

Advanced Data Visualization

Antoine Settelen, Edgar Jullien, Simon Weiss

Data Preparation

We started by analyzing our files, to check if their structures are homogeneous. Thanks to this, we have identified that the column city was missing in the file. Therefore we duplicate Column "Commune" in 2019 csv, rename it "City", remove ARR information in this column.

Then we unify the files by merging them by the column sale date and we clean up our data by :

- keeping only sale and sale before completion transactions
- keeping only apartment and house transactions
- removing transactions without prices, postal code, rooms, or living area
- replacing null values for land area with a default value (e.g. 0)
- replacing "," in price into ".", change format value for price and for other variables (geographical for postal code and city)

Afterwards, we study outliers by computing price per m2 for each city and comparing it with the data present on the internet. We also checked with a box plot, the data on the number of rooms and the size of the living area. Lastly, we remove outliers with range values and we export our data.

Data visualization

o How does your visualizations leverage at least one "pop-out effect" or "pre-attentive attribute?" Which one(s) was (were) chosen and why?

By choosing the colorization of the clusters of cities reflecting the distribution of the price per m2 per city, we managed to simplify our city mapping. The color hue of cluster was then chosen as an important pop-out effect tool in our visualization. Likewise, the color brightness and the color saturation was used in our second map to simplify the distribution of prices per m2 by district within cities. The more expensive a city is, the more important its colorization was. We have chosen to use for each city the color corresponding to its cluster.

o How does your design reflect an understanding of cognitive load and clutter?

The first cognitive load identified within our data was the number of cities and communes. How can we draw general interpretations with a diversity of information? As explained for the pop up effect, one of the responses to this cognitive load was to choose the clustering of cities.

Subsequently, when we still wanted to represent all the cities for a graph, we made two choices. The first one, to choose the colorization by city for simple graphs (e.g.: histogram of the agency's transaction volume).

The second, cut out in “view to select” for the graphs showing the evolution of prices. This selection allows the user to control the information and to choose which type of graph he wants to see. The first view that is displayed should be the easiest to understand with the least amount of information. The second view offers a more precise view with more information.

o Which metric(s) you have used? and why?

In order to choose the most relevant metrics, we inquired about what was used in the real estate market. The first and most used metric is the price per m2, which allows a quick understanding of the price for sales. Next, we looked at transaction volumes to identify in which sector our sales were taking place. We have added the cluster view described above. Finally, we used the sales volumes to find out which sales were the most profitable for us.

Design

o What is (are) the most important(s) dashboard design principle(s) you have used?

The fundamental principle behind the design of our dashboards is that each dashboard must answer a specific question in order to further our analysis. Thanks to this, we chose the graphs that we felt were most relevant to our answer. Then from a more graphical point of view, we constrained ourselves not to put more than 2 different graphics for the dashboard. This decision was made in order not to overload the message we wanted to convey. In the same spirit, each graphic on the same dashboard is linked to the other.

In the building of our dashboards we have tried to follow these principles :

- ❖ Choose the right measure : like mentioned above, we identified our key metrics in our sheets and stiked to them in our dashboards.
- ❖ Choose the right display : We have built several visualizations in our sheets in order to identify the best graphic according to our needs. For example, initially, for the distribution of the price per m2, we simply had the histogram, which we improved by adding a map. For each visualization, we make every effort not to forget the legends and labels in our dashboard and to place them in a strategic view.
- ❖ No scrolling : This could be a mantra for social networks but it has been instead our motto in our dashboards. We have taken care to automatically change the size of each construction first, resizing them so that the user does not scroll. Each dashboard has been tested in presentation view.
- ❖ Meaningless Color Coding. Variety of Elements : We have tried to have a diversity of visualizations in our dashboard so as not to tire the reader: map, histogram, box plot, lines, dynamic chart etc...

- ❖ Highlight important Data : Finally, we tried to highlight our main discoveries in our dashboards by underlining, putting in bold and pointing out the points that seemed to us to require particular attention.

o What is (are) the most important(s) data stories design principle(s) you have used?

The most important aspect of our story is that it was built for our target audience (Gerard T.). First of all, we were interested in the needs of the marketing director and the agency, in order to elaborate a problematic which will serve as a guideline for us : how to start our journey ?

In order to answer the identified question, we used several types of data story.

Within the navigation tools, we have used the caption box and put it inside the answer of the question. Our story can be read chronologically: we started with a general view and then tried to go into detail little by little. We used data story points for each view combined with annotations within the dashboards, trying to minimize the text.

As our story was originally intended for Gerard, we have tried to build a story that can be modified and manipulated by Gerard as he sees fit. This, while leaving both points of our history in suspension but also some points that the latter will be able to use in its presentation to the executive board.

o Why do you think your data story meets the audience's needs?

All the answers given above confirm our view that our data story is a good one and meets the audience's needs. We think that it is addressed first and foremost to Gerard T. our marketing director, by providing him with chapters directly addressed to which he can play with the data (e.g. dynamic views, views to choose) but also with strong leads to present to the executive board, notably Aude P, our General Manager.