Lost and Found at Train Stations in 2016

Sebastien Delolme-Sabatier* ShangNong Hu† Valentin Reymond‡

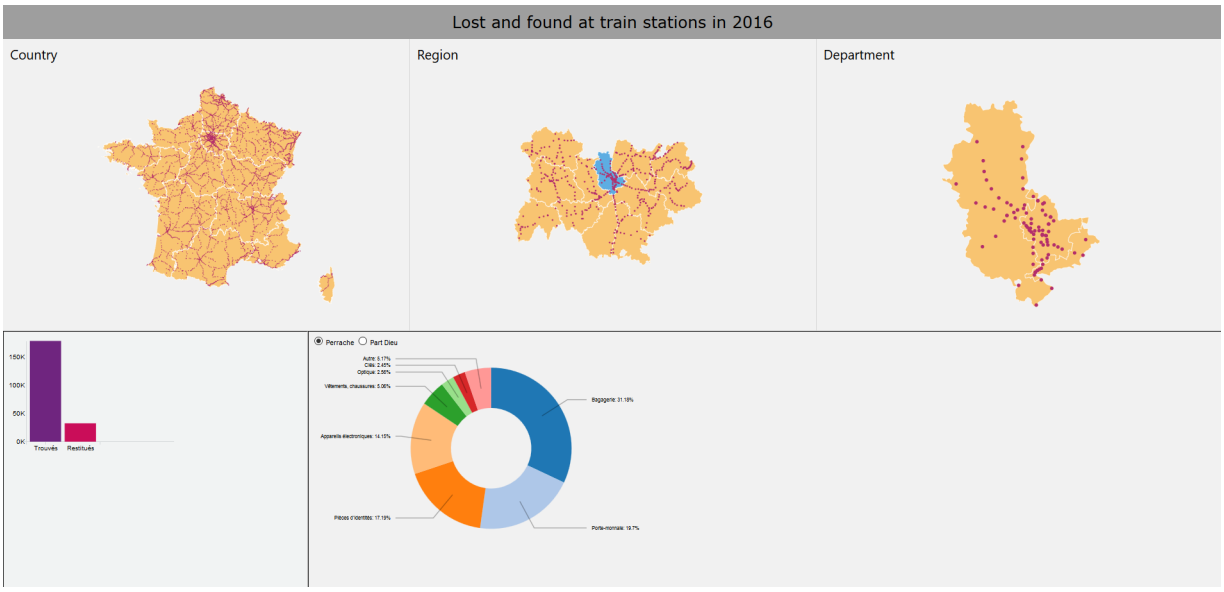

Figure 1: Screenshot of visualization page

Abstract

The Societe nationale des chemins de fer francais (SNCF) provides multiple free-access data on their activity. Among these data, we can for example find data about lost items in the SNCFs train stations. The aim of this project was to answer some questions such as how many items are lost in each train station and what are the most frequently lost. To share this project with the rest of the world, we developed a web page containing the data visualizations we made. We tried to make it simple to understand so it will be accessible and user friendly for most of people accessing it. We used D3.js JavaScript library to create our visualizations methods. Our first goal was to visualize train stations in France from three points of view. So we focused on making a national level visualization, a region one level (region divisions from 2015) and a department one. These three representations are interactive, a click allowing to display the selected region on the superior level (region or department). We also made two others representations. A histogram allowing to compare the number of lost items in train stations and the number of items returned in 2016. Finally we implemented a donut-chart allowing to compare the categories of the found items in Lyon Perrache and Lyon Part-Dieu train stations.

1 Introduction

This project was realised during the UE of Data Visualization proposed by the Universite Claude Bernard Lyon 1. This discipline is in common between the Masters Degree of Bio-Informatics, Data Sciences and Artificial Intelligence. The aim of this discipline was to raise awareness among students to the data visualisation and the discovery of the D3.js library. Aurelien Tabard (Associate Professor, Computer Science Laboratory (LIRIS)), Nicolas Bonneel (CNRS researcher at LIRIS (Lyon, France)) and Romain Vuillemot (Assistant Professor, Ecole Centrale de Lyon, France, LIRIS Lab, Lyon, France) were in charge of the Data Visualisation subject. This project was achieved by Sebastien Delolme-Sabatier, Shangnong Hu, and Valentin Reymond, all students from the second year of the Masters Degree Bio-Informatique Moleculaire: Methodes et Analyses.

In order to realise this project, we chose to create our data visualisation with D3.js library and Blockbuilder to test our code. Furthermore, this project was managed and drop it off in GitHub which is a tool to help the user to follow all the modifications made by the group.. This important platform allows a user to go back to an old version of the project source code. The theme for our project was mobility and transports. Thereby, we chose to work with the free-access data given by the SNCF, more specifically the lost, found and returned items in various train stations in France as well as data peak periods.

Those different dataset have been selected in order to answer specific questions such as : What are the train stations where we report the most lost objects ? What are the train stations where we report the most found objects ? Are lost objects and find objects the same or is there a difference between them ? In which train station we have the highest chances to find the objects that we lost? Is there a relation between the number of lost objects and the influx inside a train station ? Consequently, this project is more intended to analysts and for the data exploration than to SNCF users who do not necessarily ask themselves those kind of questions.

To allow the data visualisation, we set up several methods of data visualisation which seemed to be consistent with the questions asked. Thus, we chose to represent the French map as our starting point, that we split in regions and department, and those three representation displayed the different train stations presents. Finally, we implemented two graphical data representation giving additional

*e-mail: sebastien.delolme-sabatier@etu.univ-lyon1.fr
†e-mail: shangnong.hu@etu.univ-lyon1.fr
‡e-mail: valentin.reymond@etu.univ-lyon1.fr

information more or less precise. Those representations need to be improved and further details will be provided later in this document.

2 Related works

The loss of items is a recurrent topic that has multiple time led to different studies. Le Figaro [1] for example published the results of a study about found items during july and august in 2017. They present us some statistics on the duration of return of items after loss. They also provide some quantitatives information of found items such as the number of cuddly toy during a year. They also try to connect items return with the fee to pay if someone want to get his item back. Novedia Vseo also made an infographics [7] about this topic. To collect the multiple data we worked on, SNCF implemented different services such as a call number [2] to report the loss or the find of an item in a train station. It is also possible to fill out a form on SNCFs website but an account is needed to do it. Finally, some train station provide a lost property service directly accessible to its users. The data used in our study are all free-access data. We focused on a few datasets:

- Lost statements (661843 entries) [5]: this dataset contains information about the type and the nature of the lost items, the date of the report and the train station in which the item has been reported (with its name or its id).
- Found items (396 941 entries) [6]: Such as the preceding dataset, this dataset contains information about the type and the nature of the item that has been found, the date and the place at it was found and an indication if the object has been returned or not.
- Train stations in France (5032 entries) [4]: this dataset contains multiple information about train stations in France. In this dataset, we can find things such as the station id, the city in which it is located, its geographical coordinates, the type of trains they receive, etc...
- Influx by train station (3040 entries) [3]: this dataset contains the information about the train station (id, place) and the number of travellers from 2016 to 2014. Note that travellers and non travellers are distinguished.

These datasets are well-formatted and exist in CSV or JSON format for export. They are free-access but we have to agree terms and conditions and license of the dataset for each one we want to use..

3 Project description

As part of the second year of the Bio-informatic Master degree, we realised this project focussed on the lost and found items within the SNCF train stations during 2016. This project is interesting and needed in the sense that the information given, the percentage of lost and found items by category and by train station, remains unpublished to this day.

Those items are split in different categories as presented in the following table.

Bags	Wallets	Eltronic devices	ID
Clothes	Optical	Books	Keys
Bikes (+scooter with spelling mistake)	Bikes (+scooter without spelling mistake)	Children items	Sport items
Medical items	Musical instruments	Jewels	Umbrellas

Working on this specific set of data enables us to deliver percentage on a local scale instead of a national scale, which is the scale used by the SNCF in their charts. This representation will give us more information concerning which kind of item is the most lost in each train stations existing in France. Furthermore, we thought that it would be interesting to know which proportion of lost items is returned, that is why we added data concerning the restitution of those lost items to their owner.

In order to represent those data, we used the powerful graphic library D3.js, which offers a wide range of functionalities suitable for the data visualisation. In parallel, we also used the online tool Blockbuilder[1], a platform executing the HTML/CSS/JavaScript code as and when it is written. This platform allows us to have a critical look on our work quickly.

The project is built around two main principles presented on the visualisation: the first one is focussed on the train stations positioning on the different scale of maps of France, and the second one is the graphical representation of items lost and found in the train stations in 2016. For the first principle, we chose to give the user the opportunity to access different representations of the map to provide them a visual as well as informative comfort. That is why we divided the first part of the representation into three different blocks that are three different SVG. In the first block is represented the map of France with regional demarcations according to the regional division of 2015. The second block represents the map of a specific region with departmental demarcations, and then the third block shows the map of the department with demarcations between large cities and the rest of the department. On the map of France represented on the first block are drawn purple dots all over the surface. Those dots are the geographical position of all the train stations (Figure 2).

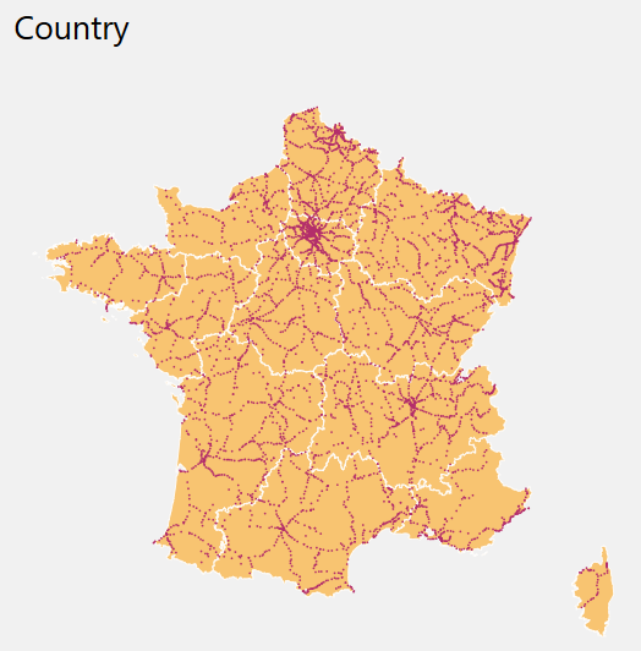

Figure 2: Screenshot of France map

By default, the user will see on his screen only the national map of France. Once he clicks on a region, the latest will be highlighted and loaded in the second block named Region. This regional map will also show the purple dots for each train stations present in this region, but those dots will be bigger than the previous map (Figure 3).

With the same logic, when the user will select a department of that region, this department will be display in the third block along

[1] www.blockbuilder.org

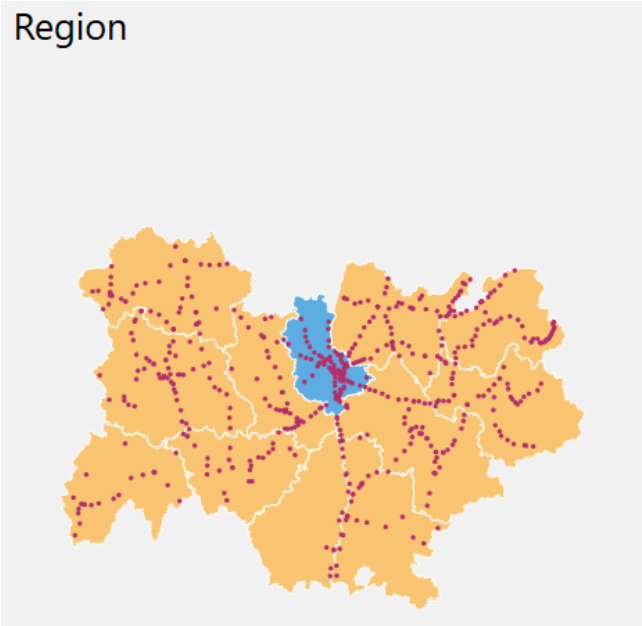

Figure 3: Screenshot of Auvergne-Rhone-Alpes region

with the purple dots (Figure 4).

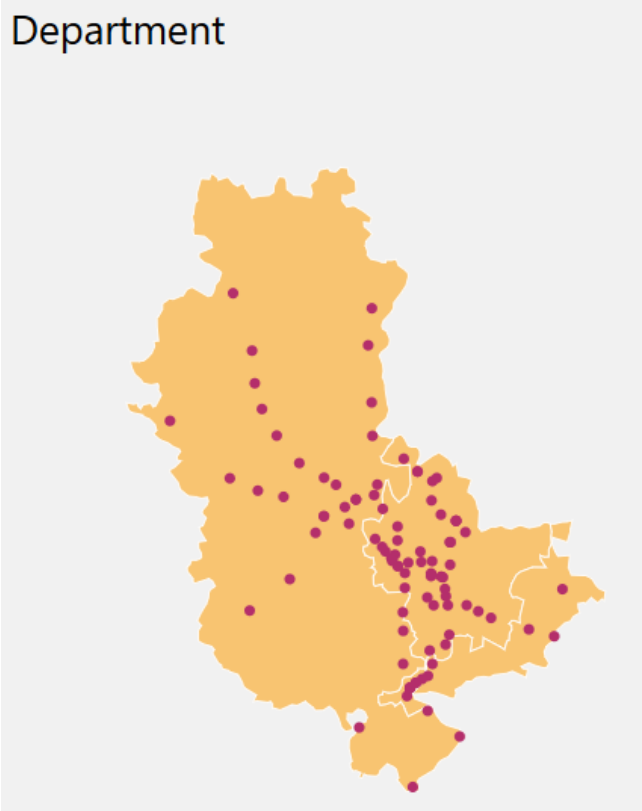

Figure 4: Screenshot of Rhone departement

Our choice of colour for those map visualisations was made to get the access to the information quickly and easily for the eye. We decided to get a pale yellow for the backgrounds map which easily outlines the information because of its neutrality. Then we used purple for the dots representing the train stations, because it is a colour often used by the SNCF (cf logo, crossbars,). Then, the colour chosen for the selection of the zones (departments, regions) is the blue. It is indeed a cold colour, gentle for the eye, and easily remarkable when placed next to a warm colour as the yellow. The second part of the visualisation is based on the display of two distinct charts offering information on lost and founds items. The first one is an histogram with two entries. It gives information about the number of items returned to their owners following a complaint in the French train stations in 2016 (Figure 5).

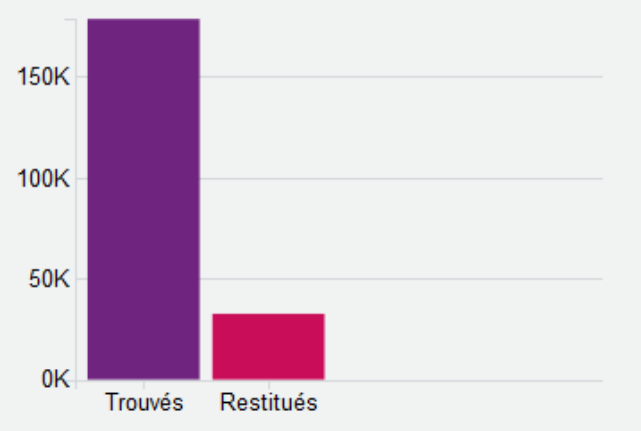

Figure 5: Histogram

We worked on the data of the year 2016 for two reasons: it is quite recent, and the number of lost and found items was superior compared with other years; the latest allows us to have more categories of different items. The histogram representation gives to the user the access to quick information on the ratio of found items on returned items which are respectively 178K and 32K. The found items are shown in purple and the returned items are shown in pink, and the choice of colour makes the two categories easily differentiable. The graphic and colour choices allow the user to have a direct access to the information. The second graphic displayed is a donut chart. This representation gives a strong quantitative aspect thanks to the repartition of the data of interests in circle along with the variety of colours. Our goal was to represent information concerning the different found items depending on their category, for each train stations of a given department. The represented donut chart gives information about items from the Lyon Perrache and Lyon Part-Dieu train stations (Figure 4).

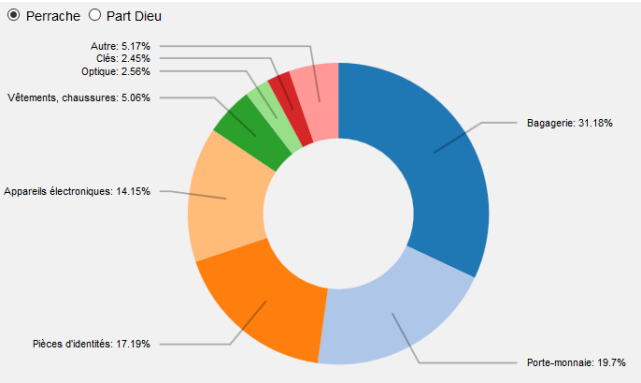

Figure 6: Donut-chart

The user can switch to one another by a radio button, and a dynamic transition of the values will appear on screen. This dynamic transition will highlight the values changes, for example wallets are more often lost in Lyon Perrache than in Lyon Part Dieu. This animation allows to distinguish the differences of found items between two or more train stations.

4 Discussion

The main interest of this visualisation is the quick access to the information, in the present case concerning found and return items in train stations through maps and charts. In our case, the innovative aspect is the fact that it gives to the user a precise idea of the distribution of the railway network, and which zones are well served, and which are not served. This information is available on a national scale, but also on regional and departmental scale.

One of the key point of this visualisation is the interactive zoom, which leave the choice to the user to display the desired region and department by clicking on the desired zone on the national map.

However, this zoom has an important fault: it can only focus the region Auvergne Rhone-Alpes and the department of Rhone. The others regions and departments are loaded in the visualisation but are not in the centre of the desired block: this constitutes the first bug to correct.

Furthermore, our first idea was to load a single map which will start as a national map. Then, by clicking on one region, a dynamic zoom would have been done on the region, and later on the department chosen. We also wanted to realise a dynamic and interactive visualisation on each dot by clicking on them. Once selected, it would load the donut chart corresponding to the selection and thereby we would have access to the full information of the found items in train stations. If no data is available for a specific train station, then the chart will not be drawn but a message will appear signalling the lack of information.

A criticism of the data set available is that we found a major mistake present in the raw data. It is a spelling mistake in the name of a categories of found items: the word trottinettes is written as trottinettes and trotinettes. Consequently, two categories exist with the same information. This kind of mistake is frequent in raw data set and it is our responsibility as a developer to pay attention. We voluntarily decided to keep the two categories to raise awareness among the public on the importance to watch out the data they are looking at.

5 Conclusion

The data visualisation discovery through some efficient informatics tools like D3.js library and Blockbuilder allows us to make the raw data, offered by the SNCF survey, speak. The major interest of this project was to deploy our entire knowledge about D3.js and JavaScript, as well as making use of examples from existing visualisations. The SNCF designed a lot of different charts in their website but they only use one dataset at a time. Our project combines multiples dataset to add a real value. It was very interesting to work as team group of 3,because we had the opportunity to discuss our personal choices of representation and to compare our ideas. Thereafter we wish to pursue the development of this project in order to fix the main bugs and to lead to our main idea : get only one map of France on which the user will browse and also to connect every train stations dot to a donut chart.

References

[1] Figaro. http://www.lefigaro.fr/actualite-france/2017/08/18/01016-20170818artfig00167-violons-oublies-doudous-perdus-et-happy-end-les-chiffres-cles-des-objets-trouves-de-la-sncf.php.

[2] Foundy. http://www.foundy.fr/sncf/?gclid=cjwkcaiapjnrbrbleiwaptgmxmike9fgr71-2844mmz6ag6-_zavyhgp0uap7wut8rmoxy1ccyiuqrocfvkqavd_bwe.

[3] SNCF. https://data.sncf.com/explore/dataset/frequentation-gares/.

[4] SNCF. https://data.sncf.com/explore/dataset/liste-des-gares/.

[5] SNCF. https://data.sncf.com/explore/dataset/objets-trouves-gares/.

[6] SNCF. https://data.sncf.com/explore/dataset/objets-trouves-restitution/.

[7] N. Viseo. http://www.sncf.com/ressources/infographie_viseo_novedia_data_science_sncf_o