Project description

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

Description of the data

A database with info on taxi rides in Chicago:
`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

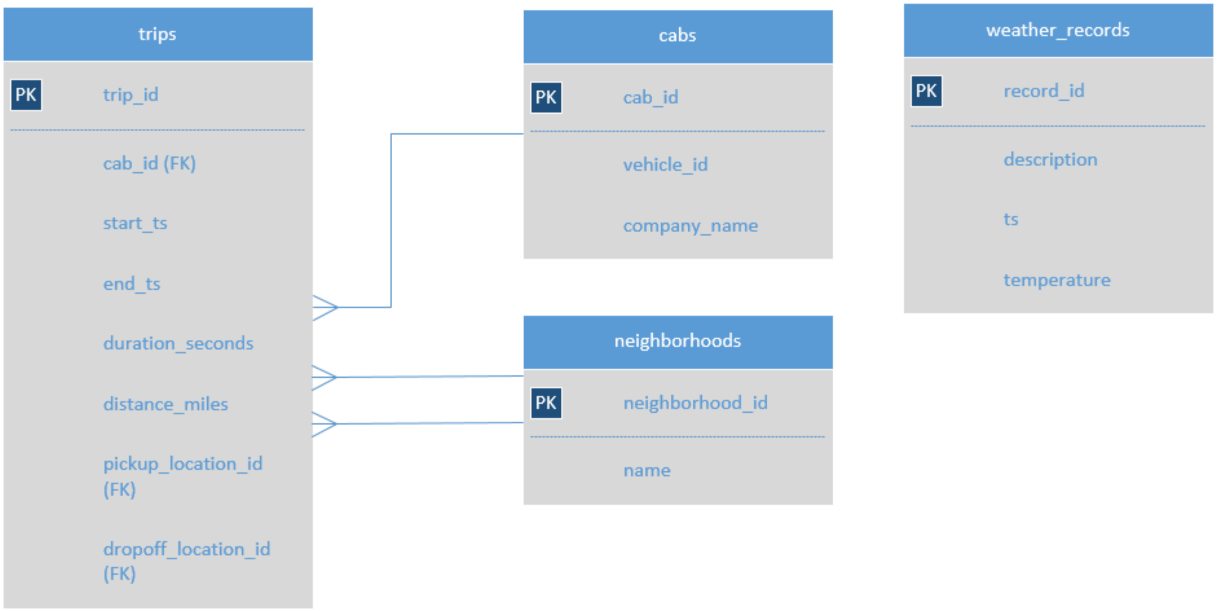
- *trip_id:* ride code
- *cab_id:* code of the vehicle operating the ride
- *start_ts:* date and time of the beginning of the ride (time rounded to the hour)
- *end_ts:* date and time of the end of the ride (time rounded to the hour)
- *duration_seconds:* ride duration in seconds
- *distance_miles:* 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



Note: there isn't a direct connection between the tables *trips* and *weather_records* in the database. But you can still use JOIN and link them using the time the ride started (*trips.start_ts*) and the time the weather record was taken (*weather_records.ts*).

Instructions on completing the project

Step 1. Write a code to parse the data on weather in Chicago in November 2017 from the website:
https://code.s3.yandex.net/data-analyst-eng/chicago_weather_2017.html
Step 2. Exploratory data analysis

1. Find the number of taxi rides for each taxi company for November 15-16, 2017. Name the resulting field *trips_amount* and print it along with the *company_name* field. Sort the results by the *trips_amount* field in descending order.
2. Find the number of rides for every taxi company whose name contains the words "Yellow" or "Blue" for November 1-7, 2017. Name the

resulting variable *trips_amount.* Group the results by the *company_name* field.

3. In November 2017, the most popular taxi companies were Flash Cab and Taxi Affiliation Services. Find the number of rides for these two companies and name the resulting variable *trips_amount.* Join the rides for all other companies in the group "Other." Group the data by taxi company names. Name the field with taxi company names *company*. Sort the result in descending order by *trips_amount*.

Step 3. Test the hypothesis that the duration of rides from the the Loop to O'Hare International Airport changes on rainy Saturdays.

4. Retrieve the identifiers of the O'Hare and Loop neighborhoods from the *neighborhoods* table.
5. For each hour, retrieve the weather condition records from the *weather_records* table. Using the CASE operator, break all hours into two groups: "Bad" if the *description* field contains the words '"rain" or "storm," and "Good" for others. Name the resulting field *weather_conditions*. The final table must include two fields: date and hour (*ts*) and *weather_conditions*.
6. For each hour, retrieve the weather condition records from the *weather_records* table. Using the CASE operator, break all hours into two groups: "Bad" if the *description* field contains the words "rain" or "storm," and "Good" for others. Name the resulting field *weather_conditions*. The final table must include two fields: date and hour (*ts*) and *weather_conditions*.

The table columns should be in the following order:

- *start_ts*
- *weather_conditions*
- *duration_seconds*

Step 4. Exploratory data analysis (Python)
In addition to the data you retrieved in the previous tasks, you've been given a second file. You now have these two 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 you 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

Step 5. Testing hypotheses (Python)
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 Loop neighborhood to O'Hare International Airport changes on rainy Saturdays."
Set the significance level (alpha) value independently.
Explain:

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