Hello Guys, as discussed we will take as final assignement the analysis on the NYC taxi dataset.

The purpose for this assignement is mainly to show the aquisition of the following skills :

- You know how to deploy a cluster spark on GCP
- You know how to use the spark dataframe API
- You know how to layout your code in spark jobs
- You know how to extract some business value using this tool

I'm expecting you to team up in teams of 2 I'm expecting you to use the job pattern layout to answer the assignement I'm expecting you to add my account hamza.senhajirhazi@gmail.com with a role editor on your GCP project see how IAM works. I'm expecting a github/gitlab/bitbucket repo with a clear README.md that i can follow easily to reproduce the results of your analysis

I'm expecting you also to report when possible the result of your analysis in your readme, and comment them with explanation and interpretation when possible, (keep in mind there is no perfect answer or interpretation)

Each section should be in a job, for example the section trip analysis, i should find a corresponding job in /src/jobs/trip_analysis

As said previously, the grades will be based on this assignement, your presence in class + participation

For final submission send me an email with object [ASSIGNEMENT EPITA 2023][] referencing the github repo and if any further explanation due

The final date tolorated for submission will be Saturday 05 August at 14h00

Good luck and it was a pleasure to have you as students, wish you the best for your careers

Assignements questions :

1. **Trip Analysis**:
    - Average duration and distance of rides: Compare these metrics by time of day, day of week, and month of year. This can reveal patterns such as longer trips during rush hours, on weekends, or during holiday seasons.
    - Popular locations: Identify the top 10 pickup and dropoff locations. This could be interesting when mapped visually.
2. **Tip Analysis**:
    - Tip percentage by trip: Do some locations tip more than others? Is there a correlation between distance and tip?
    - Tips by time: Does the time of day, week, or even year affect tipping behavior? You could cross-reference this with holidays or events.
    - Does the payment type affect the tipping
3. **Fare Analysis**:

- Can you calculate the average fare by pull & drop location ?
- Can you calculate the average fare by Passenger count ? to see if there is any correlation with passenger count and fare amount
- Can you correlate the fare amount and the distance trip ?

4. **Traffic Analysis**:
    - Trip speed: Create a new feature for the average speed of a trip, and use this to infer traffic conditions by trying to find if for similar trip (when they exist) they more or less have the same avg speed or not, try then to group the avg speed by trip then hour, day, week

5. **Demand Prediction**:

- Feature engineering: Use the date and time of the pickups to create features for the model, such as hour of the day, day of the week, etc.
- Regression model: Use a regression model (such as linear regression) to predict the number of pickups in the next hour based on the features.