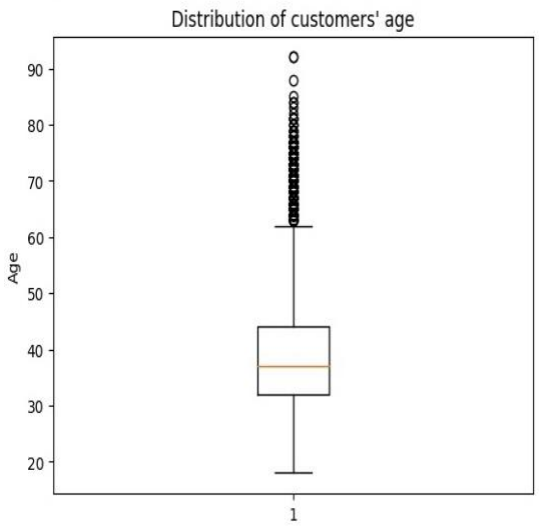
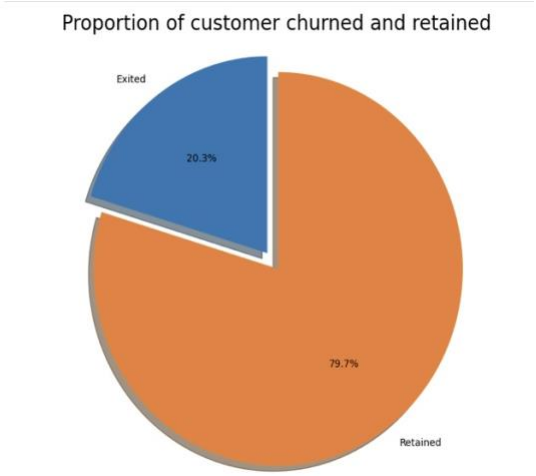
Week 5

Feature engineering is an important part of the machine learning pipeline, as it can have a significant impact on the performance of the model. It involves creating new features or variables from existing ones in a dataset. In this project, we created a new variable called ‘age_outliers’ based on existing ‘age’ feature. We kept the age outliers, or individuals with ages significantly higher or lower than a threshold, as a separate binary variable because they were likely to be predictive of churn.

To identify potential outliers in the data, we drew a boxplot (see graph below). Outliers were plotted as individual points beyond the whiskers. The outlier threshold was calculated as 1.5 times the IQR above the third quantile (Q3) of the age variable. In this case, the threshold was 62. We then created the ‘age_outlier’ column to indicate whether the age value for each row was an outlier, using the predefined threshold. If a customer’s age was greater than 62, it was considered an age outlier and ‘age_outlier” column was set to 1. Otherwise, the ‘age_outlier’ column was set to 0.

The pie chart ‘Proportion of customer churned and retained’ showed that, among outliers, about 20% of customers exited while 80% chose to stay at the same bank.





Principle Component Analysis (PCA) is a common technique to reduce the dimensions of a dataset. It is a dimensionality reduction method that finds a new set of variables that are linear combinations of the original variables. However, this dataset is simple with a low number of features, so PCA is not necessary.