DSCI 5350 Assignment-3 Satyasriharsha Buddha

Databricks Published Link: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4179776347872718/3707216431996142/222406585371466/latest.html

1. Read the provided file into Spark successfully

Created a cluster-BigQuery and Scala Notebook DSCI5350-Assignment3-on-BigQuery

A screenshot of a computer

Description automatically generated

The file was uploaded

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

Path- dbfs:/FileStore/shared\_uploads/sriharshabuddha@my.unt.edu/Steel\_industry-1.csv

Importing Required Libraries and Loading the file into a dataframe

A screenshot of a computer

Description automatically generated with medium confidence

Checking the dataframe

A screenshot of a computer

Description automatically generated with medium confidence

2)Write Scala code to display Interquartile Range of the ‘Usage\_kWh’ without using libraries.

Creating usageKWH from the dataframe and checking the values

A screenshot of a computer

Description automatically generated with medium confidence

Sorting the values in an array

A screenshot of a computer

Description automatically generated with low confidence

Finding Median of the array and q1

A screenshot of a computer

Description automatically generated with low confidence

Finding q3 and the difference and printing

A screenshot of a computer

Description automatically generated with low confidence

3.Take the mean of ‘Usage\_kWh’ round to next integer and break it into 3 whole numbers such that they sum up to the original number and count number of ways to do so and display it.

Finding the mean of ‘Usage\_kWh’ and rounding off to the next integer

A screenshot of a computer

Description automatically generated with low confidence

Finding number of ways to break the rounded mean into three whole numbers and displaying the answer

A screenshot of a computer

Description automatically generated with low confidence

Write a Spark SQL query to display the details of highest ‘Usage\_kwh’ in each

load type (top 3)

Forming a temporary view and verifying the data

A screenshot of a computer

Description automatically generated with medium confidence

Forming a subquery with load\_type and usage\_kwh partitioned by load type and usage\_kwh in descending order and limiting to top 3 highest values and displaying the result

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Apply any machine learning algorithm on data using Spark ML-lib and give description.

Perform linear regression and predict 'Usage\_kWh' based on 'Leading\_Current\_Power\_Factor' using Spark MLlib.

Importing the libraries and getting the data

A picture containing text, screenshot

Description automatically generated

Splitting the data into test and train dataset

A screenshot of a computer

Description automatically generated with low confidence

Running a linear regression



Fitting model on the training data

A screenshot of a computer

Description automatically generated with low confidence

Showing predictions

A screenshot of a computer

Description automatically generated with medium confidence

Evaluating the model and finding out the RMSE value

A screenshot of a computer

Description automatically generated with low confidence

Finding the R square value

A screen shot of a computer

Description automatically generated with low confidence

Based on the given RMSE and R-squared values, we can infer that the current model's performance in predicting 'Usage\_kWh' is not very accurate or strong. It may require further improvements, such as incorporating additional relevant features, trying different regression algorithms, or optimizing model parameters, to enhance its predictive power.