TAA 3

1 Anomaly detection

1. Preparing the dataset
   (a) Load the `gapminder_data_graphs.csv` dataset. Filter to keep only the year 2015. Remove textual columns.
   (b) Apply an anomaly detection using `IsolationForest` from sklearn.
   (c) Plot the distribution of anomaly scores (`decision_function`)
   (d) Plot the most anomalous countries
   (e) To understand why these countries are considered anomalies, print for each variable the difference from the mean. Using a relative distance (scale by variance) can be more informative. Interpret.
   (f) Do the same using the PCA approach. You can use `pca.inverse_transform`
   (g) Do the same using Gaussian Mixture (`score_samples`).

2 Imbalance

(a) In the `cars_synth_clean.csv` dataset, create a new target variable called `15k+` , which is true if a car is worth more than 15000. Remove the price column
(b) Check that there is class imbalance
(c) Using `RandomForestClassifier` perform a classification to predict this target variable. Don't forget to remove the price from the variables used for the prediction...
(d) Compute the Accuracy, the ROC_AUC. Plot the confusion matrix
(e) Using `RandomUnderSampler` from `imblearn` library, perform under-sampling. If you prefer, you can do it manually.
(f) Check the new scores for accuracy, ROC_AUC. Plot the confusion matrix
(g) Explain the difference: what is this model better and worst at doing ? Why it explains the differences in the scores ?

3 Feature selection

(a) Load the `cars_synth_clean.csv` dataset
(b) Use a `clustermap` from `seaborn` library to visualize the correlations between variables (`df.corr`)
(c) If we were using a threshold of 0.5, what variable should we keep?

4 Going Further

2. SMOTE and SMOTER
   (a) Use instead the SMOTE strategy from the same library and compare the results
   (b) When trying to predict directly the price, you are confronted to the same problem. To observe the poor predictions on rare values, you can draw a scatterplot with a relation between target values and average errors.
   (c) Search a solution to perform SMOTER. Observe how the performance is affected.