Explore Weather Trends

Udacity Data Analyst Nanodegree

Muhammad Adipurna Kusumawardana

March 25, 2020

1. Summary

In this project, we will analyze local and global temperature data and compare the local temperature to overall global temperature trends. And in addition, we will add other city for comparison.

2. Outline

2.1. Tools

Tools that we used in this project is:

1. Udacity SQL Workspace

   To extract data, we need to query from the SQL Workspace. The data will be including year, global average temperature, and desired city average temperature.

2. Excel

3. Word

2.2. Moving Average

Moving averages are used to smooth out data to make it easier to observe long term trends and not get lost in timely fluctuations. Commonly called by Moving Mean (MM) or Rolling Mean.[1]

[1] https://en.wikipedia.org/wiki/Moving_average

Data Analysis is one of Add-In in Excel to develop complex statistical or engineering analysis.[2] Instead using manual selection and average multiple times, we use Moving Average Tools to automate this calculation.

2.3. Data Visualization

The temperature data is a time series data, so we will visualize it using line chart between year vs temperature.

3. Methodology

3.1. Data Extracting using SQL Query

We have used the following query to retrieve data from the Udacity SQL Workspace. The global_data table was placed in the left. Then used 'LEFT JOIN' query with city_data table in the right. The 'ON' query was condition when we joined that two table. The 'WHERE' query is to filter the desired data. We chose Semarang because is the city most near our place. At last, in 'SELECT' query we selected year, global temperature average, and Semarang global temperature as a column that will be displayed as a query result.

```
SELECT g.year as Year,
       g.avg_temp AS Global,
       c.avg_temp AS Semarang
FROM global_data AS g
LEFT JOIN city_data as c
ON g.year = c.year
WHERE c.city = 'Semarang'
```

[2] https://support.office.com/en-us/article/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4

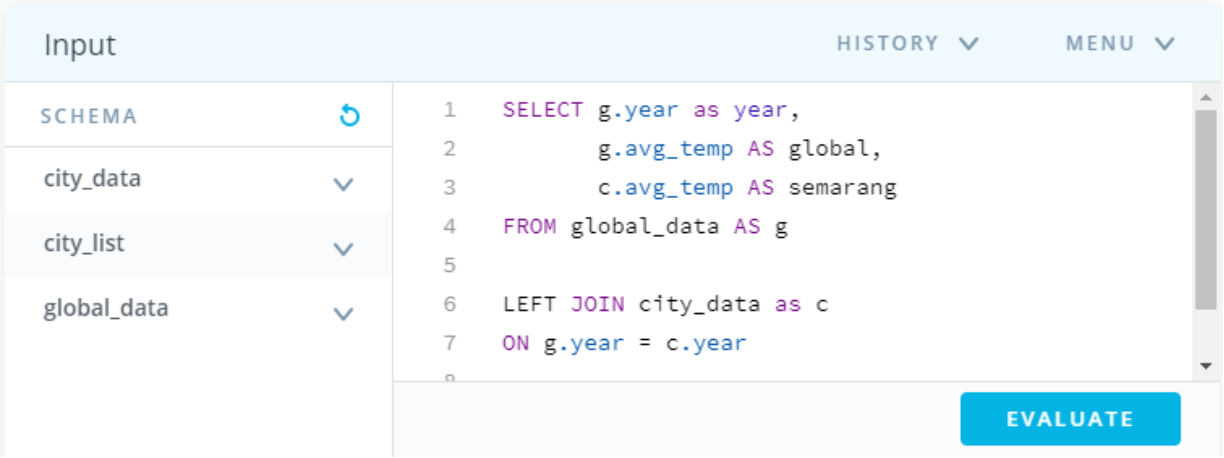


Figure 1. Querying Udacity SQL Workspace

This query was resulting table like follows. It has 189 rows and 3 columns. The column was year, global, and semarang. The rows were represented data each year from 1825 to 2013.

In the top right corner, there was Download CSV button, we used this button to save this query result. Then we proceed further with Excel. CSV stands for comma separated values.

Output 189 results Download CSV

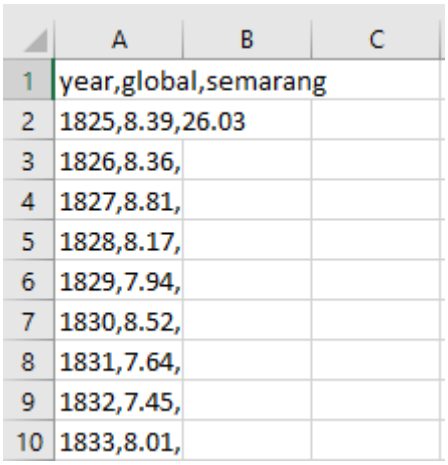
year	global	semarang
1825	8.39	26.03
1826	8.36	
1827	8.81	
1828	8.17	
1829	7.94	
1830	8.52	
1831	7.64	
1832	7.45	

Figure 2. Udacity SQL Workspace Output

3.2. Data Importing using Excel

As we stated before, we opened the CSV file using Excel. With it, the data can be placed in cells so we can process it more easily. The open method is very easy. First open your Excel desktop application, then choose 'Open' in the left bar, and finally browse your CSV files through provided explorer. Or simply right click your CSV file, choose 'open with', then select Excel as opener.

Regardless of whatever method is used to open CSV files, the file will be read as follow.



	A	B	C
1	year,global,semarang		
2	1825,8.39,26.03		
3	1826,8.36,		
4	1827,8.81,		
5	1828,8.17,		
6	1829,7.94,		
7	1830,8.52,		
8	1831,7.64,		
9	1832,7.45,		
10	1833,8.01,		

Figure 3. Imported CSV File in Excel

After the import process finished, the table is seeming bit funny. This is because the data in each row still separated by comma. So, we used Excel built-in menu to handle this.

First, block all your data or simply press Ctrl + a, then choose Data on Excel menu bar, then choose Text to Column. The wizard will be appeared and follow it until finished. You will be asked some option regarding the data that you want to show later, choose as comfortable as you wish. Then, tidier data will be like this.

	A	B	C
1	Year	Global	Semarang
2	1825	8.39	26.03
3	1826	8.36	
4	1827	8.81	
5	1828	8.17	
6	1829	7.94	
7	1830	8.52	
8	1831	7.64	
9	1832	7.45	
10	1833	8.01	

Figure 4. Tidier Data

3.3. Delete Empty Data

As we can see, some of the data from Semarang column was missing. The missing data for Semarang column were multiple range i.e. 1826 - 1838, 1842 - 1849, and 1857 - 1865, in total 30 rows was missing. The missing data was 15.87% from total 189 rows. So, we will use the data from 1866 until 2013 total (159 rows) and discard the rest. The cleaned data as follows.

	A	B	C
1	Year	Global	Semarang
2	1866	8.29	25.65
3	1867	8.44	25.65
4	1868	8.25	25.87
5	1869	8.43	25.82
6	1870	8.2	25.51
7	1871	8.12	25.45
8	1872	8.19	25.53
9	1873	8.35	25.63
10	1874	8.43	25.56

Figure 5. Cleaned Data

3.4. Moving Average

After tidying up the data, we calculated the moving average. In this project we used 15 as moving average interval, because this is the smoothest chart resulted in Excel. Usually this interval will be called short term interval in SMA (Simple Moving Average). In the chart, smoothness ratio as follows.

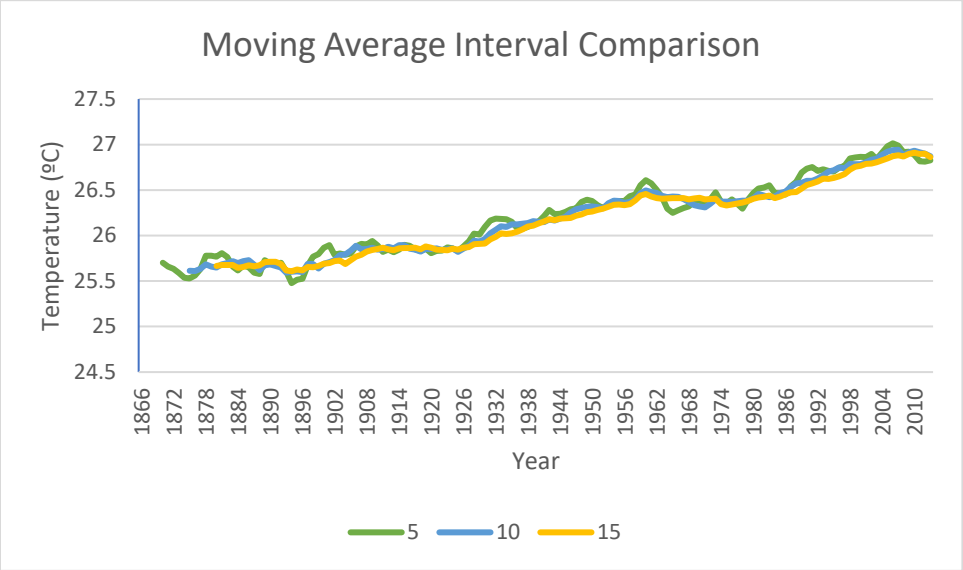


Figure 6.Moving Average Interval Comparison

Using 'Data Analyst' from Data drop-down menu in Excel, we can easily find the moving average result. We just need to define the Input Range, Interval, and Output Range from our data.

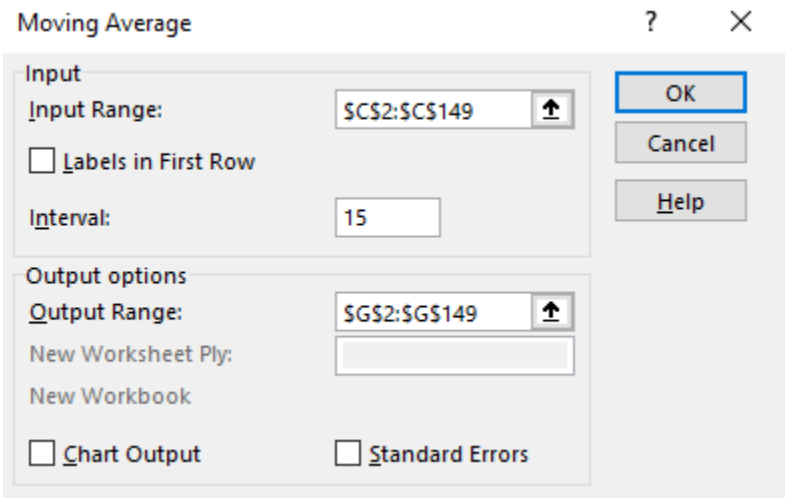


Figure 7. Moving Average Window

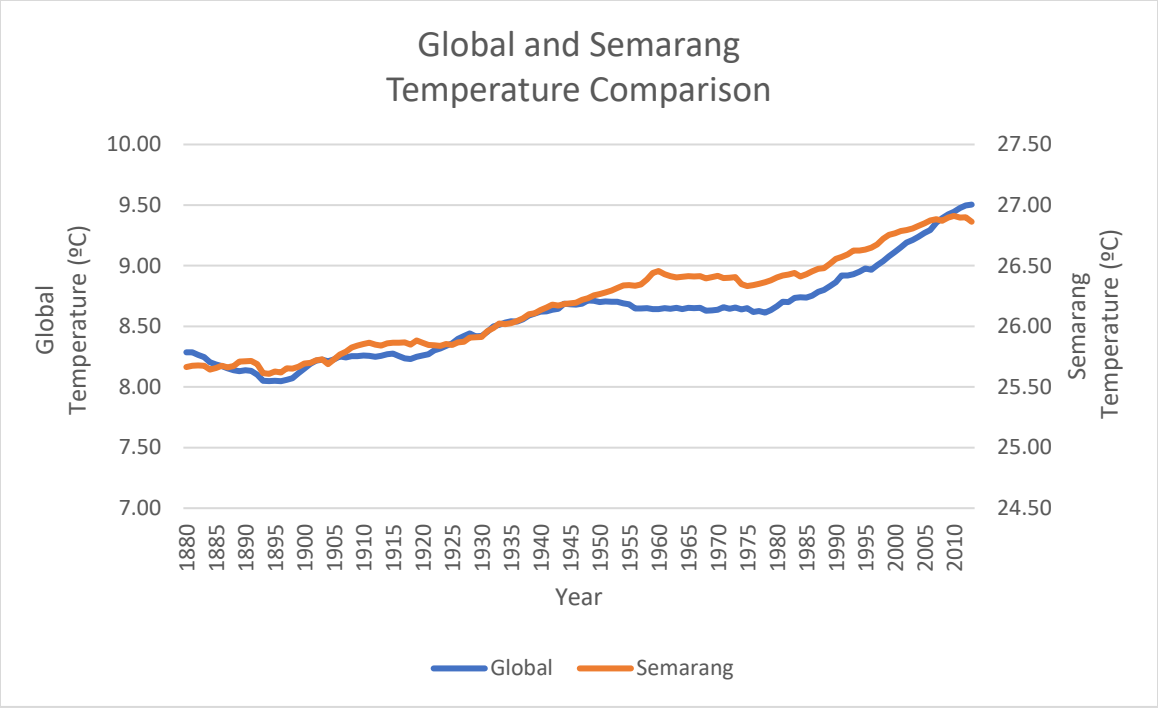
After clicked the OK button result immediately showed in desired rang, first 14 rows for each moving average column was empty because we used 15 as interval, so (n-1) rows was missing from the table. If there is error notification such as figure below (green arrow in upper left cell) you can ignore it.

	A	B	C	D	E
1	Year	Global	Semarang	Global mv_avg	Semarang mv_avg
2	1866	8.29	25.65	#N/A	#N/A
3	1867	8.44	25.65	#N/A	#N/A
4	1868	8.25	25.87	#N/A	#N/A
5	1869	8.43	25.82	#N/A	#N/A
6	1870	8.2	25.51	#N/A	#N/A
7	1871	8.12	25.45	#N/A	#N/A
8	1872	8.19	25.53	#N/A	#N/A
9	1873	8.35	25.63	#N/A	#N/A
10	1874	8.43	25.56	#N/A	#N/A
11	1875	7.86	25.47	#N/A	#N/A
12	1876	8.08	25.61	#N/A	#N/A
13	1877	8.54	25.88	#N/A	#N/A
14	1878	8.83	26.37	#N/A	#N/A
15	1879	8.17	25.57	#N/A	#N/A
16	1880	8.12	25.41	8.29	25.67
17	1881	8.27	25.81	8.29	25.68
18	1882	8.13	25.67	8.26	25.68
19	1883	7.98	25.84	8.25	25.68
20	1884	7.77	25.35	8.20	25.64

Figure 8. Moving Average Calculated

3.5. Plotting

After moving average was counted for each global and Semarang, we start to plot in line chart. We used the line chart because the data that we used is continuous. We used line chart combo type so we can easily compare the Global temperature and Semarang temperature shape yearly.



Make observations about the similarities and differences between the world averages and your city's averages, as well as overall trends. Here are some questions to get you started.

- Is your city hotter or cooler on average compared to the global average? Has the difference been consistent over time?
- "How do the changes in your city's temperatures over time compare to the changes in the global average?"
- 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?

4. Result and Observation

My local city (Semarang) is hotter than global temperature average. It can be saw in the y axis for each location, where Global temperature line is below 10^{o}C and for Semarang temperature line is between 25^{o}C and 27^{o}C. This is because Indonesia is located near equator and included in the tropics where the sun is very intense, compared to the others country that far from equator.

Overall, if we compare Semarang temperatures with Global temperature, both have nearly same fluctuation, then increased until the end, and the trend has increased over time. The two data

have same pattern. Based on the data we can conclude that Global temperature is getting hotter for the last few hundred years and the trend seems linearly positive. Same as Global temperature, Semarang temperature is sure getting hot for the last few hundred years.

Even though both temperatures are far adrift, as we can see in the chart, the difference is somewhat consistent from 1880 until 2013. It seems there is some correlation between two. To prove it we will plot each temperature data to scatter chart, show the trendline, and show the R^2 value in the chart. If the R^2 is close to 1, then this will prove that there is a relationship between the two.

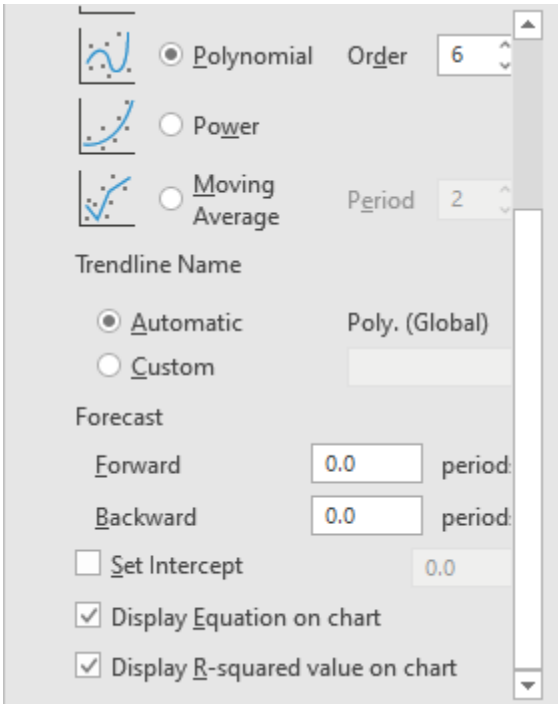


Figure 9. Chart Formatting

As we can see in the chart, Excel chart formatting can easily show to us the R^2 value. The value was 0,9604, this mean there is strong correlation between Global and Semarang temperature. The trendline (yellow dotted line) equation can be easily showed in the chart too. From the 'y' polynomial order 6 equation, we can estimate the average temperature in our city based on the average global temperature.

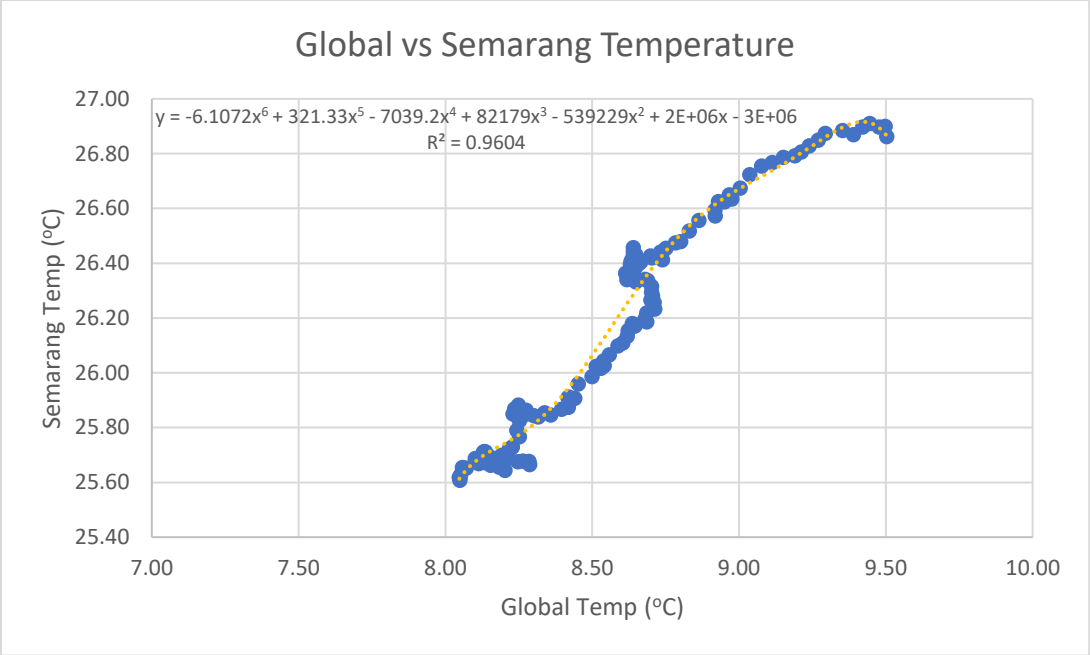


Figure 10. Trendline and R-square Global vs Semarang

4.1. Another City

To make this project more stand-up, we will add another interesting city. As before we used Semarang, Indonesia that represented tropic zone city, and now Milan, Italy will be added as representing temperate zone city.

Same method was same as before, from importing, cleaning, deleting and moving average calculation was finished in Excel.

4.2. Plotting

After used the same method to find the moving average, the data was plotted into line chart as before. Now we have three lines combo chart. X axes represent year, Y left-axes represent Global and Milan temperature, and Y right-axes represent Semarang temperature.

Same as observation before, all the three lines showing positive trends. Overall, all line is increased until the end of data.

	A	B	C	D
1	Year	Global	Tropic	Temperate
2			Semarang	Milan
3	1866	8.29	25.65	7.05
4	1867	8.44	25.65	6.8
5	1868	8.25	25.87	7.21
6	1869	8.43	25.82	6.61
7	1870	8.2	25.51	6.19
8	1871	8.12	25.45	5.98
9	1872	8.19	25.53	7.14
10	1873	8.35	25.63	7.13
11	1874	8.43	25.56	6.52
12	1875	7.86	25.47	6.45
13	1876	8.08	25.61	6.65
14	1877	8.54	25.88	6.79
15	1878	8.83	26.37	6.42
16	1879	8.17	25.57	5.91
17	1880	8.12	25.41	6.97
18	1881	8.27	25.81	6.47
19	1882	8.13	25.67	6.78
20	1883	7.98	25.84	6.05

Figure 11. Added Milan Moving Average

Then we will see if there any correlation between Milan and Global temperature. The methodology is same as before. Using versus scatter plot between Global and Milan, show the trendline, and R^2 value.

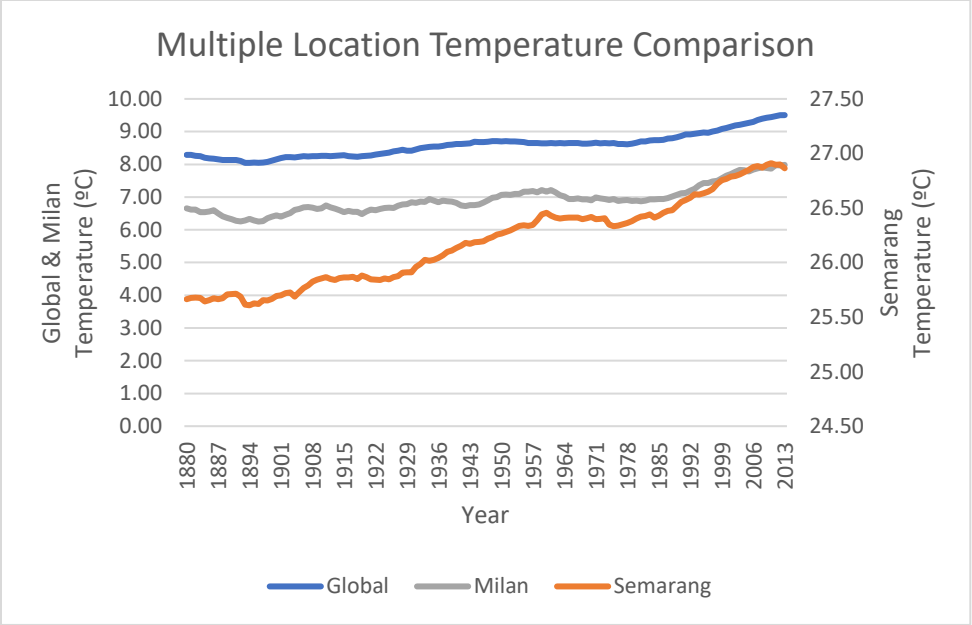


Figure 12. Multiple Comparison

There is correlation between Global and Milan temperature, the R^2 score is quite high, 0.9337. But the trendline is linear, different from Global vs Semarang which have polynomial with 6 order. And absolutely, we can estimate the average of Milan temperature using 'y' equation from the chart.

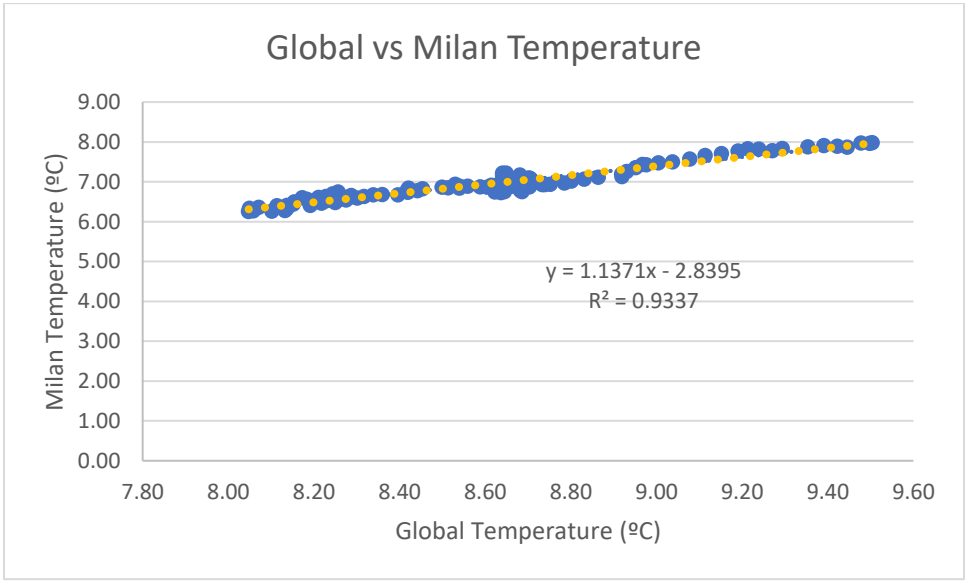


Figure 13. Trendline and R-squared Global vs Milan

5. Conclusion

This project concludes:

1. All the temperature data was fluctuating, but overall is increased until the end of data.
2. Not only local city temperature was rising, but different city in other parts of the world temperature was rising too.
3. There is strong correlation between increasing local temperature with global warming.
4. The local average temperature can be calculated from global temperature using each equation from trendline.