

### **Problem 1: SQL Querying (25 pts)**

Assume a PostgreSQL database, server timezone is UTC.

Table Name: trips

Column Name	Data Type
uuid	Integer (key)
driver_uuid	Integer (foreign keyed to driver.uuid)
city_uuid	Integer (foreign keyed to city.uuid)
status	Enum('completed', 'cancelled')
request_at	Timestamp with timezone
completed_at	Timestamp with timezone

Table Name: driver

Column Name	Data Type
uuid	Integer (key)
is_test_account	Boolean

Table Name: city

Column Name	Data Type
uuid	Integer (key)
timezone	Character varying
city_name	Character varying
country_name	Character varying

A. Provide a SQL query which returns to % of total drivers which are NOT test accounts.

```
SELECT CAST((select COUNT(uuid) as not_test from driver where is_test_account=FALSE)
AS DECIMAL(5,2))/ CAST(COUNT(uuid) AS DECIMAL(5,2)) * 100 AS drivers_not_test from
driver;
```

B. Provide a SQL query which returns the total number of trips which were completed (HINT: see status) in 2016 UTC time; please exclude all trips associated with driver test accounts.

```
select count(*) as _2016_trips from trips
JOIN driver on trips.driver_uuid = driver.uuid
where completed_at::text LIKE '2016%' and status='completed' and
driver.is_test_account = False ;
```

C. Provide a SQL query which gives the average number of trips per driver by city for trips requested in January 2017 local time (HINT: use timezone) filtered on country\_name = 'United States' and on cities where there were at least 100,000 trips during the time period.

```
with c1 as (  
    select city_uuid, city_name, count(*) as total from trips join city on  
    city.uuid = trips.city_uuid  
    where country_name = 'USA'  
    group by city_name, city_uuid  
    having count(*) > 100000  
)  
select city_name, CAST(count(city_name) AS DECIMAL(5,2))/  
CAST(count(distinct driver_uuid) AS DECIMAL(5,2)) from trips  
join c1 on trips.city_uuid = c1.city_uuid  
where trips.request_at::text LIKE '2017-01%'  
group by city_name
```

## **Problem 2: Data Quality / Data Analysis (75 pts)**

The Safety and Insurance Team is interested in understanding reported accident rate, namely total reported accidents per million miles. We have provided a mock dataset of miles and reported accidents by month, city, product (e.g. UberX, UberEATS), and segment (e.g. segmentation for drivers, riders, or trips).

Please include any code / formulas (R, Python, SQL, Excel, etc.) you wrote for the analysis in your response and delete the dataset when you have finished with the challenge.

Using the attached dataset, please do the following:

- A. Perform any cleaning, exploratory analysis, to identify any unusual or bad data. What adjustments would you make to the dataset before analyzing further?
- B. Propose charts, dashboards, or metrics to monitor to help the team better understand trends in the reported accident rate.

Based on your work in B above, provide a forecast for overall reported accident rate for Jan 2017.

\*Preferable Create a Tableau Workbook

From the above table it can be seen that the Accident rate in Jan 2017 would be about 18.08.