Task 2: Basic analysis

I created a jupyter notebook (model.ipynb) to show how I did the Exploratory Data Analysis, research, implementation and evaluate.

Task:

The task is to predict a 10-day forward looking raw purchase projection. My data comes from the parquet databass. In this basic analysis, I mainly used columns user_id, amount_cents, transaction_type and datetime.

Preprocessing:

I focused on data of user_id=5732 and transaction_type=Purchase_Activity.
Then I get date from datetime.
I aggregated the data and get sum of them, grouping by date.

Exploratory Data Analysis:

I drew some tables and plots to see the distributions of the data.
I found out the data has many missing values, for example, April only has data for 6 days, May has data only for 10 days. Then I calculated the auto-corr of April, May and April + May, I found data in May has rather high auto-corr.
Then I realized that I can't use the whole dataset to train a Linear Time Series model.
Then I calculated and drew the MA and EMA of data in May to have a sense of the baseline prediction of the amount_cents in May.
To figure out if the data is stationary, I used the ADF test and rejected H0. So, data in May is stationary. Then I was thinking "Can I use AR, MA, ARIMA models?"

Modeling:

Then I plotted ACF and PACF to decide if I can use the Linear Models, to keep it simple, I went straight to ARIMA, but I was not sure what parameters to take. I used auto_arima to help me decide the AR, I, MA parameters, and I chose to use (1,0,0).
I used ARIMA (1,0,0) and fitted it with the part of the data in May and tested it with the other part of data. Then I calculated the Root Mean Squared Error to evaluate the model.
Then I predicted the next 10 days in Jun.

Potential Improvements:

1. Use cross-validation in the test phase.
2. Find a better way to handle the missing values.
3. Find a better way to handle the outlier points.