Part 1: SQL & Data Modelling Knowledge [50 points]

Given the below subset of Uber's schema, write executable SQL queries to answer the questions below. Please answer in a single query for each question and assume readonly access to the database (i.e. do not use CREATE TABLE).

1. **For each of the cities 'Qarth' and 'Meereen', calculate average difference between Actual and Predicted ETA of all completed trips within the last 30 days.**
   The result should have 2 columns: *city_name, average_diff*
   [20 points]

2. **A signup is defined as an event labeled 'sign_up_success' within the events table. For each city ('Qarth' and 'Meereen') and each day of the week, determine the percentage of signups in the first week of 2016 that resulted in completed a trip within 168 hours of the sign up date.**
   The result should have 3 columns: *city_name, day_of_week, percentage_completed_trips*
   [20 points]

Table Name: events

Column Name	Datatype
device_id	integer
rider_id	integer
city_id	integer
event_name	enum('sign_up_success', 'attempted_sign_up', 'sign_up_failur
_ts	timestamp with timezone (for example: '2001-10-20 12:23:54+08')

Table Name: cities

Column Name	Datatype
id	integer
city_name	string

Table Name: trips

Column Name	Datatype

id	**integer**
rider_id	**integer**
driver_id	**integer**
city_id	**integer**
rider_rating	**integer**
driver_rating	**integer**
request_at	**timestamp with timezone (for example: '2001-10-20 12:23:54+08')**
predicted_eta	**integer**
actual_eta	**integer**
status	**enum('completed', 'cancelled_by_driver', 'cancelled_by_rider')**

Part 2: Uber EATS & Data Science [50 points]

Uber EATS is an online meal ordering and delivery platform. It partners with restaurants in dozens of cities around the world. It is now trying to expand their user base and wants to leverage their existing rider base (assume 1 million riders). Using a large set of anonymized features, Uber is asking you to predict which Riders should be targeted for UberEats. You are challenged to construct new meta-variables and employ machine learning techniques to resolve this problem
Marketing team wants to send out promos to stir demand and increase the number of orders. How would you prioritise whom to target based on their existing data points for RIDES business?

You need to answer following through this exercise

1) **What would be your considerations and how would you measure your success rate?**
2) **Identify at least 10 metrics that will be important to track how EATS is performing.**
3) **Detailed Summary report should include data transformations, techniques implemented, model evaluation metrics, etc End to end approach applied in solving the problem**
4) **Predictions for test data**

5) **List down the features that would have been improved the model prediction. And reasons for these features.**

Data

The dataset for this case study have rider attributes, Temporal features and sessions logs in eats mobile app[structured data]. The goal of the case study is to predict which riders should be targeted and their priority. Feel free to apply your own assumptions wherever you feel it requires.

Files Description

Most of the variable names should be self-explanatory in files.

Train.csv
Training data have 15K observations.
Each row is information about a customer. Flag for each customer indicates that rider have ordered food from UberEats.

Test.csv
For each rider in 5K test data, provide the prediction based on your model.