**Data Cleaning**

Introduction

Data Cleansing is the process of analyzing data for finding incorrect, corrupt, and missing values and Correcting it to make it suitable for input to data analytics and various machine learning algorithms. It is the premier and fundamental step performed before any analysis could be done on data.

Reasons for data corruption:

Data is collected from various structured and unstructured sources and then combined, leading to duplicated and mislabeled values.

Different data dictionary definitions for data stored at various locations.

* Manual entry error/Typos.
* Incorrect capitalization.
* Mislabeled categories/classes.

Data Cleansing Techniques

Techniques used for cleaning data depends upon the quality of the dataset. Some of the techniques are as follows:

* Handling missing values: Handling missing values is the most important step of data cleansing it can either be deleted or imputed.
* Data Normalization: Transforming data into a consistent format, like converting all text to Lowercase.
* Data Transformation: Converting data types, such as changing data fields from string to datetime format.
* Removing Duplicates: Eliminating duplicate records or rows.

**Outliers**

Outliers are values within a dataset that vary greatly from the other, they are either much larger or significantly smaller. Outliers may indicate variability in a measurement, experimental errors or a novelty. Outliers can cause anomalies in the results obtained during the process of data analysis and this requires special attention.

There are two main reasons why we give outliers speed attention which are;

* Outliers may have a negative effect on the result of an analysis
* Outliers or their behavior may be the information that a data analyst requires from the analysis.

Types of outliers

* Univarite outliers: this is the extreme value that relates to just one variables
* Multivariate outliers: is a combination of unusual or extreme value for at least two variables

How outliers end up in datasets? Here are some of the more common causes of outliers in datasets:

* Human error while manually entering data, such as a typo
* Intentional errors, such as dummy outliers included in a dataset to test detection methods
* Sampling errors that arise from extracting or mixing data from inaccurate or various sources
* Data processing errors that arise from data manipulation, or unintended mutations of a dataset
* Measurement errors as a result of instrumental error
* Experimental errors, from the data extraction process or experiment planning or execution
* Natural outliers which occur “naturally” in the dataset, as opposed to being the result of an error otherwise listed. These naturally-occurring errors are known as novelties

How to identify outliers in a dataset, methods commonly used to identify outliers are visualizations or statistical methods

How to identify outliers using visualizations: analysts create data visualizations to present data graphically in a meaningful and impactful way, in order to present their findings to relevant stakeholders. These visualizations can easily show trends, patterns, and outliers from a large set of data in the form of maps, graphs and charts.

And we have two methods of using visualizations which are the Box plots and Scattered plots