Data description

You are provided with the spend, revenue and install data associated with installs accumulated between 01-12-2021 and 15-12-2021 in different countries for the application named 'Fruit Battle'. The spend and revenue events provided in the datasets are associated with the users installed the app in different countries in the given time period.

`installs.csv` contains columns:
"user_install_id": a unique identifier of the user
"client": denotes the name of the app
"country": country of install
"country_id": id of country
"year": year of install event
"month": month of install event
"day": day of install event

`spend.csv` contains columns:
"client": denotes the name of the app
"country_id": unique id of the country
"year": year of spend event
"month": month of spend event
"day": day of spend event
"spend": dollar value of the spend event

`revenue.csv` contains columns:
"user_install_id": a unique identifier of the user
"client": denotes the name of the app
"country": country of revenue event
"year": year of revenue event
"month": month of revenue event
"day": day of revenue event
"revenue": dollar value of the revenue event

Tasks description

You are given 3 csv files named revenue, spend, installs.
Please create a local database named test and store these 3 csv files in 3 different tables.

Part 1. SQL

1. Create a database named "test" and create 3 tables named revenue, spend and installs storing the associated csv files.
2. Combine datasets using SQL queries to generate a summary table which contains columns for *Ad Spend, Installs, CPI, ARPI_D1, ARPI_D14, ROAS_D14* for each client(app), country and install date between 01-12-2021 and 15-12-2021.
   Info:
   - ARPI_D[N]: Average revenue a user has made within the first N days after the install date
   - ROAS_D[N]: ROAS after N days from the install date
   - CPI: Cost per install

Please provide the result in the form of:

1. Query - sql code
2. Summary table - csv

Part 2. EDA

- Do an analysis with few bullet points for the current state and the developments observed in Fruit Battle on the top three countries, which have the highest average ARPI_D14 between 01-12-2021 and 15-12-2021, using the metrics you have derived in the part 1
- Define recommendations for the next steps in a few bullet points for those countries you have picked. Please use visualisations to support your reasoning.

Please provide the result in the form of a Jupyter Notebook.

Part 3. LTV prediction (additional, would be a plus)

The lifetime value (LTV) of users are accounted for by the revenue they generate over their lifetime after they have installed. The data you are given has the revenue events for each user up until the first 14 days after each user has installed the app. However, users continue to generate revenue until their lifetime is reached. In the dataset, the revenue events from the 14th day until the lifetime after the install for each user are removed on purpose.

Form a regression function using ARPI values to estimate the average lifetime value (LTV) for the US. Please document how you approach the task. Hint: Assume the lifetime of the user is reached at 28 days. Please use the aggregated ARPI values on the country level during the estimation process. You can estimate the LTV for every cohort day (install date) or come up with an overall estimate for the UA activity between 01-12-2021 and 15-12-2021.

Please provide the result in the form of a Jupyter Notebook.

Part 4. AB testing

You work as an analyst at a mobile gaming company. A new feature has been rolled out to 50% of new users in an A/B test. You're tasked with evaluating its effectiveness based on the following KPIs:

- ARPI_D1 (Average Revenue Per Install)
- D1 Retention

The data below is a simplified sample of actual test results collected after 7 days.

Group	**Users**	**Revenue D1 ($)**	**Retained on D1**
Control	10,000	4,500	3,200
Test	10,000	4,800	3,500

You may assume that:

- ARPI is approximately normally distributed
- samples are independent

Please perform the following:

1. Assess whether the test group shows a statistically significant improvement over control for ARPI D1 and D1 Retention
2. Calculate the power of the test for both metrics. Decide whether the test had sufficient sample size.
3. Summarize your findings and recommendation:
    a. Should the feature be rolled out?
    b. Is there enough evidence?
    c. If not significant, what would you do next?

Please provide the result in the form of a Jupyter Notebook.