Zeppelin-Spark In Class Assignment

Name: Harshit Aggarwal

Student Number: N01550495

1.Load data into a Spark dataframe.

A picture containing text, electronics, screenshot, software

Description automatically generatedA screenshot of a computer

Description automatically generated

Print the dataframe schema

A computer screen shot of a computer

Description automatically generated with low confidence2. Print the dataframe schema

3.Filter the data frame to show units sold greater than 8000 and unit cost greater than 500 ("&&" operator can be used for multiple "AND" conditions).

A computer screen shot of a computer screen

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence4). Aggregate the data frame via group by “Region” and count.

5.) Create a separate data frame with the above group by results.

A screenshot of a computer

Description automatically generated

Save this new subset dataframe as a csv file into HDFS – make sure it is saved as a single file in HDFS.

1. Save this new subset dataframe as a csv file into HDFS – make sure it is saved as a single file in HDFS.

A screenshot of a computer

Description automatically generated with medium confidence

1. Create two views using the “createOrReplaceTempView” command.A screenshot of a computer

   Description automatically generated with medium confidence
2. View on “Salesview” from the first dataframe.

A screenshot of a computer

Description automatically generated with medium confidence

1. View on “Regionview” from the second dataframe.

A screenshot of a computer

Description automatically generated with medium confidence

1. Using SQL select all from “Regionview” view and show in a line graph.

A picture containing line, plot, text, screenshot

Description automatically generated

1. Using SQL, from the “Salesview” view, Select the region and sum of units sold, and group by region.

A screenshot of a computer

Description automatically generated with low confidence

1. Using SQL select from the “Salesview” view – the region and sum of total\_profit and group by region and display in a Bar chart.

A screenshot of a graph

Description automatically generated with medium confidence

1. Using SQL select from the “Salesview” view – show the total profit as profit, the total revenue as revenue and the total cost as cost from “Salesview”, group by region.

A screenshot of a computer

Description automatically generated

1. The client is in the process of opening up a new store and they are looking at the best location to do so - They need to see the avg profit in each region as a percentage (pie chart) compared to other regions.

A screenshot of a graph

Description automatically generated with low confidence