**ZEPPELIN-SPARK IN CLASS ASSIGNMENT**

**Name: Subhanjan Subhasis Das Humber Number: N01431473**

* Load data into a Spark dataframe

Graphical user interface, application

Description automatically generated

* Print the dataframe schema

Graphical user interface, text, application, email

Description automatically generated

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

Graphical user interface, text, application, email

Description automatically generated

* Aggregate the dataframe via group by “Region” and count
* Create a separate dataframe with the above group by results

Graphical user interface, text, application, email

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

Graphical user interface, application

Description automatically generated

* Create two views using the “createOrReplaceTempView” command
* View on “Salesview” from the first dataframe
* View on “Regionview” from the second dataframe

Graphical user interface, text, application

Description automatically generated

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

Graphical user interface, application, table

Description automatically generated

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

Table

Description automatically generated with medium confidence

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

Graphical user interface, application

Description automatically generated

* 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

Graphical user interface, text, application, email

Description automatically generated

* 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.

Graphical user interface, application

Description automatically generated