Part 1 - SQL

[2 points]

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

Assume a PostgreSQL database, server timezone is UTC.

Table Name: **trips**

Column Name:	Datatype:
id	integer
client_id	integer (Foreign keyed to users.usersid)
driver_id	integer (Foreign keyed to users.usersid)
city_id	integer
client_rating	integer
driver_rating	integer
status	Enum('completed', 'cancelled_by_driver', 'cancelled_by_client')
actual_eta	integer
request_at	timestamp with timezone

Table Name: **users**

Column Name:	Datatype:
usersid	integer
email	character varying
signup_city_id	integer
banned	Boolean
role	Enum(‘client’, ‘driver’, ‘partner’)
created_at	timestamp with timezone

1. Between Oct 1, 2013 at 10am PDT and Oct 22, 2013 at 5pm PDT, what percentage of requests made by unbanned clients each day were canceled in each city?

2. For city_ids 1, 6, and 12, list the top three drivers by number of completed trips for each week between June 3, 2013 and June 24, 2013.

Part 2 - Experiment and metrics design

[3 points]

A product manager on the Growth Team has proposed a new feature. Instead of getting a free ride for every successful invite, users will get 1 Surge Protector, which exempts them from Surge pricing on their next surged trip.

1. What would you choose as the key measure of the success of the feature?

2. What other metrics would be worth watching in addition to the key indicator?

3. Describe an experiment design that you could use to confirm the hypothesis that your chosen key measure is different in the treated group.

Part 3 - Data analysis

[5 points]

Uber is interested in predicting rider retention. To help explore this question, we have provided a sample dataset of a cohort of users who signed up for an Uber account in January 2014. The data was pulled several months later; we consider a user retained if they were "active" (i.e. took a trip) in the preceding 30 days.

We would like you to use this data set to help understand what factors are the best predictors for retention, 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 data when you have finished with the challenge.

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 observed users were retained?

2. Build a predictive model to help Uber determine whether or not a user will be retained. 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.

3. Briefly discuss how Uber might leverage the insights gained from the model to improve its rider retention (again, a few sentences will suffice).

Data description (dataset):

***city**: city this user signed up in*
***phone**: primary device for this user*
***signup_date**: date of account registration; in the form 'YYYY-MM-DD'*
***last_trip_date**: the last time this user completed a trip; in the form 'YYYY-MM-DD'*
***avg_dist:** the average distance (in miles) per trip taken in the first 30 days after signup*
***avg_rating_by_driver**: the rider's average rating over all of their trips*
***avg_rating_of_driver**: the rider's average rating of their drivers over all of their trips*
***surge_pct**: the percent of trips taken with surge multiplier > 1*
***avg_surge**: The average surge multiplier over all of this user's trips*
***trips_in_first_30_days**: the number of trips this user took in the first 30 days after signing up*
***uber_black_user:** TRUE if the user took an Uber Black in their first 30 days; FALSE otherwise*
***weekday_pct**: the percent of the user's trips occurring during a weekday*