Buckman™
Chemistry, connected.

Questions for Internship - Data Analyst

Solve all four questions:

- Explain all the steps from data pre-processing, model selection, model validation, etc. as applicable.
- Preferentially use "R" or "Python" statistical package for questions 1, 3 and 4.
- Attach and explain all the codes and formulas used for solving the problems along with your answers and reports.
- Test time: 24 hrs

1. **Problem 1: Prediction problem**

   The dataset (Dataset Problem 1) is provided for the prediction of the noise pressure level. The data set has the following attributes

   1) Frequency (Hz)
   2) Angle (Degrees)
   3) Chord Length (m)
   4) Velocity – Free-stream velocity(m/s)
   5) Displacement – Suction side displacement thickness (m)

   You are required to predict the generated noise i.e., Noise Pressure (Decibels)

   Analyse the patterns and insights from the following dataset and build a model which has the best accuracy for prediction. Explain all the steps from data pre-processing, model selection, model validation etc.

2. Problem 2: Solve using SQL

The following dataset (zipfile named Dataset Problem 2 and 4) is that of an international retail shopping group of company. The data set is separated as customer demographics, customer transaction and store details.

1) Customer_Demographics: Customer demographics details for 100000 customers.
2) Customer_Transaction: Customer-store-weekly level transaction details for last 2 years
3) Store Master: Store attribute details

You are required to

1) Write a SQL query to find out those customers and their demographics who have visited a store for more than ten times.
2) Write a SQL query to create a column called "customer rank" which will rank the customers based on their frequent visits.
3) Write a SQL Query which will create a master table (Combined all the three tables) for all the stores and customers taken into account.

Note: Attach all the codes used for solving the problems along with your answers and reports.

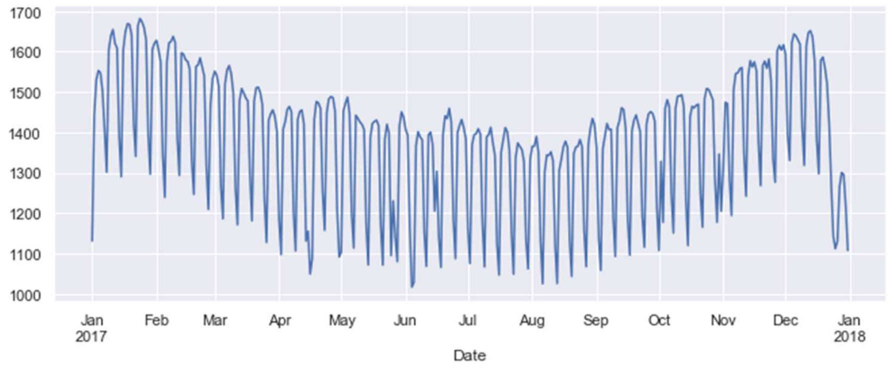
3. **Time series data Visualization: solve using R or Python**

   Attached is a file of power production from 2006-2017 (Dataset Problem 3). The data has five attributes

   1) Date – given in the format of YYYY-MM-DD
   2) Power consumption – The power consumption in a particular date given in Gigawatt hours
   3) E1 – Source of Energy 1
   4) E2- Source of Energy 2
   5) E3 =E1+E2 (Source of combined energies E1 & E2)

You are required to

1) Plot the same graph given below for 2017 instead of the consumption rate take E1, E2 in the Y-axis (Note: Separate graph for E1 & E2 with the months in a year as X-axis)
2) Aggregate the data on a seasonal basis for E1 and E2 and plot the same for the year 2017. Note: Below is the consumption rate for the year 2017 on a monthly basis. You are required to plot two graphs separately for E1 and E2 with Seasons in the X-Axis. Note: You should assume that there are at least three/four seasons in a year.
3) Does the power consumption depend on E1 and/or E2? Justify your answer using plots/tables?



4. **Problem on Statistics**

   From customer transaction data provided in dataset (zipfile named Dataset Problem 2 and 4), find the probability of each customer visiting each store. Say, for a given customer what is the probability that he will visit a store? Include your output as probability.csv and explain the formula.

Note: **Attach all the codes used for solving the problems along with your answers and reports.**