Part 1 - SQL

[2 points]

Given the following subset of Uber's schema, write executable SQL queries to answer the two questions below.

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

For 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).

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]

Uber's marketing team currently uses display advertising to acquire new drivers for the Uber platform. Display advertising spend is optimized based on the per-driver revenue during the 28 days following their sign-up.

The team wants to launch a new display partner next month and would like your help in designing a test that will allow you to tell them which display partner is performing better. The team wants to run the test in the shortest time span that will provide conclusive results.

Display Partner A ad units are priced based on a CPM model (cost per impression) and Display Partner B ad units are priced based on a CPC model (cost per click).

- What information, if any, would be needed to properly determine the sample size and length of the experiment? Which metrics would you plan to track in order to define the performance of the test?

- What method would you use to analyze the test results? Explain why you chose this method.

- Based on your method, when would you be able to say that Display Partner A is better than Display Partner B? What visual tools or key measures would you provide to illustrate the conclusion?

EXTRA CREDIT

If you have been given unlimited engineering resources, how would you make sure that we have captured 100% of the performance of channels, such as display and offline channels?

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 (link to JSON file). Please include any supporting analysis or code written.

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

EXTRA CREDIT

Build a predictive model to help Uber determine whether or not a user will be active in their 6th month on the system. 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.

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*