Report on initial Exploratory Data Analysis on a weather dataset

by Giuseppe Insana, December 2021

The dataset

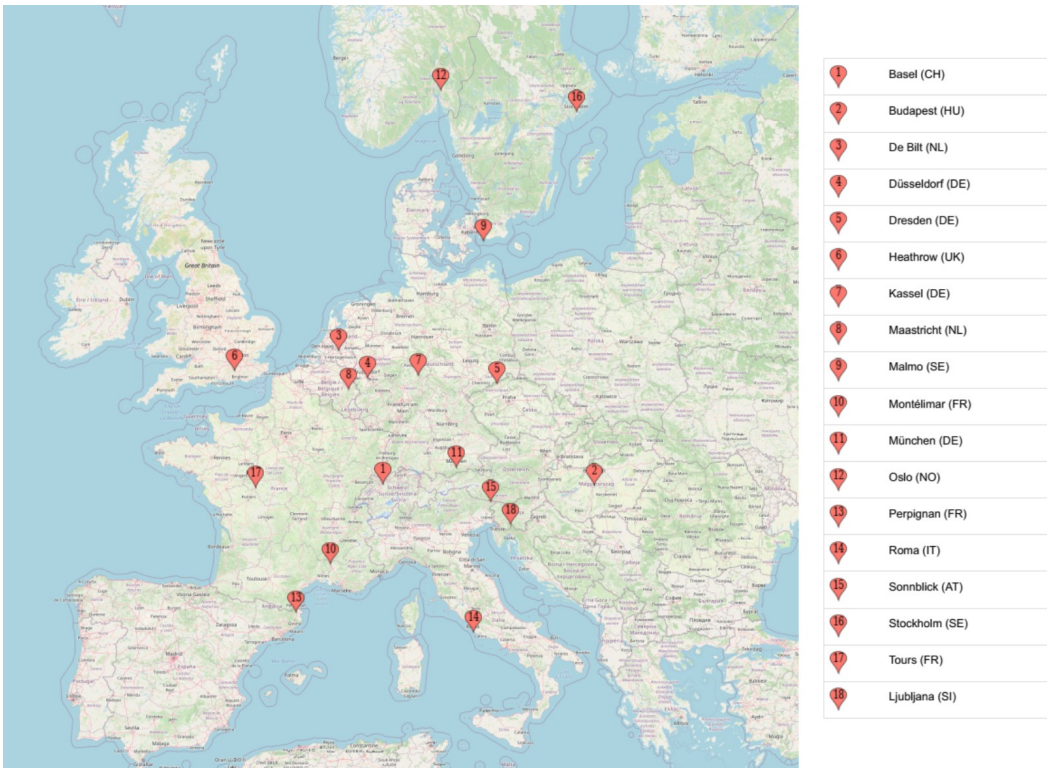
I chose to work on weather data, chiefly because I never worked on this kind of data before (having mostly dealt with biological and linguistic data in the past).

The weather data analysed in this report has been created by Huber Florian from ECA&D data (see section on Data origin at the end for references).

It contains daily weather observations from 18 different European weather stations through the years 2000 to 2010.

The description of the data set says that the minimal set of variables 'mean temperature', 'max temperature' and 'min temperature' are available for all locations. An additional number of measured variables ('cloud_cover', 'wind_speed', 'wind_gust', 'humidity', 'pressure', 'global_radiation', 'precipitation', 'sunshine') are provided, but not for all the locations.

The following map shows the locations which are included in the dataset:



Plan for data exploration

Firstly we will analyse the type and the amount of the available data. In particular, we will check what is present for each location, since the dataset description warns that not all locations include all type of weather measures. We will also check the range and distribution of the observations.

We will then proceed to check whether there is missing or obviously wrong data and talk about how we could deal with them (removing, imputing or masking).

Afterwards we'll explore the dataset visually by way of plots, in order to

- see the actual distributions of the values
- to check for presence of outliers (and discuss how they should be treated) and
- to identify possible correlations between the features.

This stage should give rise to findings and insights and lead us to formulate some hypotheses on the data. The hypotheses should be verified for statistical significance.

We will conclude by suggesting the possible next steps to take for analysing the data, discussing the quality of the data and what other additional informations could be needed.

Checking data attributes

The original data is loaded into a pandas dataframe which has 3654 rows (one per each day) and 165 columns:

- DATE, with integer values from 20000101 to 20100101, corresponding to interval from Jan 1st 2000 to Jan 1st 2010
- MONTH, integer 1 to 12
- and another 163 columns for the weather measurements at the different locations.

The measurements are labelled as LOCATION_*measure* (e.g. BASEL_cloud_cover, BASEL_pressure, OSLO_precipitation...) with the measured variables being: **cloud_cover, global_radiation, humidity, precipitation, pressure, sunshine, temp_max, temp_mean, temp_min, wind_gust, wind_speed**

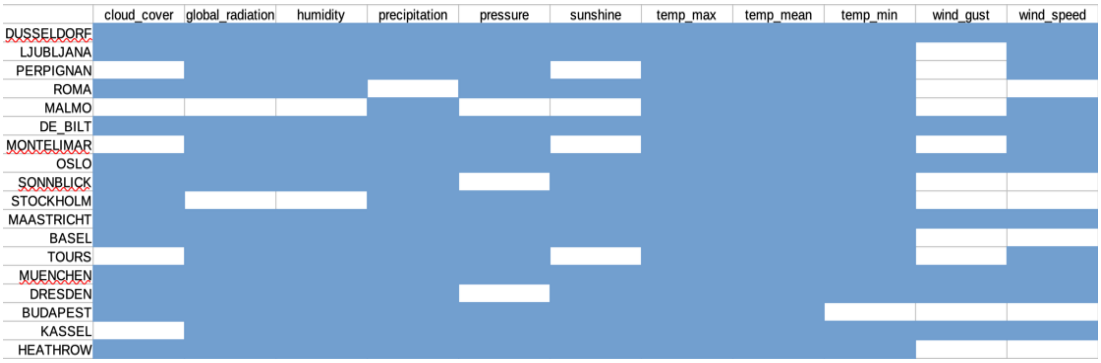
All the measurements are loaded as floating point numbers, with the exception of cloud_cover, loaded as integer.

The **physical units** for the variables are described as follows:

> **cloud_cover** in oktas; **wind_speed** and **wind_gust** in m/s; **humidity** in fraction of 100%; **pressure** in 1000 hPa, **global_radiation** in 100 W/m^2; **precipitation** in 10 mm; **sunshine** in 1 Hours; **mean max** and **min temperature** in Celsius degrees.

As mentioned in the description of the dataset, not all variables are available for all locations. In particular we found that, contrary to what affirmed in the original description, even the temp_min variable is not available for all locations (it's missing for the BUDAPEST location).

The following table shows which measures are available for which location:



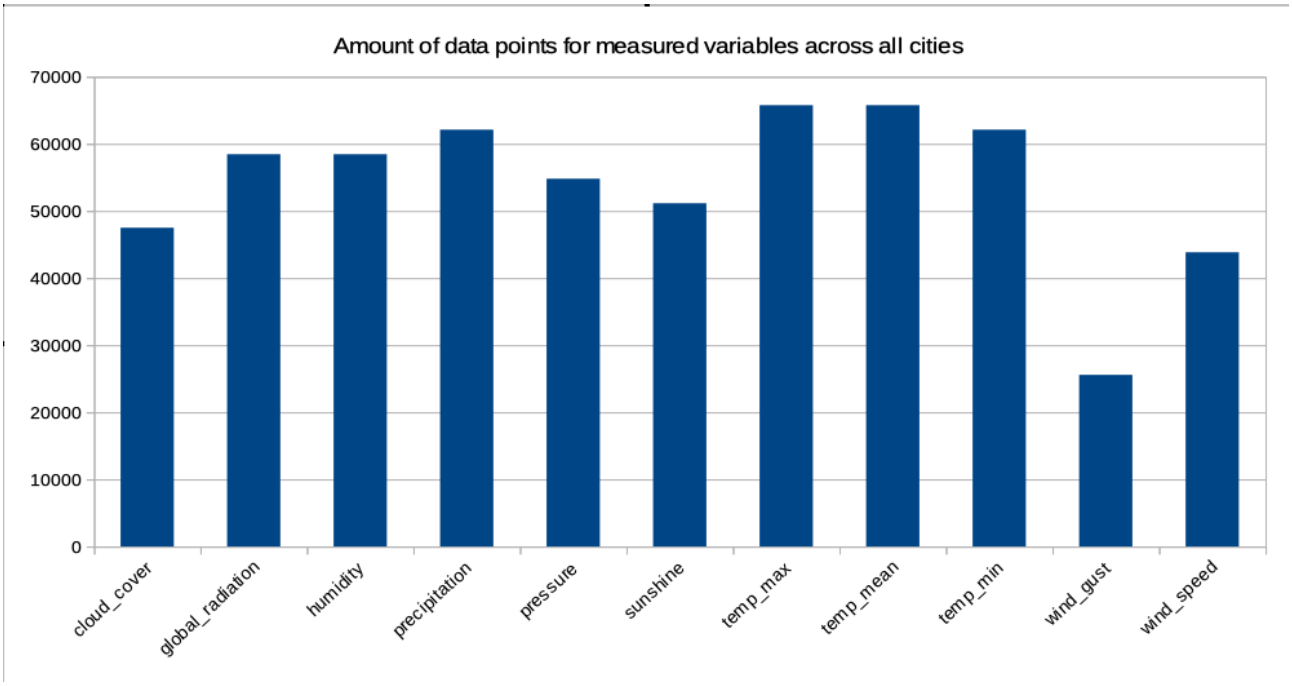
This could obviously present problems of **unbalanced data**.

Null values

The dataset does not contain missing values per se, but as just mentioned, there are many variables entirely missing for certain locations. Also we'll shortly examine out-of-range values that can be considered Null/Missing/Invalid observations.

Amount of data per variable

Stacking the same measured variable across all locations we can see the amount of data points for each variable, highlighting the unbalance in the data features.



Ranges for the variables (across all locations)

cloud_cover measures should vary from 0 (sky completely clear) to 8 (sky completely clouded) oktas. Actually the data contains two data points (both for Stockholm, 20080724 and 20090625) with a value of -99 (indicating invalid entries) and one (again Stockholm, 20031108) with a value of 9.

> These can be considered missing values and a decision will need to be taken whether removing the whole row (which implies losing valid data for the other locations), imputing them (e.g. to the average for that month in that location) or masking them (setting them all to category 9 to treat them as "no observation recorded"). Average: 5.14, Standard Deviation 2.33

global_radiation in the dataset varies from 0.01 to 4.42 (i.e. 1 to 442 W/m^2), with average 1.37 and standard deviation 0.95

humidity ranges from 0.1 to 1 (i.e. from 1% to 100%), with average (from now on ***avg***) 0.75 and standard deviation (from now on ***std***) 0.14

precipitation: 0 to 16.04 (i.e. 0 to 160.4 mm), avg: 0.23, std 0.58

pressure: the data contains three entries (again for Stockholm, 20071008, 20000124, 20070603) with value of -0.099 and one entry (Tours, 20081230) with value of 0.0003. These out of range values can again be considered invalid and a decision should be taken for them akin to those mentioned

before for cloud_cover. A part from those values, pressure ranges from 0.959 to 1.016 (i.e. 959 to 1016 hPa) with avg 1.016 and std 0.013

sunshine: the data contains 29 negative values for hours of sunshine (again for the Stockholm location) which should be treated as invalid/null and dealt appropriately (as mentioned for cloud_cover and pressure). For the location of Oslo there are 24 measurements with more than 18 hours of sunshine, with 20 of them being 24h. While the northern latitude make very long daylight possible, this is for almost 18 hours in midsummer, while these huge values are from Nov-Dec 2006. We must hence treat these as wrong invalid data as well. A part from those extremes, sunshine varies from 0 to 17.8 hours, with avg 5.00 and std 4.41

temp_min: -30.3 to 26.3 (°C), avg 6.33, std 7.58

temp_max: -24.7 to 41.1 (°C), avg 14.50, std 9.58

temp_mean: -26.60 to 33.1 (°C), avg 10.39, std 8.41

wind_gust: 1.50 to 41 (m/s), avg 10.06, std 3.88 (the fastest wind speed recorded, not related to tornadoes is 113.3 m/s, so these values appear plausible)

wind_speed: from 0 to 16.30 (m/s), avg 3.33, std 1.89

Data exploration

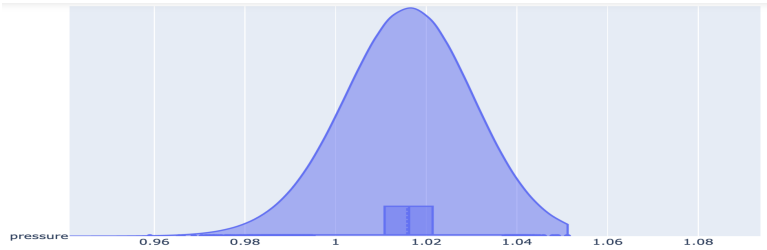
To get a better perspective on the data, which goes beyond simply calculating ranges, averages and variance, it is always a good idea to examine the dataset visually by way of plots, to see the actual distribution of values and to gather insights.

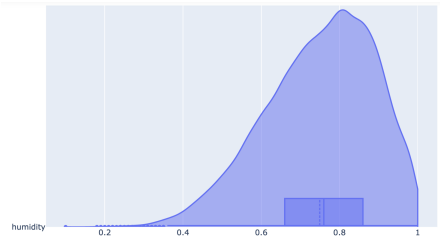
We can start by plotting distribution of the variables, at different levels

Distribution of variables across all locations

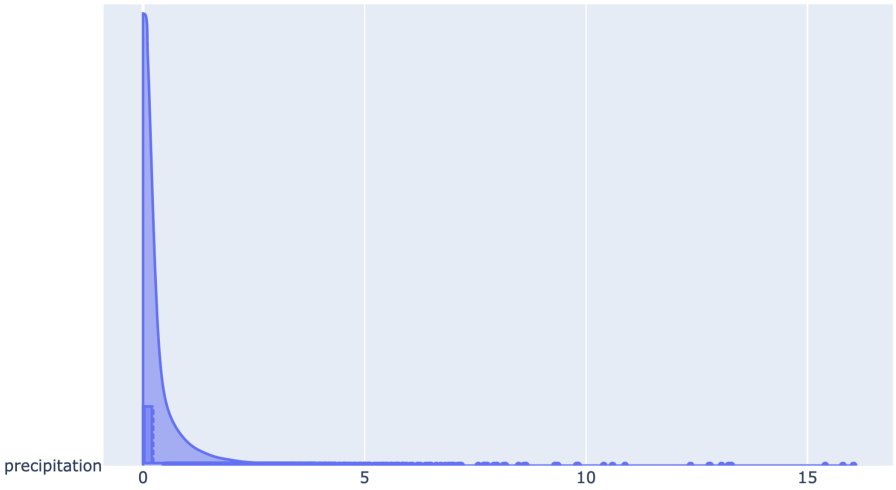
Plotting histograms, boxplots or, even better, violin plots is a good way to get a sense of the variance for each feature, beyond the aggregate values. It also enables to quickly identify **outliers**, to notice **skewing** in the data or **multimodality**.

For example while pressure appears normally distributed, humidity is skewed:

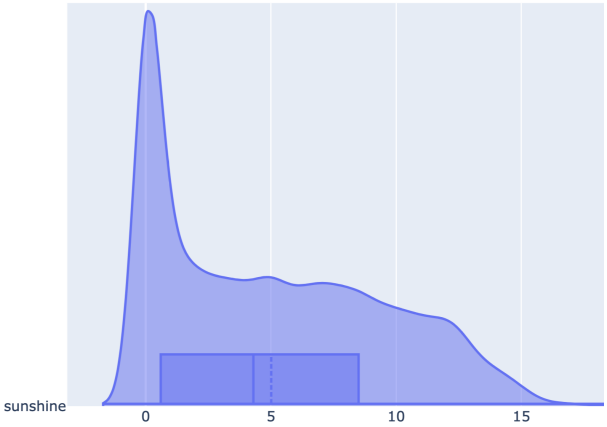


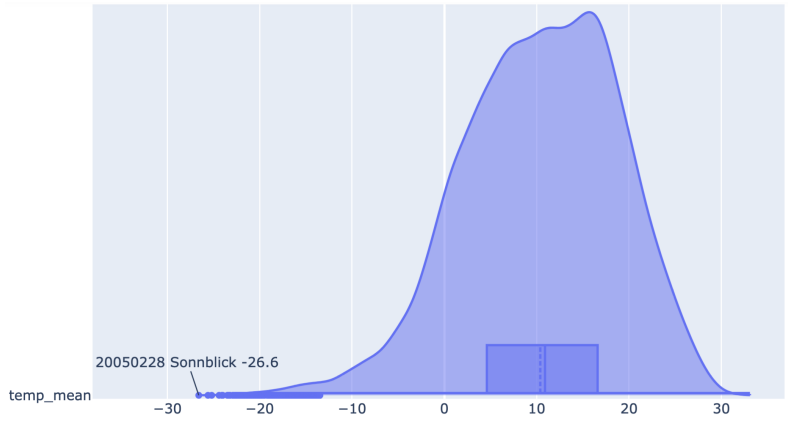


Precipitation levels are usually low but with a long tail and many outliers:



Some features appear to have multimodal distributions, like sunshine and temperatures

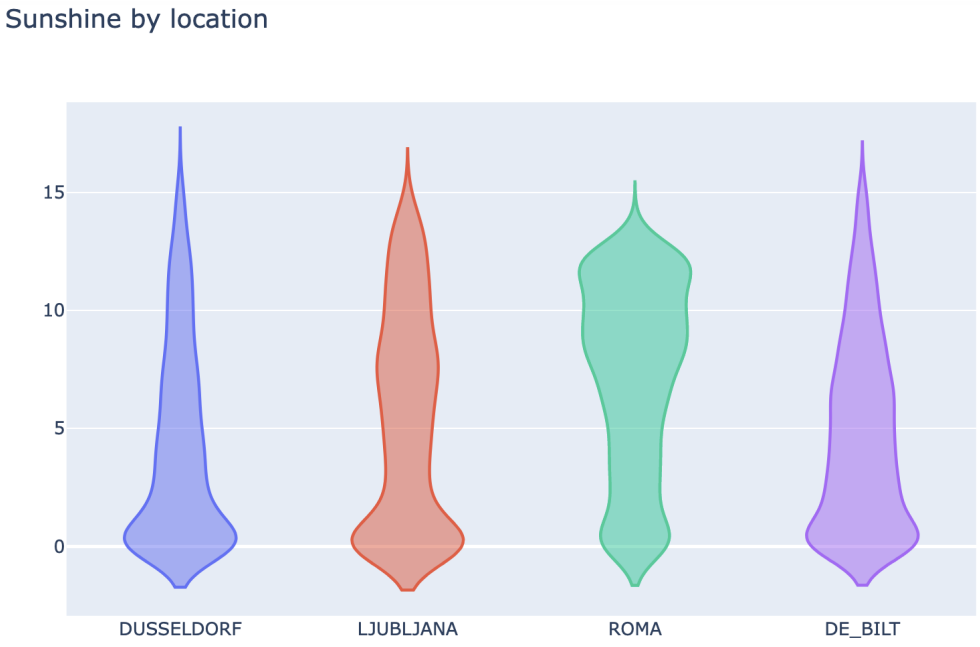




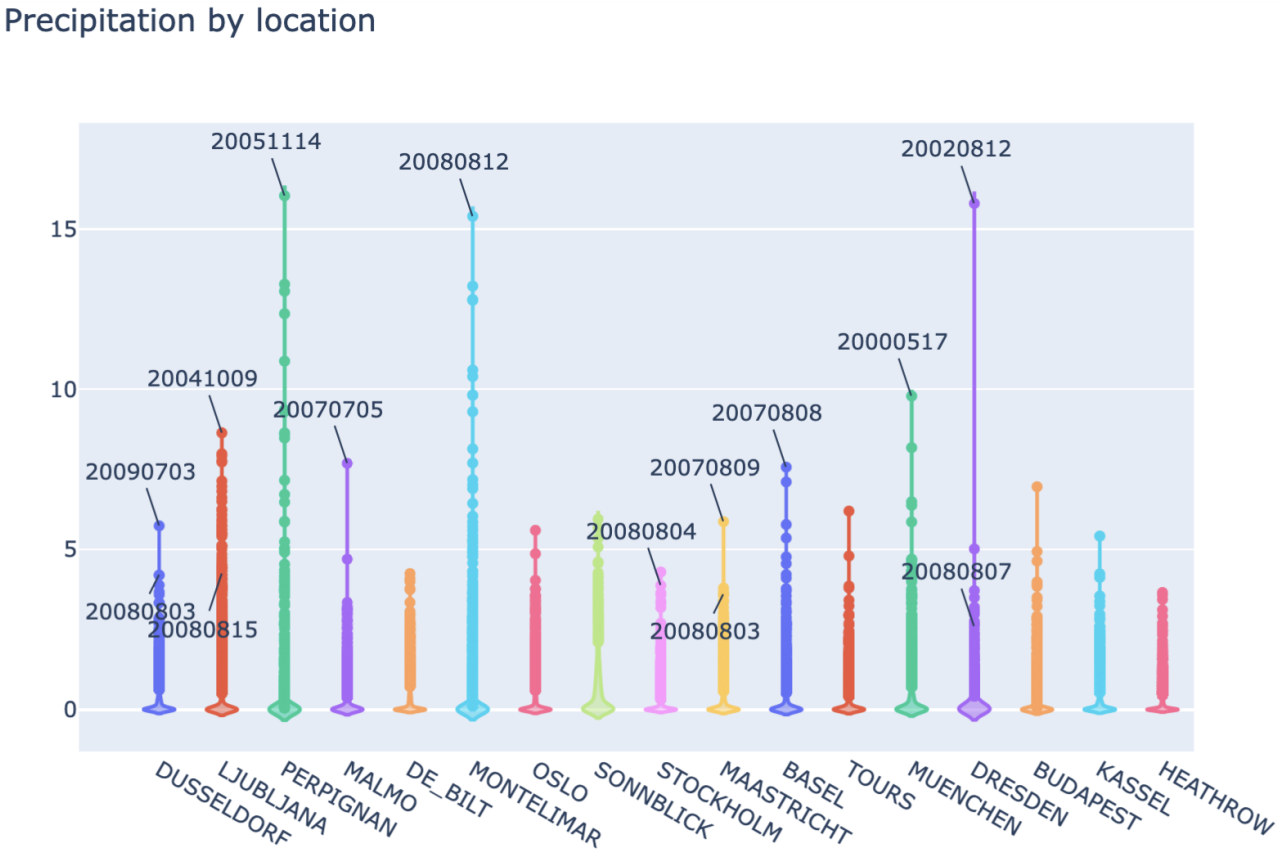
We can try to get a better understanding on the data extending the analysis in the distribution of features by **grouping the data either spatially or temporally**.

Distribution of variables by locations

For example the distribution of sunshine hours is quite different for the locations, as in the those for Rome and Dusseldorf in the following plot:



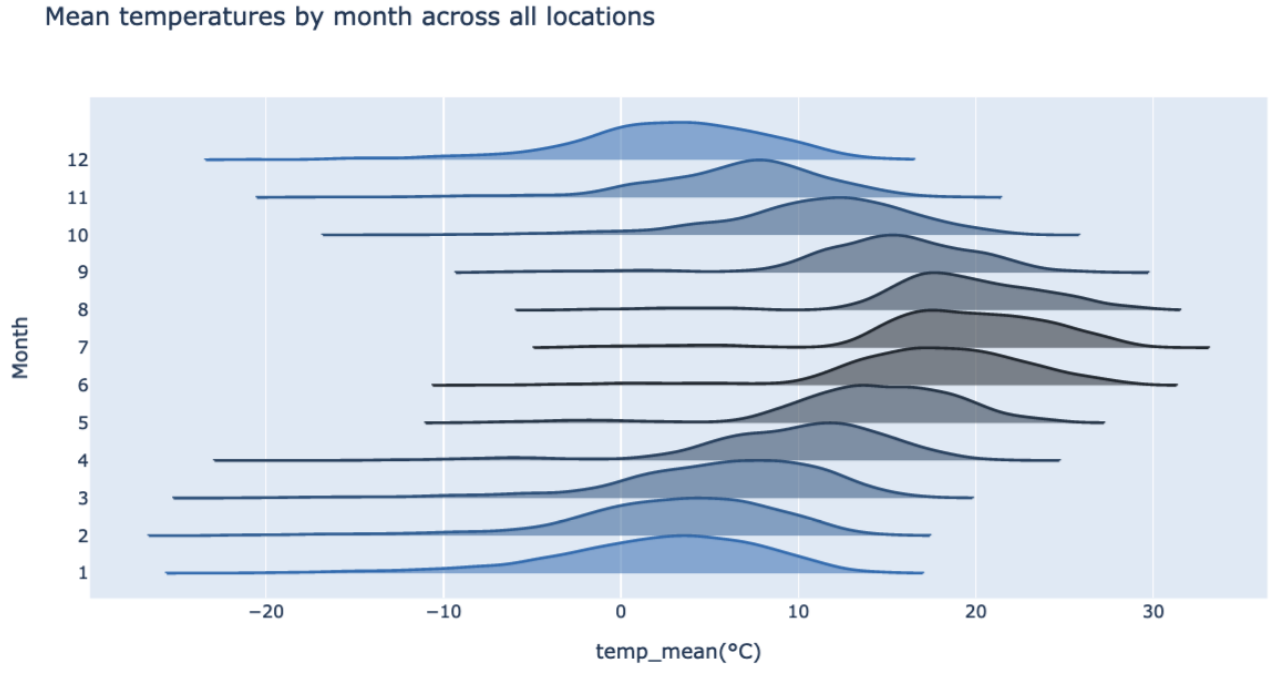
Interactively inspecting the outliers (this is facilitated by a plotting library like plotly) and labelling by date one would maybe notice a particular cluster of outliers: for several locations there are exceptional precipitation amounts for dates of August 2008 as marked in the following plot.



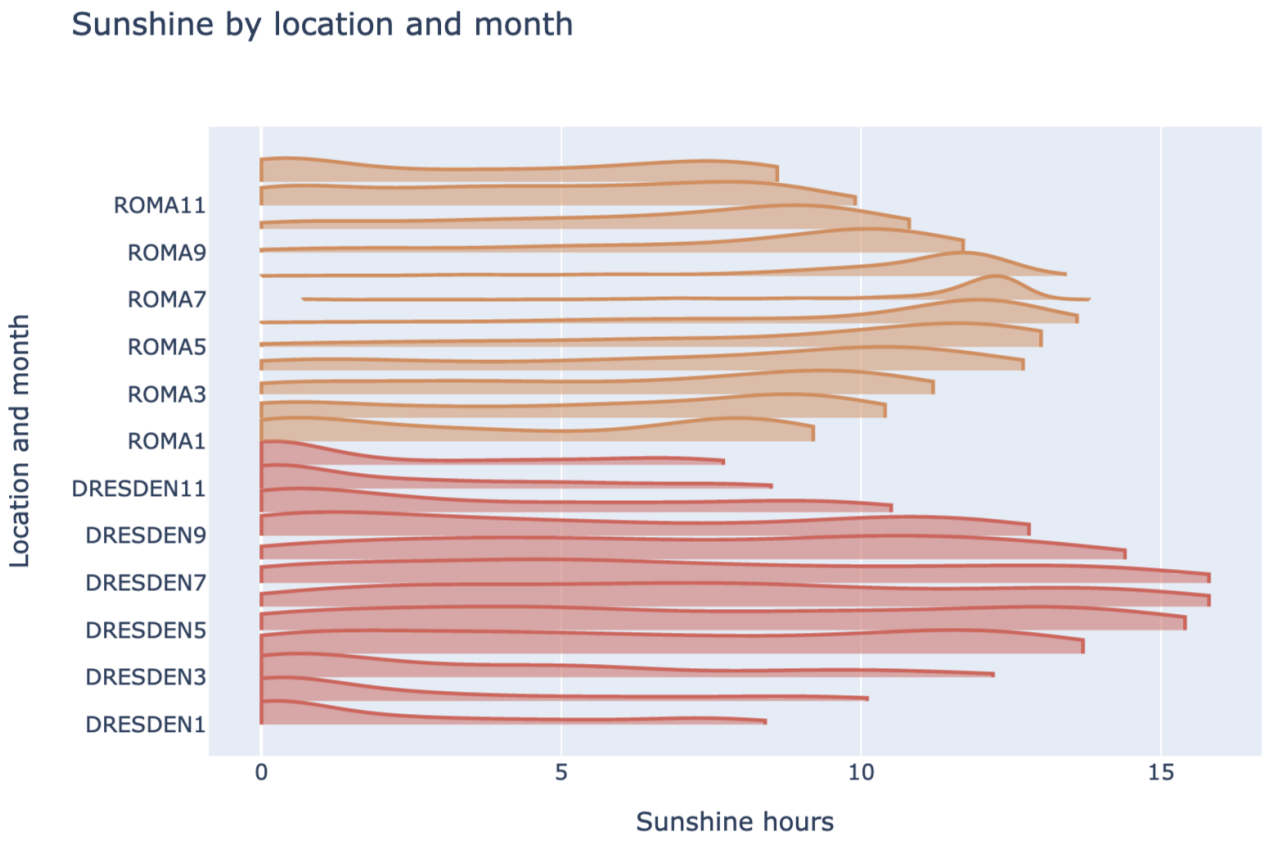
And in fact the summer of 2008 was particularly wet also in Europe.

Distribution of variables by month

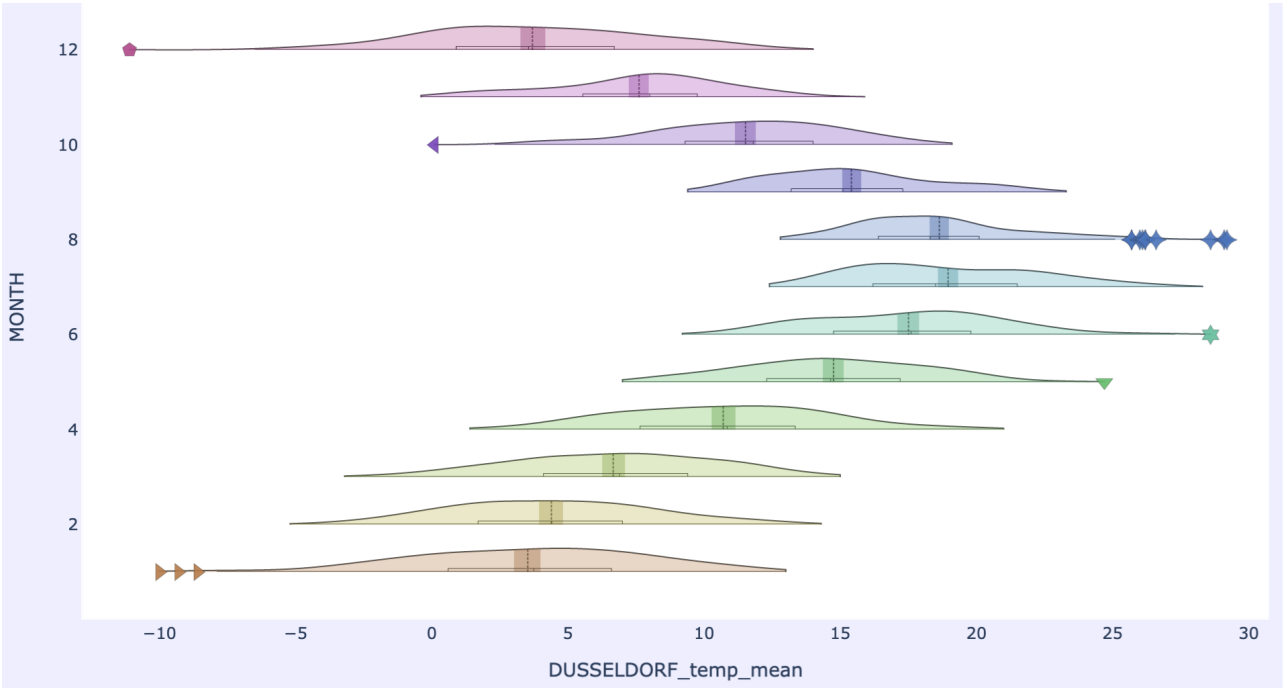
To observe seasonal variation it can be useful to group data by month, for example:



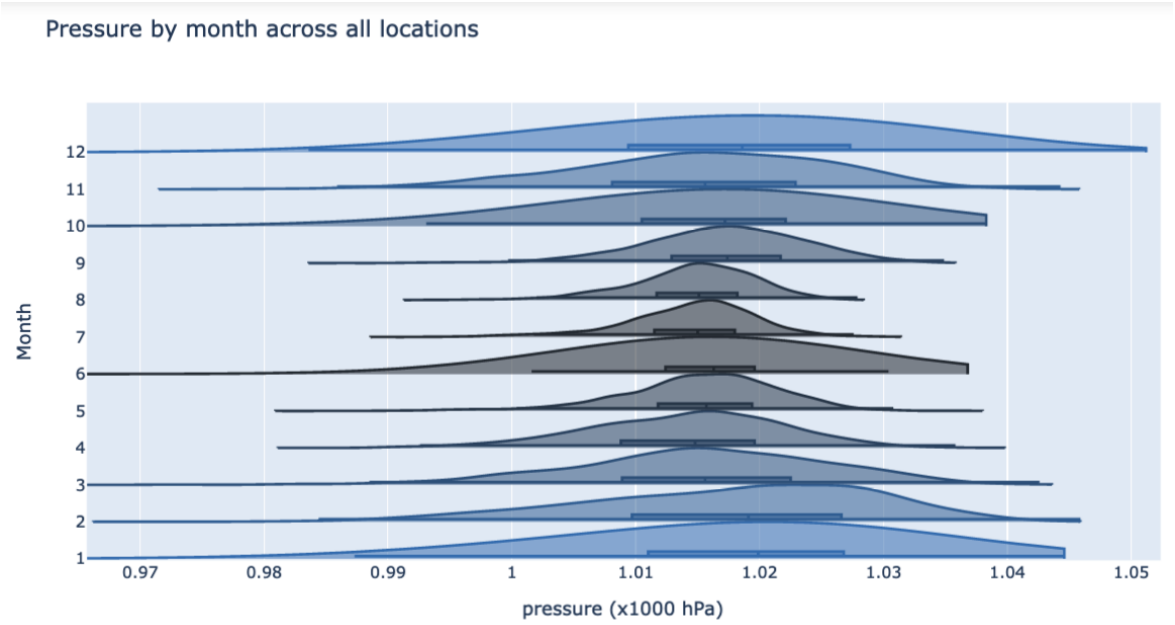
Which can be made even more informative when grouping also by location, either comparing locations, as in this plot:



or focusing on each individual location:

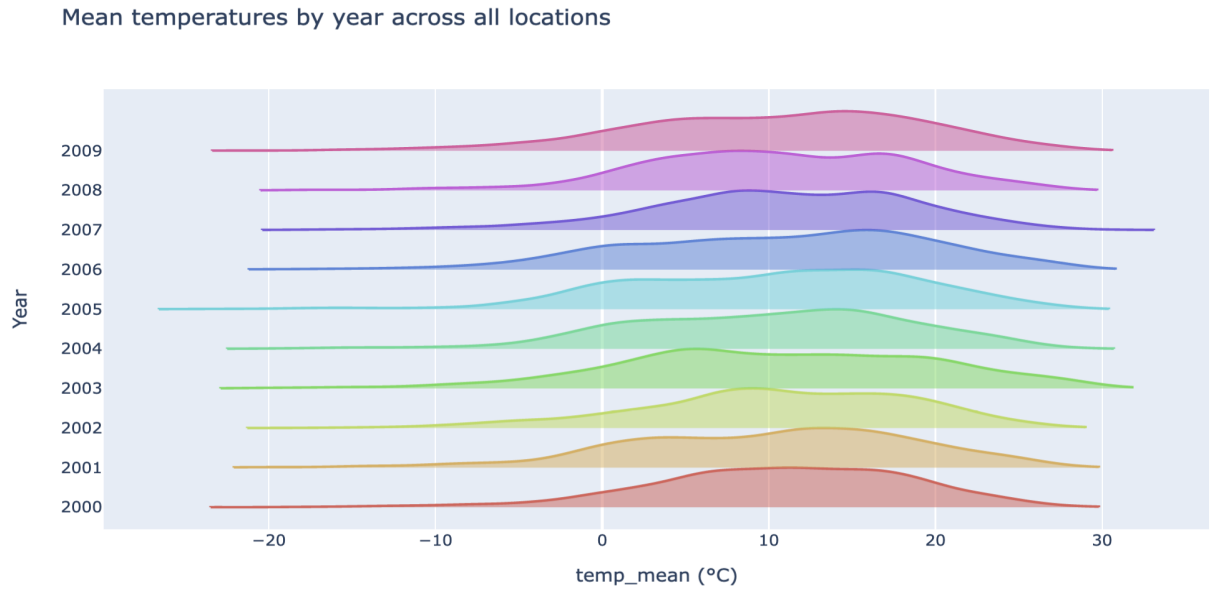


Grouping by month revealed differential distributions of the pressure feature, with smaller variance in summer months compared to winter months:

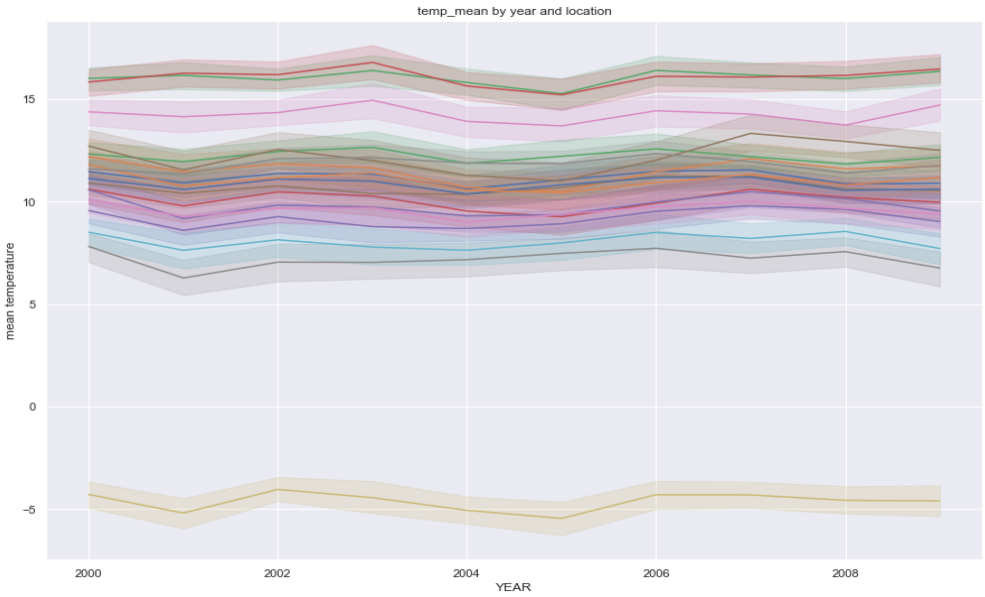


Distribution of variables by year

To explore the distributions of a measured variable for a whole year, we can group the data by year:



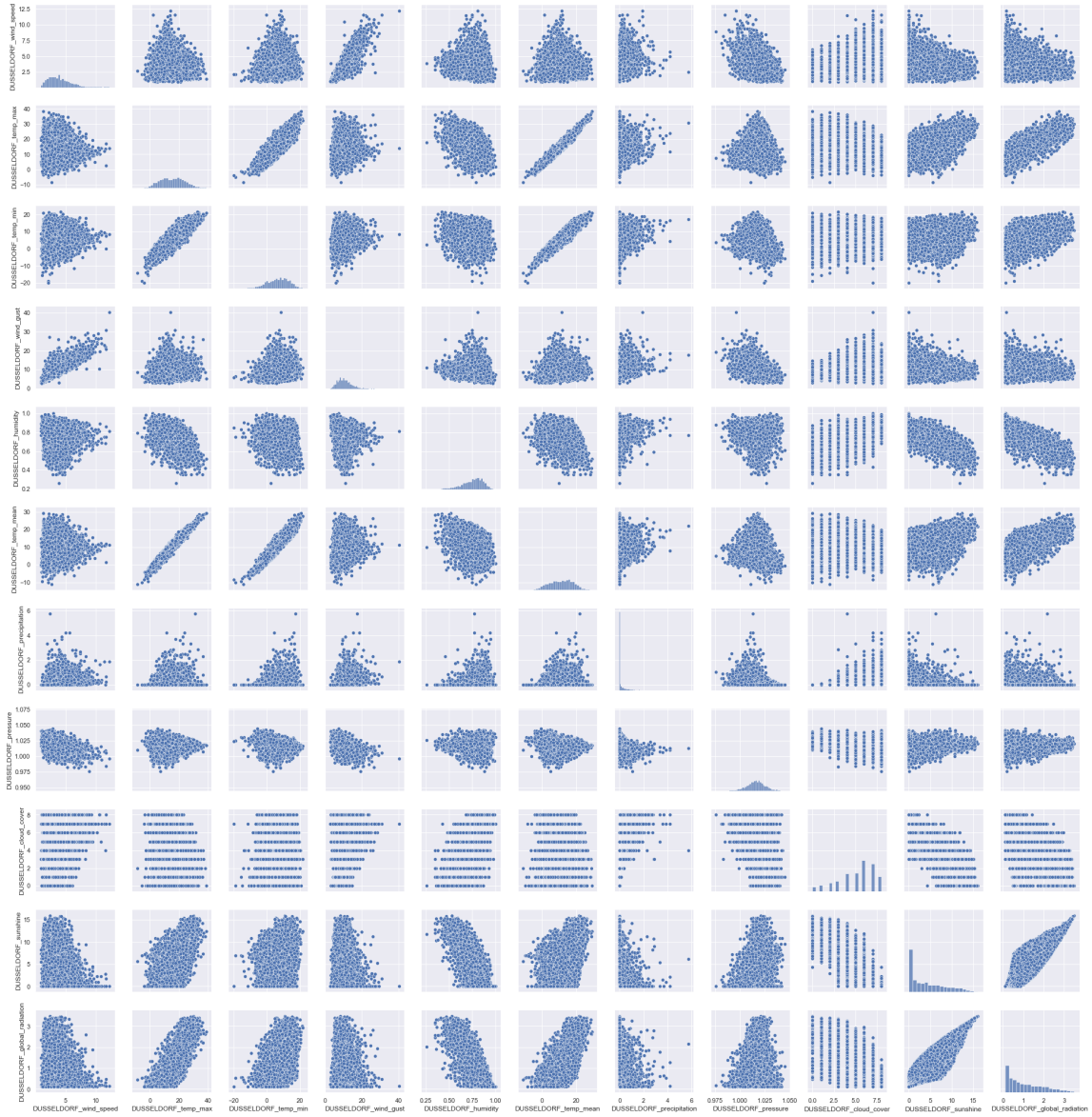
and furthermore also by location, as summarised by the following plot:



this plot also reveals how a certain location (Sonnblick) has temperature ranges very far away from the rest. In fact Sonnblick weather station is located in the Austrian Central Alps at an elevation of 3106 m.

Pairing variables

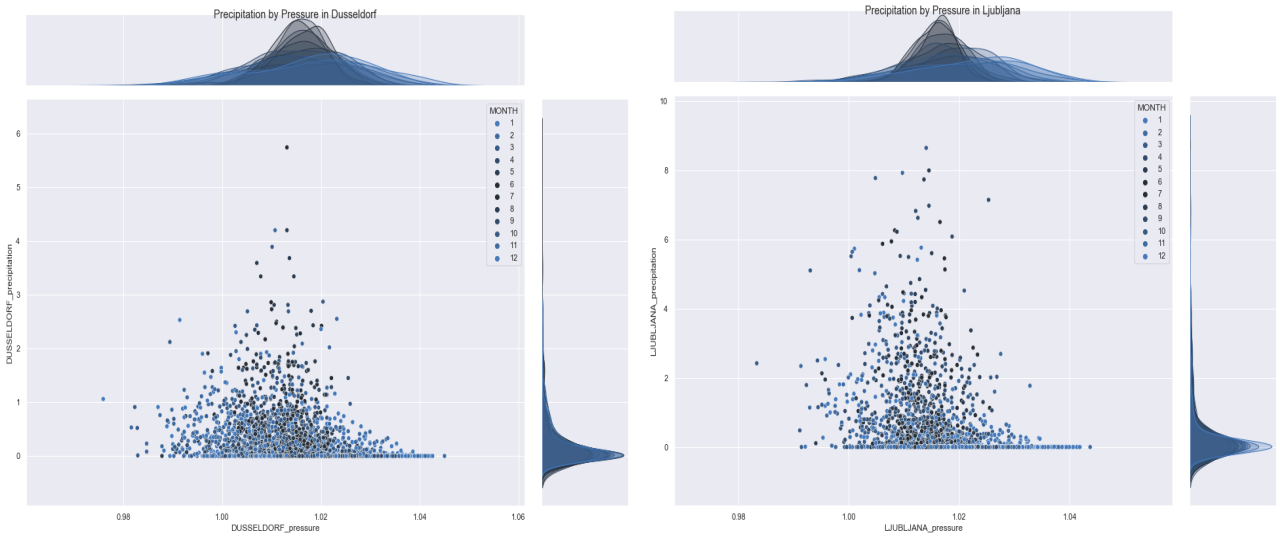
Finally, we can observe the **correlation between features**, either for a single location or across different locations. We can use pairplots to observe at a glance the pairwise relationships between features. Here's the one for Dusseldorf:



Several insights can be gathered from the above pairplot (and from those for other locations):

- related variables have (as expected) very high correlation:
  - temp_mean with temp_min and temp_max
  - wind speed and wind gust
- global radiation has high positive correlation with sunshine
- global radiation and sunshine correlate negatively with humidity
- precipitation appears to have extreme values concentrated where pressure has average values

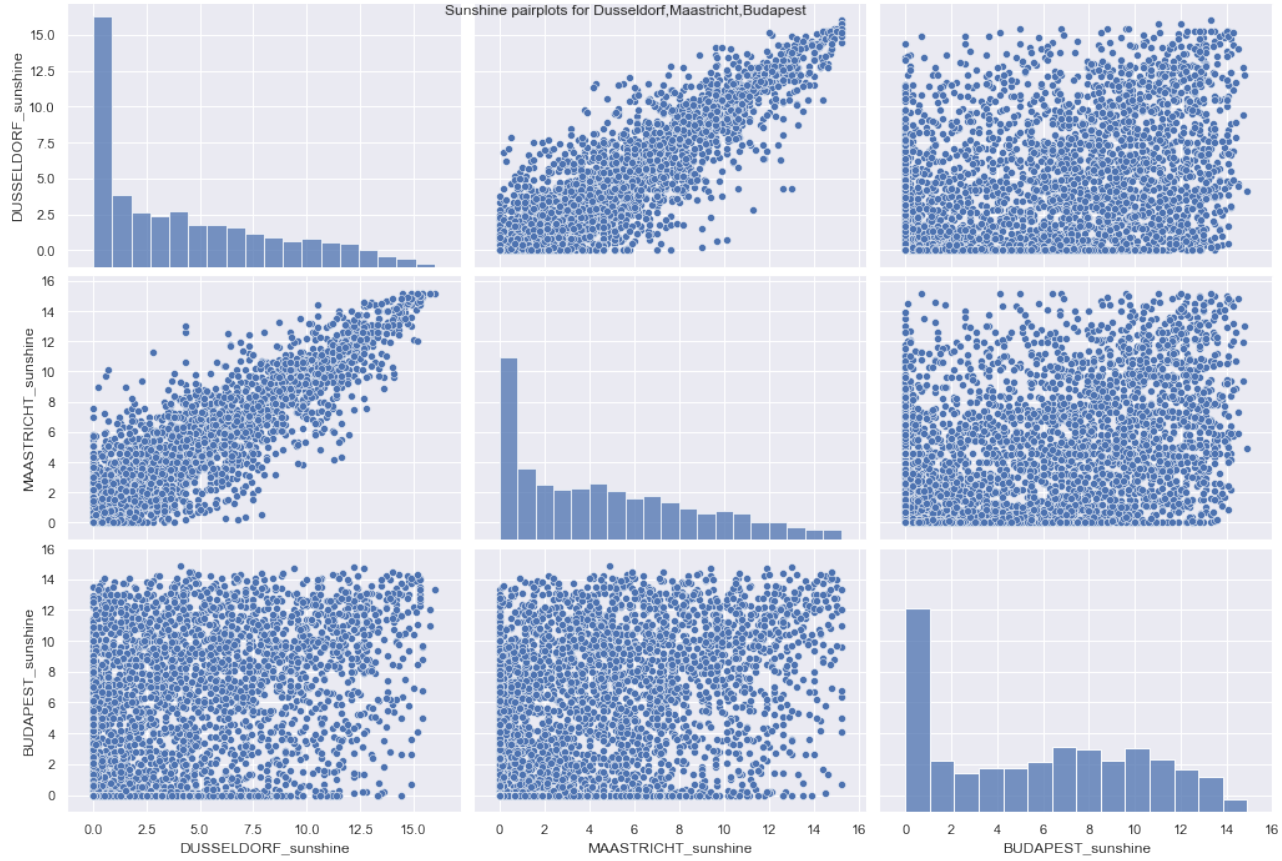
To investigate further this last insight, we can do joint plots for each location. For example:



Notice how the differential variance of pressure mentioned above is visible in the kernel density distribution by month (at the top of the above two plots, summer months in dark, winter months in blue).

Pairplots can also be used to explore the **correlation between weather measurements at two different locations**.

For example we'd expect sunshine for two nearby locations (like Dusseldorf and Maastricth) to be highly correlated compared to sunshine for two locations which are geographically distant (as Budapest to the other two):



Data cleaning and feature engineering

The first steps for cleaning this dataset are thus to deal with the **out-of-range values** identified for the cloud_cover, pressure and sunshine measures.

The number of these invalid values appear in 1.64% of the total rows but as they only affect one location at a time they constitute only 0.01% of the total values. We then **recommend to not drop** the whole days' measurements but **instead to impute** the invalid values using the average of a small window (e.g. 11 days) around the invalid value.

Secondly we need to decide how to deal with outliers. And this would depend on the influence they may have for machine learning algorithms that will be applied to the dataset.

The **main outliers to consider** would be:

- the very low values for humidity (lower than 30%), where all outliers are from Sonnblick and (in much lower proportion) from Perpignan
- the precipitation extremes
- the highest recorded wind speed measures (which are almost all from Perpignan location)

Once more, this is a choice which depends on what we would like to do with the model. If we are interested in being able to **predict weather extremes** (for example chance of floods) then these outliers would actually be the most interesting data points and should not be removed or imputed.

The original **dataset does not contain categorical data** which would need to be converted (e.g. by one-hot encoding). Nevertheless, if we wanted to use this data for **classification** (e.g. predicting the location given the weather measurements), we could reorganise the data so that each row would be the measurement for a single location and there would be a new column containing the categorical information (the location).

A simple but useful feature engineering we performed since the beginning was the addition of a YEAR feature, which allowed us to observe year-to-year variations.

Depending on the machine learning algorithm that we are going to use, we could need to **transform skewed variables**. For example the distributions of precipitation and sunshine hours are heavily skewed towards 0 with a tail towards higher values.

Similarly, depending on the machine learning algorithm we intend to use, it could be very important to do **feature scaling**: for example we can **standardize** the distributions, subtracting the mean by each value and dividing by the standard deviation.

Finally we could add new features like **interaction terms**. For example creating an interaction feature as product of precipitation and pressure.

The **biggest decision** to take would probably be whether we want to keep all the features and all the locations or whether we'd prefer to have a balanced dataset which holds only a subset of features and a subset of locations so that we would have the same features for each location. This again depends on what we'd want to do with the dataset and on the ML technique adopted.

For example for the prediction of a certain variable (e.g. precipitation) at a certain location, knowledge of the weather variables at other locations, even if incomplete (some locations missing some features) could still be important, as long as the model will be able to deal with it.

On this note, as the weather measures (*in primis* the temperature ranges) of Sonnblick are very different from all other locations, we should make a note to check the effect of including or excluding the features for that location on the accuracy of predictions in a regression model.

Key findings and insights

We summarise here what has been found so far:

- the dataset is very unbalanced in which features are available for which location. Some features are under-represented as they appear only for few locations (like wind data);
- while there are no null values per se, there is a small fraction of out-of-range values which should be considered as null values;
- in terms of outliers, the more worrying ones are found for the humidity, precipitation and wind_speed features;
- global_radiation, sunshine and especially precipitation are positively skewed (tail to the right and majority of values being 0);
- there is a strong seasonal component for all features, with either mean or standard deviation highly correlating with the time of the year (as exemplified by the month grouping);
- the seasonal component appears to be responsible for multimodality in certain features;
- there are also year to year variations, with for example some years being on average warmer or colder; not overall trend was observed but this is probably due to the scale of the dataset (only ten years period);
- some features show strong positive correlation and some show a negative correlation (like global_radiation and humidity); some correlations are less clear but worth investigating (like precipitation extremes and pressure);
- weather measures at locations close by appear to be highly correlated;
- some outliers show a relationship across locations, implying they were part of some extreme weather condition which affected major areas at almost the same time;

Hypotheses

We can formulate a series of hypotheses, for example:

- that seasonal variation exists;
- whether a feature follows a normal distribution;
- whether two features are correlated;
- whether a global trend from year to year can be inferred from our data (e.g. global warming);
- whether a positive correlation exists between weather at one location compared to weather at another location for same day (or shifted a certain number of days).

We will now discuss how to perform a significance test on the last of these hypotheses: the correlation between pressure measure at two locations which are spatially close.

Significance test

The correlation coefficient r is a measure about the strength and direction of the linear relationship between two variables. Correlations of -1 or +1 imply an exact linear relationship, the strongest possible agreement between two variables, (positive correlation imply that as x increases, so does y; negative

The reliability of the linear relationship depends also on how many data points are in the sample. We will hence perform a test on the significance of the correlation coefficient to decide whether the linear relationship observed in our sample is good enough to be used for modelling a relationship even outside of the sample data. In other words how well $\mathbf{r}$ (the correlation coefficient for the *data sample*) models $\boldsymbol{\varrho}$ (the unknown correlation coefficient for the whole *population*).

The test will let us decide whether the correlation coefficient is close to zero or significantly different from zero (values of ±1 indicate the strongest possible agreement and 0 the strongest possible disagreement).

We will hence draw one of two possible conclusions:

- There is sufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is significantly different from zero
- There is insufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is not significantly different from zero

Pearson's correlation coefficient applied to a data sample, represented by $\boldsymbol{r}$, can be computed for two paired data variables x and y as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is the sample size, x_i and y_i are the individual samples for the two variables and $\bar{x}$ and $\bar{y}$ are the sample means for the two variables.

First we set the null hypothesis H_0: $\varrho = 0$ (the population correlation coefficient IS NOT significantly different from zero);

and the alternative hypothesis H_a: $\varrho \neq 0$ (the population correlation coefficient IS significantly different from zero).

We choose a significance level α of 0.01 as the cutoff level to reject the null hypothesis.

We then carry out the test and obtain the test statistic and the corresponding P-value which allows us to interpret the test.

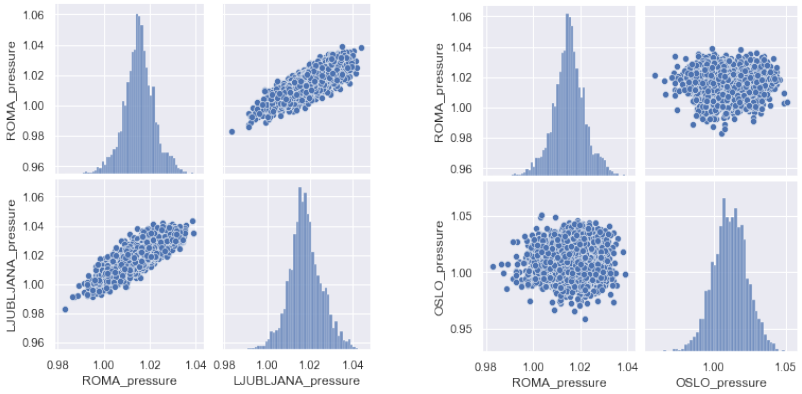
For the values of pressure at the two locations of Roma and Ljubljana we get a Pearson correlation coefficient of 0.818, with a p-value of 0.000.

The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets.

As this is lower than α, we reject the null hypothesis (we see the null hypothesis extremely improbable) and we can draw the conclusion that there is sufficient evidence that there exists a significant linear relationship between the two sets of values.

We could also test the opposite, i.e. that there should not be a significant correlation for locations far away. For Roma and Oslo, the Pearson correlation coefficient is 0.043, with a p-value of 0.009.

Pairplots:



Note: to check the effect of the sample size on this statistic test, we can randomly sample only one year (i.e. 1/10th) of the total data and repeat the test. While the correlation between Oslo's pressure and Roma's pressure is still extremely low (0.072), now the p-value returned is 0.172, much higher than the significance level cutoff chosen, making this not statistically significant.

Next steps

Depending on the objective of the machine learning study that we'd want to use the dataset for, we would need to appropriately remove or impute the outliers, perform feature scaling, choose the features to keep and engineer appropriate feature combinations.

For example, if the objective is weather prediction of tomorrow's weather given the weather measures observed today (or the measures of a window of days before the day we intend to predict) we could extract a time shifted series of observations to be used as the dependent variable **y** to use when fitting the model.

The dataset does not contain latitude and longitude or any other geographical information for the locations. It could be useful to provide this information in order to better analyse the correlation among locations in terms of their relative distance and also correlations between observed weather and latitude (or elevation, or climate zone etc).

Since we can easily compute sunrise and sunset (and moonrise and moonset) given latitude/longitude, we could consider adding the information for each location (and each day) as additional features for ML.

If we wanted to observe global trends we'd need to get data for a much larger time scale and for more locations. In particular this dataset only contains locations in Europe.

Finally it is worth noting that although the dataset's metadata mentions **wind_direction** data and informs us that it is measured in degrees, this feature was not present in the dataset. We think it would be very important to add this variable to the data set as it could prove instrumental in weather prediction, with the hypothesis that the weather at one location could influence the weather at another location situated in the direction towards which the wind was blowing.

Conclusions

The dataset looks like a very promising one to use for machine learning with limited amount of data cleaning and optionally feature scaling to be done.

The major drawback is that only two features are available at all locations (temp_mean and temp_max), but it would be too restrictive to only consider these two. If we wanted to achieve a balanced dataset it would be better to instead drop a combination of some features and some locations in order to minimise the quantity of data left. Alternatively, we could use a model robust to these unbalances.

Although the data lacks geographical information, a powerful model with enough data should theoretically be able to extrapolate geographical and temporal patterns, for example in which season which location weather at time X is a good predictor of another location weather at time X+n.

It would be interesting to train a model that would be able to extract the topology of the locations based only on the weather data. In other words: could we reconstruct the map of the different locations simply based on the correlations between columns? A possible first step in this study would be: to train a classification model that tries to reconstruct which set of measurements comes from which location.

A related line of analysis would be: rather than focusing on predictive power per se, we could instead check which are the best predictors of weather, to obtain powerful insights: examining which features and locations are more influential in determining weather at another location should reveal not only spatial distance among the locations but presumably also presence of geographical features among and along the locations (e.g. mountain chains in between) or climate macro patterns.

Methods and Materials

Data manipulation and analysis was performed on a MacOS laptop using own written code in python language, working in a jupyter notebook and taking advantage of the following python libraries: pandas, qgrid, seaborn, plotly, cmplot, scipy

Data origin acknowledgement

Dataset compiled by Huber, Florian (2021); Zenodo. https://doi.org/10.5281/zenodo.4770937

ORIGINAL DATA TAKEN FROM:

EUROPEAN CLIMATE ASSESSMENT & DATASET (ECA&D), file created on 22-04-2021 THESE DATA CAN BE USED FREELY PROVIDED THAT THE FOLLOWING SOURCE IS ACKNOWLEDGED:

Klein Tank, A.M.G. and Coauthors, 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453. Data and metadata available at http://www.ecad.eu