Data Analysis Exercise

> 📋 Candidates will be given **48 hours** to complete the exercise. We expect a candidate with the necessary skills will need approximately **4 hours** of work.

Brief

In this assignment, you will analyse the visitation of stores made by users, given GPS signals of users, and outline of the stores (referred to as *polygons*). Besides that, you will explore the demo-behavioural relationship between user affinities and specific stores they have visited.

> ✅ Data for this assignment can be downloaded here

The archive contains three data sources:

- GPS signals of users
- Berlin store polygons
- User affinity datasets

You will need a maximum of 4 GB of hard disk space to accomplish this assignment.

Data Description

GPS signals

GPS data can be found in `sample_signals.zip` and `full_signals.zip` archives. You can test your solution on sampled signals. If you can design your project such that it will handle full data, it is a plus.

Data is partitioned into chunks. Each chunk has the same format:

```
device_id        lat        lon  utc_timestamp
        1  52.676490  13.315430  1609459727000
        1  52.676490  13.315420  1609460036000
        2  52.578370  13.582170  1609459241000
        3  52.583730  13.333090  1609459502000
        1  52.676500  13.315480  1609460387000
```

Column	Data Type	Description
device_id	INTEGER	unique device identifier, i.e. unique user
lat	FLOAT	latitude of a signal
lon	FLOAT	longitude of a signal
utc_timestamp	INTEGER	timestamp in milliseconds, UTC timezone

Store polygons

Data is located in `stores.csv` and has the following format:

Column	Data Type	Description
store_id	STRING	unique store identifier
store_name	STRING	brand name of a store
wkt	STRING	"well-known text", format for storing geospatial objects

> ℹ️ If you have any questions on WKT, check out Wikipedia and Google for reference

Affinity datasets

Demo-behavioral affinity means belonging of a user to a specific demographics, interest, or consumption group.

For example: is a user a male? Is a user a female? Does a user likes McDonald's? etc.

Data can be found in the **affinities** folder. Each affinity dataset is a collection of device ids, i.e. users, that belong to a specific group.

These are the same device ids as in the GPS signals data.

Assignment

1. Analyze the store visitation by date and affinity profile of store visitors.
   a. Resolve the user visits per store, i.e. filter the GPS signals through polygons.
   b. Group the resolved visits by date (yyyy-mm-dd), store_name, and store_id.
   c. For each store_id/store_name/date provide the following metric.
      i. A total number of GPS signals per place_id/date.
      ii. A total number of **unique** visitors (i.e. device ids).
      iii. A total number of **unique** visitors belonging to each affinity group.
   d. An example of the expected format can be found in `example.csv`, you can download it here as a reference.
2. Visualize store visits
   a. Using a tool of your choice, visualize the trend of unique visits for all places.
   b. Is there any anomaly? Add a comment about your observation
   c. Add visuals and text to the README.md of the GitHub repository
3. Visualize GPS data
   a. This clause is **optional**
   b. Using a tool of your choice, visualize the GPS signals of users.
   c. Summarize what you see with a short note.
   d. Add visuals and text to the README.md of the Github repository
4. Visualize stores
   a. This clause is **optional**
   b. Using a tool of your choice, visualize the store polygons.

    c. Summarize what you see with a short note.

    d. Add visuals and text to the README.md of the Github repository

5. Commit your code to GitHub

    a. Create a public repository.

    b. Commit your code, findings, and CSV file with results you acquired.

    c. Include a short guide on how to run your script.

    d. Share your repository with us.