Click here to toggle on/off the code.

Assignment 01 - IBM ML.

Made by: Alessandro Manfredini

Analysis of temperature data: is global warming real?

Of course this is just an exercise, we all know global warming is real. However, I was interested to see how significant the effect is with a simple analysis. It turns out is very significant.

The study aims to evaluate the significance of the deviation of the average temperature, in one specific location, in the recent years with respect to a point in time. In summary can be summarized in the following parts:

- Evaluate data quality and asses a potential region where to conduct the analisys
- Define a reference time t0.
- Define a control time t1.
- H0 assumes that the temperature does not change over time. So under H0, the average temperature difference between t1 and t0 is a random variable expected to be distributed as a Gaussian centered in zero.
- Evaluate the standard deviation of that distribution.
- Evaluate the p-value of H0.

The dataset

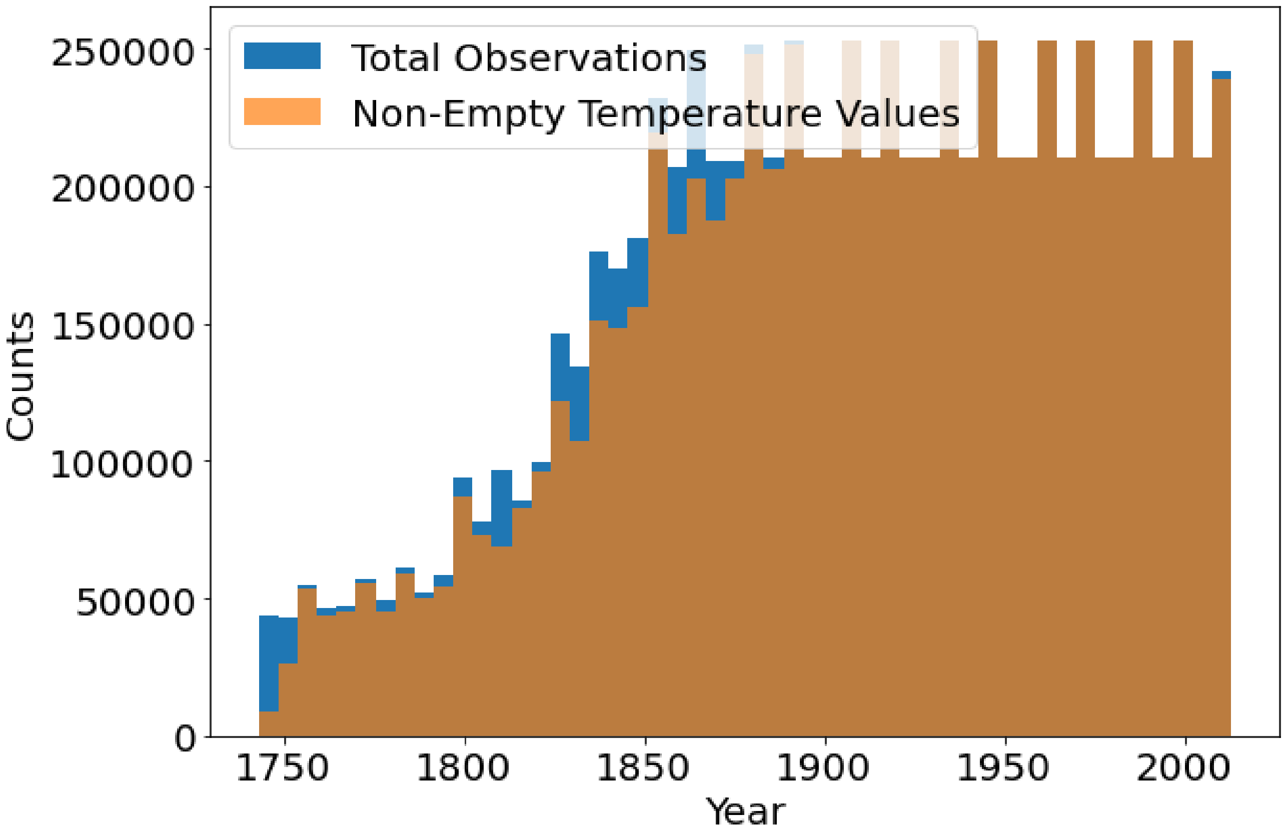
The used dataset was taken from kaggle (https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data), which is a repackaging of the Berkley Earth dataset on global temperatures. In particular I concentrated on the dataset that contains temperature reading per each month and per city since 1740. The dataset has the following features:

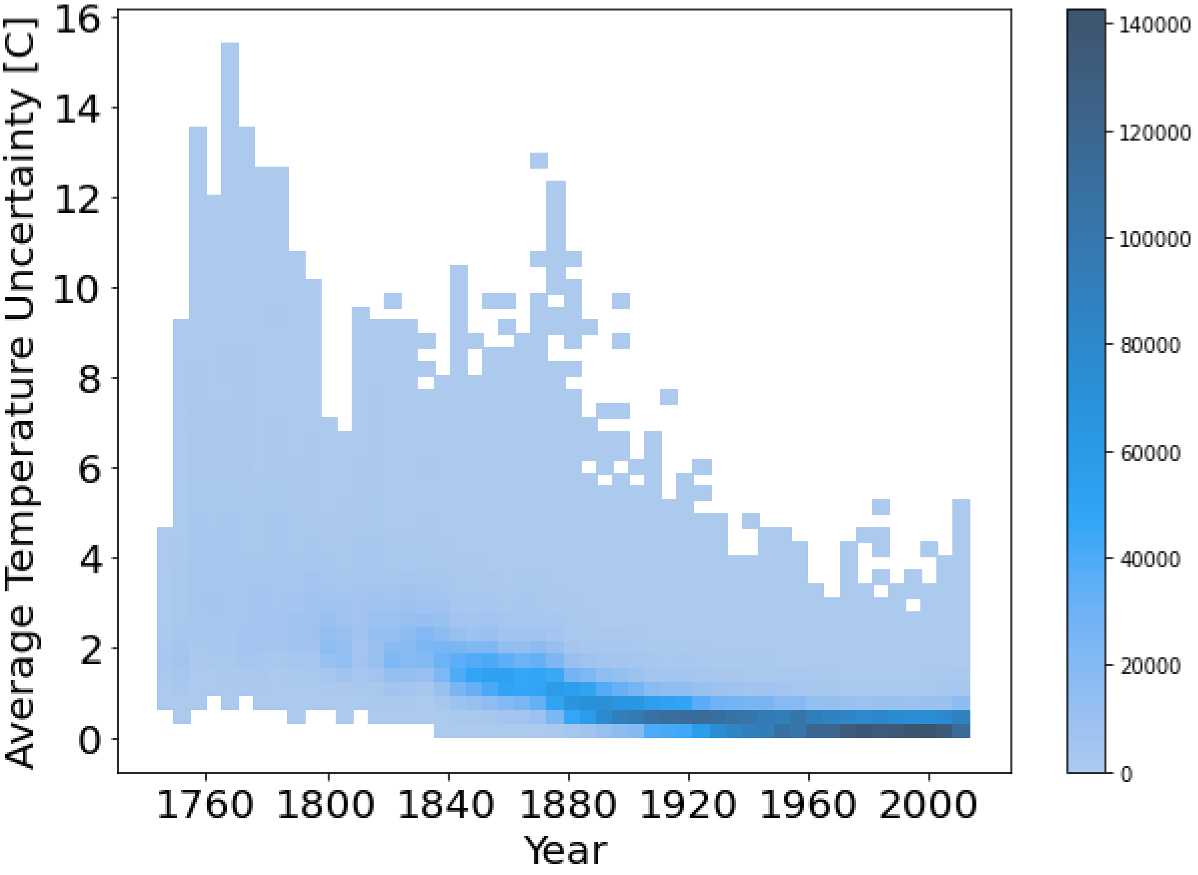
- Timestamp
- Average Temperature over a month
- Temperature measure uncertainty
- City
- Country
- Coordinates

For this analysis I considered only one location, Rome. The notebook relative to this analysis can be foun here (https://github.com/panManfredini/IBM_ML/blob/main/01_ReadingData/project_assignment_1.ipynb).

Data Quality Assessment amd feature engineering

There is a large number of NA values, especially for early years, also the number of observation reported for early years is much lower, due to the fact that there weren't measurements for certain cities, or that city perhaps not developed yet. This is shown in the figure below.


Total Observations
Non-Empty Temperature Values
250000
200000
150000
100000
50000
0
Counts
1750
1800
1850
1900
1950
2000
Year


16
14
12
10
8
6
4
2
0
Average Temperature Uncertainty [C]
1760
1800
1840
1880
1920
1960
2000
Year
140000
120000
100000
80000
60000
40000
20000
0

The uncertainty on the temperature has a large variance for years below 1900 and it is relatively higher due to tecnology difference. Now in this analysis we need a representative sample of temperatures from the past and a representative sample of temperature of the present. Since the data is by city, and data from many cities are missing in the past this can affect the outcome of the analysis (we need to compare to the same locations). Also the uncertainty of the temperature measurments is correlated with the year, beacuase of technology advancement. Because of this I decided then to only use data above year 1900 for the final hypothesis test, because has very few missing values and is more uniform in measurement uncertainty.

Data cleansing

So for data cleansing we simply drop all years below 1900, and cross check that for the chosen location Rome, all data are present.

Feature engineering

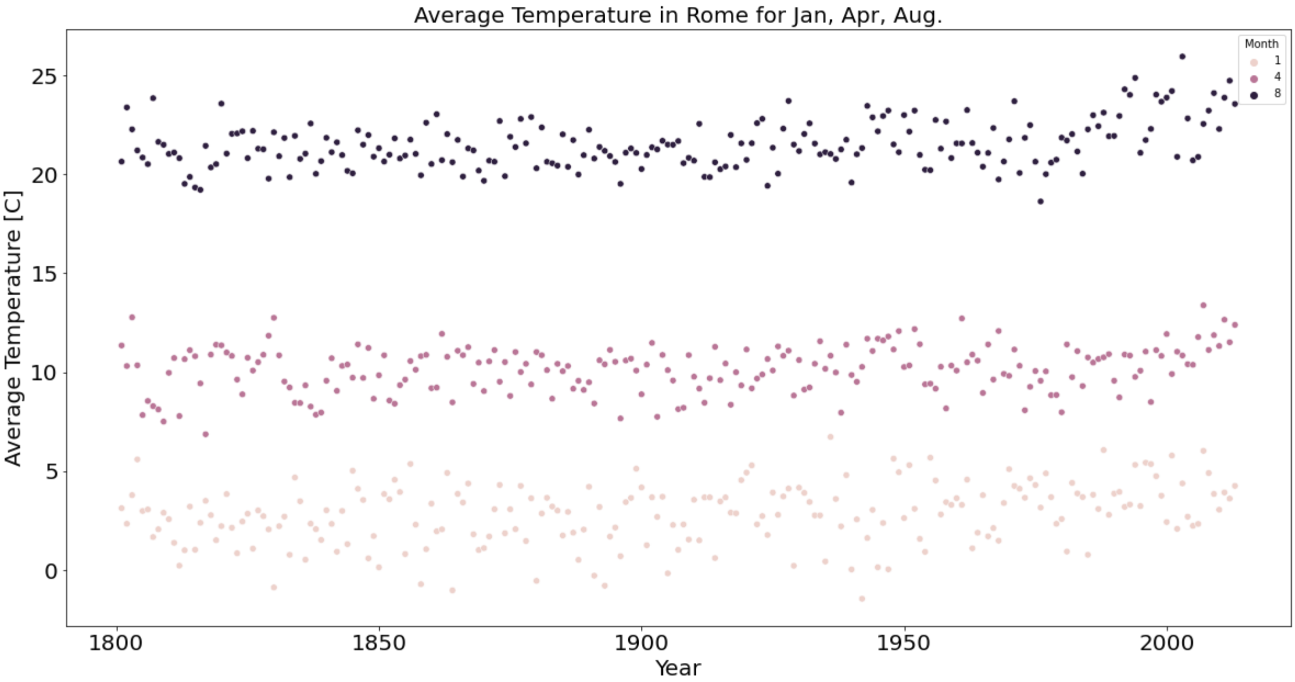
It was usefull to derive a few quantities necessary for the evaluations:

- Reduced size datasets only with Rome, Paris and Melbourne
- The average value of temperature reading per month grouped every 20 years
- Standard deviation per month grouped every 20 years for the t0

Data Exploration

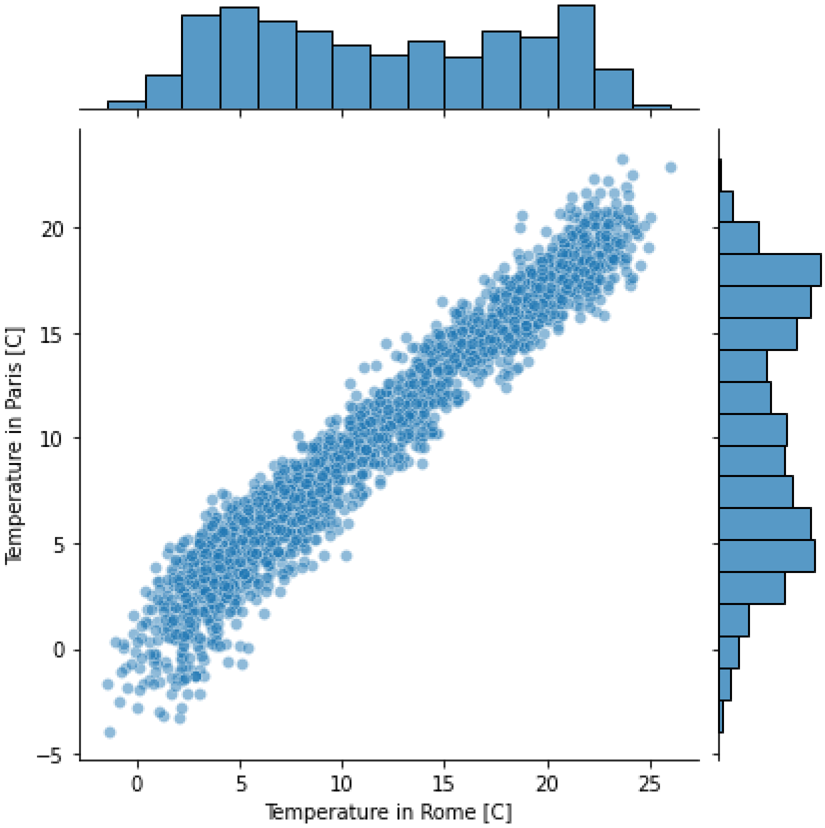
One of the aim of data exploration is to cross check few of the features that we expect in the dataset and validate the assumptions made, for then proceed with hypothesis test. As an example I crosschecked the correlation of temperature readings of neghboring cities.

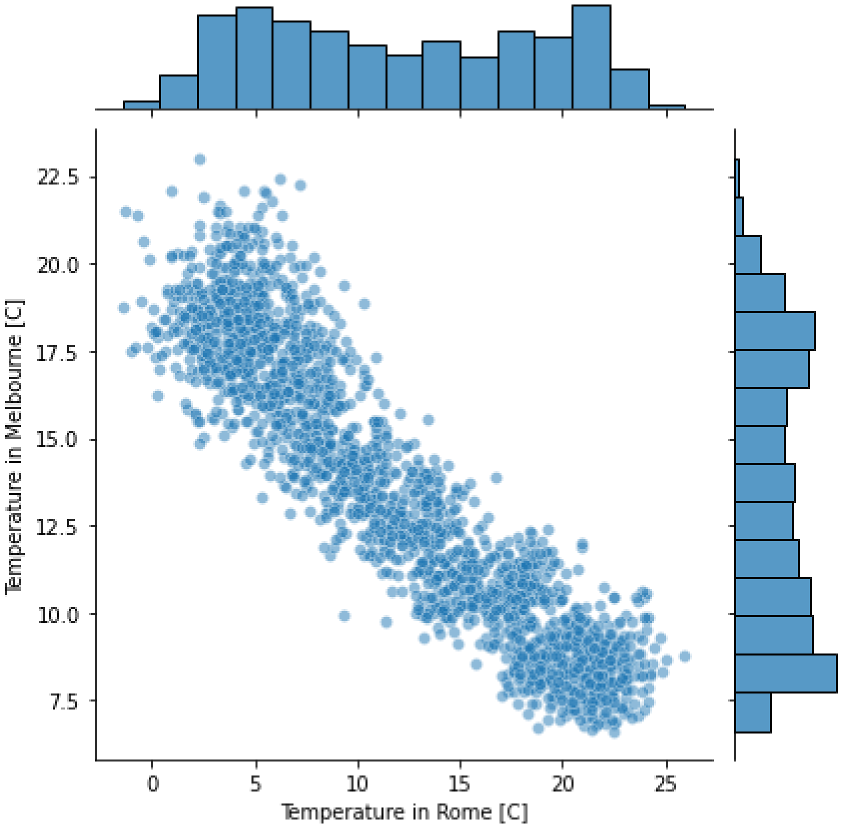
As a first study you can see below the distribution of temperature values in Rome since 1800 for 3 different months: January, April and August. This shows that temperature per month need to be treated separately, because the temperature per year is not distributed as a Gaussian.



Seasonal Correlation

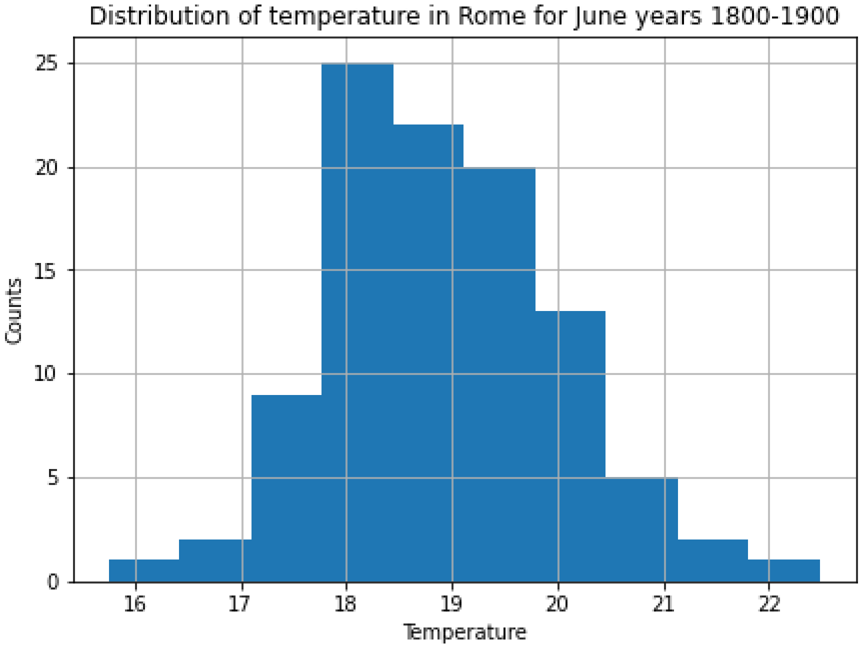
As an example here we can study the expected correlation and anticorrelation originated by the earth rotational axis orientation and position with respect to the sun (the confounding variable). In particular we expect that two locations in the same emisphere would have correlated temperature readings (because of seasonal change), whereas temperature readings in different emispheres will be anti-correlated. This is what we see in the following two plots, where we compare the temperature readings in the last 100 years of Rome, Paris and Melbourne.





Gaussianity assumption cross check

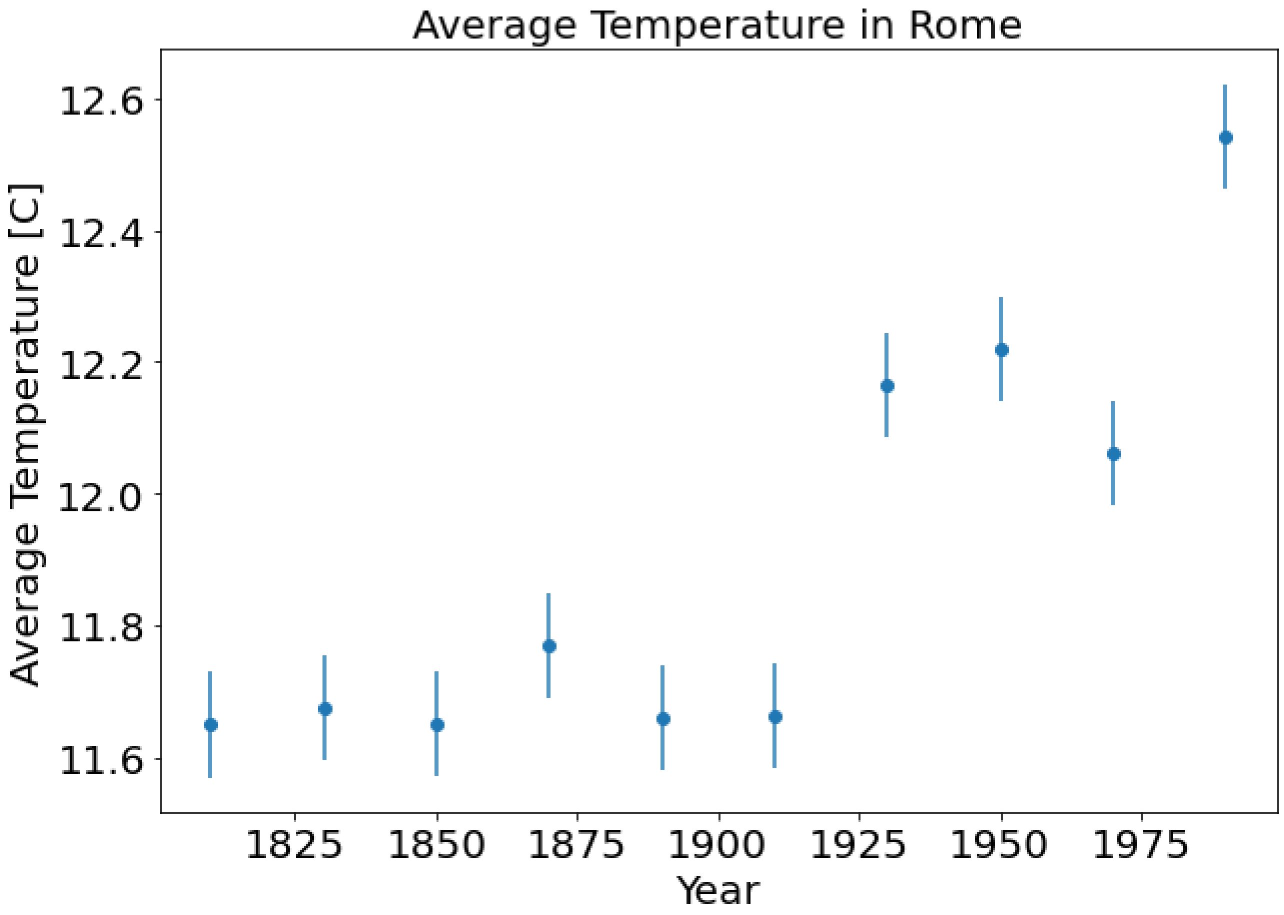
One of the main assumption for this analysis is that the temperature per single month and location has a Gaussian distribution. This is true of course for a period of time where global warming is absent. Below is shown distribution of temperature in Rome for June years 1800-1900. Considering the small amount of sample the assumption can be considered acceptable.



```
The standard deviation for the temperature sample of June is 1.1 C
```

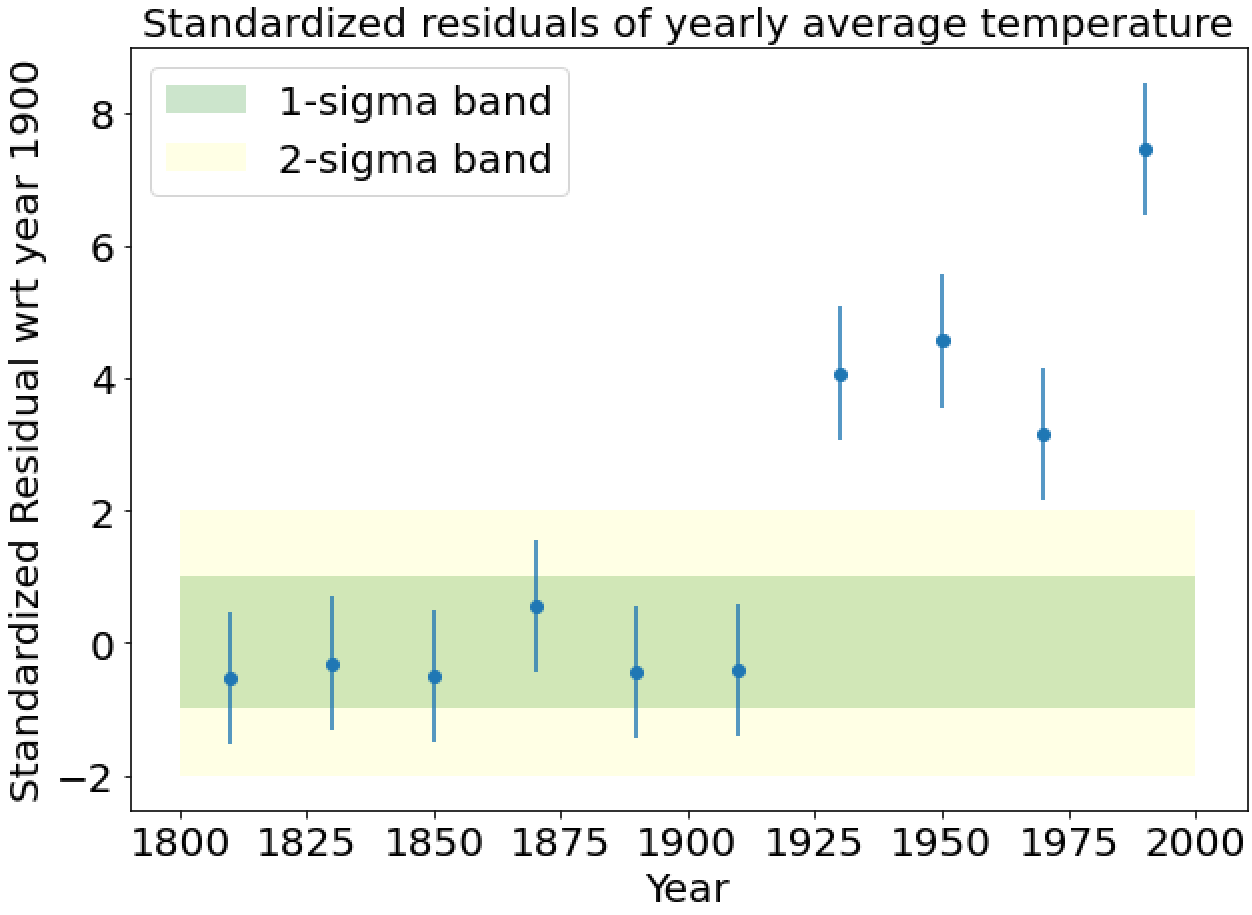
Average temperature evolution

The next plot shows the distribution of the average temperature (over the whole year) as a function of time, grouped every 20 years. It is clear that there is pertubation in the trend starting from about 1925 (is this because of WW2?).



Standardized residual with 1900 in Rome

Altough is clear that there is something going on after 1925 from the above plot, it is helpull to show it also in terms of the standardized residual with respect to a point in time. My chosen **t0** is the year range between 1900-1920 in Rome. This is the base to compute all the quantities relative to the hypothesis H0.



Hypothesis testing

There are many hypothesis one could test, as an exercise I can mention 3:

- H0: Rome is hotter than Paris
- H0: Cities on the northern emisphere are hotter than the ones in the sothern hemisphere
- H0: Average temperature does not change as a funtion of the time.

I chose to perform hypothesis test on the last hypothesis, the one on global warming, does the average temperature change with time? Or in a more quantitative way: is the sample of average temperature in time range **t1** [1993-2013] extracted from the same distribution as the one in time range **t0** [1900-1920].

To do this one needs to observe that the average temperature per month over a set of years is distributed as a gaussian with standard deviation of about 1 C. If H0 is true then the difference in average temp. per month between **t0** and **t1** is again distributed as a Gaussian and centered in zero. If one then takes the average of such differences, yhis is also gaussian distributed around zero and with standard deviation that can be calculated via error propagation. Normalizing the average of temperature **t1 - t0** by the standard deviation one obtain a test statitic that is ditributed under H0 normally with mean zero and std 1.

Once this is done I set a significance level alpha equal to 0.05 and copute the p-value for H0 as described above. I obtain:

```
The yearly Average H1 temperature is 13.0 +- 0.1 °C
The yearly Average H0 temperature is 11.7 +- 0.1 °C
```

```
the p-value for H0 is 5.44E-31
```

The p-value is extremely low, **global warming is real**.

Next steps

Apart for the missing data, the data quality of this dataset is quite high. However it lacks depth, meaning that would be interesting to have more informations and not only the temperature readings. Because of this I would suggest to join this dataset with others as a funtion of the year, for example with the population per city or CO2 emission per country as funtion of time. So that one can evaluate a model of the evolution of temperature change and see how this correlates with human CO2 emission.