[Open]

Technical Assessment (Data Analytics)

Note:

1. *Please present all answers in Jupyter Notebook / R Studio file format.*
2. *For Qn. 2(c), please present in Power BI file format*
3. *Time limit is 7 days.*

Question 1

Extract information from the sample text below using any analytics technique:

a) List of company names

b) List of dates

c) List of sentences pertaining to money

==========

THIS PURCHASE AND SALE AGREEMENT (this Agreement) is made to be effective as of October 12, 2012 (the Effective Date), by and between WESLEY VILLAGE DEVELOPMENT, LP, a Delaware limited partnership (Seller), and KBS-LEGACY APARTMENT COMMUNITY REIT VENTURE, LLC, a Delaware limited liability company (Buyer).

Deposit shall mean One Million Two Hundred Fifty Thousand and No/100 Dollars ($1,250,000.00), consisting of, collectively, the first deposit of Two Hundred Fifty Thousand and No/100 Dollars ($250,000.00) (the First Deposit), and the second deposit of One Million and No/100 Dollars ($1,000,000.00) (the Second Deposit), to the extent Buyer deposits the same in accordance with the terms of Section 2.1, together with any interest earned thereon.

==========

[Open]

Question 2

A consultant was given the spend data of a procurement department from a Multinational Conglomerate, with their main task being to produce analytical findings based on the datasets. Key challenges include data cleansing and developing accurate machine learning models for data that have yet to be properly classified by spend types.

Please assist the consultant in statistical analysis by doing the following:

a) A descriptive analysis based on the datasets.
b) Predict the segment of the spend types (Column: AD).
c) Generate Power BI dashboard of data visualisations.

(Dataset download link: purchase-order-data.csv)

Question 3

Assuming in Olympics 2020 there are 30 running athletes, your objective is to select the quickest and fastest athletes out of those 30. In each race, maximum of 6 athletes can run at the same time because there are only 6 lanes on the running track.

What is the minimum number of races required to find the 3 fastest and quickest athletes without using a stopwatch? Please provide key bullet points of the thought process involved in arriving at the final answer.

[Open]

Question 4

Complete the questions below (parts a-e) using SQL/MYSQL codes.

Table 1: Geographical_info

ID	Province_ID	Area	Postcode
1	P2	Williamstown (Carleton County)	E7K 2A6
2	P2	Williamstown (Carleton County)	E7K 2A7
3	P2	Williamstown (Carleton County)	E7K 2A8
4	P2	Williamstown (Carleton County)	E7K 3H1
5	P1	New Brigden (Acadia County)	T0J 0B3
6	P1	New Brigden (Acadia County)	T0J 2G0
7	P2	French Village (Kings County)	E5N 8C7
8	P2	Glenwood (Kings County)	E5M 1N8

Table 2: Province_info

Province_ID	Province_name
P1	Alberta
P2	New Brunswick

a. To generate desired output below:

ID	Area	Settlement	County
1	Williamstown (Carleton County)	Williamstown	(Carleton County)
2	Williamstown (Carleton County)	Williamstown	(Carleton County)
3	Williamstown (Carleton County)	Williamstown	(Carleton County)
4	Williamstown (Carleton County)	Williamstown	(Carleton County)
5	New Brigden (Acadia County)	New Brigden	(Acadia County)
6	New Brigden (Acadia County)	New Brigden	(Acadia County)
7	French Village (Kings County)	French Village	Kings County
8	Glenwood (Kings County)	Glenwood	Kings County

b. Count of number of postal codes from Kings County?

c. Count of number of postal codes from each county?

d. Rearrange the dataset to generate the following table structure:

Settlement	Postcode
Williamstown	E7K 2A6; E7K 2A7; E7K 2A8; E7K 3H1
New Brigden	T0J 0B3; T0J 2G0
French Village	E5N 8C7
Glenwood	E5M 1N8

e. Get all settlements in New Brunswick province using SQL JOIN function.

[Open]

Question 5

Feed the following paragraph into your favourite data analytics tool, and answer the following:

a) What is the probability of the word “data” occurring in each line?
b) What is the distribution of distinct word counts across all the lines?
c) What is the probability of the word “analytics” occurring after the word “data”?

==========

As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics. In that sense, it's similar in nature to business analytics, another umbrella term for approaches to analyzing data -- with the difference that the latter is oriented to business uses, while data analytics has a broader focus. The expansive view of the term isn't universal, though: In some cases, people use data analytics specifically to mean advanced analytics, treating BI as a separate category. Data analytics initiatives can help businesses increase revenues, improve operational efficiency, optimize marketing campaigns and customer service efforts, respond more quickly to emerging market trends and gain a competitive edge over rivals -- all with the ultimate goal of boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics uses. In addition, it can come from a mix of internal systems and external data sources. At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and confirmatory data analysis (CDA), which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis. Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

==========