Exploring Weather Trends

1. **An outline** of steps taken to prepare the data to be visualized in the chart, such as:
    a. What tools did you use for each step? (Python, SQL, Excel, etc)

       SQL and Python (see Appendixes)

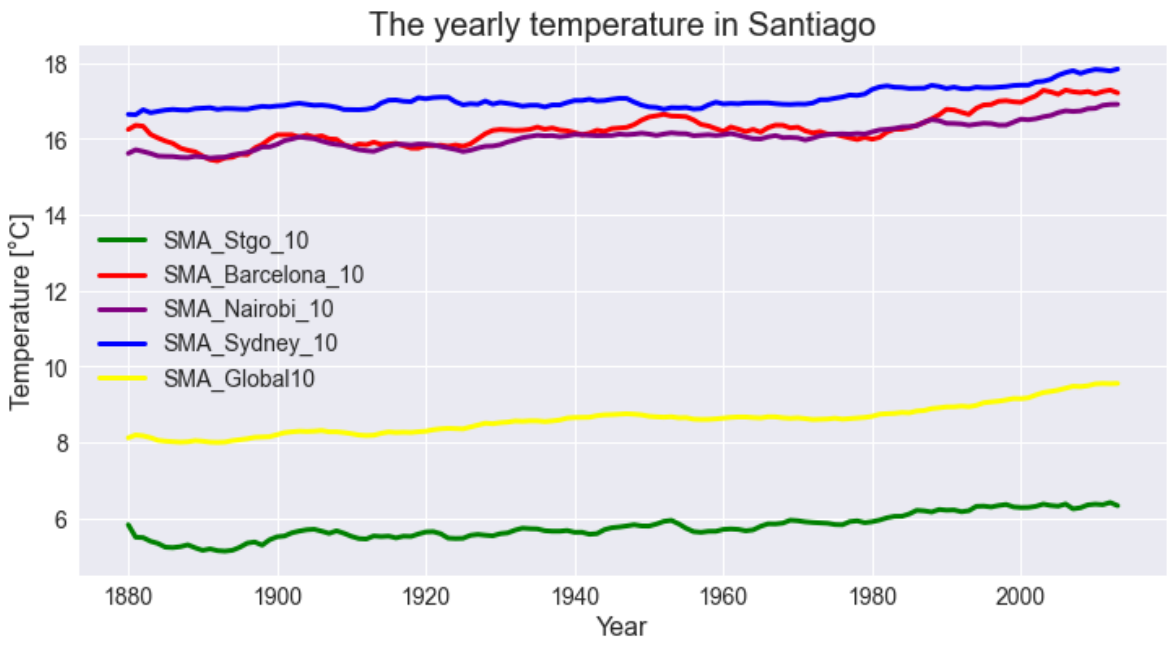
    b. How did you calculate the moving average?

       I used the pandas library. First, I tried to use a 20-year period, but finally I used a 10-year period, because the difference was less between one chart and another.

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

       To visualize the trends the most adapted graph was a linear one, because it shows the evolution of the variable under study

2. **Line chart** with local and global temperature trends



3. At least **four observations** about the similarities and/or differences in the trends

    a. a clear upward trend in temperatures can be appreciated, for different cities, as well as for the global temperature

b. Therefore, the linear regression between the years and the global temperature has a very high r2 (0,87)

c. the relationship between the temperatures of the different cities and the global one is linear (correlation table)

SMA	Stgo_10	Barcelona_10	Nairobi_10	Sydney_10	Global10
Stgo_10	1,00	0,89	0,94	0,84	0,93
Barcelona_10	0,89	1,00	0,90	0,75	0,92
Nairobi_10	0,94	0,90	1,00	0,88	0,97
Sydney_10	0,84	0,75	0,88	1,00	0,88
Global10	0,93	0,92	0,97	0,88	1,00

d. There is a linear relationship between global temperature and Santiago (Chile) temperature, when doing a simple linear regression between these variables the r2, is high (0,86)

e. Finally it is possible to conclude, that there has been an increase in temperatures at a global level, which is directly correlated to the rise in temperatures in the different cities reviewed.

4. Appendixes:

a. SQL

```
select
CD.year,
CD.T_Stgo,
CD.T_Barcelona,
CD.T_Nairobi,
CD.T_Sydney,
GD.T_global
from
        (
                select year,
                sum(case when country='Chile' and city='Santiago' then avg_temp else 0 end) as
T_Stgo,
                sum(case when country='Spain' and city='Barcelona' then avg_temp else 0 end)
as T_Barcelona,
                sum(case when country='Kenya' and city='Nairobi' then avg_temp else 0 end) as
T_Nairobi,
                sum(case when country='Australia' and city='Sydney' then avg_temp else 0 end)
as T_Sydney
                from
```

```
                    city_data
                group by
                year
            ) CD
left join (select year as ano,avg_temp as T_global
        from global_data) GD on CD.year = GD.ano

order by year asc

import csv
```

b. Python:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
plt.style.use('seaborn')

df=pd.read_csv('results.csv')

df['SMA_Stgo_10']=df['t_stgo'].rolling(10,min_periods=1).mean()
df['SMA_Barcelona_10']=df['t_barcelona'].rolling(10,min_periods=1).mean()
df['SMA_Nairobi_10']=df['t_nairobi'].rolling(10,min_periods=1).mean()
df['SMA_Sydney_10']=df['t_sydney'].rolling(10,min_periods=1).mean()
df['SMA_Global10']=df['t_global'].rolling(10,min_periods=1).mean()

colors = ['green', 'red', 'purple','blue','yellow']

df.plot(color=colors, linewidth=3, figsize=(12,6))

plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.legend(labels =df.columns, fontsize=14)

plt.title('The yearly temperature in Santiago', fontsize=20)
plt.xlabel('Year', fontsize=16)
plt.ylabel('Temperature [°C]', fontsize=16)

df.corr(method="pearson")

df.reset_index(inplace=True)
print(df.head(5))
df2=df[['SMA_Stgo_10','SMA_Global10']]
data=df2[['SMA_Global10']]
X_train = np.array(data)
y_train = df2['SMA_Stgo_10']
regr = linear_model.LinearRegression()
regr.fit(X_train, y_train)
y_pred = regr.predict(X_train)

print('Coefficients: \n', regr.coef_)
print('Independent term: \n', regr.intercept_)
print("Mean squared error: %.2f" % mean_squared_error(y_train, y_pred))
```

```
print('Variance score: %.2f' % r2_score(y_train, y_pred))
```