Exness. Test Assignment for the Data Analyst

Data

There are the following tables inside a database:

- **USER** - contains information about users (file: *data/users.csv*) with the following columns:
    - **user_id**: INTEGER - ID of user
    - **country_code**: VARCHAR - country of user
    - **registration_time**: DATETIME - date and time when user was signed up
    - **traffic_source**: VARCHAR - user acquisition channel (Organic / Search traffic / Referral link)
- **BALANCE** - contains information about how users deposit and withdraw their account balances (file: *data/balance.csv*). The columns inside are as follows:
    - **user_id**: INTEGER - ID of user
    - **operation_time**: DATETIME - date and time of balance operation
    - **operation_type**: VARCHAR - balance operation type (deposit / withdrawal)
    - **operation_amount_usd**: FLOAT - operation amount in USD
- **ORDER** - users' trading activity (file: *data/orders.csv*) containing the following columns:
    - **user_id**: INTEGER - ID of user
    - **symbol**: VARCHAR - currency pair describes which currencies we are trying to compare by opening the reading position. For example EURUSD symbol means that we are trying to compare Euro and US Dollar exchange rates against each other
    - **open_time**: DATETIME - date and time when the trading deal was opened
    - **close_time**: DATETIME - date and time when the trading deal was closed
    - **profit_usd**: FLOAT - profit / loss after closing the deal

Part 1: SQL

Task #1: Import data into database

Using any tools you want, create a database with the tables described earlier (***user***, ***balance*** and ***order***) and fill the created tables with test data presented in the corresponding CSV files.

To create the database use any tool you are comfortable with. You can use an online constructor, import data using an IDE or a familiar programming language. To check yourself that all the lines have loaded, final amount of rows should be as follows:

- User - 1,000 rows
- Balance - 6,496 rows
- Order - 117,461 rows

You are free to use any DBMS you want: MySQL, PostgreSQL, BigQuery и т.д.

Task #2: SQL queries

Only the source code of SQL queries is expected from you as a result of this task. You do not need to send the results of queries execution, only the queries itself. It is expected that there would be 3 separate SQL queries, one (and only one) for each of subtasks A, B and C.

A. *Statistics by country*. Calculate the following metrics for each country from database:

   - Total number of users from this country
   - Amount of users who made at least one deposit
   - Average amount of deposit for the country
   - Average amount of withdrawal for the country

   Provide the output sorted by number of users.

B. *Active user*. Find a user with the higher amount of profit from trading activity and calculate some metrics for him. The expected output format is as follows:

   - ID of this user
   - His country code
   - His profit
   - Total amount of deals
   - Amount of profitable deals
   - The most popular trading instrument (symbol). The position with the highest amount of opened orders for this user
   - The symbol with the highest level of profit
   - The symbol with the highest level of loss

C. *User's funnel*. Calculate the following metrics for each user

- User ID
- Country
- Registration datetime
- Date and time of the first deposit
- Date and time of the first trade (if any)
- Amount of the first deposit
- Profit / loss of the first trade
- Total deposit for the first 30 days since registration
- Total withdrawal for the first 30 days after registration
- Total profit / loss for the first 30 days after registration
- TOTAL profit / loss for the user’s lifetime

Part 2: Analytics (choose any option out of two)

This part contains two separate tasks. There is no need to do both of them, choose any you want and like most.

There are no restrictions to the choice of tools for completing the tasks. They can be done by analyzing the data simply in Excel, but for someone it will be more convenient to use Python, R or other advanced tools. If you think it is necessary to supplement the answer with graphs, calculations or attach the source code, please do it, this would be a strong advantage.

Task #1: The impact of COVID

There is a hypothesis that the uncertainties that suddenly occur make the market more volatile. One of the main reasons for uncertainty in 2020 is the pandemic, which, as many claim, has greatly affected the financial markets.

Let's imagine that the pandemic began on May 1, 2020. Please accept or reject the following hypotheses using the data from the previous task:

- After the pandemic (since May 1, 2020), the market has become more volatile compared to the period “before”;

- Market volatility makes trading in any market, including Forex, less profitable. Compare all the same periods: before May 2020 and after it

Justify your conclusions.

Comments:

- There are lots of ways to evaluate the level of volatility. To complete the task, it is enough to look at the **profit_usd** column (from the **Order** table), counting the volatility of only this indicator

- You can take all the data into account, or to consider the users of a single country and analyze them only. Your move

Task #2: Unusual users

In the last SQL task (C. User funnel) it was necessary to get some user's metrics inside the product. As usually happens, not all users are the same and there are often those of them whose behavior differs significantly from the general activity.

In this task, it is proposed to find such users whose metrics are abnormally different from the average user. It is important to be able to identify and remove such outliers, because they can significantly distort the metrics. It is possible that the indicators calculated in the SQL task will be useful, or it may be necessary to add some other ones based on the data provided.

Suggest a way to identify such anomaly users based on the data provided.

General notes

1. Please do not share the solution and do not discuss it with anyone. Let's build relationships honestly from the very beginning. Any clarifying questions can be asked to us directly.
2. The presented test data is randomly generated. Don't be surprised if some indicators will look illogical, it's totally OK.
3. To accomplish the assignment you can use any analytical tools that are familiar to you, there are no restrictions. Someone is more comfortable working in Excel, someone prefers Python or R, it does not matter.
4. The solution could be arranged in the form of a jupyter notebook or, if you have not worked with, as a presentation with the attachment of accompanying working files (scripts, sql queries, excel files, etc.)