This test is designed to give us a sense of your SQL, data analysis skills, and experience in managing large datasets. Below are some important things to remember:

- You have one opportunity only
- You will have 72 hours
- There are 3 questions on SQL, statistical analysis, and spatial analysis on larger data set (around 160M)
- You need to submit your answer together with the code used.
- There are two datasets in the exercise, both are downloadable through provided links in the exercise.

You should have Python, R, or your preferred data analysis tools installed prior to beginning. For the spatial analysis in problem 3, you will also need software or a package able to handle point-in-polygon, specifically, lat/long inside a geofence (e.g. psql/PostgreSQL). You will also want to have US State shapefiles pre-loaded before you begin.

Problem 1 SQL Querying

Consider the following database, where server timezone is UTC. Please answer the following question using SQL (it does not matter which version of SQL is used)

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.
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.
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.
D. Provide SQL queries that do the following.
    a. For each trip in the trips table, compute the driver’s historical cancellation rate prior to that trip (Hint: order by request_at).
    b. For each trip in the trips table, compute the driver’s cancellation rate in the last 100 trips prior to that trip.
    c. Exclude all trips associated with driver test accounts.
    d. Create a table with the following schema:

Column Name	Data Type
`driver_uuid`	Integer (unique identifier of driver)
`trip_uuid`	Integer (unique identifier of trip)
`pct_cancelled`	Double (cancellation rate in all historical trips prior to current trip from current driver)
`pct_cancelled_last100`	Double (cancellation rate in the last 100 trips prior to current trip from current driver)

Problem 2 Data Quality / Data Analysis

This problem uses a mock dataset (‘mock_accident_data.csv’). Please download the file and confirm you have received 7,911 rows. The dataset includes trip miles and reported accidents by month, city, product (e.g. UberX, UberEATS), and segment (e.g. segmentation for drivers, riders, or trips). The safety and insurance team is interested in understanding reported accident rate, namely total reported accidents per million miles.

Please include any code / formulas (R, Python, SQL, 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.
C. Based on your work in B above, provide a forecast for overall reported accident rate for Jan 2017.
D. Build a model (e.g., generalized linear model) to analyze the reported accident rate per mile as a function of the features in the data (i.e., Month, Segment, City, and Product). Please consider the following in your analysis:
   a) What type of distribution to choose?
   b) How to select important features?
   c) How to check that your model assumptions are supported by data?

Summarize the steps you take to build the model and report the results. Explain the findings from the model.

E. Conduct the following hypothesis tests and report your results:
   a) The accident rate per mile of Segment G is different from that of Segment A.
   b) The accident rate per mile of Segment G is different from that of Segment B.
   c) The accident rate per mile of Segment G is 40% lower than that of Segment A.

Problem 3 Spatial Analysis

All subsequent questions concern this dataset. Please download this file and confirm that you have received 3,060,528 rows of data. Please support your answer with cogent R, Python or other languages.

A. Compute the following metrics for the 'widgets' column in the dataset.
   a) the mean
   b) the median
   c) the absolute difference between the 75th and 25th percentile

B. In reviewing the 'widgets' field of the dataset, describe the rows you would exclude as unusual/errors/outliers and your rationale. How did the exclusion affect the mean and median of the field.

C. The dataset shows the number of widgets which exist at each latitude, longitude on Earth. Answer the following:
   a. Which US State(s) have the fewest total widgets?
   b. Which US State has the 4th highest total widgets? This answer should not depend on whether or not you are excluding outliers, so long as you have done so reasonably.
   c. Please give a table of total widgets by state. Bonus points if you attach a link to a choropleth.