R exerices
Plots and more complex procedures

Gilles Tredan

Abstract

1 Plots and Given Names

Exploring the Pink city

- read the table `prenoms.csv`
- inspect it
- Plot:
  - The number of births by year
  - The number of male/female births by year
  - Is your name in the dataset ?
  - Represent the 10 most given names
  - Select for each year the top 5 given names by sex and represent their evolution along years.
  - Plot the average number of letters by year
  - Plot the average number of vowels/consonants by year
  - How the number of composed names (like Jean-Baptiste or Lou-Ann
  - Define a “hype” criteria and find the hypest names

Exploring the Gray city

- read the table `prenomsParis.csv`
- repeat what you’ve done with Toulouse, rewriting as little as possible

A tale of two cities

- Combine observations made on the two cities.
- Normalise by the number of births.
- What are the most unshared names ?

A tale of many cities

- Read the table `prenomsRennesStrassNantesToul.csv`
- Inspect it. On the opendata website the description is the following:

> This file contains given names to childrens born in Rennes, Strasbourg, Nantes and Toulouse urban areas from 2002 to 2012

Is this really what you observe ?

- The cosine similarity function can nevertheless help us. Given two vectors A and B, it is defined as

$$C = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}.$$

Implement this function in R and compute the pairwise distance between all the cities.

2 Point clouds

- Read the "datacloud.csv" file. It contains the observation of the datapoints generated by 3 independent 2d random laws.
- Use `kmeans` to discover the means of each of these laws
- Plot the detected clusters

3 Clouds

Air quality is monitored in Toulouse by the *Oramip* organisation. The considered data is collected at the following stations: `JACQUI, MAZADE, BRTLOT, PERIPH, TRAFIC_TLSE, EISEN, CHAPIT`. These stations monitor the following concentrations: NO_2, O_3, PM_{10}, PM_{25}. Note that some of these pollutants are not monitored by all stations. You can download the data at the following address: http://homepages.laas.fr/gtredan/tmds/dataset.tgz. The archive contains all the data.

Each dataset is named as follows: `AEROSOL_NUMEROSERIE_STATION.csv`. Inside, a first column defines a useless line number. The second column represents the measure date, expressed as the number of seconds ellapsed since 1970 (aka unix timestamp). Last column contains the measured concentration concentration (in $g.m^3$).

1. Import the data
2. Provide a macroscopic overview of the data (number of values, average, sampling rate).
3. Which station is the biggest data producer ?
4. Present the profile of `MAZADE`, that is, the evolution of concentrations over time.
5. Are PM_{10} and PM_{25} correlated on `MAZADE` ? And on the other stations ?
6. When a station does not produce data, is it only for a single sensor, or for all ?
7. If I leave near `PERIPH` or near `TRAFIC-TLSE`, am I more exposed to NO_2 compared to somewhere else ?

Bonus

- What is the most polluted day ? (utiliser $as.POSIXct(DLdata\$t, origin = "1970-01-01")$ pour convertir en timestamp)
- Assuming a direct correlation between pollution and road use, identify the rush hours.
- How long does a sensor outage lasts (no acquired data)
- Assuming close stations provide close results, estimate the distances between stations.