Data visualization for Scientist Report

Alessandro Calvio - 0001099335 - XXXVIII Cycle

---

1 First Visualization - Italy Earthquakes Epicenter Depth

This visualization wants to show the average epicenter depth by province for earthquakes in Italy over the period 1985-2020. Two different datasets contributed to the final chart, due to the necessity to create an association between the earthquake location and the related province.

1.1 Dataset description

1.1.1 Dataset #1 - Earthquakes dataset

The dataset is downloadable at `https://www.kaggle.com/datasets/lorenzogucci/earthquakes-italy-19852020`. Below is a description of the fields and their semantics:

- Data e Ora (ITItalia): The date and time, in the format YYYY.MM.DD HH:mm:ss, at which the earthquake started.
- Magnitudo: String containing the magnitude type and level of the earthquake.
- Zona: String indicating the approximative area in which the earthquake started.
- Profondità: Depth in meters from the sea level.
- Latitudine: Number indicating the latitude coordinate.
- Longitudine: Number indicating the longitude coordinate.

Data e Ora (ITItalia)	Magnitudo	Zona	Profondità	Latitudine	Longitudine
1985-01-07 08:33:56	Mdl 2.4	Gubbio (PG)	10	43.36	12.54

The principal preprocessing operations have regarded the elimination of duplicate rows in the original dataset that could have compromised the veracity of the results.

1.1.2 Dataset #2 - Provinces map

The dataset is downloadable at `https://github.com/openpolis/geojson-italy/blob/master/geojson/limits_IT_provinces.geojson`. Below is a description of the fields and their semantics:

- prov_name: String containing the name of the province.
- prov_acr: Acronym of the province.
- prov_istat_code_num: ISTAT code of the province expressed as a number.
- prov_istat_code: ISTAT code of the province expressed as a string.
- reg_name: String containing the name of the province's region.
- reg_istat_code_num: ISTAT code of the province's region expressed as a number.
- reg_istat_code: ISTAT code of the province's region expressed as a string
- geometry: Shape of the provinces in GeoJSON format.

prov name	prov acr	prov istat code num	prov istat code	reg name	reg istat code num	reg istat code	geometry
Bari	BA	72	072	Puglia	16	16	...

No data preprocessing has been required for this dataset.

The final association was created through the geojoin operation (between the fields [latitude, longitude] and geometry) thanks to which it was possible to link each earthquake to the data of the relevant province. The last step was to calculate the average of all depths falling within the same province.

1.2 Chart choice

The final dataset type is constructed from a categorical datum (the province name) and a numerical datum (the average epicenter depth of earthquakes).

Given these assumptions, the choices, on the type of map graph to be used, involved the Choropleth and the Hexabin map. The final choice fell on the former since it was the only one that maintained a clear reference to the area and gave a clear geographical location of the trend shown. The other type, in fact, by making use of arbitrary aggregations, i.e. hexagonal shapes, would have lost information about the boundaries of the provinces.

1.3 Chart style - colors and fonts

The type of data represented is numeric and sequential, hence the choice to use a gradient of a single color hue. Specifically, the brown color, somewhat reminiscent of the idea of earth, starts from a very high brightness for low values of the epicenter depth and increases, in saturation, up to the maximum values. In addition, the use of this gradient also allows the graph to be read in black and white. As for the lettering, the choice was to use a font without graces to ease the readability of the chart. Here, also the choice to not insert the name of each province on the map.

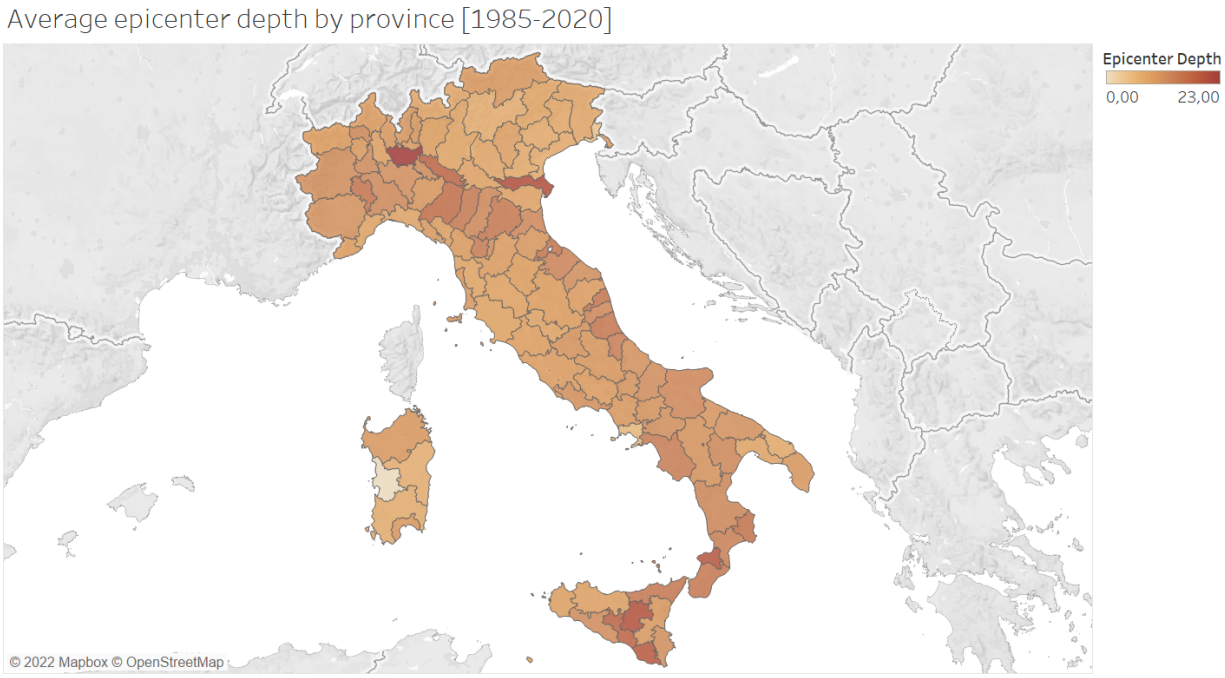

Figure 1: Average epicenter depth by province in the period 1985 - 2020

2 Second Visualization - NO2 levels for three monitoring stations in Bologna

This visualization wants to show the weekly average of NO2 levels for three monitoring stations in the city of Bologna in 2022.

2.1 Dataset description

The dataset is downloadable at `http://shorturl.at/jpqyE`. Below is a description of the fields and their semantics:

- _id: Measurement identifier string.
- reftime: Day and time of measurement, in the format YYYY.MM.DDTHH:mm:ss.
- stazione: Name of the monitoring station.
- value: Measured value of station, in $\mu g/\mathrm{m}^3$.
- agente_atm: Categoric type of pollution agent.

_id	reftime	stazione	value	agente_atm
1320113	2022-04-04T 00:00:00 +02:00	VIA CHIARINI, BOLOGNA VIA CHIARINI	36	NO2 (Biossido di azoto)

The dataset processing operations involved filtering the agent type to highlight only those entries related to the NO2 data and aggregating, by week of the year, the respective values.

2.2 Chart choice

The final dataset is presented as a set of several time series, one for each monitoring station, with which the values for detected NO2 levels are associated. Because of these characteristics, the comparison between the charts to be used mainly involved two types: the stacked area chart and the line chart.

The characteristic of the stacked area chart allows to show the qualitative trend of the three time series while providing a quantitative indication of the total and the contributions that individual stations give. In the context taken into analysis, these features would have little value since the three stations belong to independent areas and do not pertain to the same zone. Consequently, the measurement of the total would not be a useful indication to show, for example, the quality trends in a single zone. For these reasons, the choice to use a line chart is due to the desire to show the independent trends of the three stations in an accurate manner allowing, also, a comparison of the phenomena over the three zones represented.

2.3 Chart style

Involving categorical data, in which each value is independent of the others, the choice of color was oriented toward the use of three very distinct and bright hues capable of making the three stations distinguishable and readable even in black and white conditions. The specific choice of colors was dictated by aesthetic taste. As for the lettering, the choice was to use a font without graces to ease the readability of the chart.

A brief note on the choice of X-axis labels: although the aggregation is done by weeks, the X-axis labels are graded by month. This choice was primarily dictated by the fact that a label for each week would have worsened the readability of the graph without making significant improvements overall.

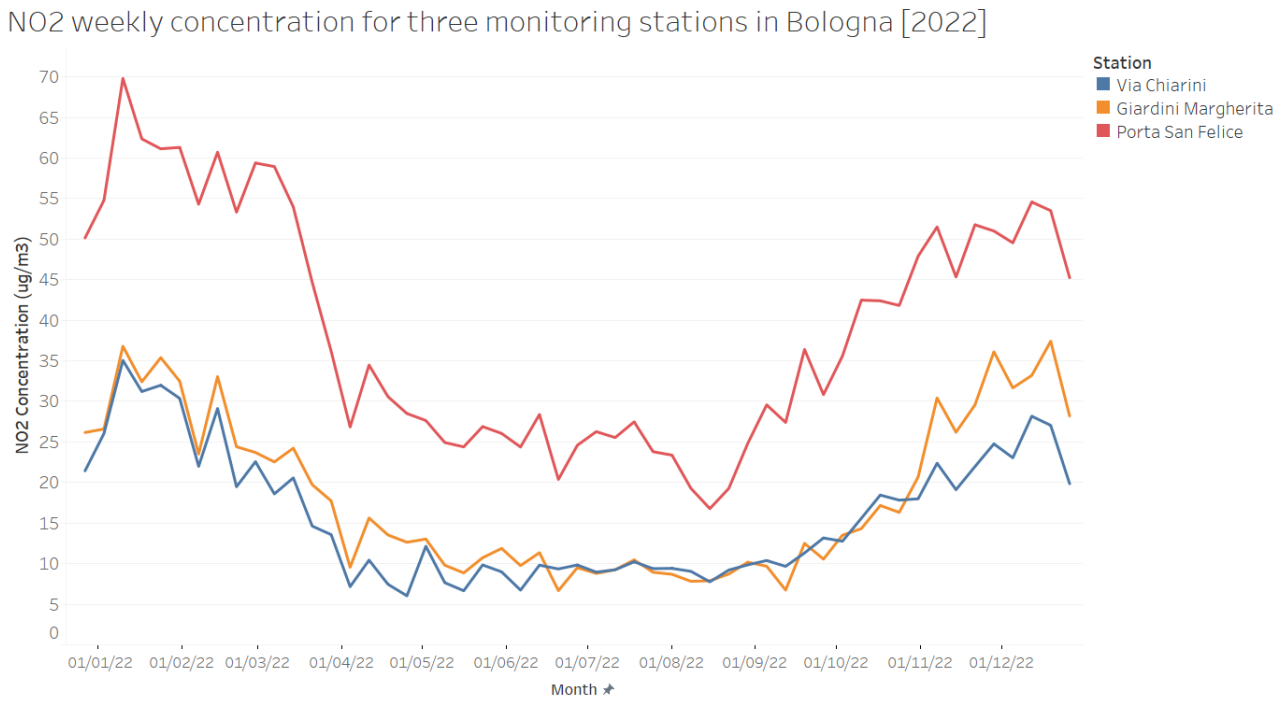

Figure 2: NO2 weekly concentration for three monitoring stations in Bologna