===

DS203 End Semester Examination
8:30 am - 11:30 am
November 21, 20204

===

All answers should be recorded using Google Form: **https://tinyurl.com/2024-11-21-endsem**
*****
In case you have problems with the Google Form, you can record your answers in a simple text file, with **rollnumber_answers.txt** as the filename, and upload it to the 'Endsem' submission point along with your other uploads.

===

This is an open book / notes / internet / tools examination.
*****
However, you should STRICTLY NOT USE any communication software / tools / devices to communicate with others during the examination.
*****
You are advised to use Jupyter Notebook (installed on your computer, on the VM, or online) to code / solve the problems.
*****
You may use one Notebook for Q1-3 and another for Q4, if need be.
*****
You ARE REQUIRED TO NECESSARILY upload the following file(s) to Moodle under the 'Endsem' submission point.
*** Completed answer document: 'Your-Roll-No_answers.txt' (in case you are NOT using Google Form)
*** All the source code files / Notebooks / Spreadsheets created during the exam

===

The submission point will be open up to 11:45 Hrs, Nov 21, 2024

===

****

SrNo: 198.0

RollNo: 23B3307

Name: Nirav Rajendra Bhattad

Allotted Company: **SOUTHBANK**

****

Note:

• The 'scipy' function **'scipy.stats.norm.ppf'** can be used to find out the quantile (x) corresponding to a given cumulative probability (p) under the normal distribution curve.

• For example, ppf(0.25, 3, 1.12) gives you the value of 'x' corresponding to the cumulative probability value 0.25 of the normal distribution curve with mean at 3 and standard deviation 1.12

This function can be used to solve some of the problems in the following question set

=============================

Q1 - 5 Marks

Using the allotted dataset answer the following:

The CFO of the company wants to create a sample of the CLOSE price of the stock by including 25% observations on either side of the median of the CLOSE price. The CFO then wants to establish the 90% CI using this sample. Help the CFO !

Required outputs:

- Mean of the sample [1 mark]
- Lower boundary of the CI [2 marks]
- Upper boundary of the CI [2 marks]

=============================

Q2 - 5 Marks
Using the allotted dataset answer the following:

A researcher in the finance team wants to do the following: using the first 100 days of the dataset as training data create an SLR model to predict the LOW price of the day based on the OPEN price of the same day. Note that the researcher does not want to account for any unknown factors / features while creating this model. Using the created model he predicts the LOW price for the 101st day and also calculates the delta change in the predicted LOW price for every unit change in the OPEN price. What are the values obtained by him?

Outputs required:

- Regression coefficients (beta_0 and beta_1) [2 marks]
- Predicted LOW price [1 mark]
- Delta value [2 marks]

=============================

Q3 - 10 Marks
Using the allotted dataset answer the following:

With the goal of establishing a model to predict 'TOTTRDQTY' based on OPEN, HIGH, LOW and CLOSE values for the same day, a VIF analysis has to be carried out.

Do the required analysis and create the following data / answer the following questions:

- VIF value for each relevant feature: (Write NA if not applicable) [0.5 marks each * 6]

- Analyze the VIF values and state your conclusions and next steps? [2 marks]

- Name the feature that you are likely to drop last based on its initial VIF [1 mark]

- Which two features remain after you progressively eliminate features based on their VIF values [4 marks]

=============================

Using the data file 'nsedata.csv', provided in the VM, solve the following problems:

=============================

Q4 - 10 Marks
Q4a

This problem refers to the nsedata.csv datafile (/home/hduser/spark/nsedata.csv) in your VM.

- First of all process only those records that have EQ in the SERIES column.

- Then find out the sample standard deviation of the OPEN price for every stock and sort the stocks in ascending order of their sample standard deviation values.

- Print the SYMBOL and the sample standard deviation values of the first 2 stocks in the list.

Required outputs:

- Symbol1 and value1 [1 mark each = 2 marks]
- Symbol2 and value2 [1 mark each = 2 marks]
- ---

Q4b

- Calculate the sample standard deviation of the OPEN price of the allotted stock.

- Identify another stock whose sample standard deviation of OPEN price is closest to that of the allotted stock (be sure to consider only those stocks that have EQ in the SERIES column)

- Print the SYMBOL of the closest stock

- Print out the count of days on which the absolute difference of the OPEN prices of these two stocks was less than 100.

Required outputs:

- Symbol of stock with the closest mean [2 marks]
- Required count of days [4 marks]

'=============================