Laboratory Session : April 9, 2023
Exercises due on : April 24, 2023

Exercise 1 - NYC bike-sharing data

The repository `https://drive.google.com/drive/folders/1NESuaJ5yGIrAli1TgrpnK5hnoxGsMi3h?usp=sharing` contains bike-sharing data provided by New York City, Citi Bike[1] sharing system. The data (in `csv` format) is structured as follows

- `Trip duration` (in seconds)
- `Start Time` and `date`
- `Stop Time` and `date`
- `Start Station ID`, `name`, `latitude` and `longitude`
- `End Station ID`, `name`, `latitude` and `longitude`
- `Bike ID`
- `User Type` (*Customer* or *Subscriber*)
- `Birth’s Year`
- `Gender` (0=unknown; 1=male; 2=female)

1) read the data and import in a `data.frame` or `tibble` structure

2) merge the five data frames in an unique structure[2]

3) check for missing data and remove it, if any

4.1) compute the average and the median trip duration in minutes

4.2) evaluate the minimum and maximum trip duration; does that sound like a reasonable value?

4.3) repeat the calculation of the average (and the median) trip duration by excluding trips longer than 3 hours. Next, evaluate the number of skimmed entries

4.4) plot the distribution of trip duration after the skimming of the previous point

5) plot the monthly average trip duration

6.1) plot the number of rides per day

6.2) plot the hourly distribution on weekdays and on weekends

6.3) plot again the average hourly distribution on weekdays but separating *customer* and *subscriber* users

[1] The official page of the service is `https://citibikenyc.com/` and the open data can be retrieved from `https://s3.amazonaws.com/tripdata/index.html`

[2] If the data is too heavy for your computing resources, you can work with a sufficiently large subsample of it.

7.1) using the latitude and longitude information[3], evaluate the average speed (in *km/h*) of a user, discarding the trip lasting longer than 1 hour

7.2) plot the average speed as a function of route length for the following group of distances $\mathrm{d} < 500\,\mathrm{m}$, $500\,\mathrm{m} < \mathrm{d} < 1000\,\mathrm{m}$, $1000\,\mathrm{m} < \mathrm{d} < 2000\,\mathrm{m}$, $2000\,\mathrm{m} < \mathrm{d} < 3000\,\mathrm{m}$, $\mathrm{d} > 3000\,\mathrm{m}$ and discarding trips longer than 1 hour

7.3) repeat the same graph, but show the results obtained separately for weekdays and weekends

8.1) find the most common start station and the least popular end station

8.2) show the distribution of start stations

8.3) find the three most common routes (start and end station) and the three least popular ones

Hint: use the `tidyverse` packages to manipulate the data frame and produce the visualization plots (i.e. `dplyr`, `ggplot2`, ...)

[3] Hint: in the `geosphere R` package, you can find the function `distHaversine` that gives you the shortest distance between two points according to the "haversine" method, which makes the assumption of spherical Earth.