Dashboard on public transport in France

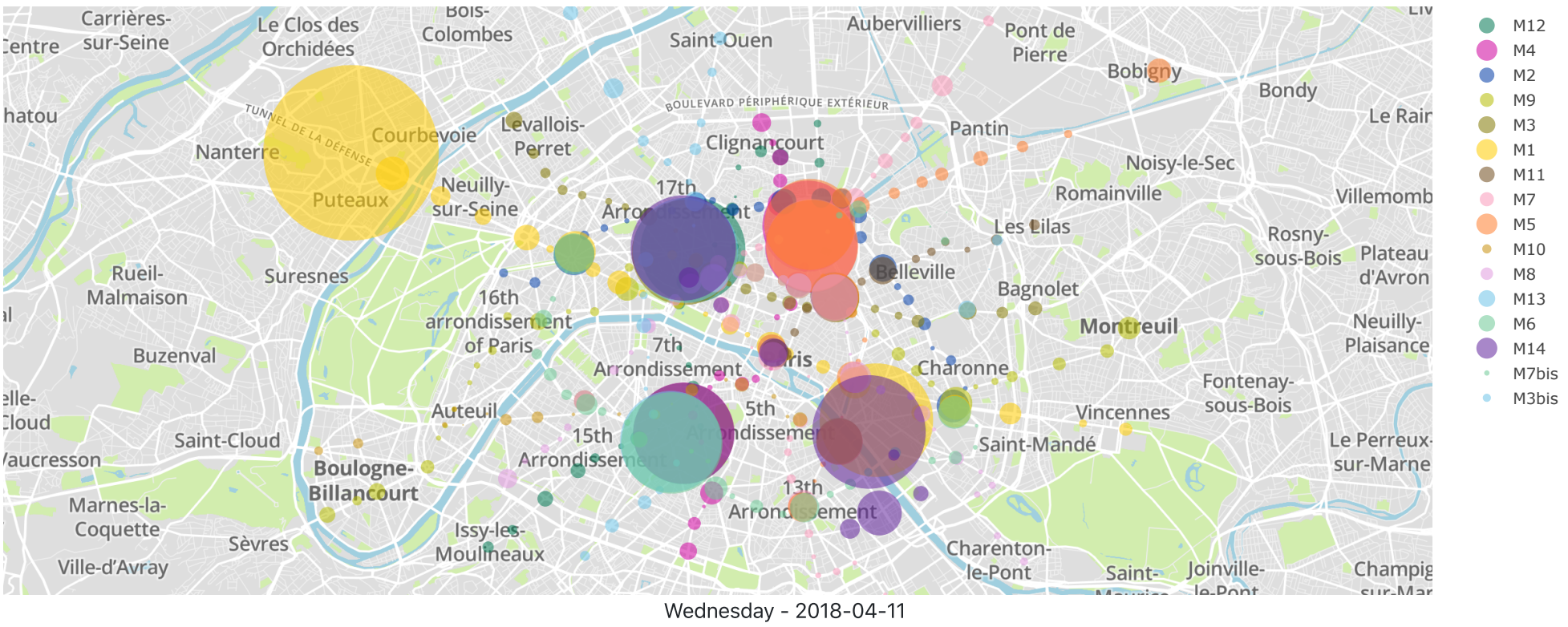
This dashboard has been created by Vincent Barbosa Vaz and Cécile Pov
This unit has been coordonated by Daniel Courivaud
Code unit : OUAP-4112

This is a case study on public transport in France.

After mastering the basics of Python in class, we were asked to produce a dashboard including data analysis and vizualisations. This dashboard uses pandas library.

The first dataset contains the number of validations for each stop, for each category of transport ticket during the first semester of 2018. It is a interesting dataset because ticket transport can reveal the category of people which takes transport at this station. For example, a big amounts validations for the « IMAGINE R » ticket can shows that a lot a students takes this station, and the « Amethyste » ticket shows that the line is attended by elder people. First, before focusing on categories of individuals, we would like to quantify those data in a more general way : how many people had attended this station on the date XX/XX/XXXX ? In order to identify high traffic areas, lets represent this in a map.

Interactive map



Wednesday - 2018-04-11

- Select a transportation system :

METRO

Select...

- Select a date from 01/01/2018 to 30/06/2018 :

Note : hovering on a station on the previous map will automatically update the data on figures below.

MAPPING : TRANSPORT USERS TRAFFIC

The dataset « emplacement-des-gares-idf.csv» contains several useful informations :

- The location (latitude, longitude) of each stop
- The different lines which the stop belongs to

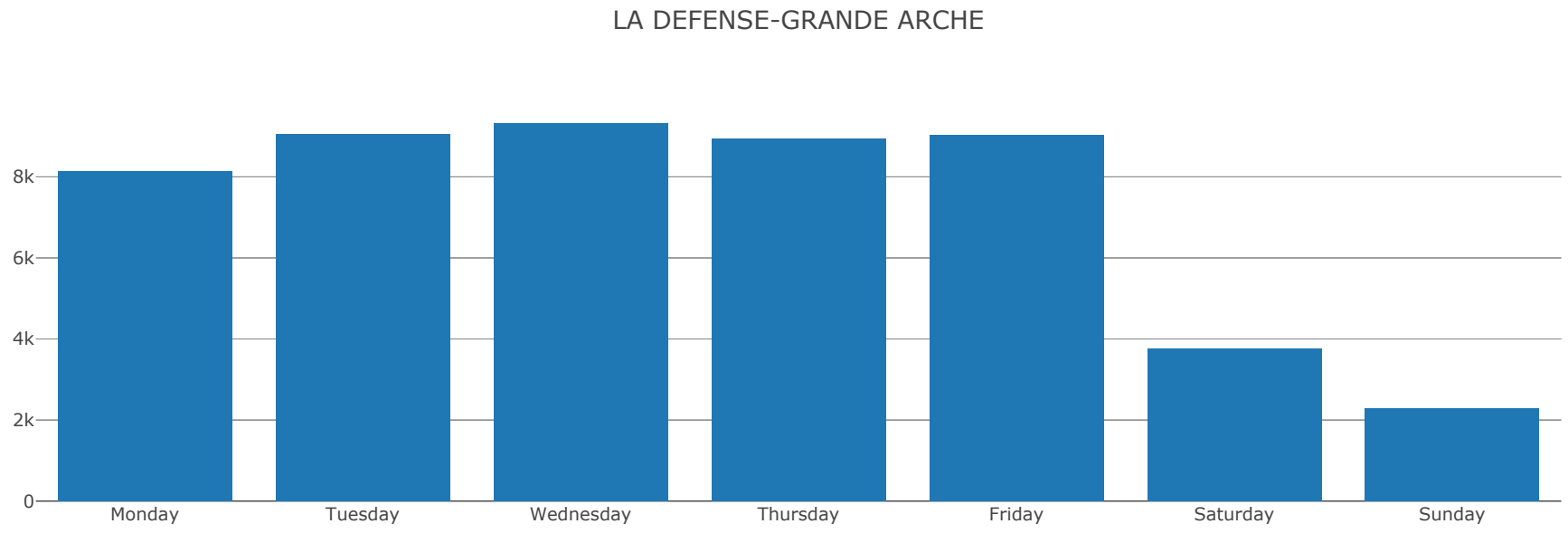
To achieve this, we use mapbox, an open source mapping platform for custom designed maps. We divide this step into 2 sub-steps :

- Plotting a map in which stations that belongs to the same line are marked with the same color ;
- Adding a time and a quantity scale to the map : the size of the marker depends on the traffic at day D. Bigger the marker is, higher was the traffic.

How did we got the total of validations per day per station ? We used the first dataset : knowing the number of validations for each transport ticket category, we group them by station and date and sum them together. The time slider allows us to change the date studied. It covers all the days of the first semester of 2018. When the date on the slider changes, it triggers a callback that update the map. The date and the weekday are printed at the bottom of the map. We can also choose which type of transportation we want to show. To hide a particular station, we can click on the legend on the right. We observe that 5 stations has a particular high traffic, independently from the day :

- La Défense
- Gare du Nord
- Gare de Lyon
- Montparnasse
- Saint-Lazare
- Châtelet-les-Halles

As a consequence, we can say that Paris proper, and particularly the center of the city is a high traffic area. La Défense, which is known as a worker place, is also one. For the weekend days, we can see that the general traffic is far less high (non-working days). To be more precise, we would like to compare the traffic for each weekday for each station.



BARGRAPH : NUMBER OF VALIDATIONS PER WEEKDAY FOR EACH STATION

How did we got traffic for each weekday for each station? To do this, we grouped the data by station and by weekday on the whole semester, and did the average.

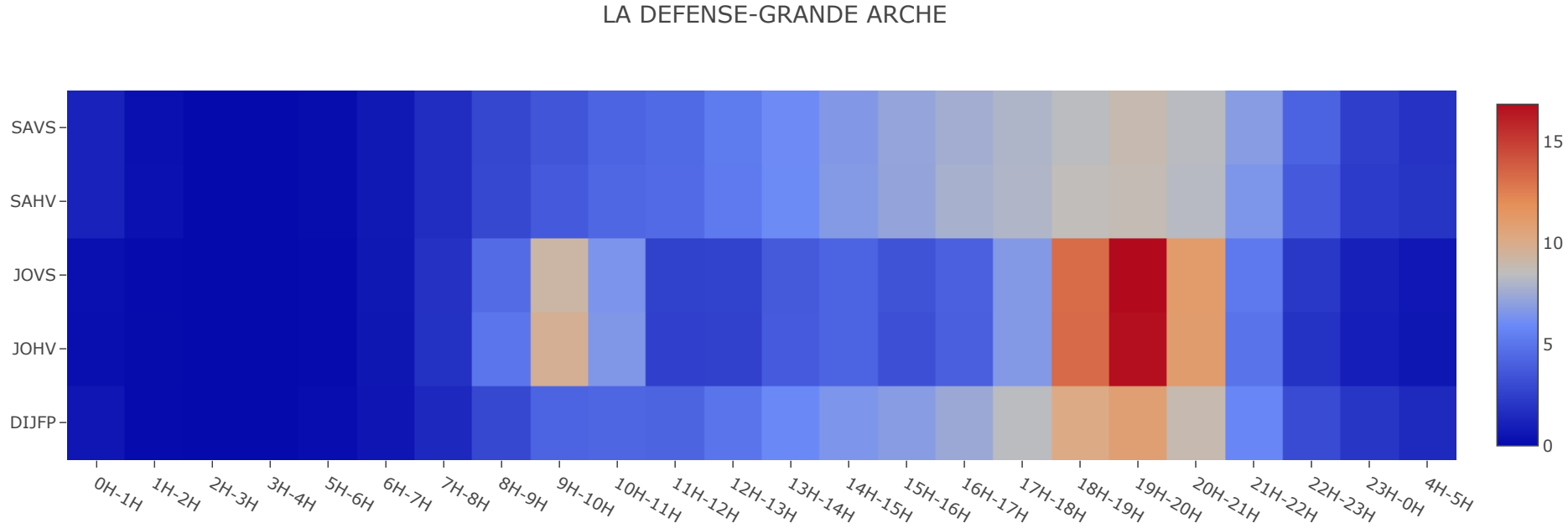
Let’s represent this information with a bargraph. On the x axis, we have the weekdays on on the y axis, we have the mean of the number of validations on the semester for the specific weekday. Note : hovering the mouse on the previous map will update the bargraph with the data of the hovered station. A big difference between the working days and the non-working days traffic show that places around this stations are mostly work or schools places. We can observe this for 2 stations :

- La Défense, which divides its attendance nearly by 4 (Wedneday : 9331 people , Sunday : 2293) -> working place
- Noisy-Champs, which also divide its attendance nearly by 4 (Friday : 3032 people , Sunday : 821) -> student campus

On the contrary, a small difference between the working days and the non-working days traffic show that places around this stations are attended not only for work, by also for entertainment and leisure. For example Châtelet-les-Halles only divide its high traffic by 2 : in fact, this station is located in the center of Paris, it is a crossing point and a touristic place. However, the Chessy-Marne-la-Vallée may be the most relevant and interesting and obvious case : the attendance mostly remain the same the whole week. The small decline during the non-working days is compensated by people who come for the Walt Disney Park. This phenomenon should be highlighted by studying the different category of transport ticket during working days and weekends : it would show that majority of transport tickets during business days are annual/longer fees (Navigo, Imagine R), while most of the tickets during the weekend are more occasionnal tickets. However, the dataset in which we are working on doesn’t take into account magnetic tickets, and that would distort our results. If we take them into account in our bargraph, the attendances for the weekend may be higher than for the rest of the week.

However, this bargraph , on its own, is not a good reference of the attendance if we want to predict, for example, the traffic at La Défense for next Thursday. Indeed, what if next Thursday is a holiday ?

Moreover, with the bargraph, we don’t have any information about the hour (the traffic would be completely different between 1AM and 6PM for example)



HEATMAP : NUMBER OF VALIDATIONS PER WEEKDAY FOR EACH STATION

If Thursday is a holiday, then the traffic at La Defense would not be as heavy as «normal » Thursday. In order to be more close to the reality, we have to categorize each day by a day-type. Depending on if it is a school break or a holiday, i twill belong to a certain category. We started on working on this using different librairies and a calendar.csv, however, it is a very long process. The RATP provides a second dataset : « validations-sur-le-reseau-ferre-profils-horaires-par-jour-type-1er-sem csv », which contains the percentage of validations per time slot per day-type, for each station. In the following, we will work with this dataset. The percentage is relative to the total attendance on whole time slots of the day-type. The dataset distinguish 5 day-types :

- JOHV : Jour Ouvré Hors Vacances Scolaires

- SAHV : Samedi Hors Vacances Scolaires.
- JOVS : Jour Ouvré en période de Vacances Scolaires.
- SAVS : Samedi en période de Vacances Scolaires.
- DIJFP : Dimanche et Jour Férié et les ponts .

Let’s represent this information with a bargraph. On the x axis, we have the time slots, and on the y axis, we have the day-types for the specific hour. We can plot this heatmap for each station.

Note : hovering the mouse on the previous map will update the heatmap with the data of the hovered station.

This heatmap allows us to know if there is more people at a time-slot A or a time-slot B for a particular day-type, but we don’t have any quantity notion. In fact, the pourcentage may has been calculated on a huge number or a small number (1\% of 10 << 1\% of 1000). We must quantify the attendance in order to have a more relevant heatmap. Moreover, in this heatmap, we can’t compare the day-type between them, that is to say read it « vertically » : a percentage is proper to a row (a day-type). In fact, if for the time-slot T, the heatmap is red for a day-type1 and blue for a day-type2, it doesn’t mean and we cannot say that the traffic is higher for the day-type1.