Report: Data Wrangling on Bank Customer Churn dataset

The Dataset is about bank customers churning and can be found on Kaggle:

https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

Disclaimer: The dataset above is simulated.

Before I started my notebook, I made sure that the dataset was in the same directory as my notebook. The first thing that I did after reading in the csv file was to check out the shape of the dataset. The bank customer churn dataset is of the shape 10000 rows and 14 columns.

I first took a glimpse into the data and saw that of the 14 columns, 13 columns are feature columns and the '***Exited'*** column is the response column.The next thing that I did was to check for null values. Fortunately, our dataset does not contain any null values and this is because the dataset was from Kaggle, and it was already very clean. This is not often the case with real world data.

I checked if there were any redundant columns we could drop in our dataframe and found the column **'RowNumber'** to be redundant. After dropping the column, the new shape of the dataframe is 10000 rows and 13 columns. The next thing that needs to be done is converting the categorical columns; 'Gender' and 'Geography' into numerical values. This is done because during modelling, some actions can not be performed on categorical values. I converted the **'Geography'** column into 3 numerical values and the **'Gender'** column into 2 numerical values.

For visual purposes, I moved the response variable column **'Exited'** to the left side of the table. I find it quicker to view the data this way, and also makes splitting the dataset into train/test sets easier at a later stage. The last thing I did was to check for outliers in the data. In this case, we look for any extreme values in the min and max fields of the columns of the dataframe. For our data, there seems to be no outliers.