**DSCI 5350 BIG DATA ANALYTICS**

Assignment-1

**NAME: YAMINI MANGARAI**

**EUID:11535594**

**Question-1**

James want to invest in Ford company stock and needs help in analysing data.

A) Data is prvided in “txt” format. Load the data into a Hive table.

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

B) James wants to compare pricing across months by viewing percentage change across months. Write the SQL statement that allows James to do that.

Text, letter

Description automatically generated

Text

Description automatically generated

A picture containing text, newspaper, screenshot

Description automatically generated

Text

Description automatically generated

C) James want to hold the stock for at least four consecutive months. Which period is most suitable for maximizing the profit?

**A black and white photo of a document

Description automatically generated with medium confidence**

**Table

Description automatically generated**

**Text

Description automatically generated**

**Question-2**

Michael has an online business that sells electronic items.

A) Process data using your VM and store in Hive table(s).

Graphical user interface, text, application

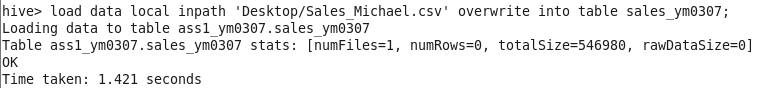
Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated



Text

Description automatically generated with medium confidence

Table

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

B) Michael wants to view a list of top three items sold in every quarter. Write a SQL query that allows Michael to see the results.

Text

Description automatically generated

C) Help Michael segregate products into three categories: *expensive*, *mid-range* and *cheap*.

Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

A picture containing text, newspaper, receipt

Description automatically generated

A picture containing text, newspaper, receipt

Description automatically generated

A close-up of a document

Description automatically generated with medium confidence

**DSCI 5350 BIG DATA ANALYTICS**

Assignment-2

**NAME: YAMINI MANGARAI**

**EUID:11535594**

**Question 1:** Print a list of prime numbers between 100 to 200.

Text

Description automatically generated

**Explanation:** On the above screenshot we are showing the prime numbers from 100 to 200.

A prime number is a whole number greater than 1 whose only factors are 1 and itself.

It has exactly two factors and is a positive integer.

**Question 2:** You have transformed the Ford company share data in assignment 1. Use the table(s) saved in Hive and import the data into Spark. In Spark, using RDD transformations, create new column ‘Quarter’ in the file and list the top 3 prices in each quarter.

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

**Explanation:** On the above screenshot I’m defining a variable for ford.txt table as ford and the view of the table.

Text, email

Description automatically generated

**Explanation:** I have defined the variable as ford2 and creating a new column as Quarter in the file ford2 and below the view of the output.

**Table

Description automatically generated**

**Explanation:** Here defining another variable as ford3 creating the quarter rank column and partitioning the quarter column to get the output of top 3 rank of the stock prices.

**Bonus 2:** Compute Average Price of every quarter and rate of change to next quarter.

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, email

Description automatically generated**

**Explanation:** Here defining the variable as ford4 and getting the average price of the every quarter of ford3 as average stock price and defining and expression as ford5 and then the ford6 if showing the column created as rate of change of the each average quarter.

**Question 3:** Harsha, a Cricket commentator, wants to know which players have played over 200 matches and have a highest score of 150 or above. Write a Spark SQL query using Cloudera or use data bricks?

Graphical user interface, text, email

Description automatically generated

**Explanation:** Here defining the table as cricket and the below screenshots shows the players played over 200 matches and have a highest score 150 and above are shown in the output.

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, application, table

Description automatically generated**Question 4:** Create a column named “Category” based on runs scored. Divide the cricketers into 5 (5000-5999 , 6000-6999, 7000-7999, 8000-8999, 9000 & above) categories. Display Cricketer’s name, strike rate, and Average. With minimum number of matches played and maximum number of matches played in each category?

Graphical user interface, text, application

Description automatically generated

**Explanation:** Here on the below screenshots a column name category is created and divided the player into these categories as (1=5000-5999 ,2= 6000-6999, 3=7000-7999, 4=8000-8999, 5=9000 & above) 1,2,3,4,5 and showing the min and max matches played in each category.

Graphical user interface, text

Description automatically generated

**Question 5:** Create any two visualizations on given data and explain importance? You can use any column or parse the data in column to make your visualization. Example: count of players in each category in the above query.

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated with medium confidence

**Table

Description automatically generated**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated

Chart, bar chart

Description automatically generated

**Explanation: Here showing the 2 visualizations on the data max and mix matches played in each category, where 100 -199 it is 37%, 200-299 is 44%, 300-399 is 14%, 400-499 is 6% as per pie chart.**

**The highest match range is in 200-299.**

**Bonus 3:**

You can come up with your own problem statement and predict using any model available in MLlib. You can use any statistical or machine learning model for your prediction.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Explanation:** Defining the value as cricketa in which the columns are Runs, BF, SR and the output of the table. And in the below screenshot changing the column as lable for runs and for th BF&STR as features and using the liner regression created the column as residuals.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

**Please find the below links of notebook:**

[**https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1271484530417332/1629884796575712/7114283325994783/latest.html**](https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1271484530417332/1629884796575712/7114283325994783/latest.html)

[**https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1271484530417332/1293763238243076/7114283325994783/latest.html**](https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1271484530417332/1293763238243076/7114283325994783/latest.html)