Data Cleaning and Analysis (Pre Model Application)

1. Importing the required packages



1. Importing the Dataset
2. Peeking your data:

Dimensions of Data: Shape

DataType:

Descriptive Stats:



Class Distribution:

Data.groupby(‘class’).size()

Correlation:

Data.corr(‘pearson’)

Skewness/Kurtosis:

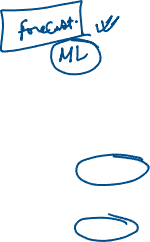
Data.skew(),

Visulization:

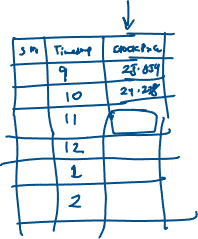
1. Univariate: Histogram, BoxPlot
2. Multivariate: Scatterplot, linechart, Heatmap (Correlation)



Data Preprocessing:



1. Rescaling Data – MinMaxScaler
2. Standadization: StandardScaler
3. Label Encoding: One Hot Encoding
4. Missing Value :



Detection:

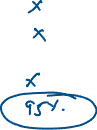
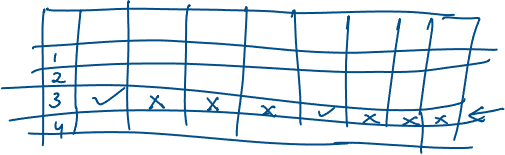
Df.isna().sum()



Imputation:

1. Mean, Median or Mode imputation
2. Backward/Forward fill
3. Forecast the missing value
4. Remove rows with missing values (if multiple empty values are there in that row)

Remove the entire column (majority of the value is missing in that column



1. Outliers: Identify and treat

Identifying: BoxPlot, Z-score, IQR Method



Data Split : Train Test Split

Train Model



Test the model



Prediction

**Evaluation Matrix (Evaluate our Model):**

Regression: R-Squared, Adj-R Squared, MSE, RMSE

Classification:

Clustering

….

Extra:

Pipelines: Deployment

Spark:

Visualize your findings: Tableau/PowerBI

EOF>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>