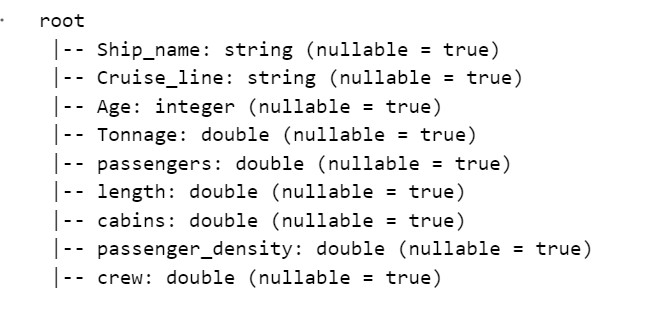
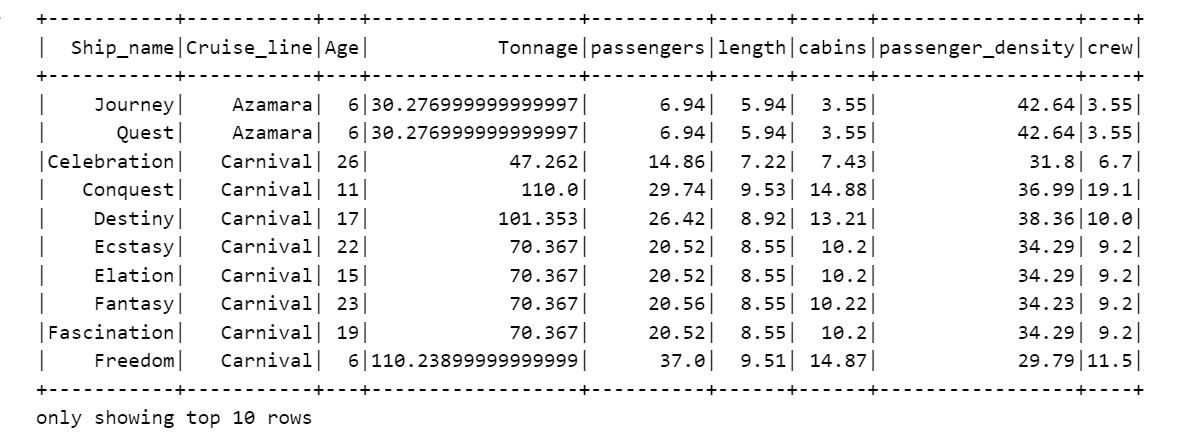
BDA Lab Internal – 2

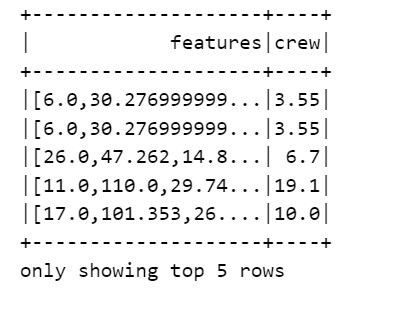
1. Linear Regression using Spark:

OUTPUT:

Displaying Data and Schema:



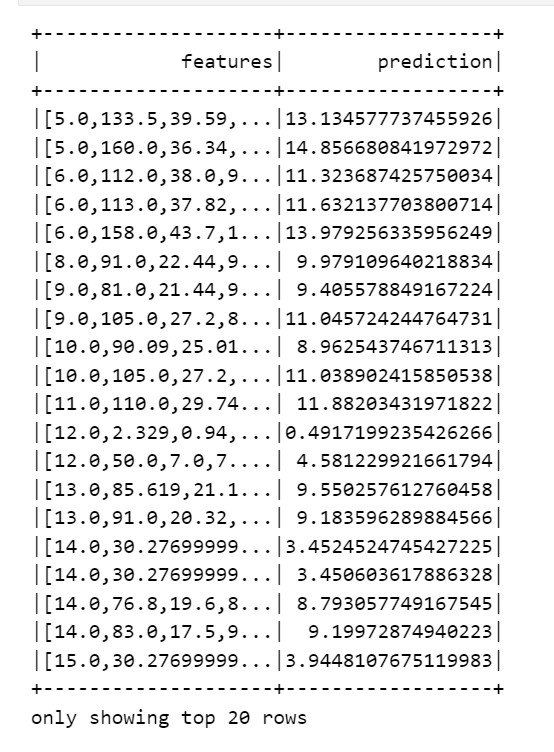
Preprocessing Data:



Rsquared error of the model after training:



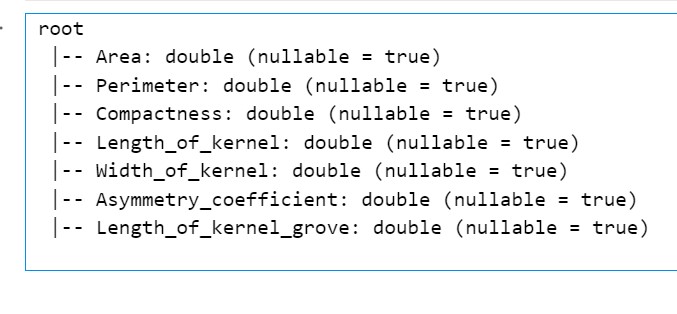
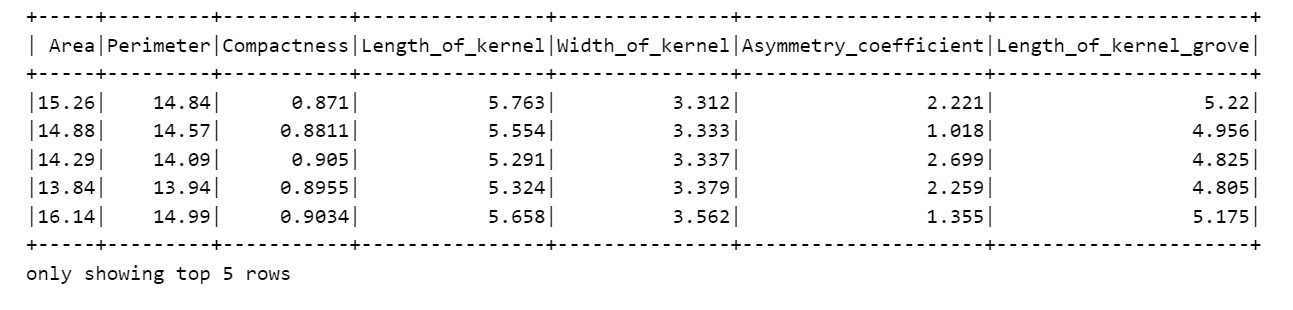
Result with the testing Dataset:



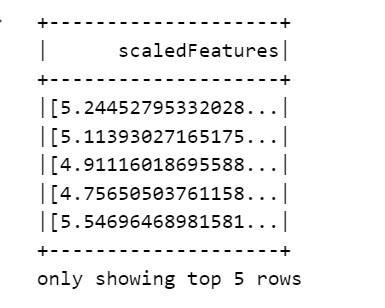
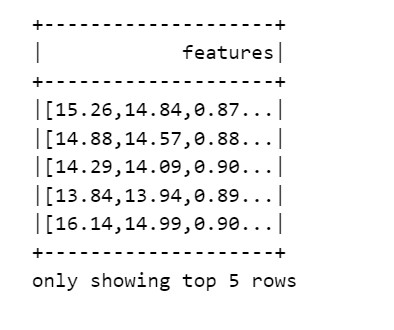
1. KMeans clustering using Spark:

OUTPUT:

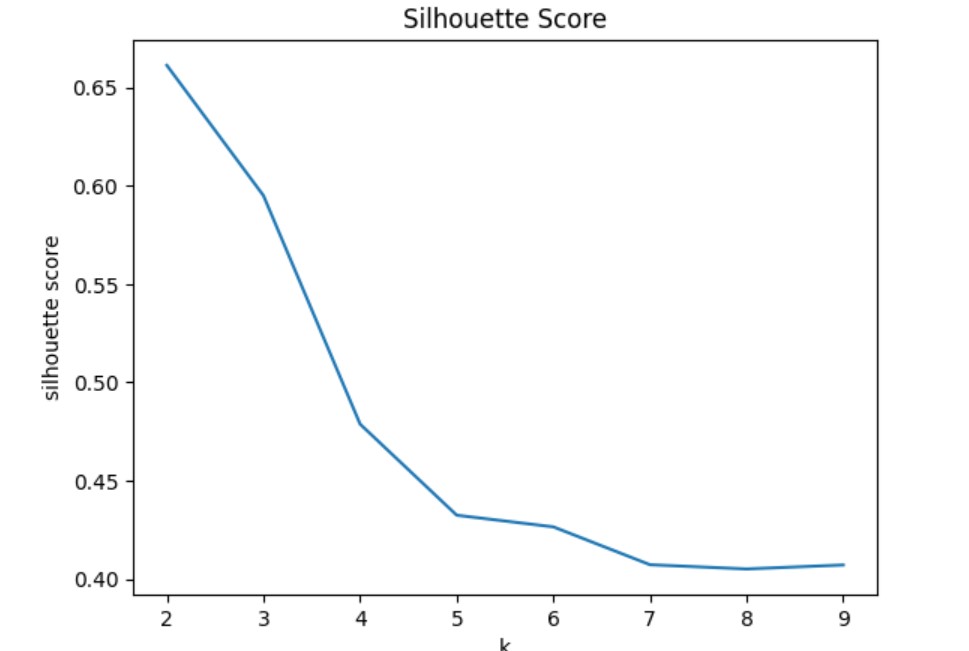
Displaying Data and its Schema:



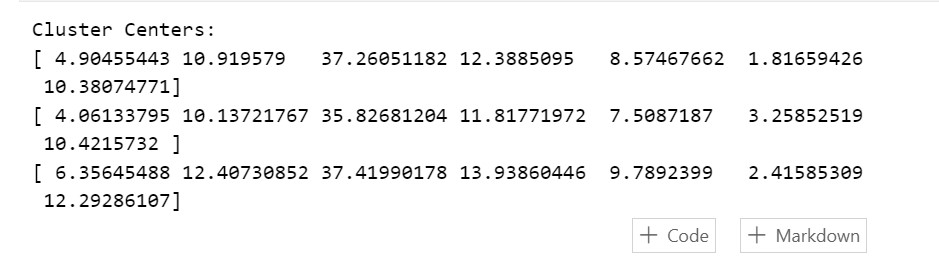
Preprocessed Data before and after scaling:



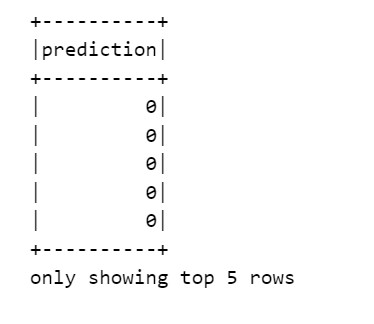
Silhouette Score graph for different k values:



Displaying Cluster centers:



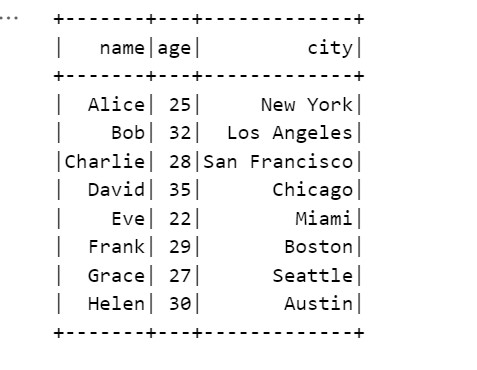
Displaying predictions for test data:



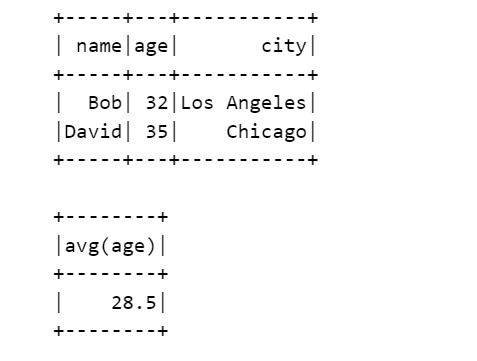
1. Dataset processing using Spark:

OUTPUT:

Printing Data:



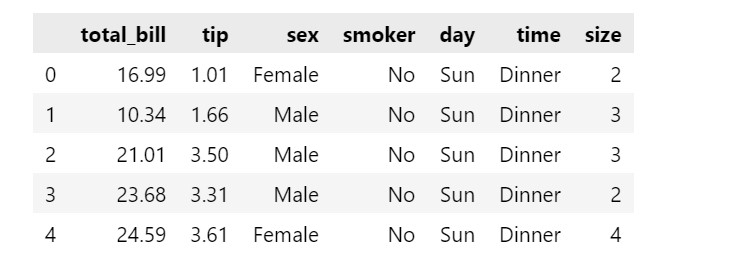
Displaying filtered data:



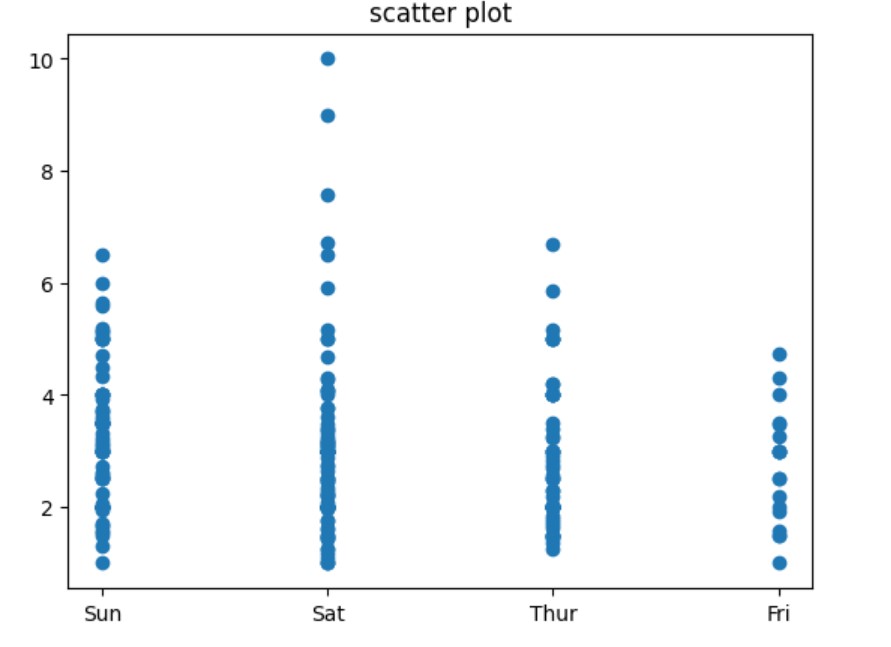
1. Data Visualization techniques:

OUTPUT:

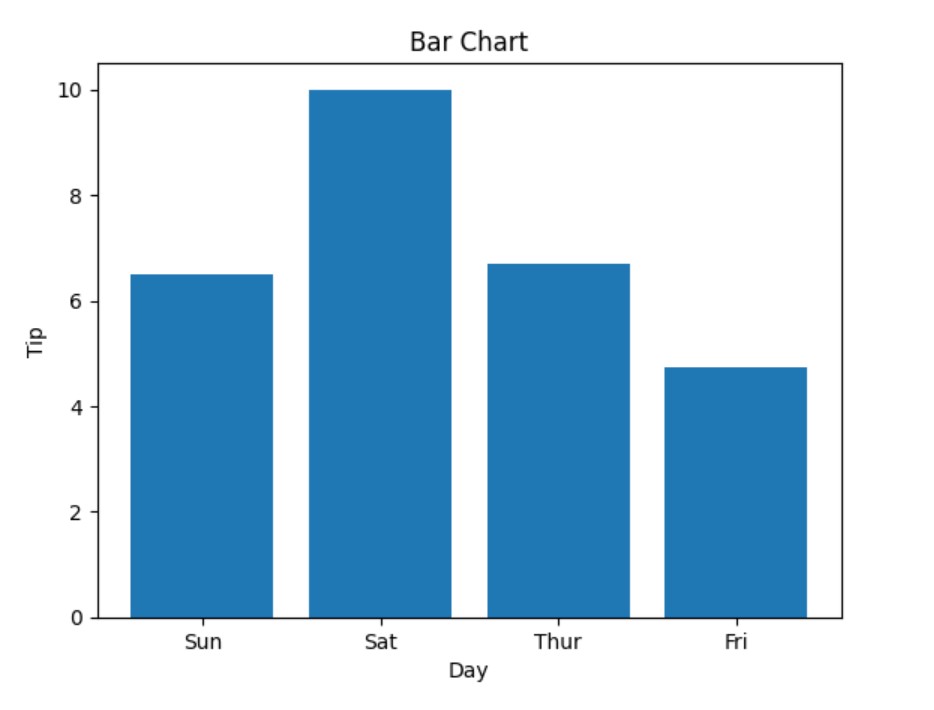
Displaying data:



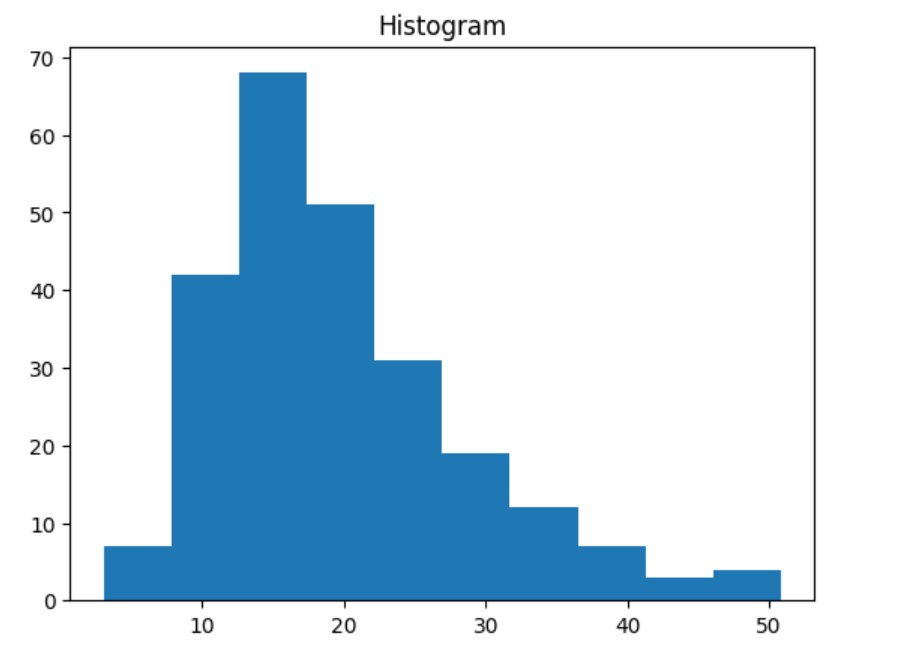
Scatter plot:



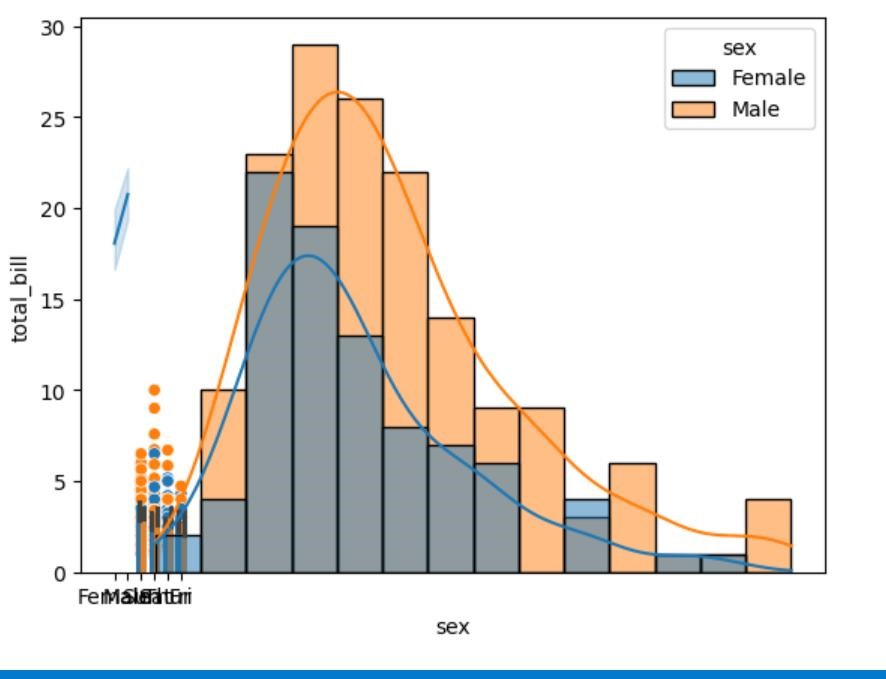
Bar chart:



Histogram:



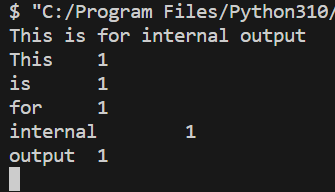
All plots together using seaborn:



1. Word count using Mapreduce:

OUTPUT:

**Mapper.py :**



**Reducer.py :**

