While importing the data I set the special attribute is personal loan as label (which is for target and predictable variable) and it's type is Binomial (0 or 1)

And the other 13 variables are regular attributes their types are

- age: integer
- ID: integer
- experience: integer
- income: real
- zip code: integer
- family: integer
- CCAvg: real
- Mortgage: real
- securities account: binomial (0 or 1)
- CD account: binomial (0 or 1)
- online: binomial (0 or 1)
- credit-card: binomial (0 or 1)
- education: nominal (categorical with three levels - high, primary, secondary)

The first step was to filter the customers' age to remove any values that are less than or equal 10 or greater than or equal 70 years old.

There are several ways to identify outliers in a dataset. Here are some commonly used techniques:

1. Boxplot: A boxplot is a graphical method used to display the range, median, and quartiles of a dataset. Outliers can be identified as points beyond the whiskers of the boxplot.

2. Z-score: Z-score is a statistical method used to determine how far away a data point is from the mean in terms of standard deviations. Points that fall outside a certain number of standard deviations from the mean can be considered outliers.

3. Interquartile range (IQR): The IQR is the range between the first and third quartiles of a dataset. Points outside the IQR can be considered outliers.

4. Density plot: A density plot can be used to visualize the distribution of a dataset. Outliers can be identified as points that fall outside the main bulk of the distribution.

To smooth data, some common techniques include:

1. Moving average: A moving average is a technique used to smooth out fluctuations in a dataset by taking the average of a subset of the data over a specified window size.

2. Exponential smoothing: Exponential smoothing is a method used to smooth out time series data by giving more weight to recent observations.

3. Kernel density estimation: Kernel density estimation is a non-parametric method used to estimate the probability density function of a dataset. It can be used to smooth out noisy data.

Outliers are data points that lie far away from the majority of the data points and can have a significant impact on the results of a predictive model. It is essential to deal with outliers before applying a predictive model to the data. Here are some ways to deal with outlier:

1 .Remove the outliers: This is the simplest way to deal with outliers. In this method, we simply remove the data points that are identified as outliers. However, this method should be used with caution, as it can significantly reduce the size of the dataset

2 .Winsorization: This method involves replacing the extreme values with the nearest value that is not an outlier. For example, if a data point is identified as an outlier, it can be replaced with the nearest value that is not an outlier

3 .Transformation: This method involves transforming the data to make it more normally distributed. Common transformation methods include logarithmic, square root, and Box-Cox transformations.

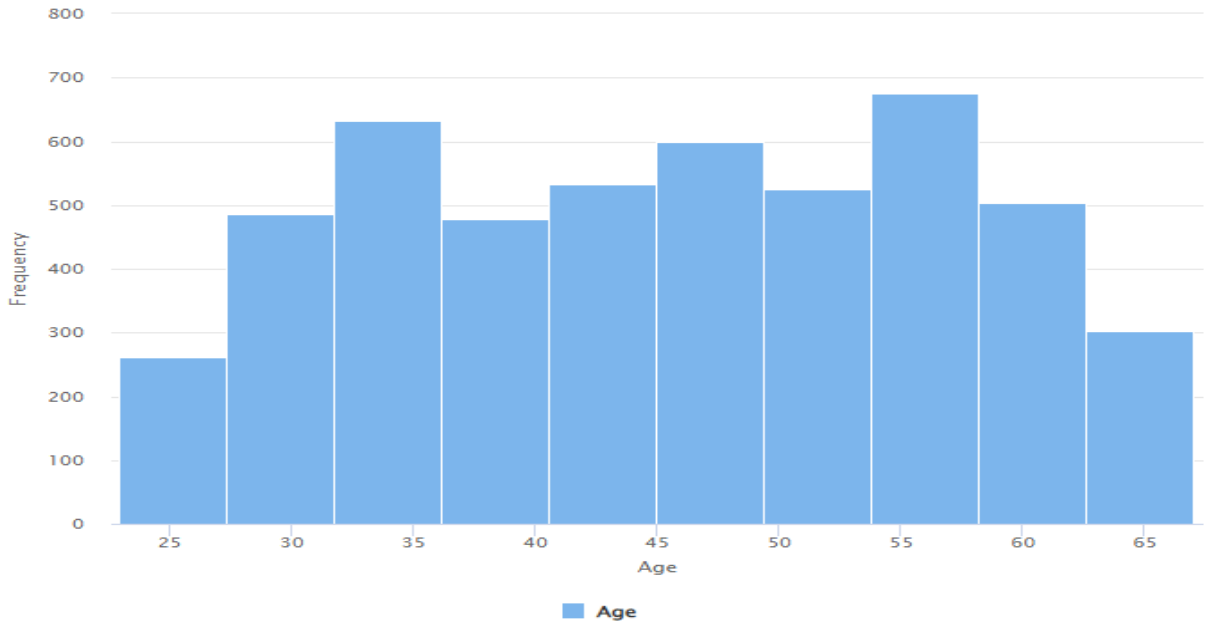
4 .Binning: This method involves grouping the data into bins or categories. This can help reduce the impact of outliers by reducing the range of values in each bin.

5 .Robust statistical models: Robust statistical models are less sensitive to outliers and can be used to model the data without removing the outliers. Examples of robust statistical models include median regression and robust regression.

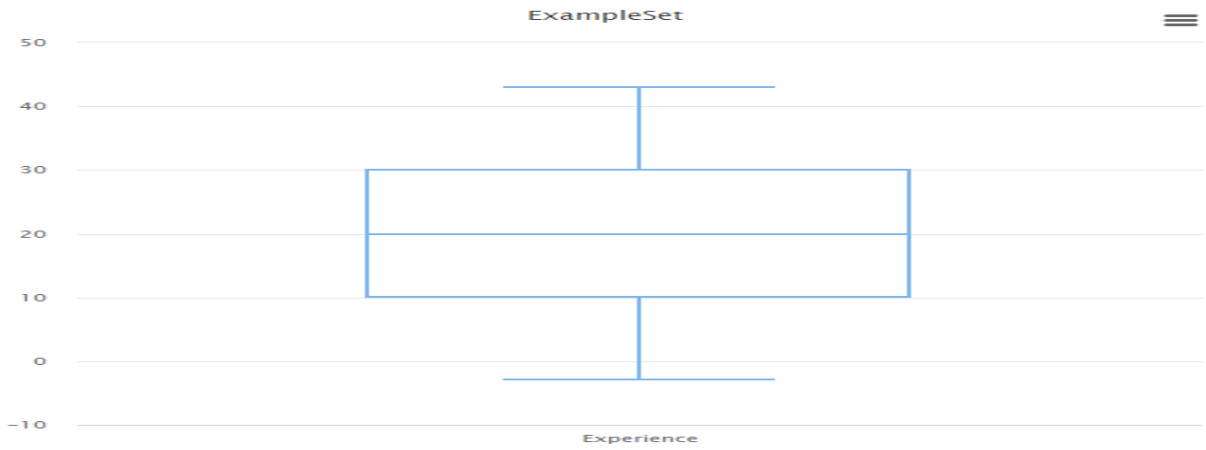
It is important to choose the appropriate method for dealing with outliers based on the nature of the data and the requirements of the predictive model.

By checking the outliers in our data I found that:

there is outlier in age column so, I make filter of age to be > 10 not to be >= 10 to remove 10 so it will be as follows



and in experience there are negative values (-3, -2, -1) I considered them as outliers



after removing them

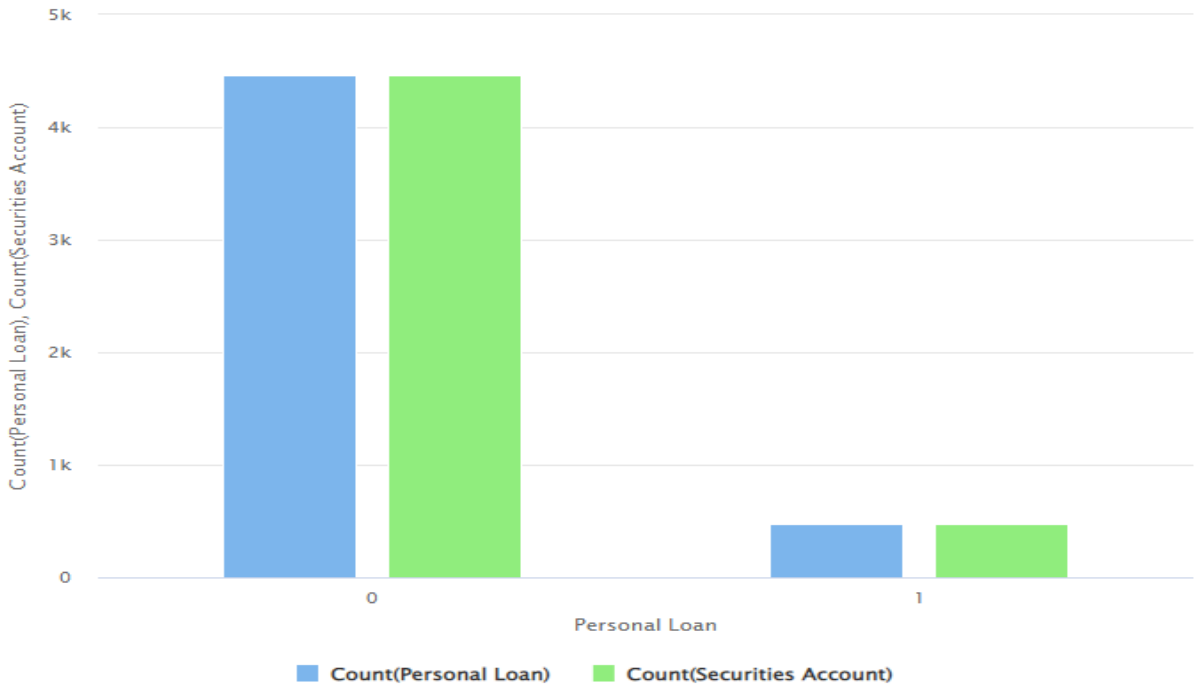
so, the minimum value is zero instead of -3

There are several methods that can be used for dealing with missing values in a dataset. Here are some common techniques:

1. Deletion: One way to handle missing values is to simply delete the rows or columns that contain them. This can be done using the "Filter Examples" or "Select Attributes" operators in RapidMiner. However, this approach can lead to loss of valuable information if there are too many missing values.
2. Mean/Median/Mode Imputation: This involves replacing the missing values with the mean, median, or mode of the respective variable. This can be done using the "Replace Missing Values" operator in RapidMiner.
3. Hot-Deck Imputation: This involves randomly selecting a value from a set of similar cases to replace the missing value. This can be done using the "Impute" operator in RapidMiner.
4. Regression Imputation: This involves using a regression model to predict the missing values based on the non-missing values. This can be done using the "Impute" operator in RapidMiner with a regression model (such as linear regression or decision trees) as the imputation method.
5. Multiple Imputation: This involves creating several imputed datasets based on different methods and combining them to get a more accurate estimate of the missing values. This can be done using the "Multiple Imputation" operator in RapidMiner.

It's important to choose the appropriate method based on the type and amount of missing data, as well as the analysis goals.

The best way to deal with missing values in “personal loan” and “Securities Account” columns is to replace the missing with the mode (as they are categorical variables)



So, by checking the mode for both of them we found that "0" has more counts than "1", that’s why I replace the missing with zero

Redundant columns are those that do not provide any useful information to the model or are highly correlated with other columns.

In this dataset, the columns "ID" and "ZIP Code" can be considered as redundant as they do not provide any useful information for predicting the personal loan variable. Therefore, we can remove them from the dataset.

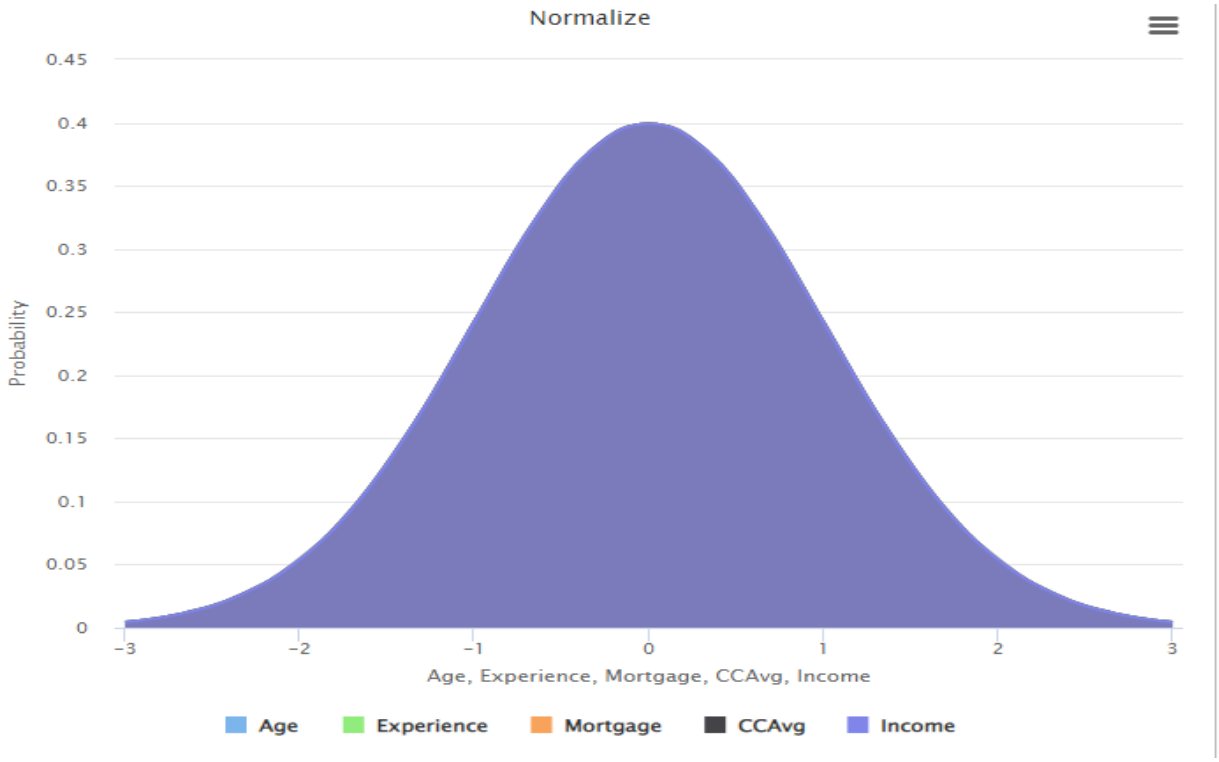
The categorical column in this dataset is "Education", which has three levels (High, primary, and secondary).

To use this variable in our model, we need to convert it to numerical values. We can do this by applying a one-hot encoding transformation, which will create three new columns for each level of the "Education" variable.

To apply the suitable feature scaling for the needed variables:

The variables "Age", "Experience", "Income", "CCAvg", and "Mortgage" are continuous and have different scales, which can affect the performance of some machine learning models. Therefore, we need to apply feature scaling to these variables to normalize their values to ensure they have the same range and contribute equally to the model.

There are different methods to apply feature scaling, but the most common ones are Min-Max scaling and Z-score scaling. Min-Max scaling transforms the values to a range between 0 and 1, while Z-score scaling transforms the values to have a mean of 0 and a standard deviation of 1.



After completing these steps, we will have a preprocessed dataset ready for modeling.