Project description

We're working as an analyst for Zuber, a new ride-sharing company that's launching in Chicago. Our task is to find patterns in the available information. We want to understand passenger preferences and the impact of external factors on rides.
Working with a database, We'll analyze data from competitors and test a hypothesis about the impact of weather on ride frequency.

We have been provided 4 datasets. Description of the data are as follows:

`neighborhoods` table: data on city neighborhoods

- `name`: name of the neighborhood
- `neighborhood_id`: neighborhood code

`cabs` table: data on taxis

- `cab_id`: vehicle code
- `vehicle_id`: the vehicle's technical ID
- `company_name`: the company that owns the vehicle

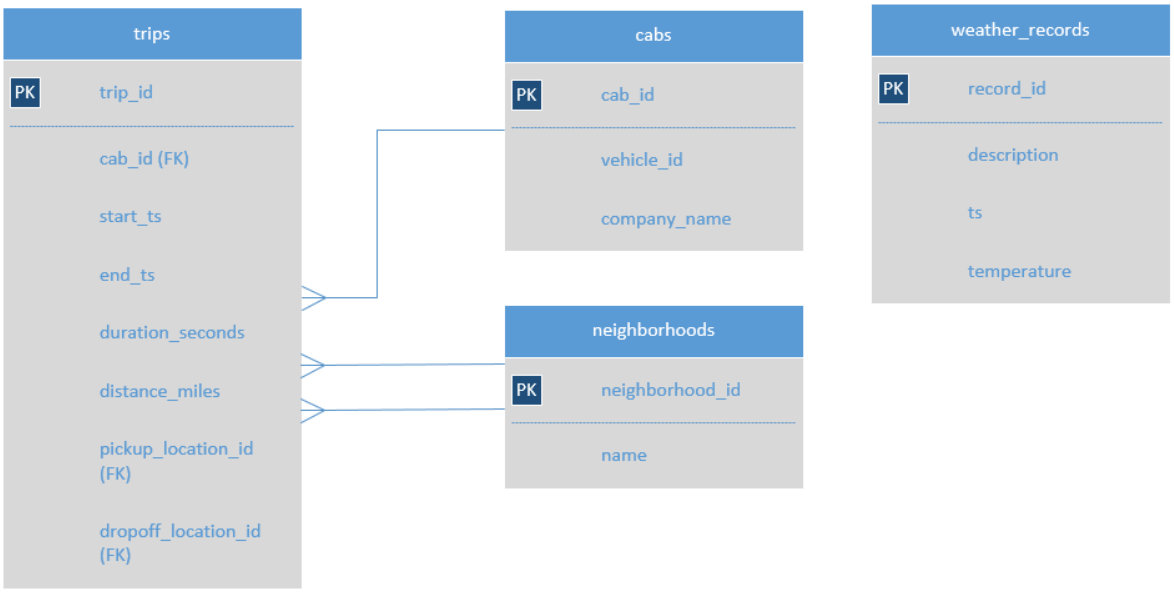
`trips` table: data on rides

- `trip_id`: ride code
- `cab_id`: code of the vehicle operating the ride
- `start_ts`: date and time of the beginning of the ride (time rounded to the hour)
- `end_ts`: date and time of the end of the ride (time rounded to the hour)
- `duration_seconds`: ride duration in seconds
- `distance_miles`: ride distance in miles
- `pickup_location_id`: pickup neighborhood code
- `dropoff_location_id`: dropoff neighborhood code

`weather_records` table: data on weather

- `record_id`: weather record code
- `ts`: record date and time (time rounded to the hour)
- `temperature`: temperature when the record was taken
- `description`: brief description of weather conditions, e.g. "light rain" or "scattered clouds"

Table scheme



In addition to the previous data, We've been given a second file which contains following CSVs:
`project_sql_result_01.csv`. It contains the following data:

- `company_name`: taxi company name
- `trips_amount`: the number of rides for each taxi company on November 15-16, 2017.

`project_sql_result_04.csv`. It contains the following data:

- `dropoff_location_name`: Chicago neighborhoods where rides ended
- `average_trips`: the average number of rides that ended in each neighborhood in November 2017.

  For these two datasets we now need to:

- import the files
- study the data they contain
- make sure the data types are correct
- identify the top 10 neighborhoods in terms of drop-offs
- make graphs: taxi companies and number of rides, top 10 neighborhoods by number of dropoffs
- draw conclusions based on each graph and explain the results

`project_sql_result_07.csv` — the result of the last query. It contains data on rides from the Loop to O'Hare International Airport. Remember, these are the table's field values:

- `start_ts` — pickup date and time

- `weather_conditions` — weather conditions at the moment the ride started
- `duration_seconds` — ride duration in seconds

Test the hypothesis: "The average duration of rides from the Loop to O'Hare International Airport changes on rainy Saturdays."
Set the significance level (alpha) value on your own.
Explain:

- how you formed the null and alternative hypotheses
- what criterion you used to test the hypotheses and why