Exploring Weather Trends

Outline

For this Project I used **SQL** to extract the data, in particular I used the Workspace provided from Udacity and I carried out one query. In this query I left joined the global_data table into the city_data table by using the column "year" which allowed me to put together the average temperature of all the cities of the world along with the global average temperature. Below you can find the query:

```
select

cd.year
,cd.city
,cd.country
,cd.avg_temp
,gd.avg_temp as global_avg_temp

from city_data cd

left join global_data gd
on  gd.year = cd.year
```

Next I used **Excel** and **Python** to get to know the data. In Excel I created a Pivot Table in which I filtered for the city I wanted to work with, London, and the country, United Kingdom, and I carried out an analysis based on Histograms and Trend Analysis of the two datasets.
I also used Python to try to use the decomposition of the time series in Trend, Seasonality and Residuals which I did not succeed and I would leave this for future work.

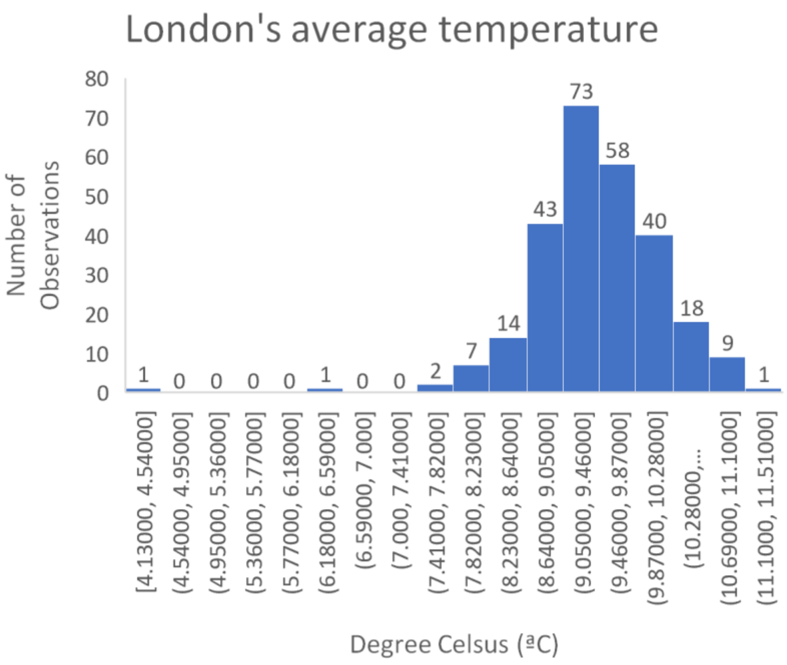
For the sections two and three, which cover the main part of this project, I used mainly **Python** but I also relied on Excel to get the linear regression formula.

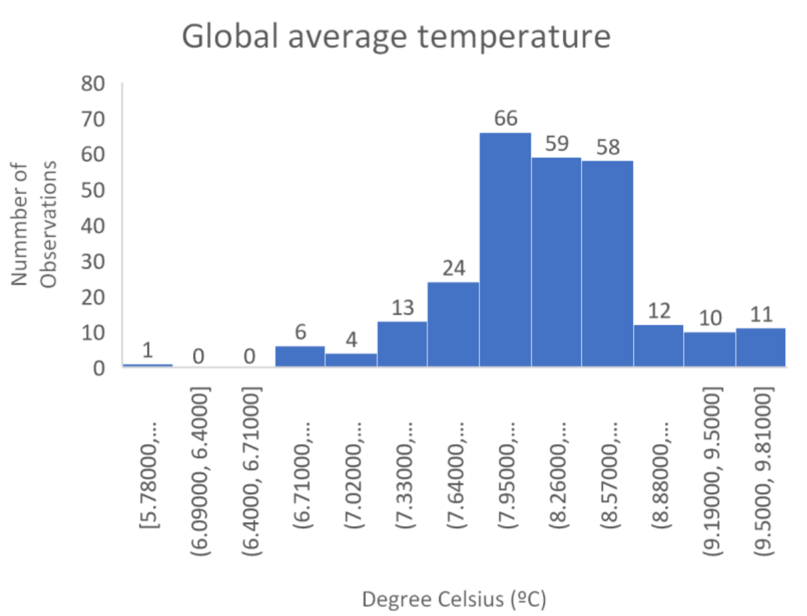
Section 1: Getting to know the data

Histograms:

An histograms gives us insights about the data distribution. Below it can be found the histograms for both London's average temperature and global average temperature. From them we learn that:

- The data for both histograms has a Gaussian distribution with a center in the interval (9,05,9,46) for the London's average temperature. For the global average temperature the center will be near (8,26,8,57) interval.
- The global average temperature is more variable than the London's average temperature
- Neither of the variables present a significant number of Outliers





Trend Analysis:

This subsection is based [1] and [2]. I will not go into detail but if more information is needed, please refer to the provided references.

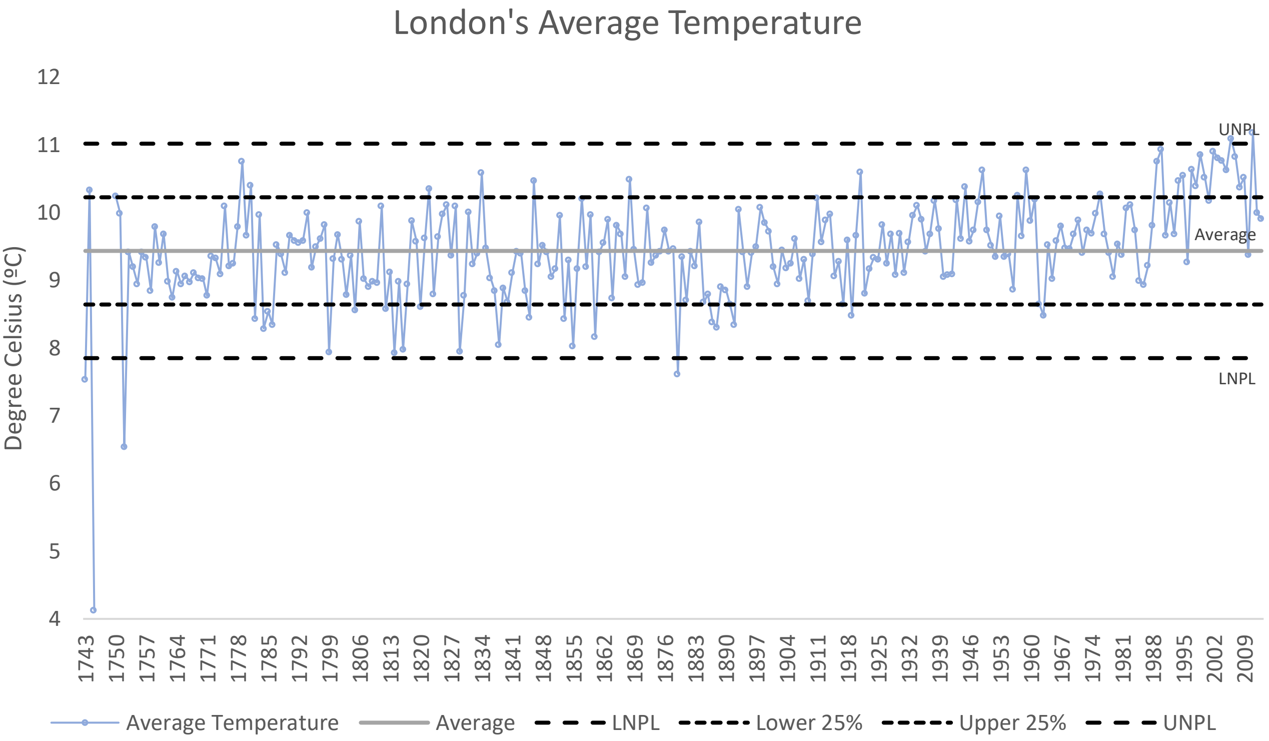
The two graph below show the average temperature overtime and are known as Process Behavior Charts. The LNPL and UNLP stand for Lower Natural Process Limit and Upper Natural Process Limit.

To extract conclusions from the process behavior charts we will use the following Trend Rules:

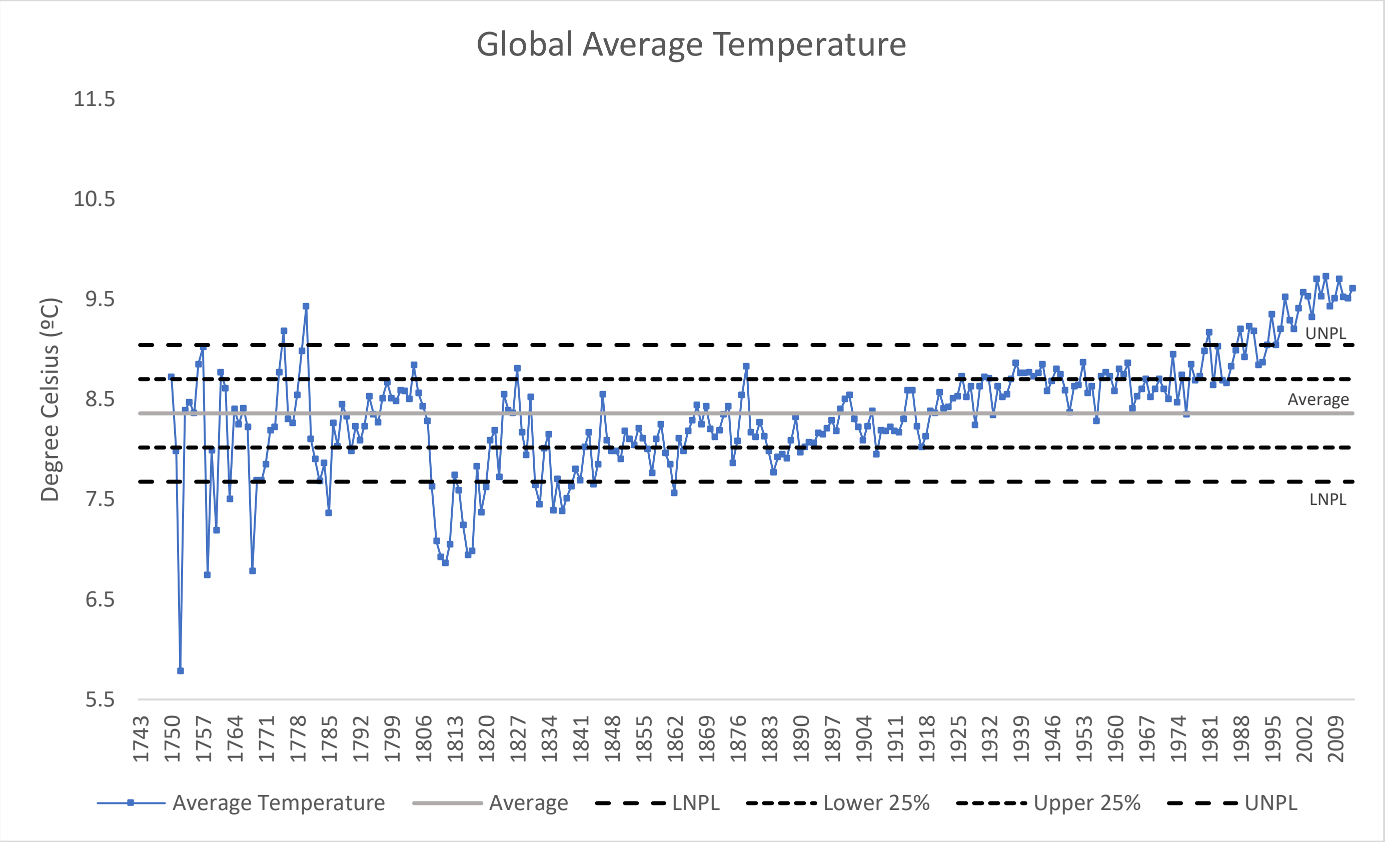
1. Points Outside the Limits:
    i. Single point outside the limits is an indication of a *dominant effect* that needs investigating.

2. Runs About the Central Line:
    i. Eight successive values on the same side of the central line is an indication of a *weak sustained effect*
    ii. We might want to investigate
3. Runs Near the Limits:
    i. Three out of four successive values within the upper 25% of the region between the limits or within the lower 25% region between the limits.
    ii. Could be indicative of *moderate sustained effect*.

From the process behavior chart below I only highlight that the points that accounts for the most recent years are within the upper 25% of the region between the limits and almost all of them from 1991 are on the same side of the central line. This implies that there is a tendency in London's temperature being hotter in recent years.



Regarding the global average temperatures. The process behavior chart indicates that the we have a clear and consistent upward trend, which is being stressed on the right hand side of the chart. In particular from 1999 on all the datapoints are outside the Upper Limit and with a stepper gradient than the rest of the data.



Section 2: The line chart

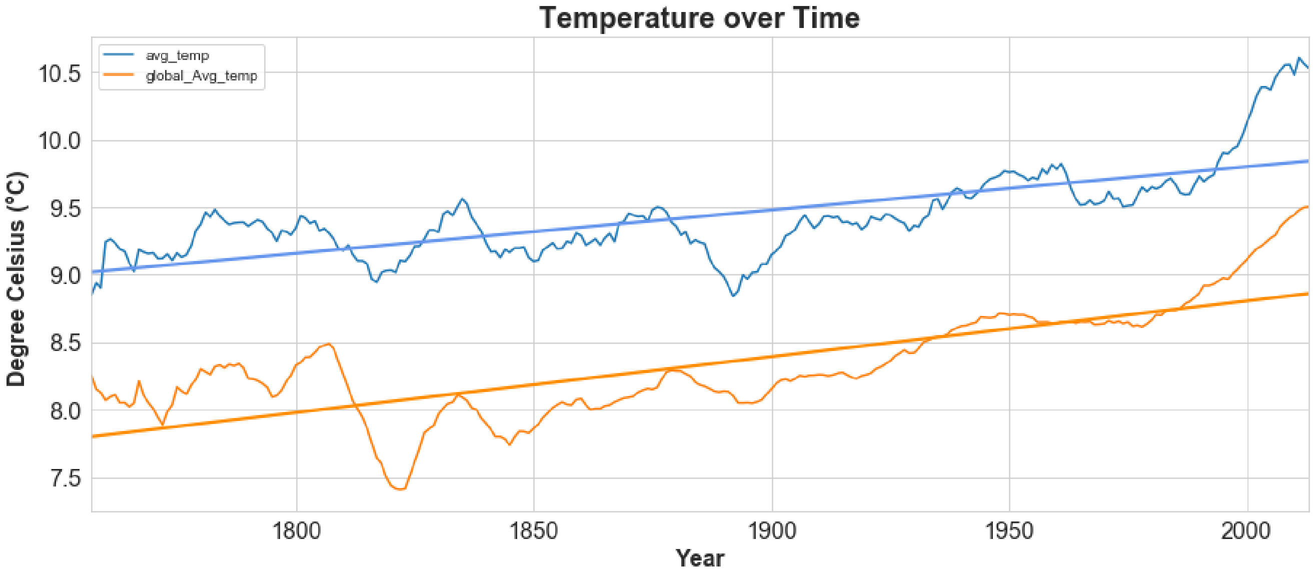
The prerequisites for this section's line chart were:

- To take care of missing values
- To find a proper window and calculate the moving averages

For *taking care of missing values* I used [3] and applied the K-nearest neighbors algorithm to my dataset. I think that this is the right thing to do as it takes into account the neighbors of the missing data points, which in my case are located at the beginning of the dataset, and then it applies the algorithm.

For the *moving averages* windows, I finally chose 15. I tried different windows and I finally decided that 15 was providing me a good picture of the trend and fluctuations overtime were also seen in a proper manner. Then I took advantage of the pandas.Series functionality to apply the .rolling,mean() to a pandas Series.

Finally we get the line graph below:



In the line graph above it can be appreciated in blue the London's average temperature and in orange the global average temperature overtime. These lines are complimented with the linear regression line, which emphasize an upward trend for both variables.

We can also see interesting patterns overtime, for example: The two variables seem to have positive correlation. In the next section we will get more into the details of the graphical observations.

Section 3: Observations

Is my city cooler or hotter on average compared to the global temperature?

As an average and taking the temperatures from 1750, the average temperature of London over the last 253 years has been 9,43ºC whereas the average global temperature has been 8,36ºC from which we are safe to say that as an average London temperature is slightly hotter than the global temperature

But this concept doesn't give us in-depth knowledge about the trend overtime and the relationship of the two variables. Thus, we need to take a deeper look at the trend of both London's and global temperature overtime and solve the following questions.

Which has been the trend over the last 253 years for the two London's and global average temperature?

We have an upward trend, in other words, the average temperature of both variables is increasing. In fact, the linear regression's formula for the London average temperature is

$$\mathrm{y} = 0{,}0033x + 8{,}96$$

whereas the linear regression formula for the global average temperature is

$$\mathrm{y} = 0{,}0042x + 7{,}72.$$

As we can see, the slope, which is the coefficient that depends of the time, is positive in both cases. Moreover, the slope for the global temperature is bigger than the one from the London's temperature. We can then conclude that both the world and London are getting hotter but the global average temperature grows at a faster pace.

Are the two variables stationary? Why?

Stationarity means that the statistical properties of a process generating a time series do not change over time [4]. The most important statistical properties are mean, variance and covariance so we need to figure out if these are independent of time or not. First we are going to analyze how numbers are spread out from their average value. This will give us hints about the variance of our variables. To identify this we can simply look at the line graphs created in the previous section and evaluate if they have a trend or a seasonality. In our case, there is a consistent upward trend effect that implies that the mean depend on the time index. For the London's average temperature, the mean does change depending on time due to the upward trend. Regarding the variance, it seem to remain similar across the period analyzed. Indeed, we can see that beside the period from 1880 and 1900 and 1995 until 2013 the line graph seem to be adjusted quite well to the linear regression line.

For the global average temperature it can be appreciated the same trend upward seen in the London's average temperature, hence its mean will be dependent on time. In terms of variance, this variable seems to have the same pattern as the London's one, but the residuals are stressed. That is, when on the right hand side of the line graph the London's average temperature seem to not match the regression line, the global temperature is even further away from it. And when later on the London's average temperature seem to follow the regression line, the global average temperature is even closer to it. This give us two intuitive conclusions:

- The variance from the lobal average temperature will be more dependent on time than the variance from London's average temperature
- The variables express high correlation factor

Putting all the above together, we have a clear indication that the data for both variables is non-stationary.

Another way to intuitively look into the stationary of the variables is to split our time series into different partitions and compare the mean and variance of each group. For example, we have:

Data	Years	Mean	Variance
Global Avg Temp	Before 1900	8,07	0,24
	After 1900	8,74	0,16
London's Avg Temp	Before 1900	9,21	0,58
	After 1900	9,74	0,36

As we can see, the difference in the mean is quite significative, which reaffirms that the data for both variables is non-stationary.

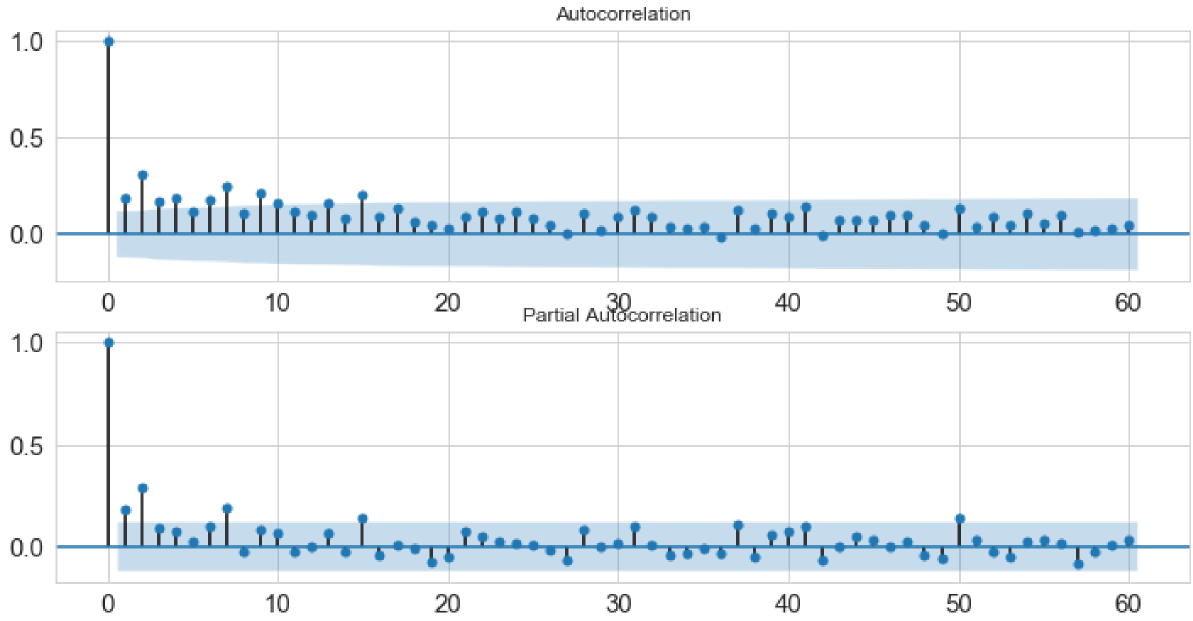
How has the London's average temperature changed overtime? And the global average temperature? Have them been consistent or not? Think about putting some Source here… You tend to get lost along the way
Source: https://towardsdatascience.com/almost-everything-you-need-to-know-about-time-series-860241bdc578

To solve these questions we need to ask ourselves whether the data is autocorrelated or not. That is, we need to understand the similarity between observations as a function of time lag between them. To do this we will plot the autocorrelation plot (ACF) and the partial autocorrelation plot (PACF) for each of the variables separated.

The first plot represents the autocorrelation of the series with lags of itself. The second plot represents the amount of correlation between a series and a lag of itself that is not explained by correlations at all lower-order lags. Colloquially, instead of finding correlations of present with lags like AFC, it finds correlation of the residuals, where residuals accounts for what remains after removing the effects which are already explained by the earlier lag(s).

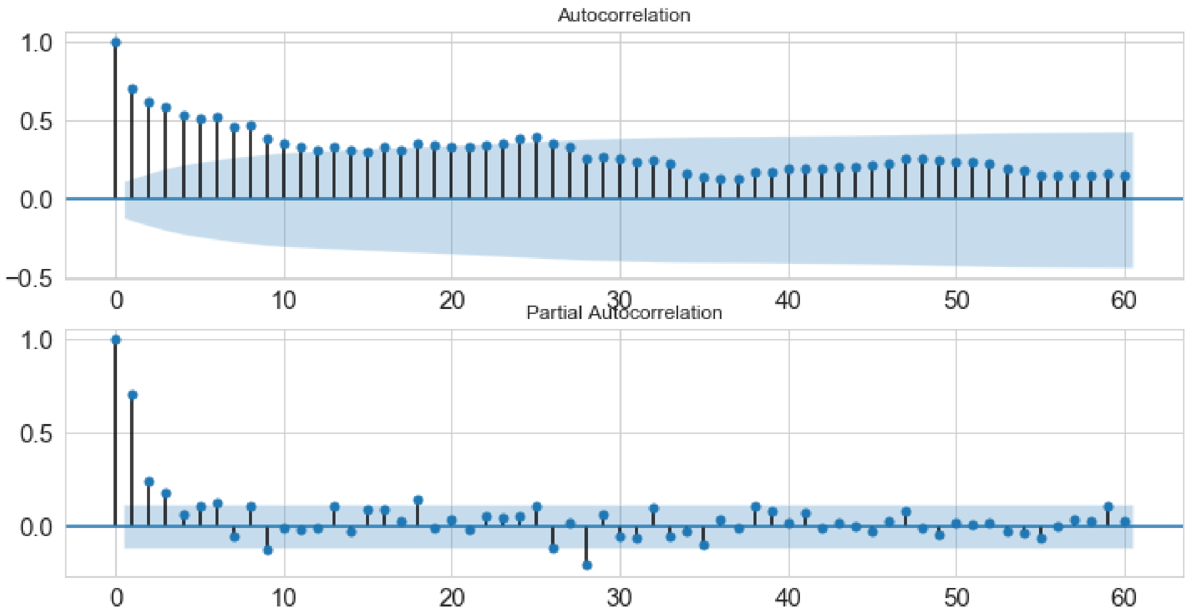
Note that by the nature of the definition of the plots, for the ACF plot the lag 0 observation will be always 1 because obviously one observation will correlate with itself.

For the PACF plot, we recall that this graph describes the correlation that results after removing the effect of any correlations due to the terms at shorter lags. The autocorrelation for an observation and an observation at a prior time step is comprised of both the direct correlation and indirect correlations. These in direct correlations are a linear function of the correlation of the observation, with observations at intervening time steps. This indirect correlations are the ones the partial autocorrelation plot seeks to remove.[5]
All the observations that do not fits into the blue region are autocorrelated and the x axis represent the autocorrelation lag. In particular, the blue region indicates whether the correlation at that lag fits into the 95% CI or not.
Note that the plots below are based on the original data once we have taken care of missing values, but without moving averages being applied. Below we can find the London's average temperature AFC and PACF plots:



From the ACF we get that the 1st value, the 2nd, the 3rd, the 4th, the 6th, the 7th and the 9th are highly correlated. This means that if the London's average temperature rises, it tends to continue rising, at least for the first nine observations.
If we are wondering whether it also means that we will find a very similar value at every one, two, three, four, six, seven or nine units of time, it doesn't. The main reason is that we cannot find a pattern that says, for example, that every 7 units of time the observation has the same correlation. In other words, the autocorrelation factor does not follow a pattern over the lags and so this is a clear example of a time series with non-seasonality.
Finally, what we also appreciate from the ACF plot is that the majority of the lags are correlated positive, this is another indication of the positive trend that we have already speak about.

Now, below we can find the global average temperature AFC and PACF plots:



The ACF plot represents a positive trend, due to the positive correlations factors, which emphasize that for the first 11th values they are highly autocorrelated but the autocorrelation factor trails off from the 12th on. Regarding seasonality, we conclude that there isn't as we cannot find neither a significant pattern nor a sinusoidal shape, which is always indicative of seasonality, on the ACF plot.

Has London's average temperature and global average temperature changed similarly overtime? In other words , are these variables correlated?

The answer to this question relates to the correlation factor between variables. In our case we already anticipated when we were getting to know the data that the variables where highly correlated. Now, if we calculate the correlation coefficient of London's average temperature and global average temperature we get that the coefficient correlation is 0,89. This means we have positive correlation and as it is close to 1, we have high correlation between variables.

The stepper curve on the right hand side of both line graph, what does it mean?

This "stepper curve" only represents that the global temperature and the London's one is increasing in a faster pace as we get closer to recent years. In other words, the slope of the linear regression that would approximate the right hand side of the graph would be higher than the rest of the graph.

References

[1] Donald Wheeler, Making Sense of Data

[2] Business Data Analysis with Excel

[3]The use of knn for missing values.

[4] Time Series Data Stationary Python. Machine Learning Mystery

[5] https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/