Product DS Exercise 2018 H2

Note to candidate: please consider the readability of both the SQL and any other code you write.

Question 1 - SQL

Part A

You have a table populated with trip information (named uber_trip) table with a rider_id (unique per rider), trip_id (unique per trip), trip_timestamp_utc (the UTC timestamp for when the trip began), and trip_status, which can either be ‘completed’ or ‘not completed’.

```
rider_id , trip_id, begintrip_timestamp_utc, trip_status
```

Write a query to return the trip_id for the 5th completed trip for each rider. If a rider has completed fewer than five trips, then don’t include them in the results.

Part B

You are given three separate tables (named trip_initiated, trip_cancel, and trip_complete) of the form:

```
trip_initiated | trip_id, rider_id, driver_id, timestamp
trip_cancel | trip_id, rider_id, driver_id, timestamp
trip_complete | trip_id, rider_id, driver_id, timestamp
```

Each trip_id in these tables will be unique and only appear once, and a trip will only ever result in a single cancel event or it will be completed. Write a query to create a single table with one row per trip event sequence (trip initiated ➔ cancel/complete):

```
dispatch_events | trip_id, rider_id, driver_id, initiated_ts, cancel_ts,
complete_ts
```

There should only be a single row per trip with a unique trip_id.

Part C

Write at least one test query to validate the data in the resulting table. Indicate what you would expect the query to return if the data were valid.

Question 2 - Experimental Design

Part A

The Driver team is planning to test a new incentive structure in which they will offer drivers an extra $5 per hour if they choose to drive during times of peak demand (4pm - 8pm) to increase the available supply. Drivers will be required to complete at least 5 trips during this window to qualify for the new incentive. As the data scientist on the team:

1) Propose and define the primary success metric of this test. In addition, propose and define 2 or 3 tracking metrics that will be important to monitor in addition to the success metric you have defined.
2) Outline an experimentation plan to evaluate the effect of this incentive, according to the metrics you outlined.
    a) What would be the rollout schedule, and how would you balance this with statistical rigor?
    b) What type of data analysis would you perform? Please explain why you chose that method over possible alternatives.

Part B

A marketing team is planning a campaign to attract new riders to Uber in which they will put billboards up across a city, and they'll be up for several weeks. As the data scientist on the team, what metrics would you be interested in when analyzing the impact of this campaign and how would you go about quantifying any effect on these?

Question 3 - Modeling

Uber's Driver team is interested in predicting which driver signups are most likely to start driving. To help explore this question, we have provided a sample dataset of a cohort of driver signups. The data was pulled a some time after they signed up to include the result of whether they actually completed their first trip. It also includes several pieces of background information gathered about the driver and their car.

We would like you to use this data set to help understand what factors are best at predicting whether a signup will start to drive within 30 days of signing up, and offer suggestions to operationalize those insights to help Uber.

See below for a description of the dataset. Please include any code you wrote for the analysis and delete the dataset when you have finished with the challenge. Please also call out any data related assumptions or issues that you encounter.

Data set description

id: driver_id
city_name*: city that this user signed up in*
signup_os*: signup device of the user*
signup_channel*: what channel did the driver sign up from*
signup_timestamp*: timestamp of account creation*
bgc_date*: timestamp when driver consented to background check*
vehicle_added_date*: timestamp when driver's vehicle information was uploaded*
vehicle_make: *make of vehicle uploaded*
vehicle_model: *model of vehicle uploaded*
vehicle_year: *year that the car was made*
first_trip_date*: timestamp of the first trip as a driver*

Part A

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 driver signups took a first trip within 30 days of signing up?

Part B

Build a predictive model to help Uber determine whether or not a driver signup will start driving within 30 days of signing up. 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.

Part C

Briefly discuss how Uber might leverage the insights gained from the model to generate more first trips (again, a few ideas/sentences will suffice).