Data Visualization: Process Book

Arnaud Robert, Quentin Müller, Aleandro Eccel

3rd of June 2022

1 Introduction

Public Houses, or pubs for short, are an essential component of the British culture. They are socializing hubs, be they in a city or in the country side, and the go to destination to celebrate many an event.

Our ambition through this project was to showcase the presence of these central places in the UK, through space and also time. We also wanted to showcase some peculiar facts about some of these pubs, and to explore their diversity. To do, we present multiple maps of Great Britain, with various options and interactive functionalities that enable the user to smoothly explore the data.

2 Data

The first step of the project was the data gathering and data cleaning. We looked through multiple website proposing open source datasets such as *Google Dataset Search* or Open Swiss Data. The first dataset we searched for was related to bird migration, but no datasets that could be accessed was large scale.

We decided that we would go in another direction and chose the subject of bars/pubs. More specifically, pubs in Great Britain as they are a big cultural component in this country.

2.1 Pubs dataset

We went with a dataset coming from Kaggle with some projects already done around it, some of which were visualisations, to get an idea of the possibilities offered by the dataset. It represent a list of pubs in Great Britain with their name, address and so on. Here is the list of its attributes:

- fas_id: a unique ID linked to the pub
- name: The name of the pub
- address: The full address
- postcode: The postcode linked to the address
- easting: measure of the projected coordinates on the x-axis
- northing: measure of the projected coordinates on the y-axis
- latitude: the latitude
- longitude: the longitude
- local_authority: the district governing this particular address

This dataset needed some cleaning as two entries had their attributes shifted because of quotation marks found in the name of the pub. A simple erasure of the quotation marks was enough to solve this problem. Furthermore, some pubs did not have a latitude and longitude, for simplicity, we decided not to use these pubs in our visualisation.

2.2 Districts dataset

We wanted to complement the previous dataset with another comprised of the districts of Great Britain. This was the first challenge that we encountered, indeed there was no existing list of districts and their population thus we had to scrap the data from Wikipedia. This was not the end of the story as the districts were separated into countries (England, Wales and Scotland) which we had to merge together.

To join the two datasets, we still had to perform another preprocessing step in the form of renaming some local_authorities in the first dataset. Indeed, some districts changed name or were merged with others since 1995, the pubs dataset still uses some of those names, therefore they had to be replaced. Some names also did not match, we can find for example St. Helens which did contain the dot in one dataset but not the other.

This dataset was not used directly in the end, but the processing of the district came of use later on, when handling the history dataset.

2.3 Pub counts through the years

Searching through the national british website Office for National Statistics (ONS), we found a dataset showcasing the evolution of the number of pubs per district from 2010 to 2017, we call it the history dataset.

Our resulting idea was to do a choropleth map with a slider on which one could see the evolution of the number of pubs through time. Hence, the next required element was to obtain a geojson with all the district boundaries. As mentioned earlier, we faced again the issue of synchronising the districts names and shapes. The history dataset would use the old administrative division, thus we needed to obtain boundaries corresponding to the old divsion as well. We found such data on the Open Geography Portal website of the ONS, and performed some additional name preprocessing to standardise both data sources.

3 Visualisation

3.1 Index

The index is the first page of our visualisation, which the user first sees upon connecting to the website. We inspired ourselves from the profile page of Netflix as can be seen in Figure 1, a simple and minimalist layout with only some squares that the user would click on to get to a given page. At start the number of squares was unknown therefore we searched for a way to put them in a row and possibly also in columns. In the end, we only have two different pages that the user can go to, so a single row was needed.

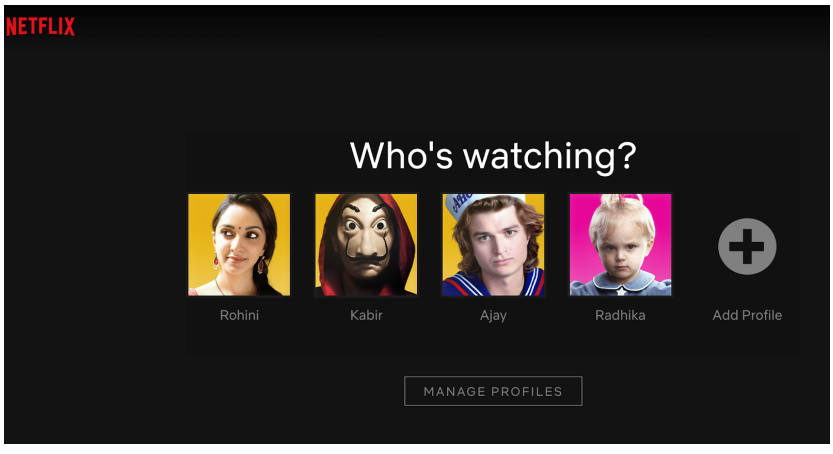

Figure 1: Netflix profile page

3.2 Original map

The original map that we added to the visualisation was a leaflet map centered around London. We filled the map with the pubs from the first dataset. We encountered a problem there with the version of D3.js we used which forced us to directly feed a function to the "CSV" method, as simple change to the newest version resolved this issue. We then filled the map with markers representing the pubs found in the first dataset and clustered them using "Markercluster", a plugin for leaflet which enables, as its name indicates, the clustering of markers. With this plugin, the problem arose when the visualisation was deployed on GitHub as the files contained some characters it did not interpret correctly. This issue was solved adding a ".nojekyll" file to the repository.

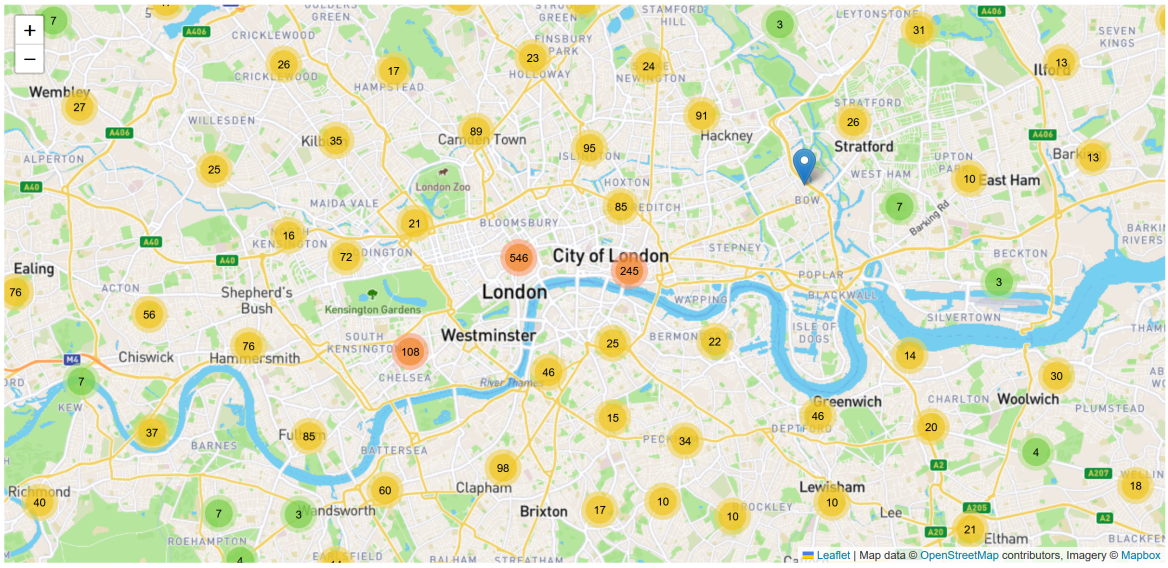

Figure 2: Original Map with clusters

3.3 Spicy Facts

Because of the redundancy of the pages, we decided to include the first map into the spicy facts page. Therefore, this page contains a map with all the pubs with some facts written on the right panel. One can click on the fact on the right and a little explanation will appear underneath the clickable fact. It also "flies" on the map to show where the corresponding bar is.

Instead of just a table, we found that it was prettier to make a side panel appear with the table within.

The less trivial fact to find was the point furthest away from any bar. We need to constrain this problem because otherwise we would have a point at the other side of the globe. We used a Monte-Carlo approach to find this point but it comes with some drawbacks. First, the method only works well when the surface contains only "valid" points (we didn't want the point to just lie in the sea). We therefore decided to only consider England's main island. We also need to create the boundaries where we randomly sample the points. We didn't find a usable file with the precise boundaries of England so created ourselves a very rough boundary of England (using gps coordinates). Then we do the following:

1. We generate 5'000 random points within the selected boundaries.
2. Then for all random point:
3. Compute the minimal distance between the random point and all bars
4. Take the random point with the maximal minimum distance (this is our random point that is the furthest away from any bar).

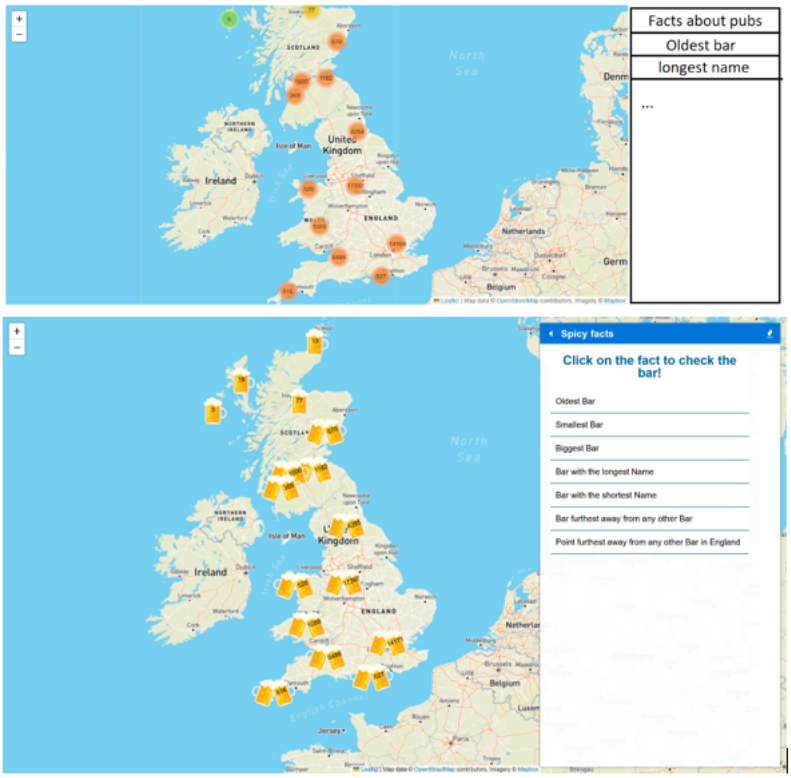

Figure 3: Draft page designed during milestone 2 against the final product

5. Create a circle around this point and re-generate 5000 random points within this circle and run step 2 to 4 a second time.

As this method is random, we ran the algorithm 10 times and took the point that appears the most.

3.4 Beer icons

As a supplement, we decided to add some custom icons for the clusters and markers on the leaflet map. To fit the theme, we used beers as our cluster icons, if there were more than 100 pubs in the cluster, the icons becomes two beers cheering instead of a single one. The markers became small houses to represent pubs. The main difficulty in this case was to find a way to make the clusters beers. Originally, we though about using an image, but quickly realised that using a css class would be more fitting to the problem. Therefore, we copied a beer made in css from a website about the Oktoberfest (https://scriptcodes.co/css-beer-single-div-octoberfest-rzbpma).

3.5 Pub History

The second page displays an interactive choropleth map representing the evolution of the number of pubs in each of Great Britain's district over the course of eight years. A time slider is present on the bottom right corner which the user can use to update the map information, either by clicking or using the keyboard arrows. A color gradient is associated to the absolute number of pubs in a district, with the corresponding legend in the bottom left corner. Upon hovering over a district with the mouse, information relative to this district is displayed in the top right corner. Such information includes the name of the district and the actual number of pubs its hosts, as well as the percentage change in this value compared to the previous year (except if the year is 2010 since there is no past data in this situation).

The map allows to quickly see what is the distribution of pubs in the country in a given year, but is of little help when it comes to appreciating the evolution of this quantity through time. Indeed, since the number of pubs can be very different from one district to another (e.g. less than 5 vs more

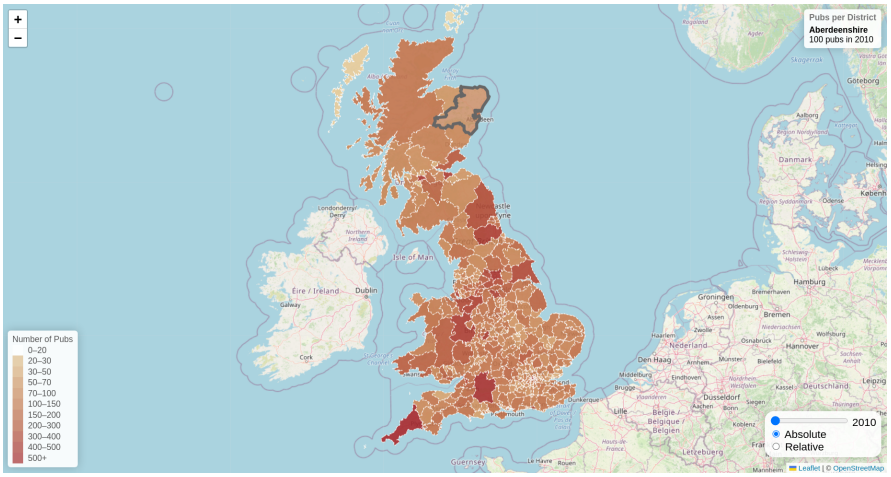

Figure 4: Choropleth map, pub counts per districts

than 500), the color scale is relatively stretched, and thus for a given district, the change of shade is not so obvious when using the slider. Thus, we added a "relative mode", which displays a choropleth map focusing instead in the percentage change relative to the previous year, and uses a diverging color scale. The legends update accordingly when switching mode through the radio form buttons, and naturally such map can not be displayed for year 2010 for the aforementioned reasons (which defaults to 2011 when attempted).

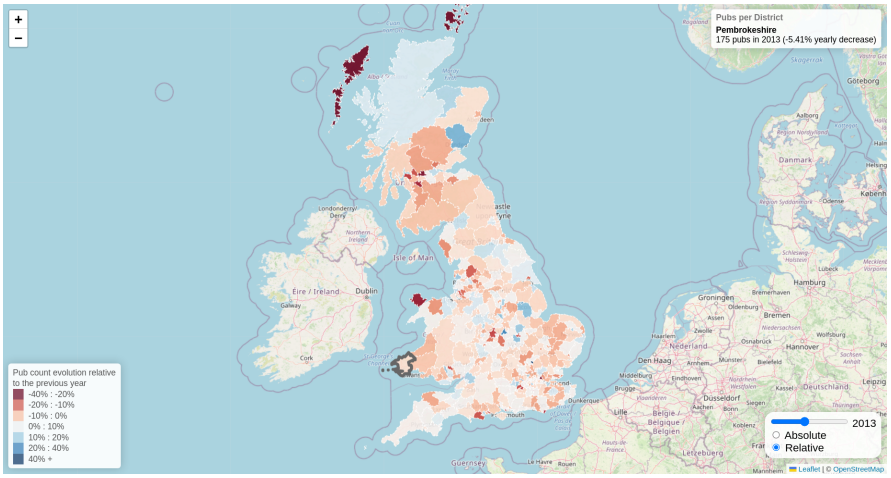

Figure 5: Choropleth map, relative yearly change in pub counts per districts

4 Task breakdown

The data gathering and data cleaning was performed by everyone.
The index, the original map and beer clusters was done by Quentin.
The evolution of pub distribution was done by Arnaud.
The actual map, otherwise known as spicy facts, was done by Aleandro.
It brings the workload to, more or less, one third of the total per person.