Milestone 1

Dataset

We will use multiple datasets, the first one is composed of the list of all the pubs in Great Britain, not England or the UK as the name of the dataset suggests, and can be found here : https://www.kaggle.com/datasets/rtatman/every-pub-in-england. A second one will be scrapped from wikipedia, it is a list of all the districts in Great Britain and their population. This dataset needs to be scrapped from multiple wikipedia pages as there exists no single page containing every district.

The columns of the first dataset are :

fas_id (int)	name (str)	address (str)	postcode (str)	easting (int)	northing (int)	latitude (float)	longitude (float)	local_aut. (str)

The second dataset will only be composed of District (string) and Population (int).

The two datasets will be merged, thus it will need some data preprocessing done to match the districts. This should not be too time-consuming as most districts already have the same nomenclature. Another preprocessing step to take only concerns two rows where some shift in the columns entries occured, this can also be solved easily.

Another dataset that is being considered to complement the visualization concerns happiness levels throughout the UK. It is recovered from the office of national statistics (https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/measuringnationalwellbeinghappiness). Its subdivision by counties and unitary authorities make it easily linkable with our previous data.

A dataset of pub counts by counties over 7 years will show the dynamics of pub creation/closures throughout the UK.

We would like to find other datasets to see if there are some correlations with alcohol-related accidents.

Problematic

The topic of our project is “Pubs in Great Britain and their influence”. The main axis of our work will therefore be on the distribution of all the bars in Great Britain and we want to see how it is proportional to the population. We also want to add the possibility to select any point on the map of Great Britain and tell which bars are the closest to you. This can lead to a map where we show a kind of heatmap of the

distance between any point and a pub (showing the one that is the furthest away too). Additionally, we wish to showcase the potential influence of pubs distribution on socio-economic factors/events, such as well-being, and alcohol related-deaths.

Our motivation is to showcase the status and impact of an important part of British culture. It can be an interesting fact to know about for any tourists or British people.

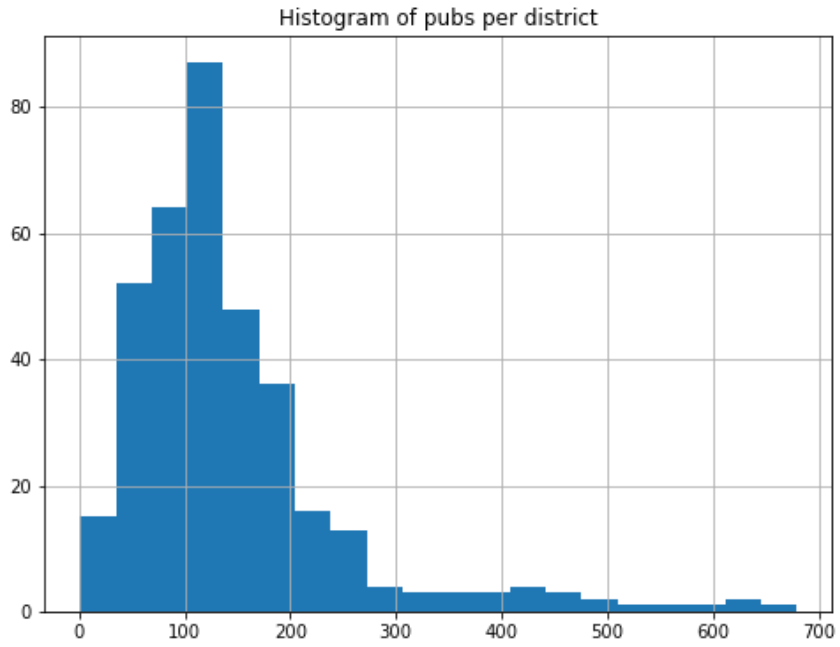
Exploratory Data Analysis

How many bars do we have, how many districts, how many bars per district.

After doing the small preprocessing steps explained in the dataset section (matching of districts), here are some statistics :

- There are 359 districts.
- There are 51566 different pubs.

Here is an histogram of the number of pubs per district :



The dataset does not contain any NaNs.

For the data fetched from Wikipedia, it is trivial as we just have the population per district.

Related work

As the main dataset is taken from Kaggle, there are some related projects, the main one being :
https://www.kaggle.com/code/gpreda/map-with-every-pub-in-england/report.

Our addition to the existing projects using this data is to put it in relation with other factors such as happiness or accidents, and to have a starting point to assess the potential impact of pubs distribution over the population.

We will take inspiration from the VoteInfo app that gives precise information on the swiss votation by canton and by town, in an interactive format that allows both global point of view and progressive granularity. Multiple interactive maps of the UK will be available, depending on the the subject being showcased (e.g. map with a slider to see evolution of pub counts, heatmap to see the happiness to pub count ratio … etc)