero, indicating a nearly symmetrical distribution. Finally, the "Age" column has a skewness value of 0.9944826745950124, which is also positive but closer to 1, indicating a moderately skewed distribution.

In summary, the skewness values suggest that the "Months.since.Checking.Acct.opened" column has the most extreme values on the right side of the distribution, while the "Residence.Time.In.current.district" column has a nearly symmetrical distribution, and the "Age" column is moderately skewed.

Here are the steps to handle the skewness for each column:
Months.since.Checking.Acct.opened:

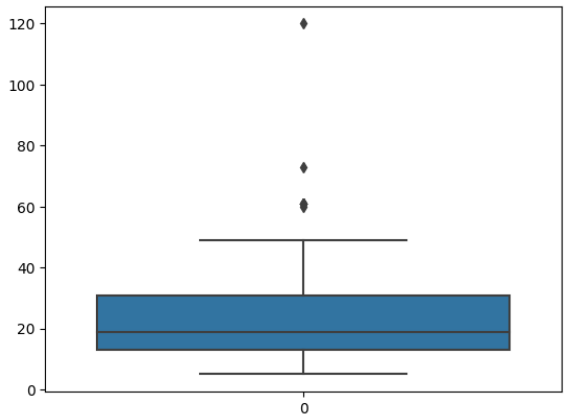
- Since the skewness value is greater than 1, this means the distribution is highly skewed to the right. In this case, we can apply a log transformation to reduce the skewness. After applying the log transformation, we can check the skewness value again to see if it has reduced.
- Residence.Time.In.current.district: Since the skewness value is close to 0, this means the distribution is approximately symmetric. In this case, we don't need to apply any transformation.
- Age: Since the skewness value is close to 1, this means the distribution is slightly skewed to the right. In this case, we can apply a log transformation to reduce the skewness. The log transformation can be applied.

And these are the results after the log transformation:

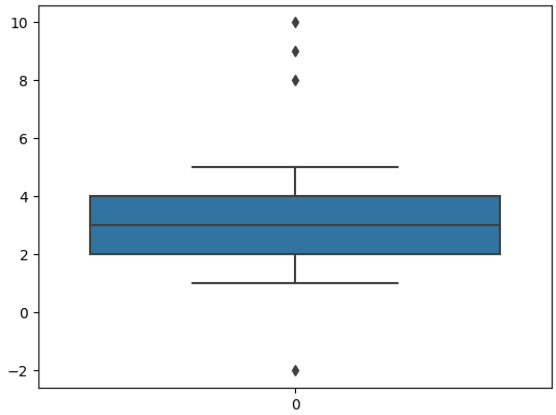
Months.since.Checking.Acct.opened	-0.11947624636158087
Residence.Time.In.current.district	0.2396128529730903
Age	0.21714904219380676

It seems that there are some outliers in the dataset.

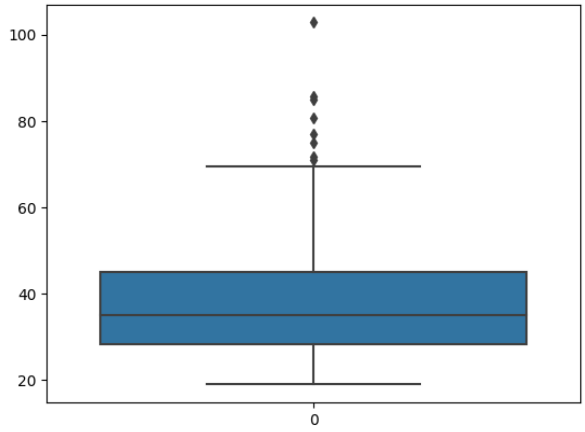
For the variable "Months.since.Checking.Acct.opened", there is 1 outlier with a value of 0.475885.



For the variable "Residence.Time.In.current.district", there are 4 outliers with values of -2, 10, 9, and 8.



For the variable "Age", there is 1 outlier with a value of 4.63264907.



It's important to investigate these outliers to understand if they are valid data points or if they are errors. Outliers can have a significant impact on data analysis, so it's important to handle them appropriately. I have dropped the rows with outliers using the drop method with the index of the rows containing outliers.