Insights Analyst – technical exercise

This technical exercise is part of the interview process for the Insights Analyst role at Lexer. In this exercise, you'll look at a synthetic dataset with some sample customer order data.

There are **two** key deliverables to send through within **7 days** of getting the exercise:

1. **The Python or R script you used to analyze the data.** Here, the goal is to show how you work with data, what checks you do to ensure the data is good, how you organize code, and how you do exploratory data analysis.
2. **A slide deck (5-10 slides) which shows key plots or tables you generated as a result of the analysis.** Here, the goal is to show how you would present results to a client (e.g. the marketing team or executive team at a company), so please design this deck as if you were going to present it to a marketing team or executive team.

If you have any questions about expected outputs, please get in touch (Thomas Larsen, in the data science team at Lexer, thomas.larsen@lexer.io). Also, we know it can be tricky to organize around other commitments, so please Thomas know if the 7-day deadline is a challenge. Finally, the data is sample data, so let us know if anything looks strange!

For a couple of the questions there is no "right answer" – you'll need to make a decision, and summarize your thinking behind that decision.

Data overview

Firstly, let's talk about the data. There are two files. The Order Details file has order, product, and spend data but no Customer ID. This means you will need to get the Customer ID from the Customers-to-Orders mapping file, by connecting the tables together using order ID as the link.

Customers-to-Orders mapping table [customer_orders.csv]

This file has the following headers:

- customer_id - the ID of the customer making the order

- order_id - the ID for an order made by the customer

Order Details [order_details.csv]

This file has the following headers:

- order_id - this is an ID that uniquely identifies a customer order
- order_date - the date of a customer order
- price_paid - how much the customer spent on a particular product category in an order
- product_category - Shirt, Jacket, or Pants

How to load the data

There are two preliminary steps:

1. You will need to load the datasets (customer_orders.csv and order_details.csv) into Python or R.
2. You will need to connect the two datasets (customer_orders.csv and order_details.csv) together to form one dataset with the following 5 columns: customer_id, order_id, order_date, price_paid, product_category. You can connect these datasets using 'order_id' as the link between them.

Please do all of the steps below in the same script. When you are done, please send: (1) the completed script, (2) a slide deck (5-10 slides) that is "presentation-ready."

Using the connected dataset, you can then proceed to the questions.

Question 1

Note: Please plot these metrics at a **weekly** level, rather than daily. The raw data is at a daily level, so you will need to change it to look at the data by week.

Use time series plots to show:

1. **Total spend per week** across the customer base over time.
2. **Number of unique customers per week** across the customer base over time.
3. **Number of unique customers, by category, per week** across the customer base over time, broken out by category (Shirt / Jacket / Pants) to compare purchasing

patterns

Please include plot #3 in the slide deck.

Question 2

There are at least two significant issues in the data which could affect analysis. Please comment on what these are, and then make any fixes as appropriate before continuing, justifying your assumptions.

Question 3

According to the data, how many Shirt buyers were there in 2019? (i.e. customers who had a Shirt in an order in 2019)

Question 4

Find customers who meet BOTH the following criteria ('2019 First-Time Shirt Buyers'):

- purchased for the first time in 2019, AND
- had a Shirt in their first order (you can use order_date to find the first order).

Then, using this set of customers ('2019 First-Time Shirt Buyers'):

- how many customers have only ever made one order (one-time-only buyers)?
- how many customers went on to make at least one more order (repeat buyers)?
- how many customers went on to buy Pants in 2020?

Please include a plot or table summarizing these results in the slide deck.

Question 5

The client you are working with asks, "We're trying to develop our marketing strategy for January–March next year. We're not sure whether to promote Jackets, Pants, Shirts, or some combination of these. Based on the data, especially any seasonal patterns, what would you recommend?"

Please include at least 2 slides in the slide deck which show your preliminary findings and what you would suggest. Outline your assumptions and any further questions you would have

for the customer.

Note: Usually there is no “right answer” for these questions. Typically Lexer will present the customer with a range of different options to consider. The important thing is to explain the options, the reasoning behind them, and the pros and cons of each option.

Final outputs

Congratulations! When you are done, please send:

- the completed Python or R script
- a slide deck (5-10 slides) with the "presentation-ready" outputs (plots, tables, etc.)

to Thomas Larsen at thomas.larsen@lexer.io.