

Datascientist Junior technical test

Technical Test Data Scientist

Presentation

You have two datasets in the data directory and 10 questions to solve for the test.

This exercise is a chance for the candidate to demonstrate the ability to provide well organised and reusable code.

- Use Python to complete the assignment.
- No action should be needed before running your code. Otherwise, please provide documentation to initialise your project.
- The code should be written
- Your code is well documented and organised.
- GitHub repository is appreciated

Datasets

Transactions

- `client_id`: ID of the client
- `store_id`: ID of the store
- `chain_type_key`: Type of the store
- `transaction_id`: Transaction ID, represents a customer's shopping cart
- `product_id`: Product ID
- `quantity`: Quantity of product purchased by the customer during this transaction
- `spend_amount` = amount
- `transaction_date` = Date of the transaction

Primary key: (`transaction_id`, `product_id`)

Products

- product_id: ID of the product
- brand: The brand of the product
- category: The category of the product

Questions

1. What checks can you think of to audit the quality of this data? Do these checks allow you to identify anomalies?
2. We wish to study the performance of the products of CATEG_1 in 2018. Calculate the turnover per week on the scope concerned.
3. We know from a business and merchandise perspective that CATEG_1 is a seasonal product category. What is a seasonal product category? Can you give us some concrete examples? how should this seasonality manifest itself on the turnover? what other exogenous characteristics can affect the sales performance of a product? can you give us concrete examples?
4. By exploring the data, can you identify the bias that explains why the turnover per week remains stable despite the seasonality of CATEG_1? What solution can you imagine to correct this bias?
5. We now want to focus our analysis on stores 1, 2, 3, 4 and 5 and on the products of BRAND_2 and CATEG_2. How do we perform the necessary filters?
6. The desired indicators are those of what is called "the commercial equation", calculated week by week. the commercial equation allows to break down the turnover into several levers. Its indicators are: the total turnover on the category, the number of customers buying the category, the total quantity of products purchased, as well as the number of baskets containing at least one product of the category. Can you calculate the indicators of the commercial equation for each week ?
7. The first breakdown of the sales equation is: $\text{sales} = \text{number of customers} * \text{average sales per customer}$
The other levers are: average revenue per checkout, average revenue per quantity purchased, average number of checkouts per customer, average number of quantities purchased per checkout
From the results of question 6, calculate these levers. What relationship can you establish between all these indicators? You should arrive at a breakdown such

as:

turnover = number of customers * average turnover per customer

average revenue per customer = ... x ...

...

8. CATEG_2 is not seasonal; store managers are concerned about the performance decreases observed in weeks 36, 37, 38 and 39 (= period called "P2"). By comparing the performance of this period with the four weeks preceding it (= period called "P1"), can you highlight the lever(s) that most explain the decrease in turnover?
9. A CRM action has been decided in order to revitalize the BRAND_2 customer base within CATEG_2. the actions will consist in sending a coupon on BRAND_2 products. a first target must contain the customers lost between P1 and P2 (= customers who have consumed BRAND_2 within CATEG_2 in P1, but not in P2): it is a reactivation target. A second target must contain customers who are exclusive to BRAND_2 within CATEG_2 (= customers who have made at least two separate transactions within CATEG_2, but only bought BRAND_2): this is a loyalty target. Identify the customers of both targets, count the number of customers per target and calculate the average turnover per customer (based on all transactions) on the BRAND_2 within CATEG_2. A customer can belong to both targets, no need to correct.
10. Describe the methodology that would allow you to measure the impact of these CRM campaigns.