Technical Test

Junior Data Analyst

1. Predictive Feature Analysis for Missing Data

Description :

Identify patterns in missing values across the dataset by writing an SQL query. Specifically:

- Find columns with more than 20% missing values.
- Calculate and return the following metrics for each such column: Table name, Column name, Total rows in the table, Number of missing values, Percentage of missing values
- Order the results by the percentage of missing values in descending order.

Table Definitions

Table : customer

Column Name	Column Type	Key
id	int	PK
name	varchar(255)	
email	varchar(255)	
phone	varchar(128)	
address	varchar(255)	

Table : orders

Column Name	Column Type	Key
id	Int	PK
customer_id	int	FK
order_date	date	
delivery_date	date	
tracking_number	varchar(128)	

Table Sample Data

Table : customer

id	name	email	phone	address
1	Rizaldy Uto	uto@example.com	1234567890	NULL
2	Caemila	NULL	NULL	NULL
3	NULL	caem@example.com	9876543210	Elm Street
4	NULL	NULL	NULL	NULL
5	Johan Chris	jo@example.com	5555555555	Pine Street

Table : orders

id	customer_id	order_date	delivery_date	tracking_number
1	1	01/01/2025	03/01/2025	123-ABC
2	1	02/01/2025	NULL	456-DEF
3	2	NULL	NULL	NULL
4	3	03/01/2025	04/01/2025	NULL
5	4	NULL	NULL	NULL

Expected Result

The query should return this :

table_name	column_name	total_rows	missing_value	missing_percentage
orders	delivery_date	5	3	60.00%
orders	tracking_number	5	3	60.00%
customer	address	5	3	60.00%
customer	name	5	2	40.00%
orders	order_date	5	2	40.00%
customer	phone	5	2	40.00%

2. Dealer Performance and Product Analysis

Scenario :
You are working as a data analyst for a company that tracks dealer performance and product sales. The management has requested a comprehensive report to gain insights into dealer performance, product profitability, and forecast accuracy. This report must include detailed segmentation and ranking based on various metrics.

Task :
Create a **detailed SQL-based** report that answers the following questions about dealer and product performance :

1. Identify the top-performing and least-performing dealers in terms of total sales. Include dealer ID, dealer name, total sales, and their rank.
2. Group dealers into sales deciles based on their total sales and identify the decile for each dealer.
3. Segment products into quartiles based on total sales and list each product's quartile.
4. Rank dealers based on their income percentile and list the percentile rank for each dealer.
5. Categorize dealers into 5 age groups based on their age and display the group for each dealer.
6. Rank products by sales performance within each dealer and identify the top-performing product for each dealer.
7. Segment monthly sales into deciles and identify the decile for each month.
8. Analyze subscription types and sales performance to categorize dealers into 5 groups.
9. Evaluate forecast accuracy for each dealer by calculating the percentage difference between forecasted and actual sales. Group dealers into 5 categories based on their forecast accuracy.
10. Rank products into deciles based on their profit margin.

Dataset : https://drive.google.com/file/d/1DYjIql-2hfDKDfAGipJWUWNyIoU_4UVK/view?usp=drive_link

3. Telco Company Customer Behavior Analysis

Problem Statement :
A telecommunication company wants to analyze its customer's behavior. Given the following dataset, build a comprehensive analysis that provide meaningful insights to the telecommunication company.
https://docs.google.com/spreadsheets/d/1DbcPxy7XRlA7Xj8AbYYzcZlNafcZV5Hi/edit?usp=sharing&ouid=104635612541063243651&rtpof=true&sd=true

Dataset :
Dataset shows the usage of telco service in one period. There are several columns provided :

- **customer_id** : Identifier of the customer
- **brand** : What brand do they use from the company's product
- **segment** : What Segment are they
- **area** : Territory based
- **regional** : Territory based
- **total_revenue** : How much they spent their balace in one month
- **data_usage_group :** Classification of their internet usage
- **tenure** : How long they have stayed with the company
- **app_user** : Area they App User or not
- **f_purchase_product_a** : Whether they buy this product or not
- **f_purchase_product_b** : Whether they buy this product or not
- **f_purchase_product_c** : Whether they buy this product or not
- **risk_segment** : risk segmentation of the customer

Task & Scoring:
You will be focusing on ***Who are subscriber of product a, b and c,*** where price of product_a is 15000, product_b is 23000 and product_c is 10000

Score from this task will be combination of

1. What ***problem statement*** can you generate from this dataset
2. How can you explain the ***EDA*** using ***data analyst tools***.
3. Result and ***conclusion*** from the EDA.
4. What ***solution(s)*** can you offer for the company?

Submission

The mandatory for these 3 tasks is ***PDF***, all answers must be written in one PDF file (prefered PPT file that converted into PDF file).
Format name : **JDS_TA - [Your Name]**
but supporting files are also accepted such as : ***excel, pbix, ipynb, py, etc.***
Additional / Supporting files can be ***compressed into zipped file***.

You can send your answers and files through this link
https://docs.google.com/forms/d/e/1FAIpQLScrPllMSi8VArAqO4ML7hMR5EN52htYnlwsD18EGFZLeer_dQ/viewform?usp=dialog
by **Friday January 17, 2025 before 08:00 AM GMT+7**