CAValli Team: Report 3

- Alessandro Longo – 5697430
- Vittorio Bartolomeo Secondin – 4798279
- Christian Dagnino – 4663694

Assignment 3 – Timelines and distributions

1. Data Preprocessing [in Python]

This part involved activities related to data import and rearrangement of the three datasets containing temperatures (minimum, maximum and average). Specifically:

- we imported the data and **replaced temperature values** which were set to -99.9°F, occurring in November and December 2023, **with instances of NaN, since the data had not been collected at that time**;
- we converted temperature values from Fahrenheit to Celsius degree scale;
- from codes given in the datasets we retrieved **information about year and reference state** for each observation;
- in the end we subdivided **minimum, maximum and average temperatures into different .csv files, one for each state**.

See *Temperatures.ipynb* in our repository for further details.

2. Website setting [in JS, HTML, CSS]

We designed the requested data visualisations, specifically:

- one line chart displaying the monthly temperature (min, avg, max) for a selected state over the chosen years;
- three distinct radar charts for the same purpose;
- a ridgeline chart with two data densities (respectively related to min and max temperatures), whose heights and shapes illustrate the concentration and spread of data on the given year for a selected state.

More in detail:

- **technical choices**, in particular:
    - we decided to **set Celsius degree scale as the default scale** on the *y* axis of the line chart and on the circles that are drawn within the boundaries of each radar chart. One extension could consist in letting the user possibly personalise this choice;
    - **the full range of years listed in the dropdown menu is not available for all states, specifically Alaska misses some of them**. Therefore, when Alaska is selected, previously checked years are no longer plotted in visualisations and checkboxes are disabled;
    - the ridgeline chart is **the result of a kernel density estimation (KDE) aimed at reconstructing the probability density function of minimum and maximum temperatures over each year**. The estimation involves kernel smoothing, which is a statistical technique that can smooth a function on the basis of a weighted average of neighbouring observed data, where weights are returned by some *kernel* (a window function). In our case it was applied to the histogram of yearly temperatures, discretised in bins and computed using the twelve

monthly temperatures. In specific terms we applied **the Epanechnikov kernel, which is known to be optimal in the sense that it minimises the mean integrated squared error** and it's a standard option according to D3.js website examples. Note that the function is fitted beyond the actual limits given by the minimum and maximum temperature values over all years in the selected state, highlighted in green. This is a common approach in several examples. A variant may possibly truncate the function at these two boundaries.

- **stylistic choices**, they include the most aesthetic decisions:
    - in the line chart, the colour of lines/dots associated with each year is **selected in a list of 50 colours** that will repeat (since the dropdown menu presents more than 50 years), but are **placed in such a way that contiguous years are less likely to be linked to similar or even identical colours**;
    - **two mutually synchronised and interactive legends** appear next to the line chart and the rightmost radar chart: a click on each of the rectangles contained in these legends allows to add only lines and dots associated with the corresponding year, so that **the final visualisation can be displayed step by step**, prior to any overlap;
    - **neither the line chart nor radar charts instantiate dots whose tooltips would show NaN temperature values** (e.g. November and December 2023), because these values are filtered in JS code as a further preprocessing before the actual visualisation being formatted;

- the above-mentioned tooltips show temperature in both Celsius and Fahrenheit degrees;

- in the ridgeline chart, the two data densities shaping minimum and maximum temperatures over one year **are respectively filled with blue and red, with the same colours kept persistent over different years** in order to make the visualisation less confusing.