Part 1 - SQL Syntax

[2 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 read-only access to the database (i.e. do not use CREATE TABLE).

1. For each of the cities 'Qarth' and 'Meereen', calculate 90th percentile difference between Actual and Predicted ETA for all completed trips within the last 30 days.

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.

Assume a PostgreSQL database, server timezone is UTC.

Table Name: **trips**

Column Name:	Datatype:
id	integer
client_id	integer (Foreign keyed to events.rider_id)
driver_id	integer
city_id	integer(Foreign keyed to cities.city_id)
client_rating	integer
driver_rating	integer
request_at	Timestamp with timezone
predicted_eta	Integer
actual_eta	Integer
status	Enum('completed', 'cancelled_by_driver', 'cancelled_by_client')

Table Name: **cities**

Column Name:	Datatype:
city_id	integer
city_name	string

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_failure')
_ts	Timestamp with timezone

Part 2 - Data analysis

[5 points]

Uber's Driver team is interested in predicting which driver signups are most likely to start driving. To help explore this question, we have provided a sample[1] dataset of a cohort of driver signups in January 2015. The data was pulled a few months after they signed up to include the result of whether they actually completed their first trip. It also includes several pieces of background information gather about the driver and their car.

We would like you to use this data set to help understand what factors are best at predicting whether a signup will start to drive, and offer suggestions to operationalize those insights to help Uber.

See below for a detailed description of the dataset. Please include any code you wrote for the analysis and delete the dataset when you have finished with the challenge. Please also call out any data related assumptions or issues that you encounter.

1. Perform any cleaning, exploratory analysis, and/or visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the driver signups took a first trip? *(2 points)*

[1] Please note that this data is fake and does not represent actual driver signup behavior

2. Build a predictive model to help Uber determine whether or not a driver signup will start driving. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance. *(2 points)*

3. Briefly discuss how Uber might leverage the insights gained from the model to generate more first trips (again, a few ideas/sentences will suffice). (1 point)

Data description:

id: driver_id
city_id*: city_id this user signed up in*
signup_os*: signup device of the user ("android", "ios", "website", "other")*
signup_channel*: what channel did the driver sign up from ("offline", "paid", "organic", "referral")*
signup_timestamp*: timestamp of account creation; local time in the form 'YYYY-MM-DD'*
bgc_date*: date of background check consent; in the form 'YYYY-MM-DD'*
vehicle_added_date*: date when driver's vehicle information was uploaded; in the form 'YYYY-MM-DD'*
first_trip_date*: date of the first trip as a driver; in the form 'YYYY-MM-DD'*
vehicle_make: *make of vehicle uploaded (i.e. Honda, Ford, Kia)*
vehicle_model: *model of vehicle uploaded (i.e. Accord, Prius, 350z)*
vehicle_year: *year that the car was made; in the form 'YYYY'*