# SQL TEST:

Background: schema, datatypes, and schema relationships:

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 |
| request\_device | Enum(‘android’, ‘iphone’, ‘sms’, ‘mobile\_web’) |
| status | Enum(‘completed’, ‘cancelled\_by\_driver’, ‘cancelled\_by\_client’) |
| predicted\_eta | integer |
| actual\_eta | integer |
| request\_at | timestamp with timezone |

Table Name: users

|  |  |
| --- | --- |
| Column Name: | Datatype: |
| usersid | integer |
| email | character varying |
| firstname | character varying |
| lastname | character varying |
| banned | Boolean |
| role | Enum(‘client’, ‘driver’, ‘partner’) |
| creationtime | timestamp with time zone |

Problem Statement: It’s common at X rideshare company to want to know various business metrics about recent trips. Given the above subset of X rideshare company’s schema, write executable SQL queries to answer the following questions:

1. For request times between 12/1/2013 10:00:00 PST & 12/8/2013 17:00:00 PST, how many completed trips (Hint: look at the trips.status column) were requested on iphones in City #5? on android phones?

2. In City #8, how many unique, currently unbanned clients requested a trip in October 2013 that was eventually completed? Of these, how many trips did each client take?

# Business Questions Part A

**Experiment / Metrics Design**

A product manager on FinTech has proposed a new feature. Instead of having drivers manually fill out earnings on tax forms at year end, drivers will get 1 tax form with all the information automatically filled out.

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 B

**Data Analysis**

Revenue Products is interested in predicting expenses for a particular financial line item as a % of Gross Bookings. To help explore this question, we have provided a sample dataset of vendors associated with this line item, the expenses paid out, and the total Gross Bookings attributed to that vendor and product.

Please include any code you wrote for the analysis and delete the dataset when you have finished with the challenge.

Using the attached dataset, please do the following:

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).
2. Build a predictive model to help X rideshare company determine the % of Gross Bookings that will be spent in May 2016. 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.