project-1-notebook

March 22, 2021

1 Data Analyst Nanodegree - Project 1 - Weather Trends

1.0.1 Summary

In this project, you will analyze local and global temperature data and compare the temperature trends where you live to overall global temperature trends.

1.0.2 Instructions

Your goal will be to create a visualization and prepare a write up describing the similarities and differences between global temperature trends and temperature trends in the closest big city to where you live. To do this, you'll follow the steps below:

- Extract the data from the database. There's a workspace in the next section that is connected to a database. You'll need to export the temperature data for the world as well as for the closest big city to where you live. You can find a list of cities and countries in the city_list table. To interact with the database, you'll need to write a SQL query.
- Write a SQL query to extract the city level data. Export to CSV.
- Write a SQL query to extract the global data. Export to CSV.
- Open up the CSV in whatever tool you feel most comfortable using. We suggest using Excel or Google sheets, but you are welcome to use another tool, such as Python or R.
- Create a line chart that compares your city's temperatures with the global temperatures. Make sure to plot the moving average rather than the yearly averages in order to smooth out the lines, making trends more observable (the last concept in the previous lesson goes over how to do this in a spreadsheet).
- 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?

1.0.3 Submission

Your submission should be a PDF that includes:

- An outline of steps taken to prepare the data to be visualized in the chart, such as:
  - What tools did you use for each step? (Python, SQL, Excel, etc)
  - How did you calculate the moving average?
  - What were your key considerations when deciding how to visualize the trends?
- Line chart with local and global temperature trends
- At least four observations about the similarities and/or differences in the trends

1.0.4 Rubric

A Udacity reviewer will assess your project based on the criteria in the project rubric. Use the rubric as a guide while you complete the project, then give yourself a quick self-assessment before you submit it.

1.1 Project starts

1.1.1 Tools and methods used

- JupyterLab and JupyterNotebook to import, explore and visualizate the data in python
- Calculate the moving average using pandas
- matplotlib e seabor libraries to visualizate the data
- scatter plot and spearmanr correlation coeficient to check correlation between 2 numericals variables

1.1.2 Import libraries

[23]:
```
import os
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
%matplotlib inline
```

1.1.3 Queries used to export the data

Below, included the queries used to export the data from Nanodegree database:

Export city data

select * from city_data where lower(country) = 'brazil' and lower(city) = 'rio de janeiro'

Export global data

select * from global_data

1.1.4 Import data

[2]:
```
df_city = pd.read_csv('C:
↪\\Users\\mayke\\Documents\\Cursos\\Data-Analyst-Nanodegree\\projects\\project-1-explore-wea
↪csv')
```

```
df_global = pd.read_csv('C:
↪\\Users\\mayke\\Documents\\Cursos\\Data-Analyst-Nanodegree\\projects\\project-1-explore-wea
↪csv')
```

[3]:
```
df_city.head()
```

[3]:
```
   year            city country  avg_temp
0  1832  Rio De Janeiro  Brazil     23.05
1  1833  Rio De Janeiro  Brazil     24.11
2  1834  Rio De Janeiro  Brazil     23.27
3  1835  Rio De Janeiro  Brazil     22.73
4  1836  Rio De Janeiro  Brazil     22.91
```

[4]:
```
df_global.head()
```

[4]:
```
   year  avg_temp
0  1750      8.72
1  1751      7.98
2  1752      5.78
3  1753      8.39
4  1754      8.47
```

[5]:
```
df_city.dtypes
```

[5]:
```
year          int64
city         object
country      object
avg_temp    float64
dtype: object
```

1.1.5 Calculate the moving average of last 5 years

[6]:
```
#creating moving average of last 5 years
df_city['moving_average_5y'] = df_city.iloc[:,3].rolling(window=5).mean()
df_global['moving_average_5y'] = df_global.iloc[:,1].rolling(window=5).mean()
```

[7]:
```
df_city.tail(10)
```

[7]:
```
     year            city country  avg_temp  moving_average_5y
172  2004  Rio De Janeiro  Brazil     24.24             24.734
173  2005  Rio De Janeiro  Brazil     24.79             24.820
174  2006  Rio De Janeiro  Brazil     24.57             24.736
175  2007  Rio De Janeiro  Brazil     24.78             24.656
176  2008  Rio De Janeiro  Brazil     24.26             24.528
177  2009  Rio De Janeiro  Brazil     24.98             24.676
178  2010  Rio De Janeiro  Brazil     24.95             24.708
179  2011  Rio De Janeiro  Brazil     24.32             24.658
180  2012  Rio De Janeiro  Brazil     24.84             24.670
181  2013  Rio De Janeiro  Brazil     25.19             24.856
```

[8]:
```
df_global.tail(10)
```

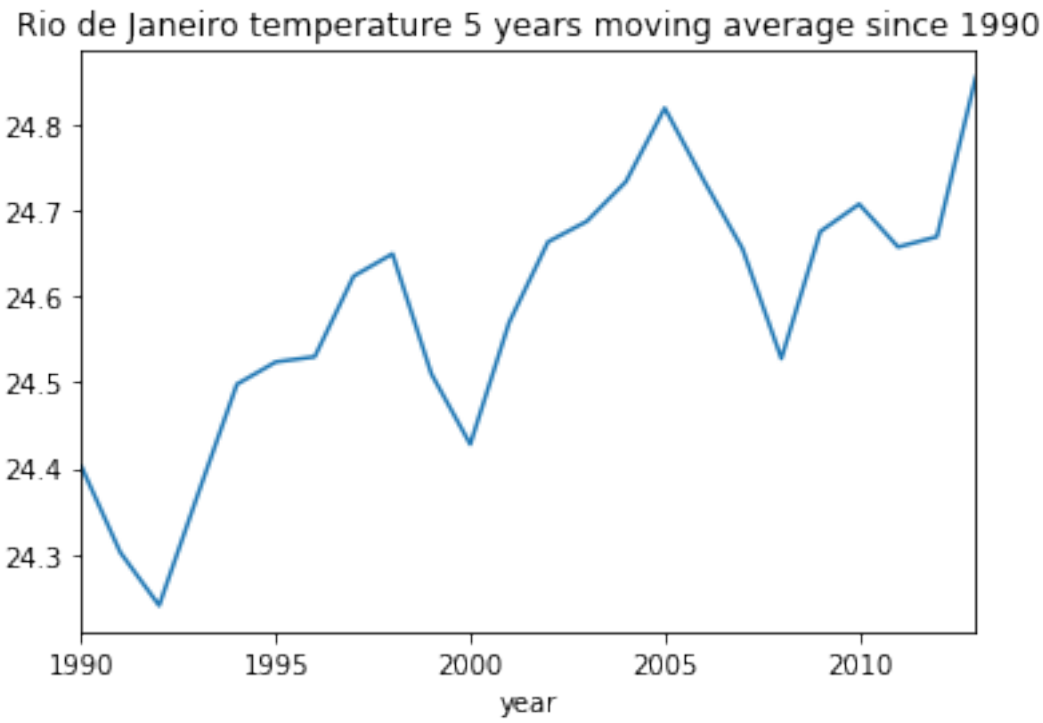
[8]:

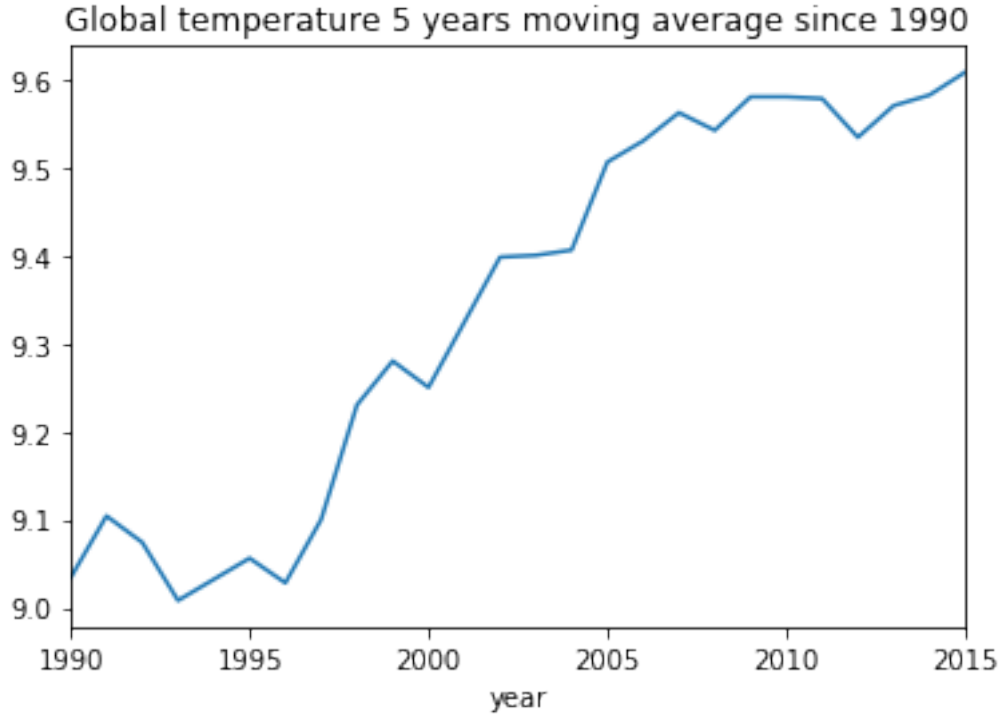
	year	avg_temp	moving_average_5y
256	2006	9.53	9.530
257	2007	9.73	9.562
258	2008	9.43	9.542
259	2009	9.51	9.580
260	2010	9.70	9.580
261	2011	9.52	9.578
262	2012	9.51	9.534
263	2013	9.61	9.570
264	2014	9.57	9.582
265	2015	9.83	9.608

1.1.6 Comparing city and global temp ma - Since 1990

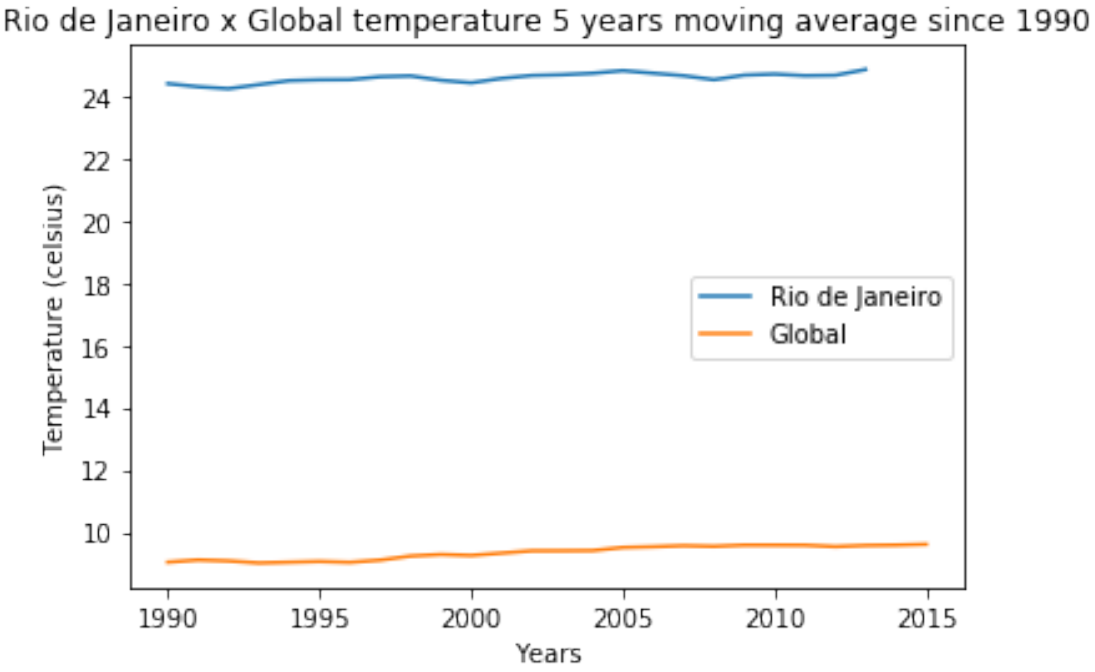
[9]:
```
#filter last years
df_city_filter = df_city[df_city['year'] >= 1990]
df_global_filter = df_global[df_global['year'] >= 1990]

df_city_filter2 = df_city[df_city['year'] >= 1875]
df_global_filter2 = df_global[df_global['year'] >= 1875]
```

[10]:
```
#To compare, I decided to select the line graph to identify the curves for each␣
 ↪base
title1 = 'Rio de Janeiro temperature 5 years moving average since 1990'
title2 = 'Global temperature 5 years moving average since 1990'

df_city_filter.plot(x="year", y="moving_average_5y", kind="line", title =␣
 ↪title1, legend = False);
df_global_filter.plot(x="year", y="moving_average_5y", kind="line", title =␣
 ↪title2, legend = False);
```


Rio de Janeiro temperature 5 years moving average since 1990
24.8
24.7
24.6
24.5
24.4
24.3
1990
1995
2000
2005
2010
year

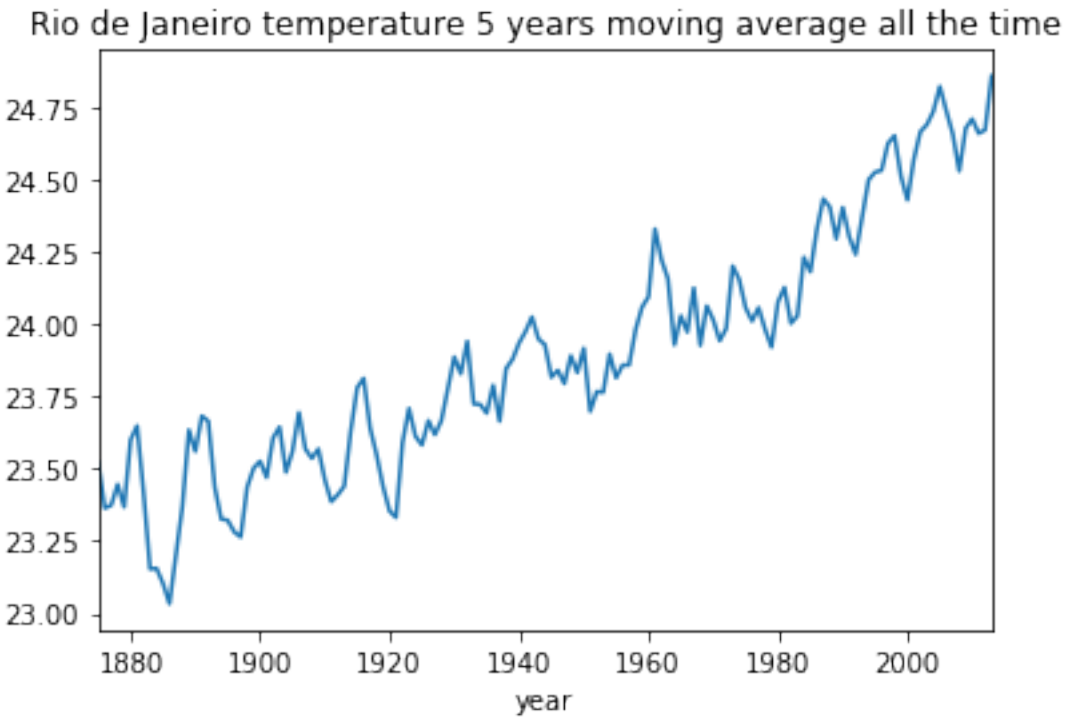

Global temperature 5 years moving average since 1990
9.6
9.5
9.4
9.3
9.2
9.1
9.0
1990
1995
2000
2005
2010
2015
year

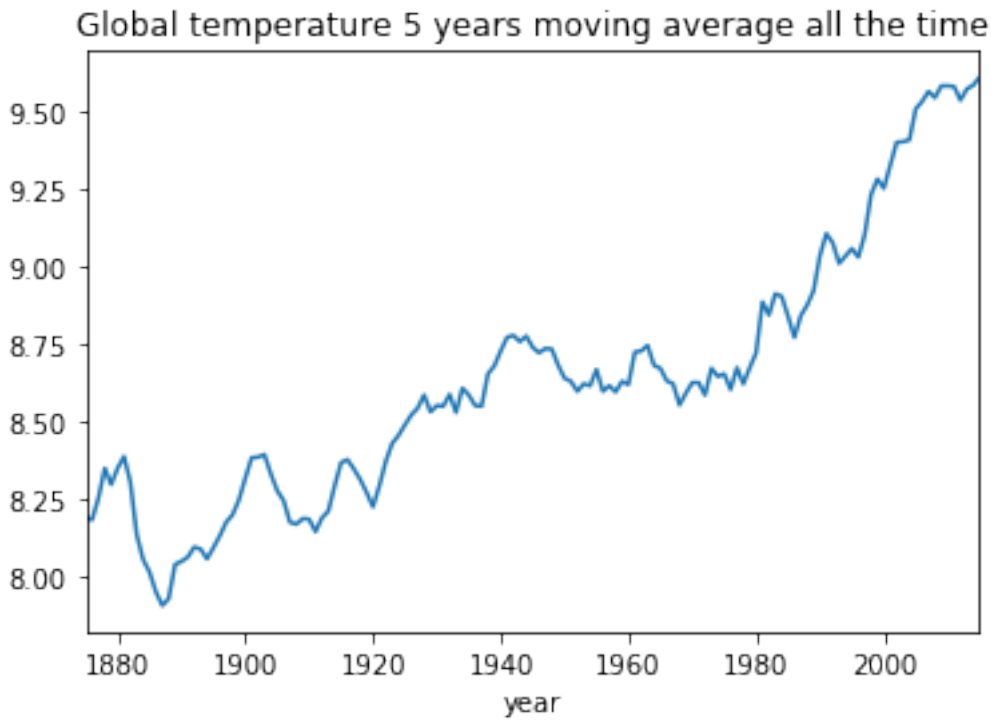
[12]:
```
#both at the same chart

x1 = df_city_filter['year']
y1 = df_city_filter['moving_average_5y']
plt.plot(x1, y1, label = "Rio de Janeiro")


x2 = df_global_filter['year']
y2 = df_global_filter['moving_average_5y']
plt.plot(x2, y2, label = "Global")

# Set the x and y axis label
plt.xlabel('Years')
plt.ylabel('Temperature (celsius)')


# Set a title of the current axes.
plt.title('Rio de Janeiro x Global temperature 5 years moving average since␣
 ↪1990')
# show a legend on the plot
plt.legend()
# Display a figure.
plt.show()
```



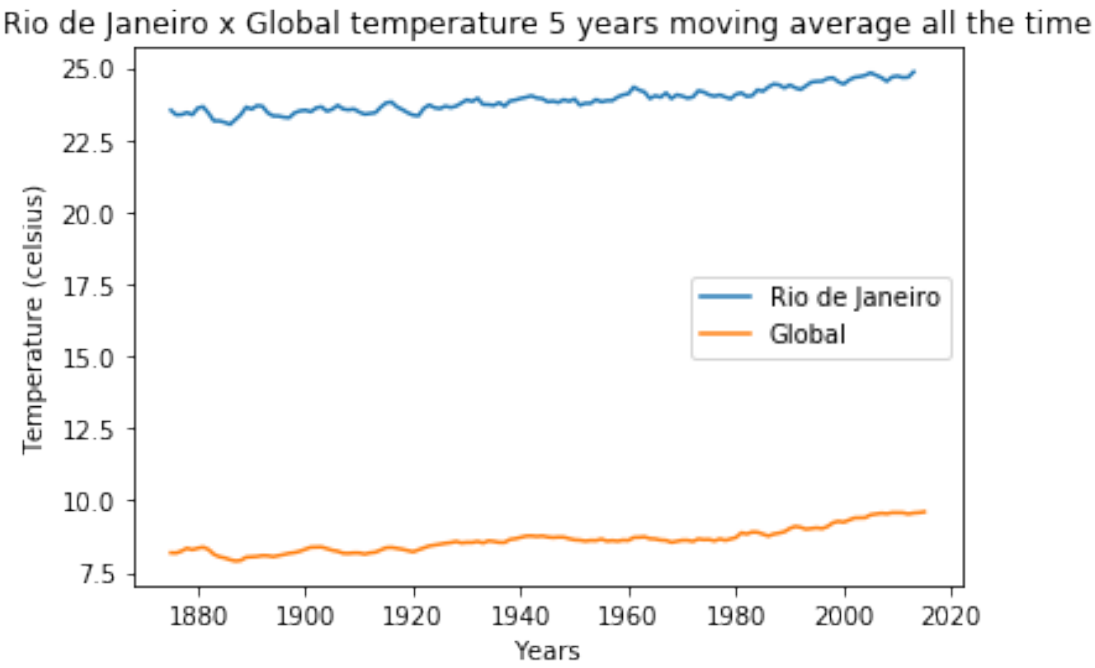
1.1.7 Comparing city and global temp ma - All the time

[13]:
```
#To compare, I decided to select the line graph to identify the curves for each base
title1 = 'Rio de Janeiro temperature 5 years moving average all the time'
title2 = 'Global temperature 5 years moving average all the time'

df_city_filter2.plot(x="year", y="moving_average_5y", kind="line", title =title1, legend = False);
df_global_filter2.plot(x="year", y="moving_average_5y", kind="line", title =title2, legend = False);
```





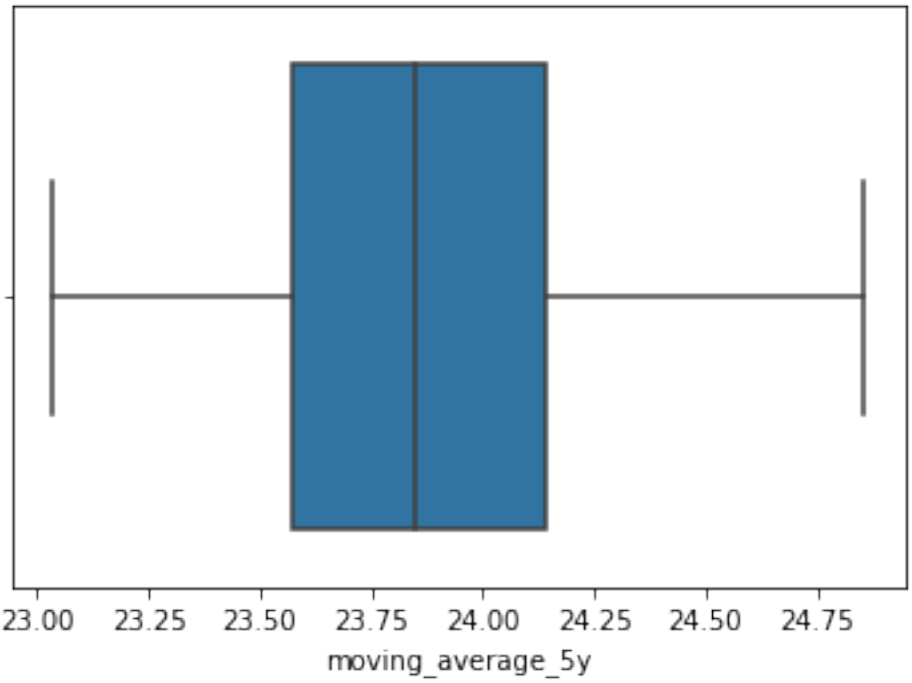
[15]:
```
#both at the same chart

x1 = df_city_filter2['year']
y1 = df_city_filter2['moving_average_5y']
plt.plot(x1, y1, label = "Rio de Janeiro")


x2 = df_global_filter2['year']
y2 = df_global_filter2['moving_average_5y']
plt.plot(x2, y2, label = "Global")

# Set the y axis label of the current axis.
plt.xlabel('Years')
plt.ylabel('Temperature (celsius)')
# Set a title of the current axes.
plt.title('Rio de Janeiro x Global temperature 5 years moving average all the␣
 ↪time')
# show a legend on the plot
plt.legend()
# Display a figure.
plt.show()
```



1.1.8 Checking correlation

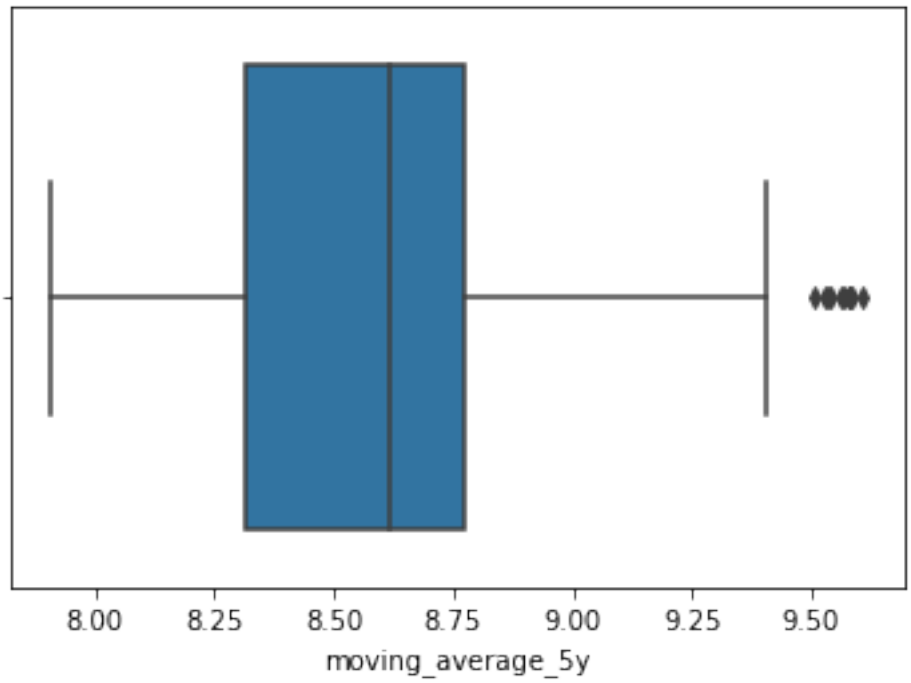
[16]:
```
#Check bloxplot for city to identify outliers
sns.boxplot(x=df_city_filter2['moving_average_5y'])
```

[16]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x4175170>
```



[17]:
```
#Check bloxplot for global to identify outliers
sns.boxplot(x=df_global_filter2['moving_average_5y'])
```

[17]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x4248fb0>
```



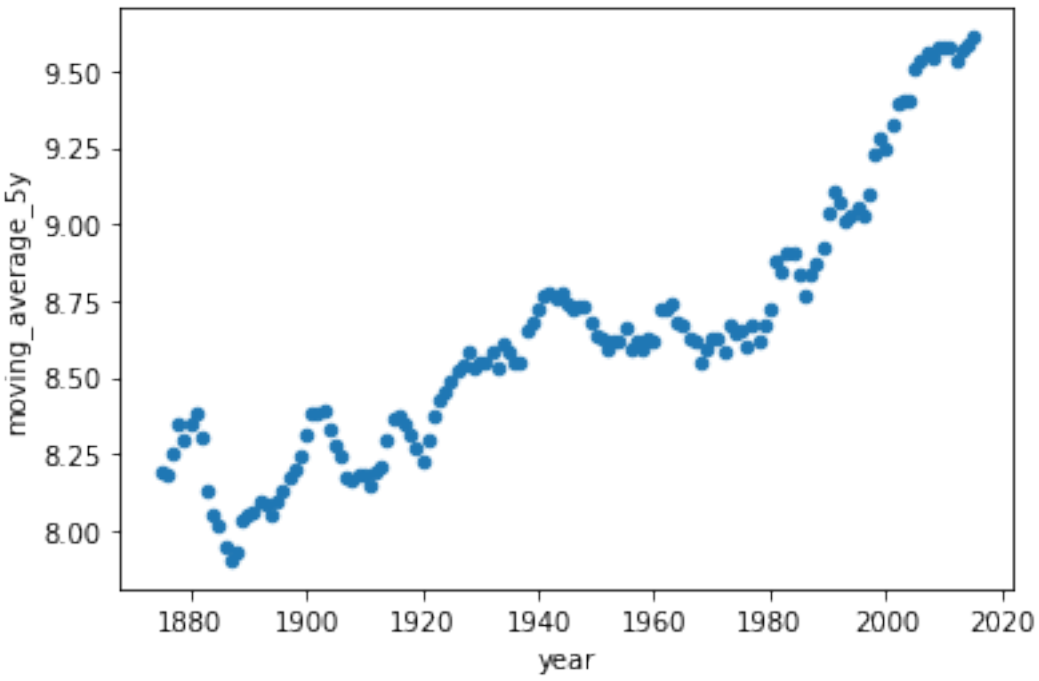
[18]:
```
#Calculate the IQR
Q1 = df_global_filter2['moving_average_5y'].quantile(0.25)
Q3 = df_global_filter2['moving_average_5y'].quantile(0.75)
IQR = Q3 - Q1
IQR
```

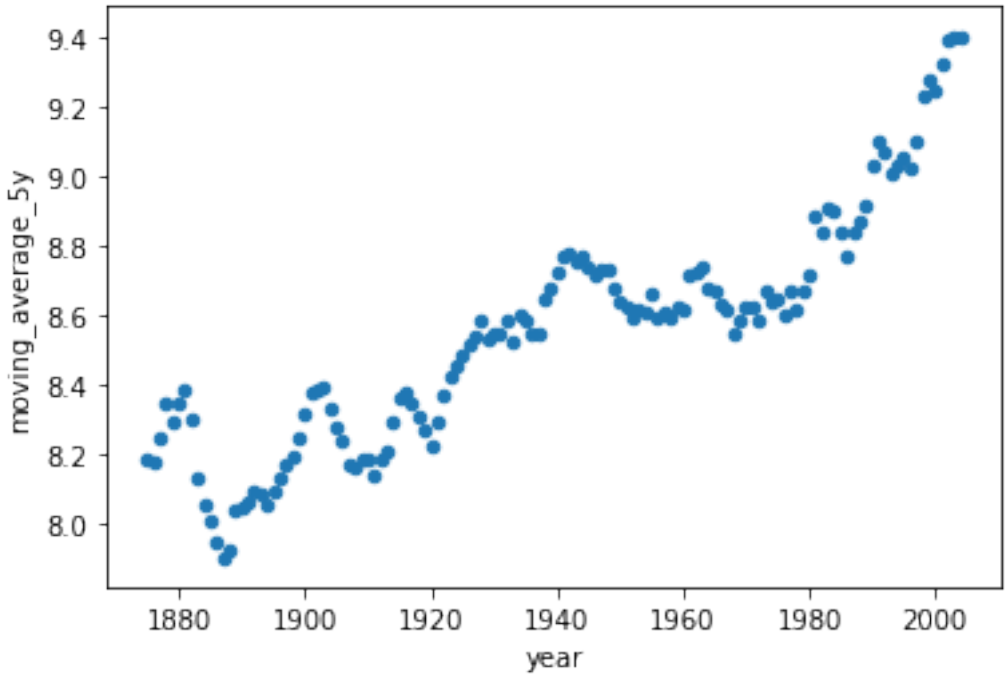
[18]:
```
0.4620000000000015
```

[19]:
```
#Create dataframe for global without outliers
df_global_clean = df_global_filter2[~((df_global_filter2['moving_average_5y'] <␣
↪(Q1 - 1.5 * IQR)) |(df_global_filter2['moving_average_5y'] > (Q3 + 1.5 *␣
↪IQR)))]
df_global_clean.shape

#11 outliers
```

[19]:
```
(130, 3)
```

[20]:
```
#checking scatter plot for global dataframes, to see the difference
df_global_filter2.plot(x="year", y="moving_average_5y", kind="scatter", legend␣
↪= False);
df_global_clean.plot(x="year", y="moving_average_5y", kind="scatter", legend =␣
↪False);
```





[27]:
```
#Calculate the correlation coefficient for city, using spearmanr
stats.spearmanr(df_city_filter2['year'],␣
 ↪df_city_filter2['moving_average_5y'])[0]
```

[27]:
```
0.9425329216454152
```

[28]:
```
#Calculate the correlation coefficient for global, using spearmanr
stats.spearmanr(df_global_clean['year'],␣
 ↪df_global_clean['moving_average_5y'])[0]
```

[28]:
```
0.9097670023839941
```

1.2 Conclusions

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

My city, Rio de Janeiro, is hotter when compared to the global average. The difference has remained constant over the years.

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

The temperature in my city has followed global temperature curves over the years. Even if we evaluate a more recent period, last 5 years for example, the behavior is very similar.

1.2.3 - 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 world has gotten hotter. In 1880 the global temperature was approximately 8.24 degrees and that number grew to 9.50 in 2015. It is an increase of about 15% degrees in temperature.

1.2.4 There is a correlation between time and temperature ?

When we look at time (years) and temperature, we can identify a strong positive correlation. Although this correlation exists, we cannot consider causality as it does not mean that the temperature increases just because time advances. Other factors affect this relationship, such as increased co2 emissions, melting glaciers, etc. To further study this correlation, we need other variables.