Udacity

Nanodegree

Data Analyst

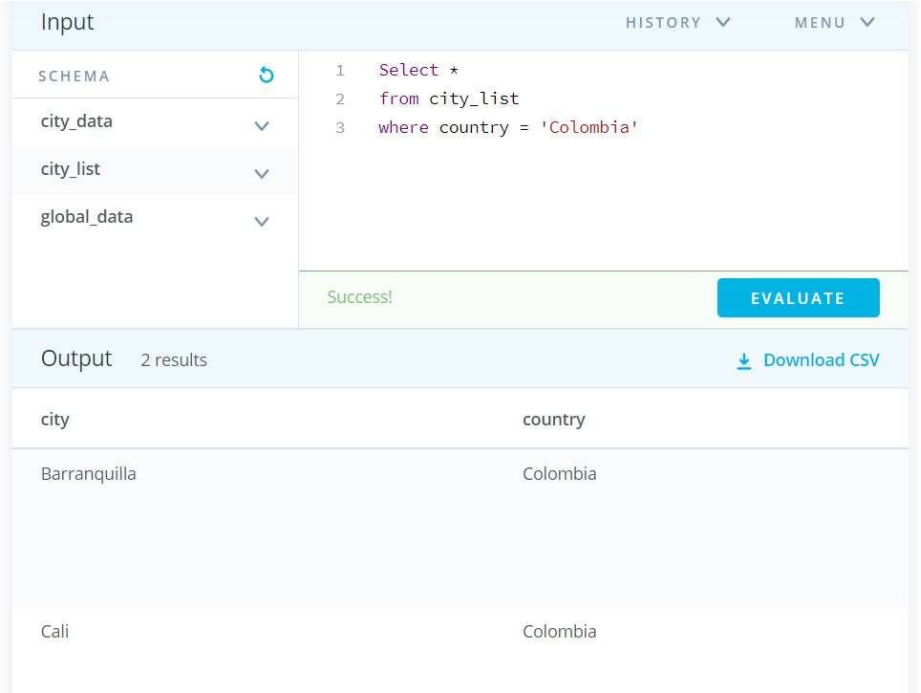
Project 1

Julian David Suarez Florez

- **Extract the data**

For extracting the data, we used the platform that contains the information of cities and global temperature. the next images show the queries that we used to extract the data of nearest city that where I live in Colombia and global.

First, we must identify which cities of the country where I live it would be in the table. The cities are Barranquilla and Cali the nearest to Bogota where I live so Cali is near, so we select the information of this one.



Second, to select information from Cali, we run the next query that allows me to get the data we needed.

In finally, I created a query to get the data from global temperature that was

- **Tool use**

With the information already in our hands we decide to work in Excel because it was not so much information to be process in Python or R also, we can create graphics and mathematical operation to work with our data in addition we want to use the information from the lesson.

- **Data exploration**

Before to make a comparation between Cali and Global temperature average and the other requirements, we begin with find anomalies in the data. First, we look the city data. We identified that CVS file has information from 1825 to 2013, the data has missing values since 1830 until 1851 that means 22 years of missing that of 189 years that has the table. That missing represent 11% of the data, in this case, in our opinion the missing values is significative.

However, we can take a decision what to do with missing values they are few options, one we can take de mean and used in this time that the data is missing, two we can interpolate the information between missing values, third discard this

information and work with the data that is available and fourth try to find the information in the principal source. Some analysts prefer option one or two because argue that the standard deviation is not going to change and that means you could have more data. In our opinion, we prefer third or fourth option because we do not alter the mean of the data, also I try to fix the information use the principal source. So, we decide third option and use the data since 1852.

Now, we present the descriptive statistics of the temperature in Cali, Colombia

Statistics	
Mean	21.79882716
Error	0.040460635
Median	21.765
Mode	21.23
Standard deviation	0.514979811
Variance	0.265204206
Kurtosis	-0.574654412
Coefficient of skewness	0.249706693
Rango	2.33
Minimum	20.74
Maximum	23.07
Sum	3531.41
Count	162

With this table we can observe that the information does not present other anomalies besides we explain before. They are not any outliers, mean, median and mode are similar so we can assume that de data is normal distributed.

Continue with the analysis of the data, in the global information we have information since 1750 until 2015, even though they are not any anomaly in the information. As we need to make a comparation between Cali and Global we took the same period that we use in city information in means since 1852 until 2013, we present the descriptive statistics of the global temperature.

Statistics	
Mean	8.564444444
Error	0.036005499
Median	8.54
Mode	8.73
Standard deviation	0.458275189
Variance	0.210016149
Kurtosis	-0.045107718
Coefficient of skewness	0.597225921
Rango	2.17
Minimum	7.56
Maximum	9.73
Sum	1387.44
Count	162

As before with the city information the table shows that the information does not present anomalies. They are not any outliers, mean, median and mode are similar so we can assume that de data is normal distributed. With all that we can start to make our moving average and comparation between the data that we have.

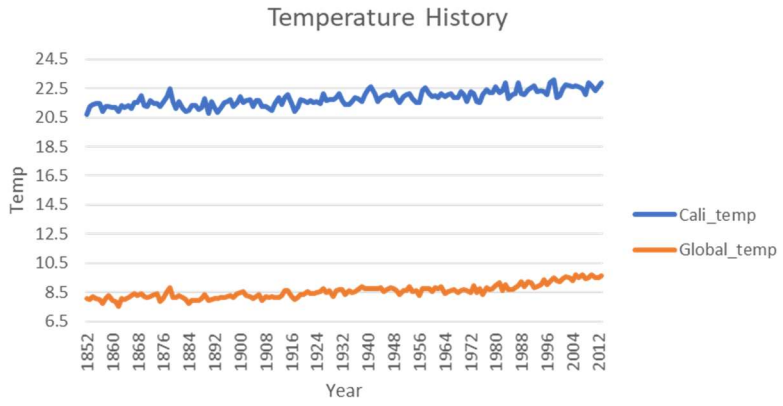
- **Observations**

In the observations part we have some question that we will answers and explain everything that we did and conclusion that we have about weather trends.

- Is your city hotter or cooler on average compared to the global average? Has the difference been consistent over time?

Yes, Cali is hotter than Global temperature because Cali average temperature is

21.79 C° and global average is 8.56 C° also Cali has a consistent difference in temperature over time as show this graphic.

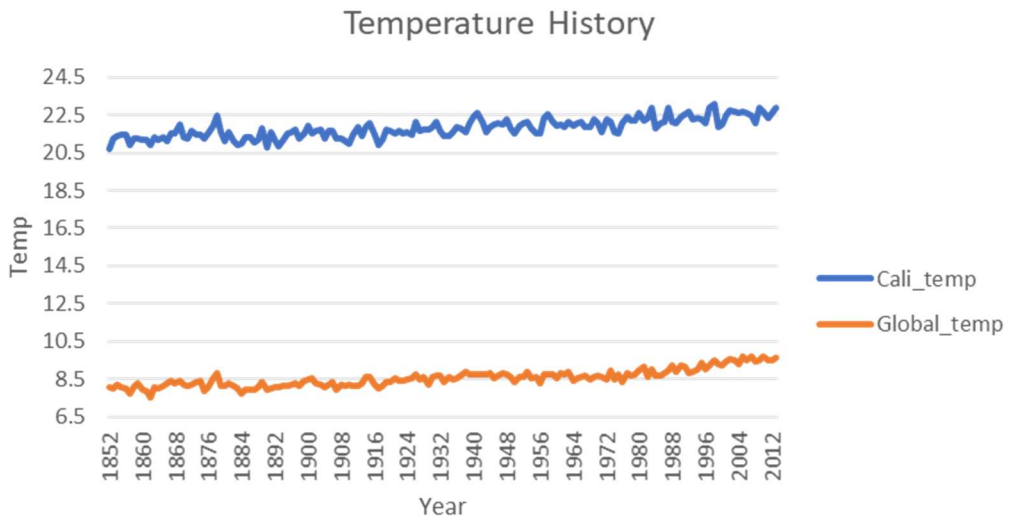


Even though, we think that we cannot conclude the assumption than one is hotter o cooler it is because global information contains the information all of city in the

planet and some are hotter and others cooler if we done this comparation between two cities it would be easier and certain this affirmation.

- How do the changes in your city's temperatures over time compare to the changes in the global average?

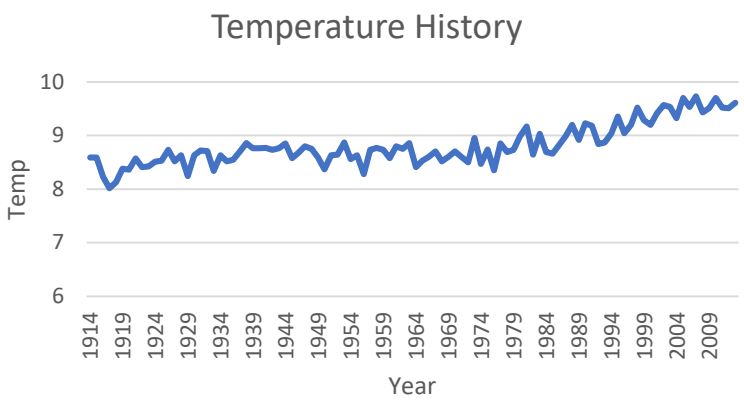
Looking at the next graphic we can see that both information has grown over the years we can conclude that both series have a correlation and between them that mean if one goes up the other too, it does mean that both increase in the same scale. If we see the minimum and maximum of both series, we see how temperature has change, for example Cali MIN is 20.74 and MAX is 23.07, global MIN is 7.56 and MAX 9.73 showing as how temperature has evolve.



- What does the overall trend look like? Is the world getting hotter or cooler? Has the trend been consistent over the last few hundred years?

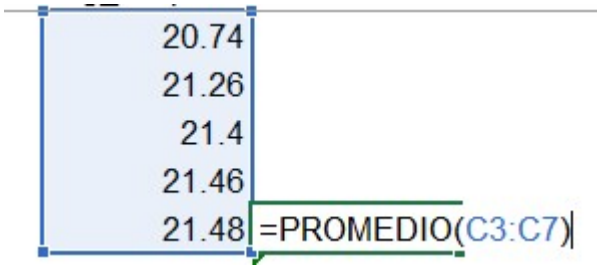
  The next graphic shows the evolution of the temperature in the world we can see that the temperature has been increase in the last hundred years in 11.87% that is a significate amount that means that our planet is getting hotter, looks like the trend is positive also in the session of "How did you calculate the moving average?" we can see how strong the trend in the information is.

  We can look how the temperature of Cali has changed in the last hundred year is only 4.43% is less than the change that it has the planet that show how strong the trend is how we talk before.



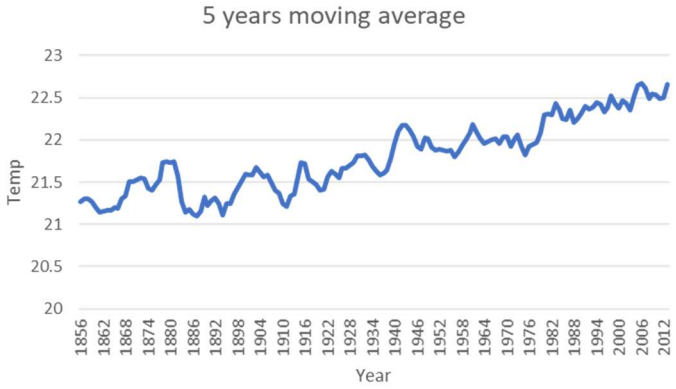
- How did you calculate the moving average?

To calculate the moving average for Cali and Global we use the function “Average” in Excel as show in the image.



We calculated different moving average every 5 years until we reach 30 year of moving average and made graphics for Cali and Global individually for all cases. We do this to reduce de volatility to a minimum even that we reduce so much data, but we reach our goal that is reduce de noise in the series.

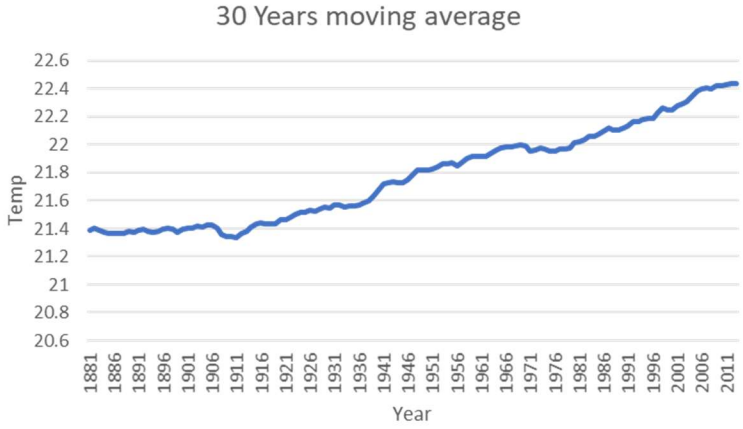
- **Cali**




10 years moving average
Temp
Year
23
22.5
22
21.5
21
20.5
20
1861
1867
1873
1879
1885
1891
1897
1903
1909
1915
1921
1927
1933
1939
1945
1951
1957
1963
1969
1975
1981
1987
1993
1999
2005
2011

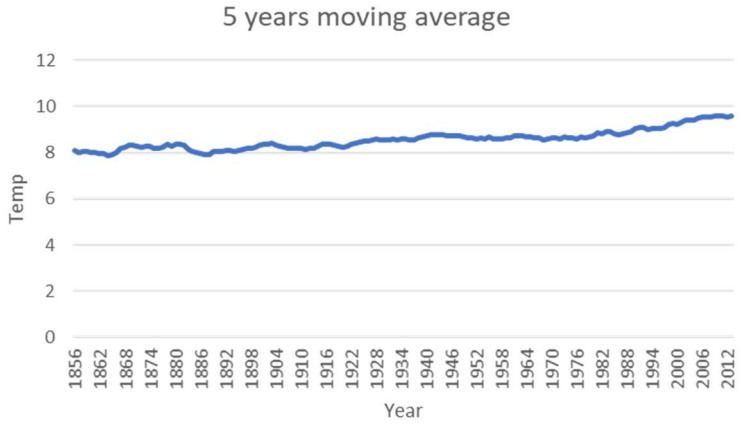

15 years moving average
Temp
Year
23
22.5
22
21.5
21
20.5
1866
1872
1878
1884
1890
1896
1902
1908
1914
1920
1926
1932
1938
1944
1950
1956
1962
1968
1974
1980
1986
1992
1998
2004
2010


20 years moving average
Temp
Year
22.8
22.6
22.4
22.2
22
21.8
21.6
21.4
21.2
21
20.8
20.6
1871
1877
1883
1889
1895
1901
1907
1913
1919
1925
1931
1937
1943
1949
1955
1961
1967
1973
1979
1985
1991
1997
2003
2009



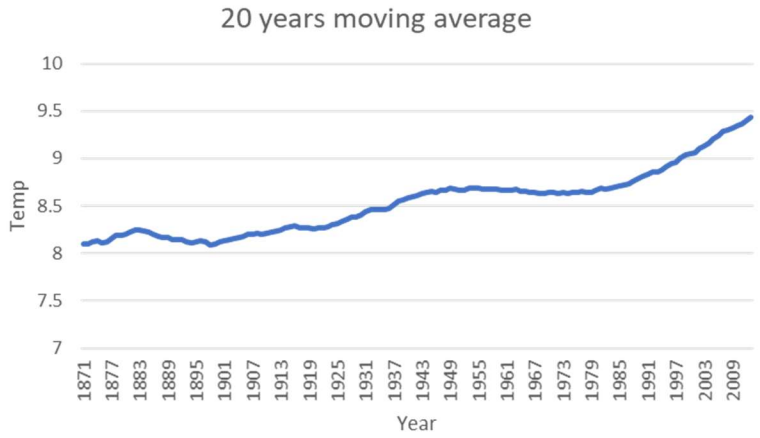
We can see that when we increase the moving average the data star to show a smooth volatility if we compare 5 years moving with 30 years moving the first show more noise in the series but the second show a small noise, both have a positive trend, we can make this exercise with all graphics and have the same behavior.

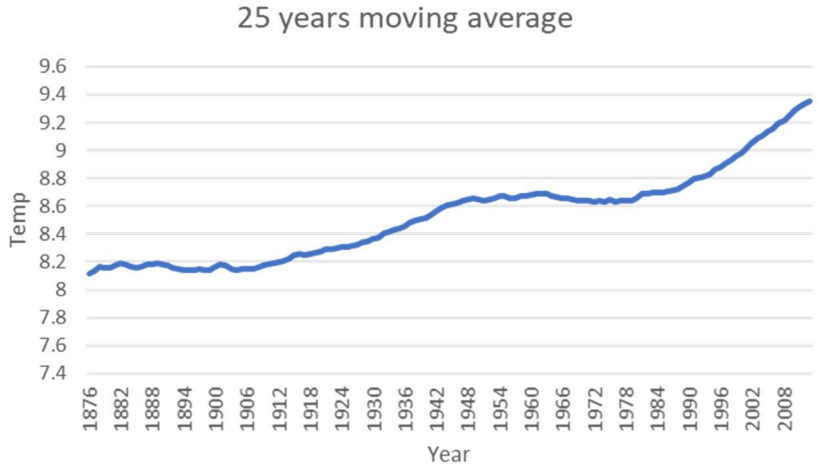
- **Global**

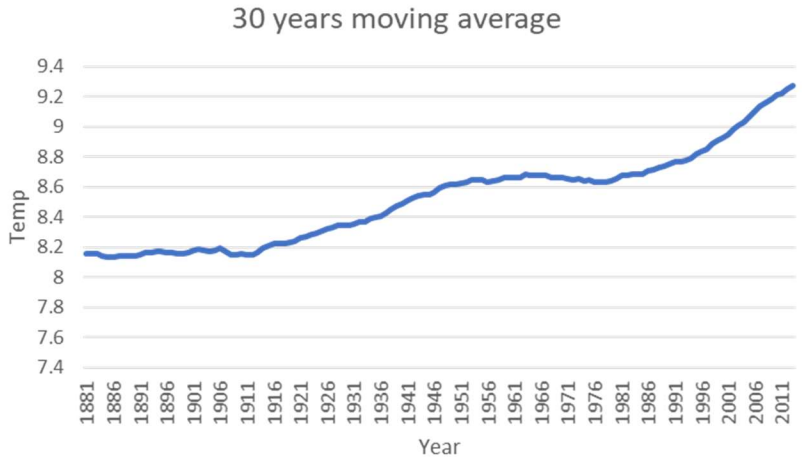



10 years moving average
Temp
Year
10
9.5
9
8.5
8
7.5
7
1861
1867
1873
1879
1885
1891
1897
1903
1909
1915
1921
1927
1933
1939
1945
1951
1957
1963
1969
1975
1981
1987
1993
1999
2005
2011


15 years moving average
Temp
Year
10
9.5
9
8.5
8
7.5
7
1866
1872
1878
1884
1890
1896
1902
1908
1914
1920
1926
1932
1938
1944
1950
1956
1962
1968
1974
1980
1986
1992
1998
2004
2010

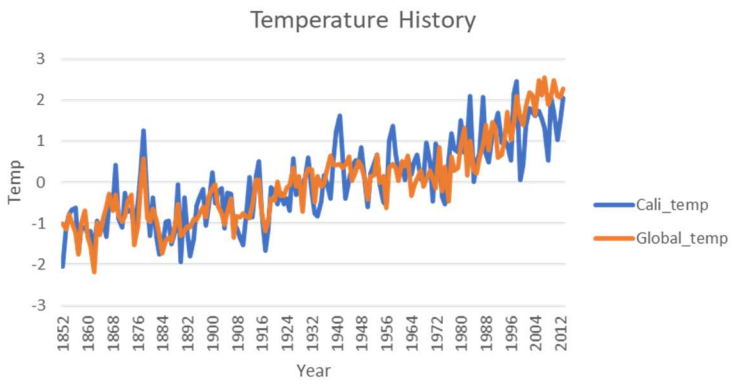

20 years moving average
Temp
Year
10
9.5
9
8.5
8
7.5
7
1871
1877
1883
1889
1895
1901
1907
1913
1919
1925
1931
1937
1943
1949
1955
1961
1967
1973
1979
1985
1991
1997
2003
2009

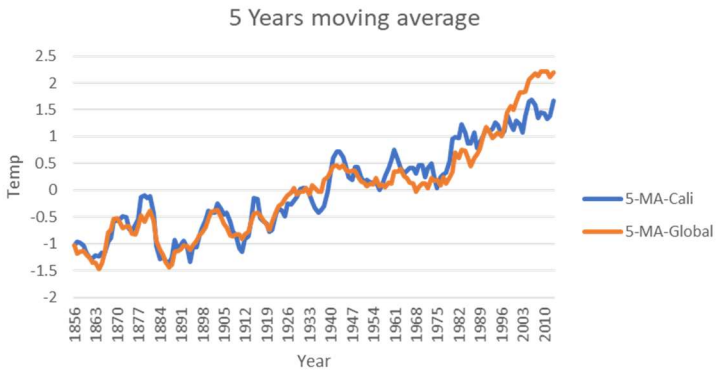


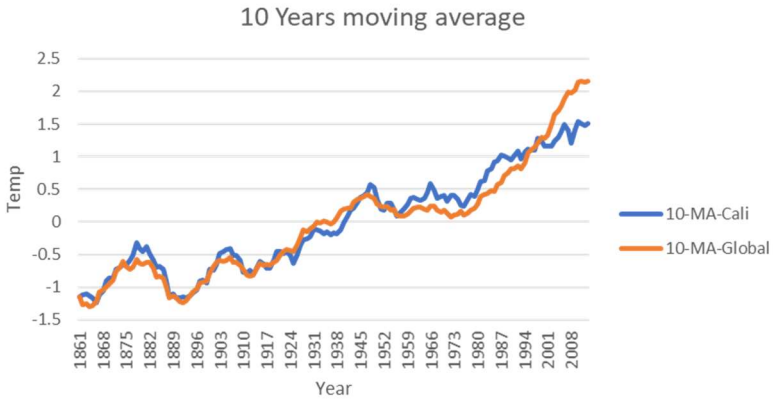


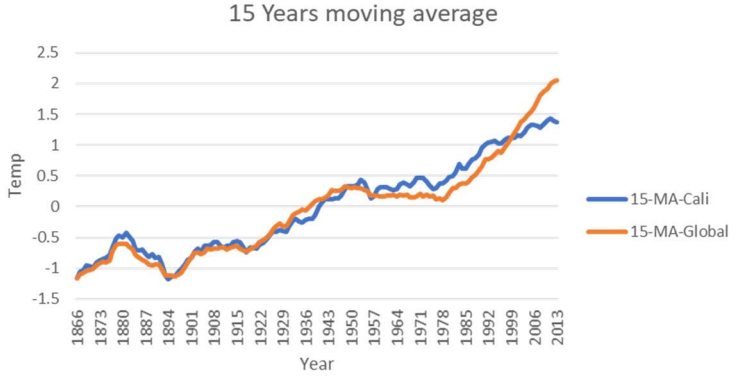
In this case, the data of global temperature show a super small volatility in the series so when we look at 5 years moving with 30 years moving the first show noise but it is not the same noise that show the information from Cali, in this series the noise is minimum but we look at 30 years shows a littler noise, also both have a positive trend, we can make this exercise with all graphics and have the same behavior.

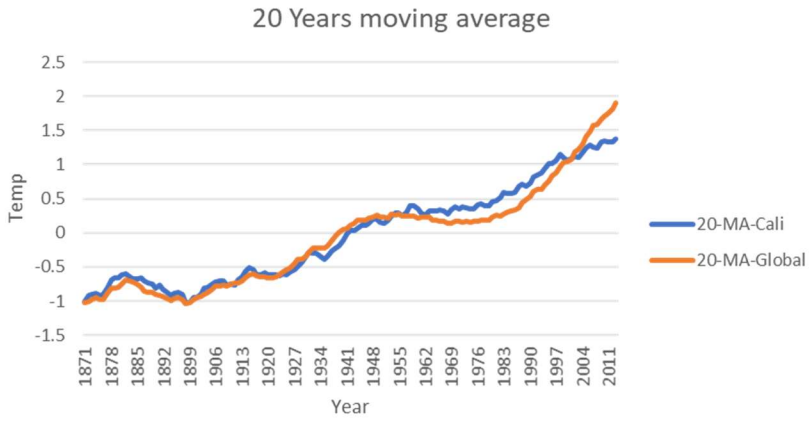
- **Comparative**

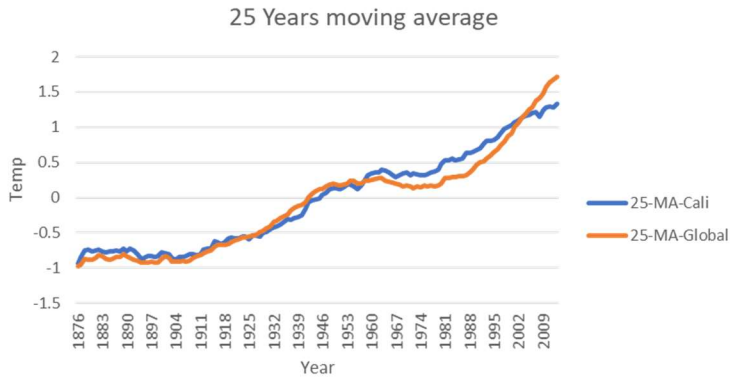

Temperature History
Temp
Year
Cali_temp
Global_temp

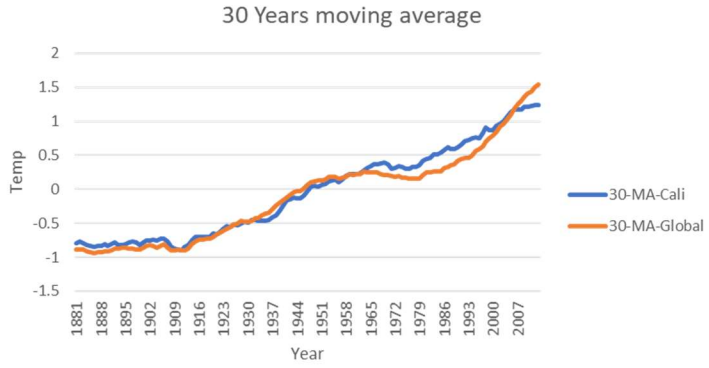

5 Years moving average
Temp
Year
5-MA-Cali
5-MA-Global


10 Years moving average
Temp
Year
10-MA-Cali
10-MA-Global


15 Years moving average
Temp
Year
15-MA-Cali
15-MA-Global







In this part, we can see how both series are too similar in trend even though Cali has more noise in all the graphics in the moving average that noise is minimum in 30 years graphic and both trends go to the same point and slope.

- What were your key considerations when deciding how to visualize the trends?

For this case, we decide to use a normalization formula that allows make a standardization(Ilker, 2020) of the data putting both series in the same scale, with this we can make a better comparison between Cali and Global. This formula is:

$$z = \frac{x - \mu}{\sigma}$$

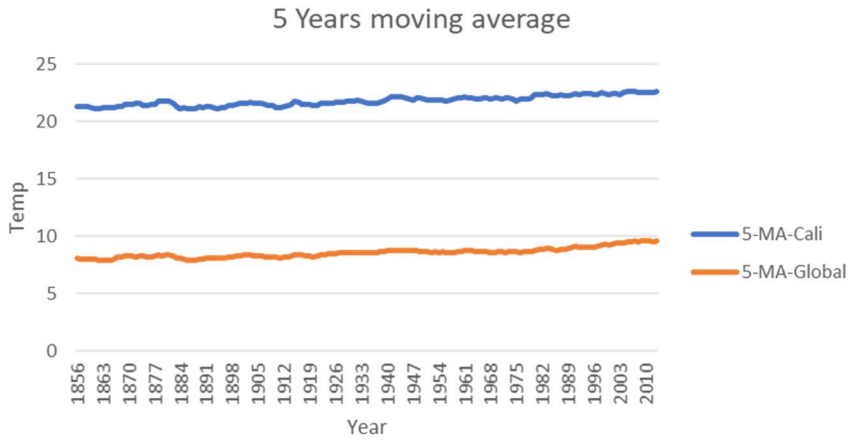
Z= Value standardization

X = Value

μ = Mean

σ = Standard Deviation

With both series in the same scale, it was easier to make a comparation of the trends because if we use the data its original form it would look like this:



This graphic does show anything clear even though we reduce the year of observation the gap between both series we have the same result of visualization but use standardization we solve this problem and the trend it does not affect.

In conclusion, we see that the global trend is stronger than the Cali's, we can see more volatility noise in Cali series, also that we make moving average in more periods of time we reduce significative the noise and have better trend with both series, finally that the planet is hotter every year and some cities like Cali has not change too much in temperature.