جامعة محمد الأول بوجدة
UNIVERSITE MOHAMMED PREMIER OUJDA
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵎⵓⵃⴰⵎⵎⴷ ⴰⵎⴰⵣⵡⴰⵔ ⵏ ⵡⵓⵊⴷⴰ

Data Cleaning Workshop:

Club Data Science And Cloud Computing

Causes & Impact of Missing Values

Important Concepts :

- Handling Missing Values.
- Handling Outliers.
- Learning some Important Data Manipulation Functions useful for Cleaning the Data.

Missing values :

- Missing values occur when there is no data or value stored for the variable in an observation.
- Missing data are a common occurrence and can have a significant effect on the conclusions drawn from the data.
- There can be several causes for the occurence of missing values in a data set.
- Most statistical procedures require a value for each variable.
- The missing data can cause a bias in the estimation of parameters and the accuracy of our machine learning model can be affected.

Types of Missing values :

MAR - Missing At Random

● The missing values in this category have some association with other features of the dataset.
● The variable which has missing values can be linearly related to any other variable of the dataset.

MAR - Missing At Random

● The missing values in this category have some association with other features of the dataset.
● The variable which has missing values can be linearly related to any other variable of the dataset.

MNAR - Missing Not At Random

● These types of missing values are the values which are missing with some specific reasons.
● And we will have a clear understanding and logic for the missing value.

Imputing Missing Values using Mean/Median/Mode :

Mode :

●Having some missing values in a categorical variable called Gender.
● Two Types of Values: Males and Females.
● To Impute or Replace the Missing Values from such a Column we can use the Mode Function.
● It returns the maximum occurring value. Which in turn can be used to Replace the Missing Values.

Median :

● We should impute the missing values in numerical variables using the Median function.
● If there are outliers present in the data.
● As the median function is not sensitive towards outliers.

Mean :

● We should use the “Mean” Function when the data does not contain any Outliers.

● Mean Function is very sensitive towards Outliers.

Dealing with Outliers

Outliers :

●Outliers are extreme values that fall a long way outside of the other observations.
●Outliers are those Values which are very very different from most of the Values.
●Example: There is a Column called “Age” for B.tech College.
○Where students generally have age around 17 to 24.

Types Of Outliers :

1.Univariate Outliers.
2.Bivariate Outliers.

●***Univariate Outliers*** are the points which are beyond the normal values in a single variable.

●***Bivariate Outliers*** are the points which lie far from the expected values when two variables are plotted against each other.