**Data preprocessing steps**

**Memory optimization:**

* In machine learning, memory is used for storing various components such as data, model parameters, intermediate results, and temporary variables.
* Efficient memory management techniques are employed to minimize memory consumption and optimize memory usage, which can significantly impact the performance and efficiency of machine learning algorithms.

**Null value Treatment:**

* In machine learning, handling null or missing values in your dataset is an important step in the data preprocessing phase. Null values, also known as missing values, are data points that are not available or are marked as "NaN" (Not a Number) in the dataset.
* Dealing with null values is crucial because they can negatively impact the performance and accuracy of machine learning models if not handled properly.
* Here are some common approaches for null value treatment in machine learning:

**Removal of Null Values:**

* This approach involves removing rows or columns with null values from the dataset.
* If the number of rows or columns with null values is relatively small and removing them does not significantly impact the size or integrity of the dataset, this approach can be used.
* However, caution should be exercised to ensure that important information is not lost by removing rows or columns.

**Imputation:** Imputation is the process of filling in missing values with estimated or imputed values. There are several techniques for imputing null values, including:

* **Mean/Median/Mode Imputation:** Null values are replaced with the mean, median, or mode of the respective feature. This approach is simple and can work well when the null values are missing at random and do not introduce any bias into the dataset.
* **Forward or Backward Fill:** Null values are filled with the previous or subsequent value in the same column. This approach is useful when dealing with time series data or datasets with a natural ordering.

**Regression Imputation:**

* Null values are predicted using regression techniques based on other features in the dataset. Regression imputation can be more accurate if there are strong correlations between the features.

**K-Nearest Neighbors (KNN) Imputation:**

* Null values are imputed based on the values of their k-nearest neighbors in the dataset. This approach can be effective when there are patterns or clusters in the data.

**Outlier Treatment**

**Definition of the outliers**

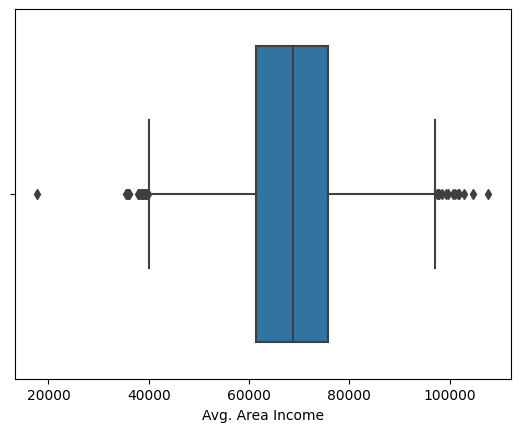
The outliers are the extreme values within the dataset i.e. the data points which vary greatly from the expected values—either being much larger or significantly smaller. And the outliers are removed if they exist

There are various ways to predict the outliers like

a) describe method (if mean value is affected much)

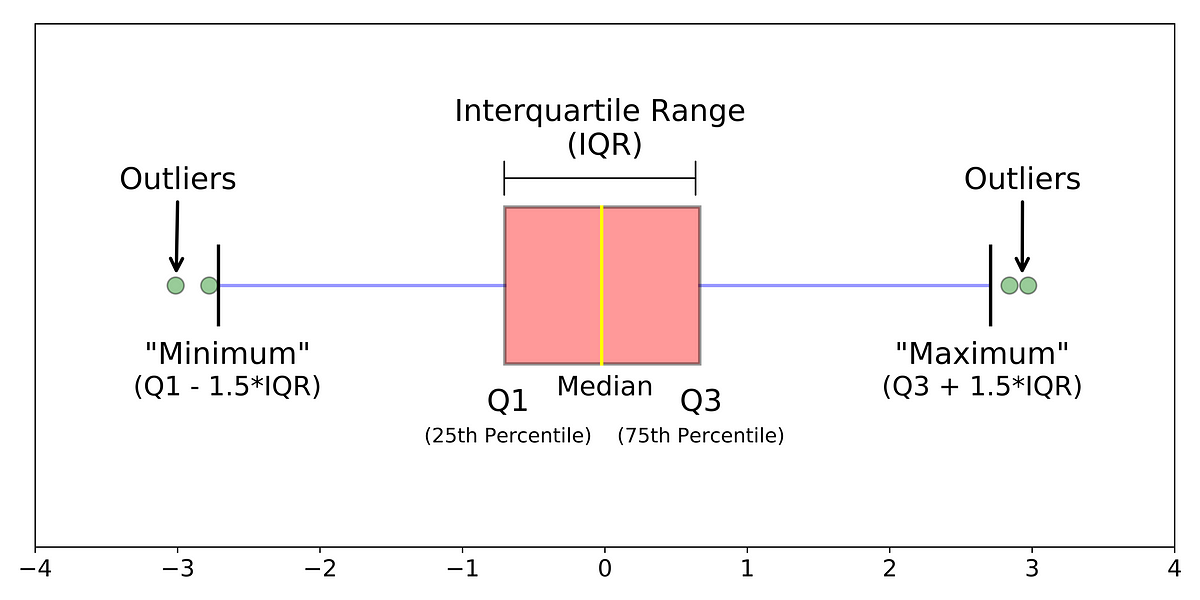
b) \*visualization method \*(like box plot, scatter plot, histograms)

Box plot example:



we can see the outliers in the box plot.

c) using the statistical method called interquartile range (IQR) --- For data that follows a normal distribution, the values that fall more than three standard deviations from the mean are typically considered outliers.



d) z-score--- (x-mean)/standard deviation

**Feature Scaling -----Normalization/Standardization?**

* Feature scaling is a preprocessing step and method used to normalize the range of independent variables or features of data.

**Why it is necessary:**

* Consider an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.
* **Normalization** is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data, like K-Nearest Neighbors and Neural Networks.
* **Standardization**, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.