Bike-sharing stations: profiling and availability prediction

FOSDEM2018 - Bruxelles

Raphaël Delhome(*), Damien Garaud

Introduction

Shared-bike services

Velo'v (Lyon)

- Shared-bike rental service in large cities
- Small-duration rents
- Stations and availability

Major challenges

- Is it possible to classify bike-sharing stations according to bike availability?
- What if there are no bike when we want one?
- What if there are no available bike station when we have to drop our bike off?
- ...
- How to build a complete ETL framework to analyze data and report results?

Outline

(Part 1) Handle open geospatial data

(Part 2) Bike-sharing station unsupervised classification

(Part 3) Bike and station short-term availability prediction

(Part 4) Demo of an API

Data overview

Open geospatial data

Data GrandLyon

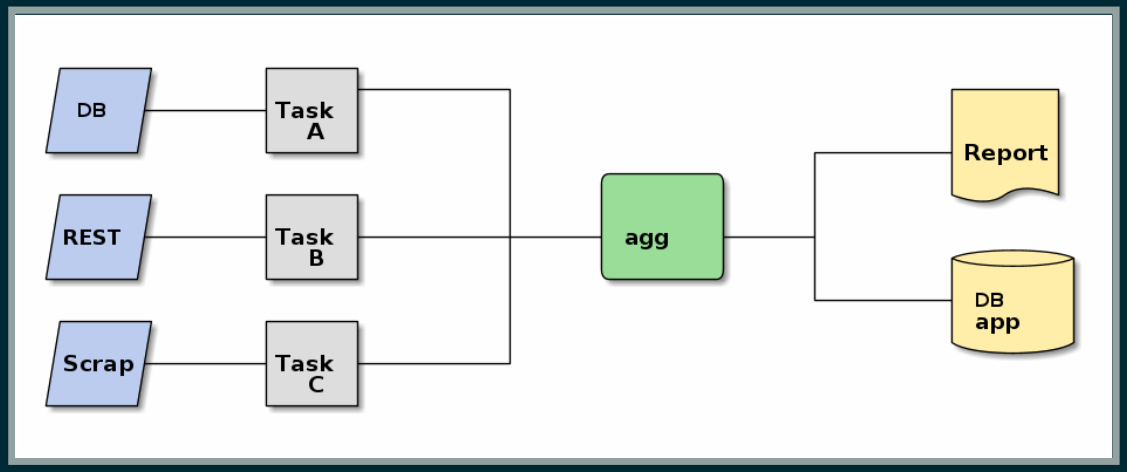
Opendata Bordeaux

Data presentation

id	last_timestamp	bs	abs	ab	bonus	status
10063	2017-07-08 23:49:09	34	10	23	Non	OPEN
10021	2017-07-08 00:30:12	19	0	0	Non	CLOSED
8038	2017-07-08 23:49:26	20	6	14	Non	OPEN
7045	2017-07-08 23:52:43	20	13	7	Non	OPEN

Data pipeline

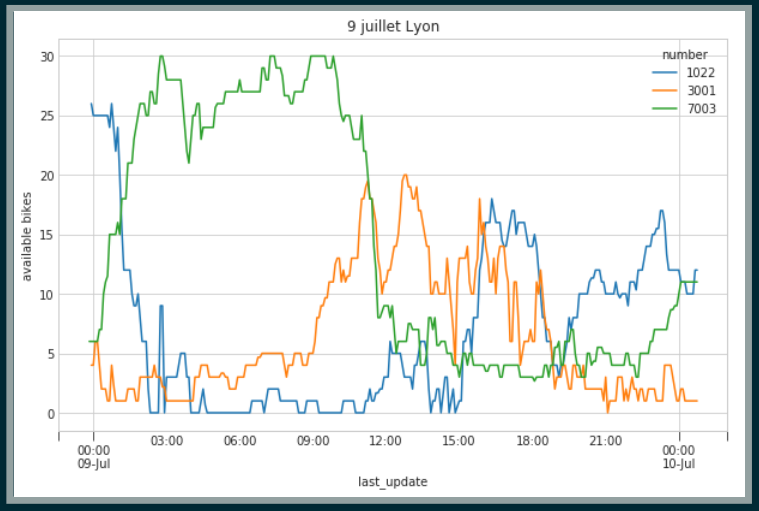
- Build a Python data pipeline thanks to Luigi
- Get, transform and store the data
  - gather data every ten minutes (`json`, `xml`, `shp`)
  - in-base storage (`postgreSQL`, `postgis`)
  - feature engineering and ML treatments



Bike-sharing station classification

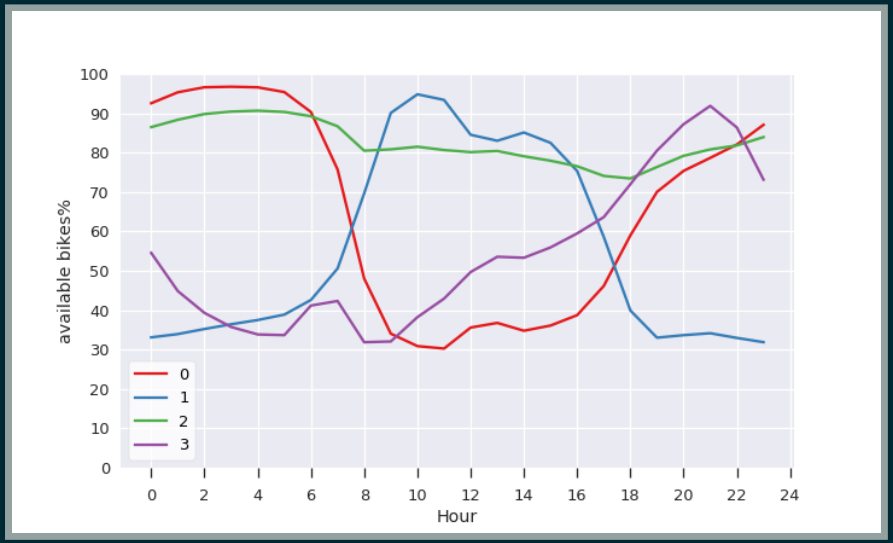
Objective

- Classify bike-sharing station according to the way their are used by customers
- *Main idea* = group stations that looks similar
- ... *What does it mean?* => Focus on the time series



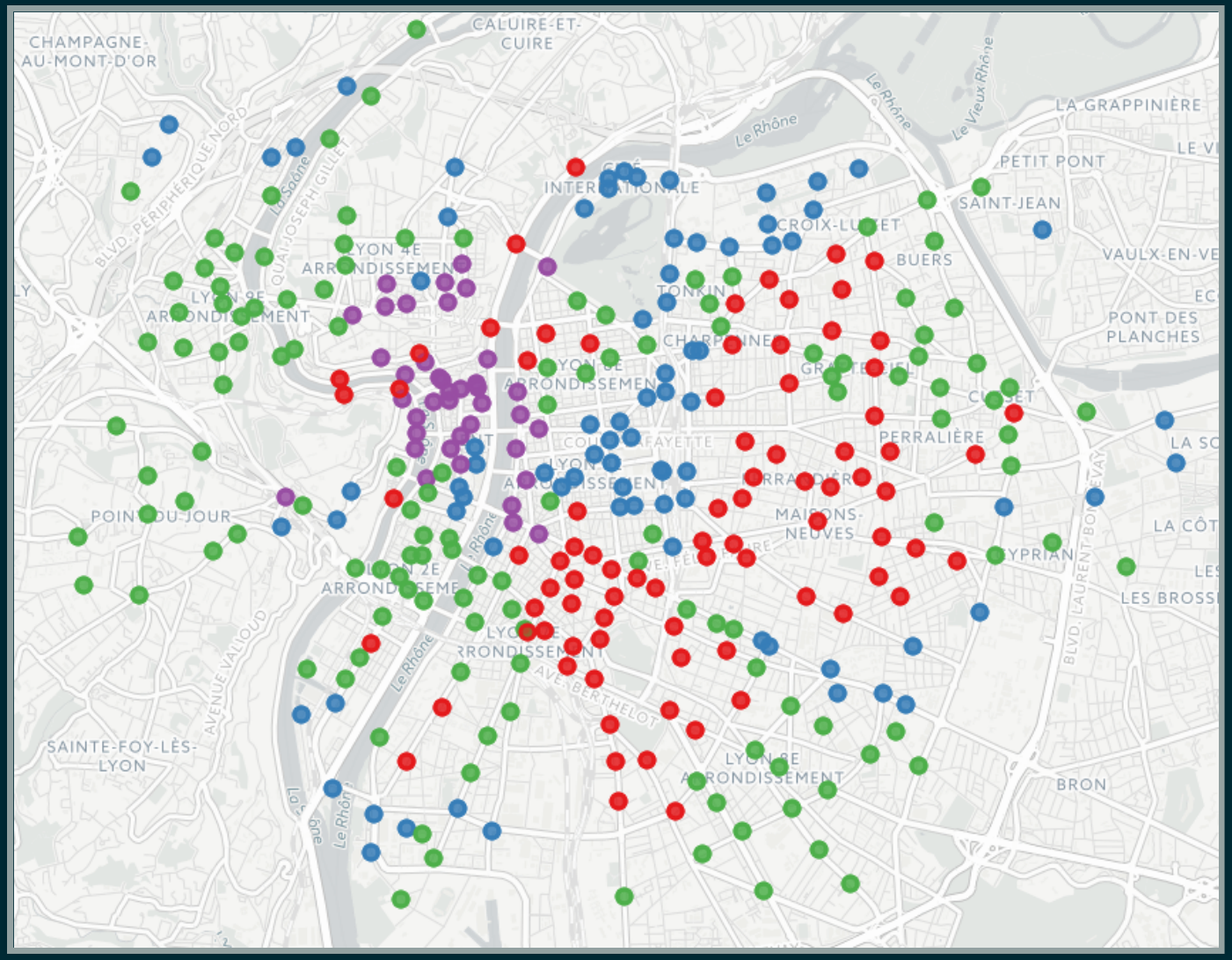
K-means clustering

- Inspired from a similar work of James Lawlor
- One profile = one individual
- Group similar individual together
- Deduce stations profiles



4 clusters have been identified

Clustered station mapping



Shared-bike availability prediction

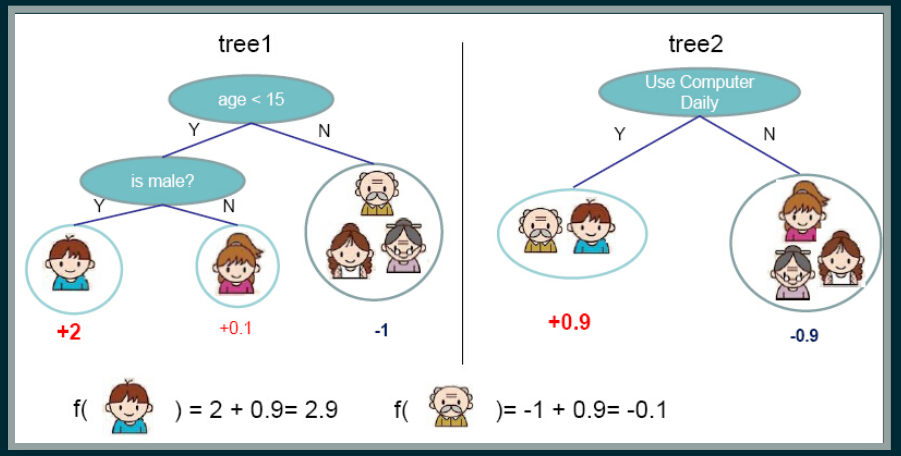
Objective

- Know if some bikes (*resp*. stations) will be available in the next few minutes
- *Main idea* = Use available information to predict availability
- ... *What does it mean?* => Supervised learning to learn an availability probability

XGBoost method

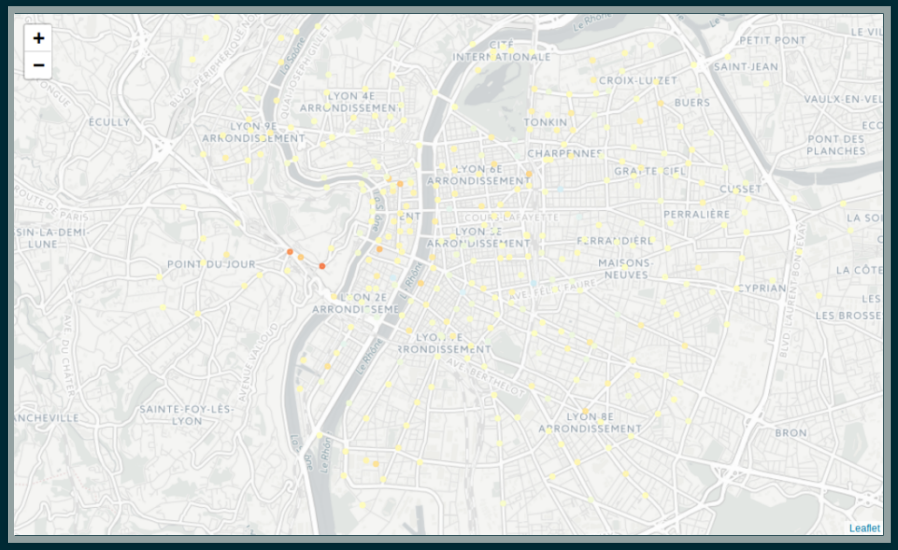
Use a boosting tree method :

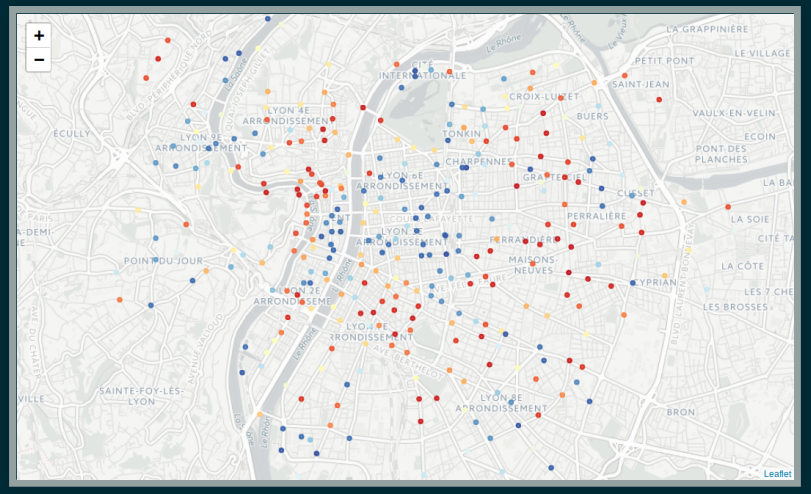
- to predict **Y** (availability probability at H+1)
- starting from **X** (hour, day, available bikes at H, ...)

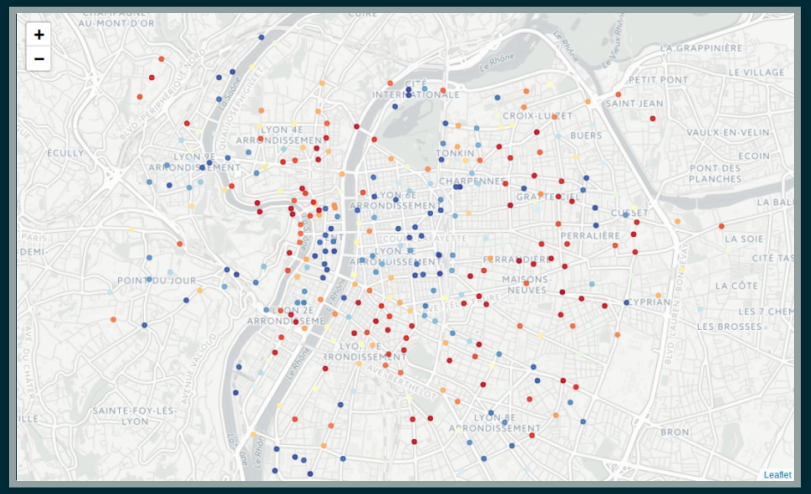


Results

Without tuning features,
RMSE = 0.095







Demo

Bike-sharing open-data API: http://ns3044290.ip-37-59-21.eu:7122/

API documentation: http://ns3044290.ip-37-59-21.eu:7122/doc/

Demo: main page

Bicyle-sharing data analysis

Get and visualize some bicyle-sharing Open Data.

See the project page on Github.

Summary

Available for the following cities

- Bordeaux (France)
- Lyon (France)

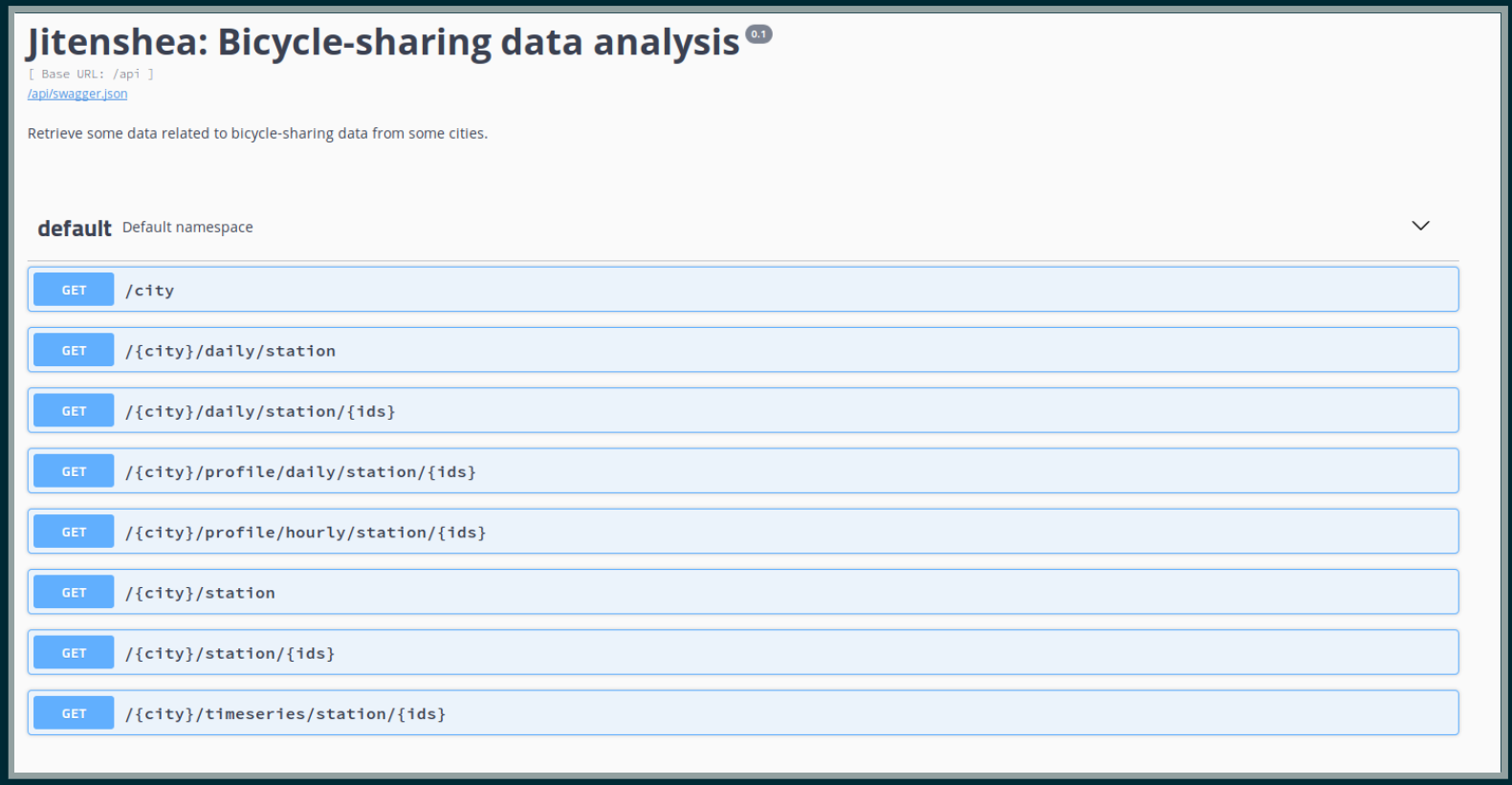
Access

Examples:

- Get some information from a city: `URL/api/bordeaux/station`
- Retrieve details for a given bicycle-sharing station: `URL/api/lyon/station/1009`

You also can read and try the REST API generated by Flask-RESTPlus and Swagger.

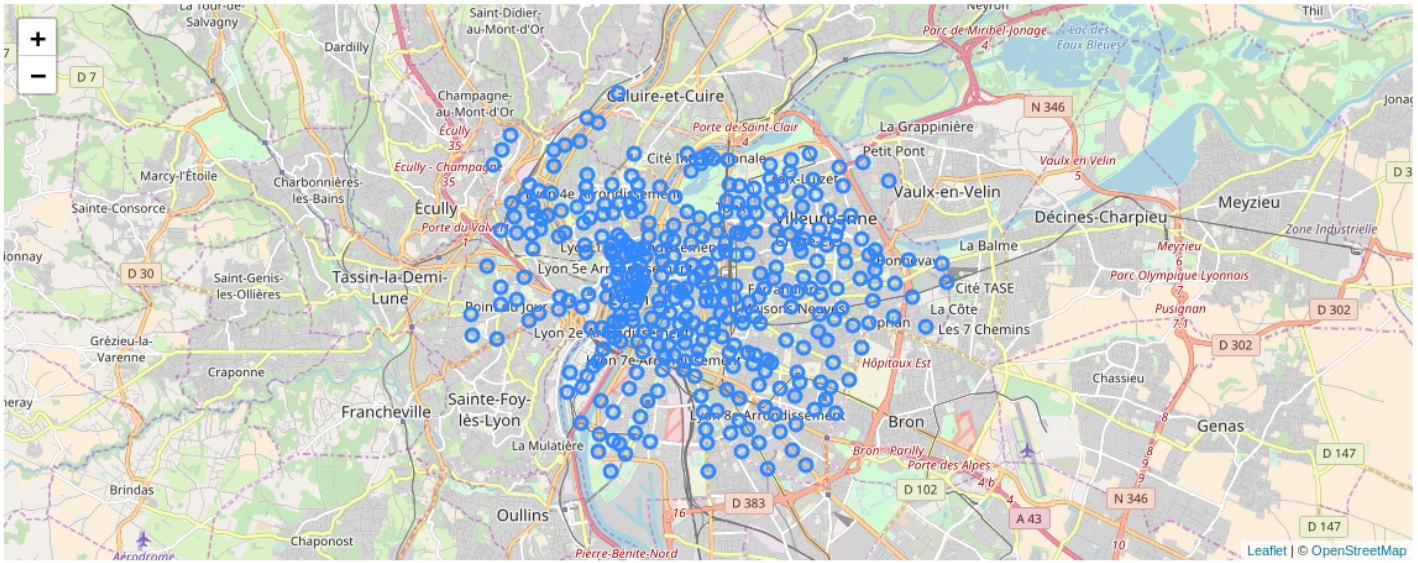
Demo: documentation



Demo: Lyon page (1/2)

Bicyle-sharing data analysis for Lyon

Map



Stations

Search:

Id	Name	City	Bikes	Address
1001	Terreaux / Terme	lyon 1 er	16	Rue Terme
1002	Opéra	lyon 1 er	22	Angle Place de la comédie, rue Joseph Serlin
1003	Place Sathonay	lyon 1 er	16	4 Place Sathonay

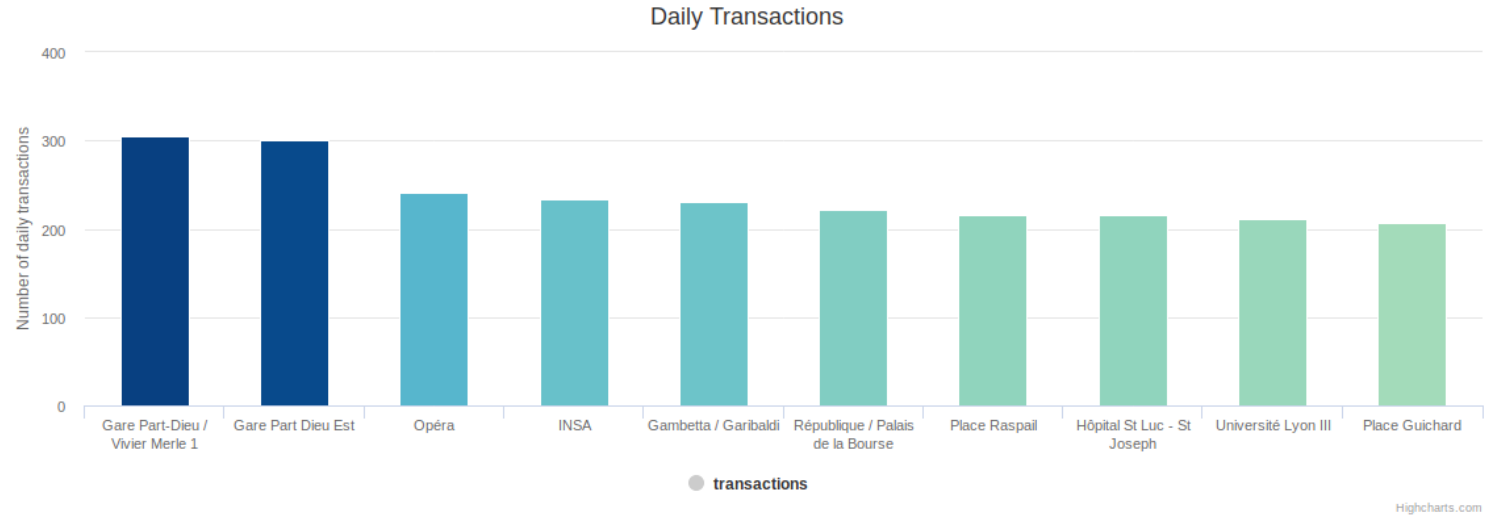
Demo: Lyon page (2/2)

Id	Name	City	Bikes	Address
3087	Part-Dieu / Deruelle / Garibaldi	lyon 3 ème	20	4 Bd Eugène Deruelle
3088	Guichard / Mazenod	lyon 3 ème	16	5, Rue Moncey
3089	Trarieux / Lacassagne	lyon 3 ème	24	221 avenue Lacassagne
3090	Vinatier	lyon 3 ème	20	95 bd Pinel
3091	Hôpital Neurologique	lyon 3 ème	30	59, Boulevard Pinel

Showing 1 to 348 of 348 entries

Charts

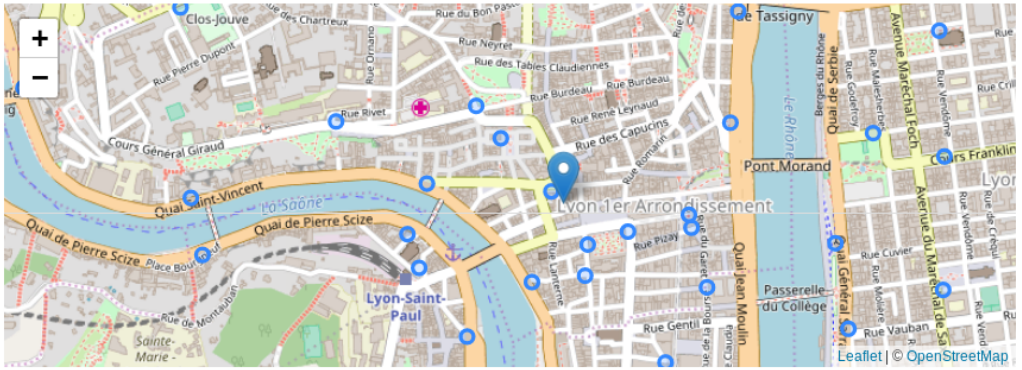
Daily transactions for yesterday. Just the most important ones.



Demo: Terreaux page (1/2)

#1001 Terreaux / Terme in lyon 1 er

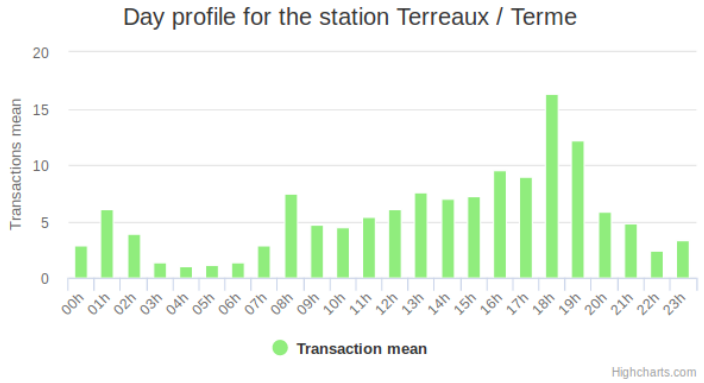
- **ID**: 1001
- **Name**: Terreaux / Terme
- **Bikes**: 16
- **Address**: Rue Terme
- **City**: lyon 1 er



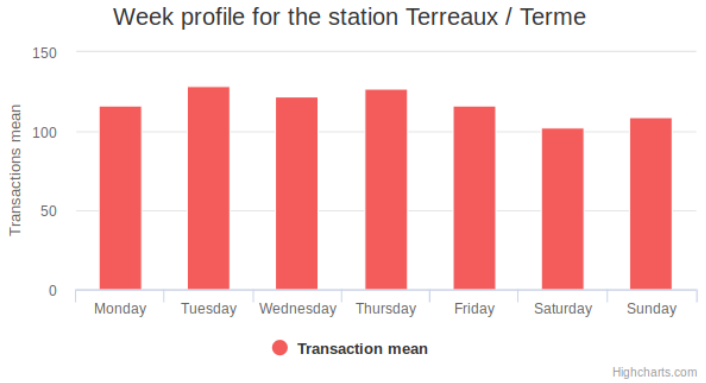
Profiles

Station profiles: transactions mean for each day of the week (resp. hour of the day) for a period of 30 days.

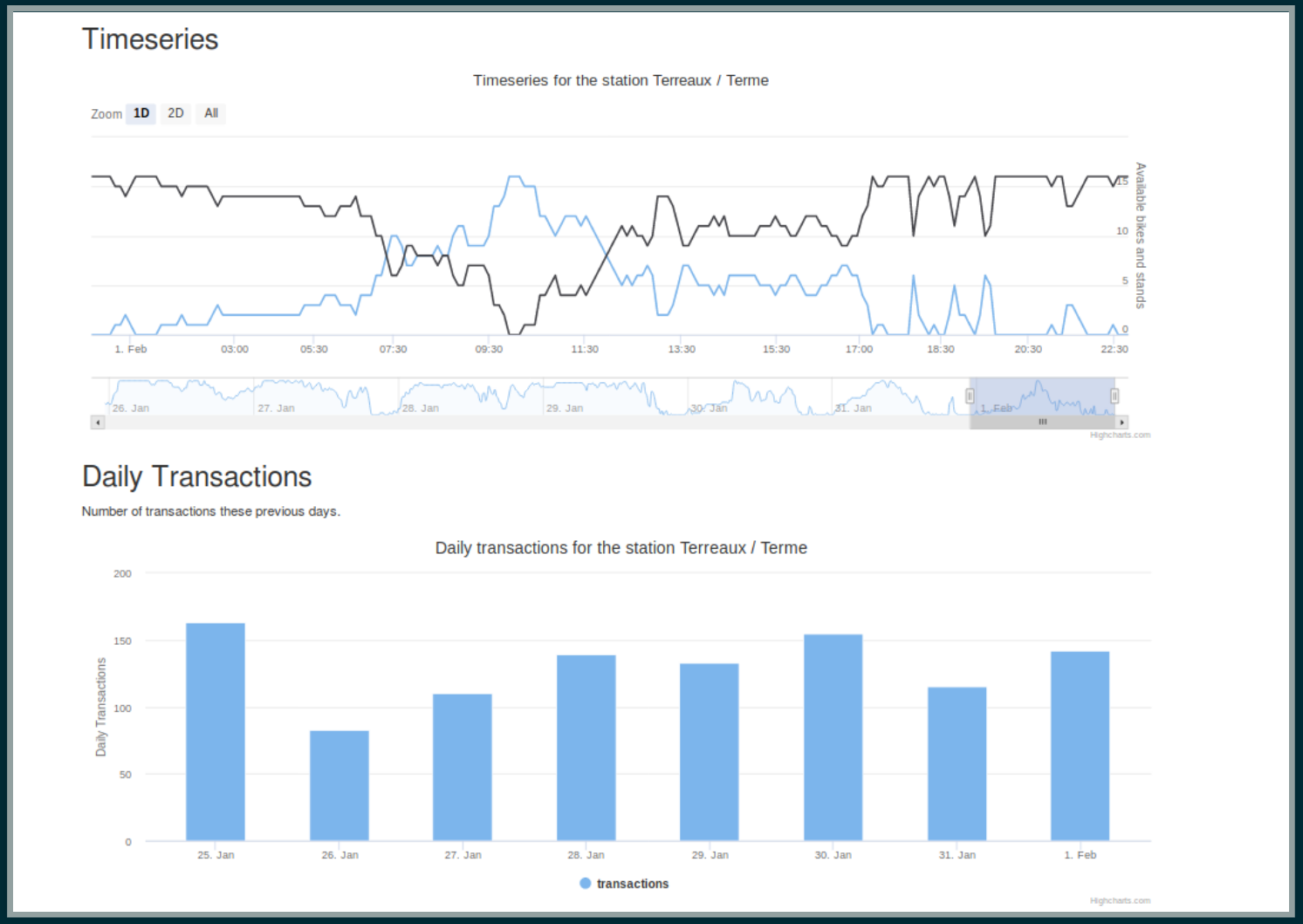
Hourly



Daily



Demo: Terreaux page (2/2)



Conclusion

Conclusion and perspectives

- Addressing some simple research questions with some open geospatial dataset
- From data source to database (ETL-like) with Luigi
- Production of an API to visualize data => towards production?
- Online learning: keep on gathering data, and learn continuously

Thanks for your attention!

Questions?

damien.garaud@oslandia.com
raphael.delhome@oslandia.com

See more on Oslandia's blog and on
github.com/garaud/jitenshea