Data Analyst Assessment – Paack

ROSE ALCOLEA

General Questions

1. *Could you enumerate the most common problems you usually encounter while performing a data analysis?*
    a. Not enough data
    b. Too much irrelevant data
    c. Lack of understanding of the subject matter
    d. Unclear goals

2. *Which are the most common questions or the best approach you should follow before building a dashboard?*
    a. Deeply understanding the needs of the person/role the dashboard is designed for
        i. What is the most important piece of information this person should have?
        ii. What are the key metrics/KPIs this person needs to look at to perform their job optimally?
        iii. What key comparisons & results could I provide that this person doesn't currently have access to but that would be beneficial?
        iv. What does this person need to know that they don't know they need to know?
        v. What information can I remove from the dashboard to avoid over-clutter?

3. *Which are the best practices or techniques for data cleaning?*
    a. Explore the data:
        i. Learn about the context of the data
        ii. Check data types and adjust
        iii. Check for missing values and adjust
        iv. Check for outliers
            1. Understand their meaning and keep or,
            2. Delete if they seem to be an oddity or mistake
        v. Unpack columns that have several pieces of data (ex: many different categories fitted in one line – having them together will make them harder to classify)
        vi. Potentially deleting unnecessary columns in function of the analytical goal for easier access to the relevant information

4. *Explain what should be done with suspected or missing data?*
    a. Identify why this data is missing, i.e. whether it's a mistake or it has a logical explanation. From that deal with it one of these ways:

i. Replace the missing data with an alternative (ex: 0)
ii. Delete the rows with missing values (only if appropriate!)

SQL Exercise

Note: I had some complications importing the data into my SQL Workbench that I couldn't fix in time, therefore I couldn't check the following queries. In a real scenario I would check them and fix them if necessary (typos, information I omitted by mistake, errors, etc.).

1. *Write a query showing for every driver that delivered any order how many orders they delivered.*

```
SELECT driver_id, SUM(Order Status) AS "total_orders_delivered"

FROM orders_table

WHERE Order Status = 'delivered' AND SUM(Order Status) > 0

GROUP BY driver_id;
```

2. *Write a query showing the drivers that do not deliver any order.*

```
SELECT driver_id

FROM orders_table

WHERE Order Status = 'delivered' AND SUM(Order Status) = 0

GROUP BY driver_id;
```

3. *Write a query showing the different fleets ranked by number of orders delivered.*

```
SELECT drivers_table.vehicle,

        orders_table.SUM(CASE WHEN Order Status = ‘delivered’ THEN 1 ELSE 0 END)

FROM orders_table

JOIN drivers_table ON orders_table.driver_id = drivers_table.id

GROUP BY orders_table.vehicle

ORDER BY orders_table.SUM(CASE WHEN Order Status = ‘delivered’ THEN 1 ELSE 0
END);
```

4. *Write a query showing the percentage of orders delivered and the percentage of orders attempted on time per country.*

```
SELECT SUM(CASE WHEN Order Status = ‘delivered’ THEN 1 ELSE 0 END)/

        SUM(Order status)*100 AS Percentage Orders Delivered,

        SUM(CASE WHEN Attempted time > Delivery Start AND Attempted Time <
Delivery End THEN 1 ELSE 0 END) )/

        SUM(Order status)*100 AS Percentage Orders Attempted on Time

FROM orders_table

GROUP BY country;
```

5. *Write a query showing the percentage of orders delivered by green vehicles (E-Van, G-Van, motorbike and bicycle) per retailer.*

```
SELECT orders_table.Company,

        drivers_table.SUM(CASE WHEN vehicle = ‘E- Van’ OR vehicle = ‘G-Van’ OR
vehicle = ‘Motorbike’ OR vehicle = ‘E- bicycle’ THEN 1 ELSE 0 END)/SUM(vehicle)*100
AS Percentage of Orders Delivered by Green Vehicles,

FROM drivers_table
```

```
JOIN orders_table ON orders_table.driver_id = drivers_table.id

GROUP BY Company;
```

6. *Write a query showing the productivity of each driver per day. Productivity is the number of orders that have been attempted per hour (orders attempted/hour).*

```
SELECT driver_id,

    SUM(Order Status)/ DATEDIFF(hour, Delivery Start, Delivery End)

FROM orders_table

GROUP BY driver_id;
```

Python Questionnaire

GitHub link:
https://github.com/rosealcolea/Technical_Assessment_Paack/blob/master/Python%20Questionnaire.ipynb

Business Case

In this exercise I have analyzed the data in a Jupyter Notebook linked below and created a presentation summarizing my main observations, also linked below.

GitHub link:
https://github.com/rosealcolea/Technical_Assessment_Paack/blob/master/Business%20Case.ipynb

Google Slides link:
https://docs.google.com/presentation/d/1sB_s4AxmqUAevYbNjXOHfJQO4GtErCFL5Eb66aP6gtk/edit?usp=sharing