**Big Data I**

**Zeppelin – Spark Assignment 1**

Mona Shah

N01479948

1. Show how you loaded the data into a data frame.

Graphical user interface, text, application, email

Description automatically generated

1. Show the dataframe

Graphical user interface, text, application, email

Description automatically generated

1. Print the dataframe schema

Graphical user interface, text

Description automatically generated

1. Filter the dataframe to show units sold greater than 8000 and unit cost greater than 500

Graphical user interface, text, application, email

Description automatically generated

1. Show the dataframe in group by “Region” and count

Graphical user interface, text, application, email

Description automatically generated

1. Create a separate dataframe with the group by results

Graphical user interface, text, application, email

Description automatically generated

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

A picture containing text, screenshot, monitor

Description automatically generated

1. Create two views using the “createOrReplaceTempView” command
2. View on “Salesview” from the first dataframe

Graphical user interface, text, application, email

Description automatically generated

1. View on “Regionview” from the second dataframe

Graphical user interface, text, application, email

Description automatically generated

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

Graphical user interface, chart

Description automatically generated

1. Using SQL select from the “Salesview” view – the region and sum of units sold and group by region and display in a data grid view

Graphical user interface, text, application, email

Description automatically generated

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

Graphical user interface, application

Description automatically generated

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 – The client wants to see this data in a line graph so as to see the correlation between cost ,revenue ,profit between regions.

A screenshot of a computer

Description automatically generated

1. The customer is in the process of opening up a new store an 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, please use both views created to demonstrate answer also point out the region where it is most profitable.

Graphical user interface, text, application

Description automatically generated