

JetBrains Data Analyst Test Assignment

Introduction

In this assignment, your task is to analyze user activity data aggregated over a three-month period (March 2025 to May 2025). This dataset is synthetic and centered around user interactions with a system, specifically focusing on how users interact with different large language models, features and license types. The dataset contains anonymized user-level and activity-level information, including requests and spending.

The goal of this assignment is to extract meaningful insights by exploring the dataset, understanding user behavior, identifying patterns, comparing models and features and proposing recommendations based on findings.

You can use any tool that is most convenient for you to solve this assignment. Please make sure you will be able to share the results with us in the form of a report. If needed, we can provide you with a Datalore license for the sake of this assignment.

Link to the dataset: [Google Drive](#)

Dataset Description

- uuid – user id
- day_id – day of the user activity (data is daily aggregated)
- license – user licence type
- model – used LLM type
- feature – used functionality type
- requests_cnt – number of requests done within the day
- spent_amount – amount of units (credits) spent within the day

Deliverables

- Include all code written for the analysis with clearly marked sections.
- Annotate your code with comments explaining your logic and approach where you think it is necessary.
- Create clear charts to illustrate findings.
- Create a report summarizing your key findings, insights, forecasts and recommendations.

Assignment Questions

Section 1: Data Exploration

- Dataset Overview
 - Provide a descriptive summary of the dataset, including the number of unique users, unique license types, models, features and so on.
 - Identify the total number of rows per user and describe the general behavior. On average, how many rows are generated per user per day?
- License Analysis
 - Explore the relationship between license type and spending. Which license type have users with higher expenses?
 - Analyze the average number of requests per license type. Are users with more powerful licenses associated with higher activity?
- Usage Trends Over Time
 - Analyze the number of requests and spending across all users over the 3-month period. Are there visible patterns in activity?
 - Identify which days generated the highest and lowest spending. What might explain these trends?

Section 2: Model and Feature Usage

- Model Usage Behavior
 - Compare the number of requests across the 5 models. Which model is the most frequently used? Which model is the least used?
 - Analyze spending patterns for each model. Are certain models associated with higher spending than others?
- Feature Usage Behavior
 - Conduct a similar analysis of feature usage.
 - Investigate whether certain features are more commonly used with specific models. Do any combinations stand out as being particularly popular or particularly rare?

Section 3: User Behavior Analysis

- User Activity
 - Identify the most and least active users based on the number of requests, spending and active days.
 - Analyze the spending patterns of users over time. For example, do certain users spend much more on specific days or exhibit clear patterns of spending?
 - Assess whether there are any missing or anomalous values in the dataset. If so, describe how you would handle them.

Section 4: Forecasting

- User Spendings
 - Create daily forecasts of the total spendings by model and grand total for all models for the period of next 4 weeks (June 2025).
 - According to the forecast, propose recommendations regarding the number of credits that should be prepared for the next 4 week period to lower the chance that actual value of credits for June 2025 will surpass the predicted one.
 - Evaluate the quality of the forecast. Please, make sure that the results of forecasts are interpretable. In this specific case it is more important than the quality of the forecast itself.