Project 1: Exploring Temperature Data

Lauran Hazan
lhazan@gmail.com

Overview & Notes

General notes about data and analysis presented here.

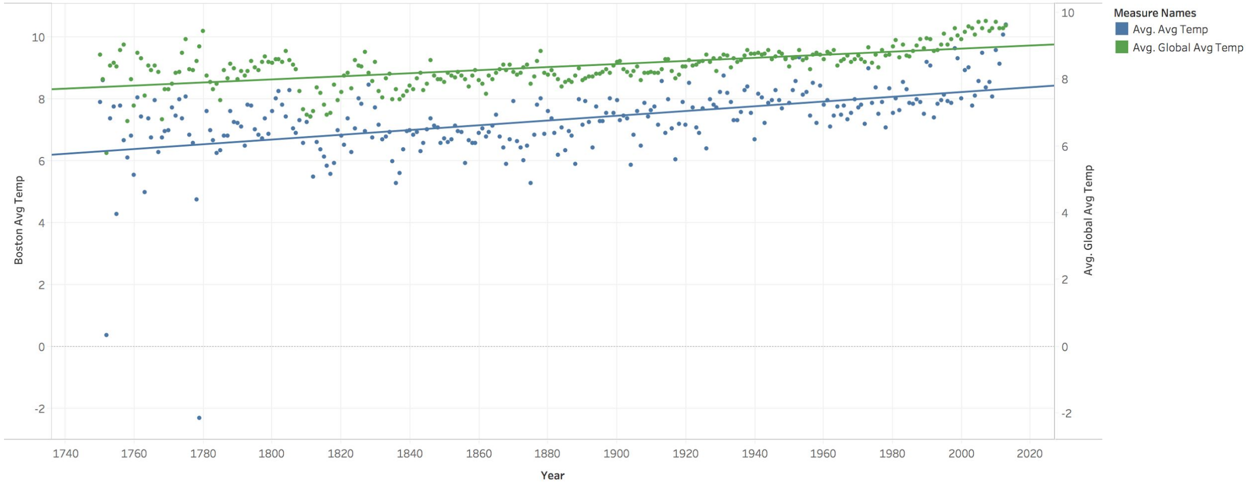
- Data manipulation and visualization done using Tableau Desktop v. 10.4: I love this tool especially for EDA because it’s quick and relatively easy to get a feel for the data and produces great visualizations.
- Reproduced calculations in Python 3.5(validate version) - to practice!
- In addition to project requirements, I included my preliminary EDA steps in the process.

1. EDA: Scatter Plot of Boston and World Temps

About This Step:

Wrangling	Query for Global Data:
	SELECT * FROM global_data;
	
	Query for city data for Boston:
	SELECT * FROM city_data
	WHERE city = 'Boston’

Manipulation	None (yet)
Observations	Boston’s measurements start in 1743, the world measurements start in 1750, so I we’ll filter all the data out from earlier than 1750. The trend lines have similar slopes, but Boston’s variability appears to be noticeably higher (which makes sense).
Notes	



Boston Average Temperatures - Scatter Plot

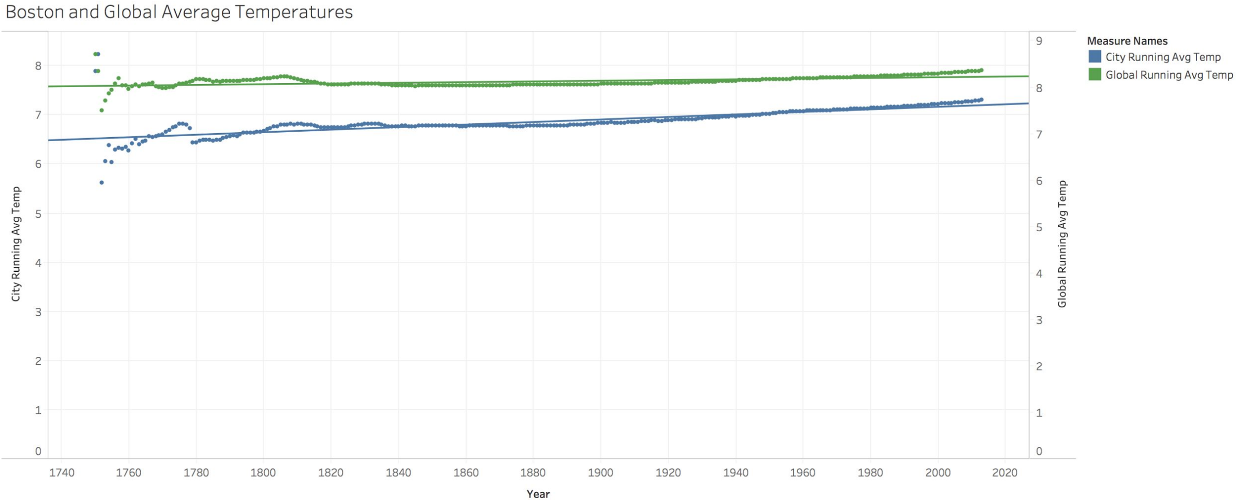
The plots of Avg. Avg Temp and Avg. Global Avg Temp for Year. Color shows details about Avg. Avg Temp and Avg. Global Avg Temp. The data is filtered on City and Year (global). The City filter keeps Boston. The Year (global) filter ranges from 1750 to 2015.

2. Line Graph of Boston and World Temps

Wrangling	Same as previous.

Manipulation	Running Average Calcs:
	RUNNING_AVG(Sum([Global Avg Temp]))
	RUNNING_AVG(Sum([Avg Temp]))

Observations	Boston's higher variability is basically invisible now. Trend is pretty clearly a linear increase for both, though Boston's is slightly steeper.
Notes	The data before 1780 looks strangely different than the later data.

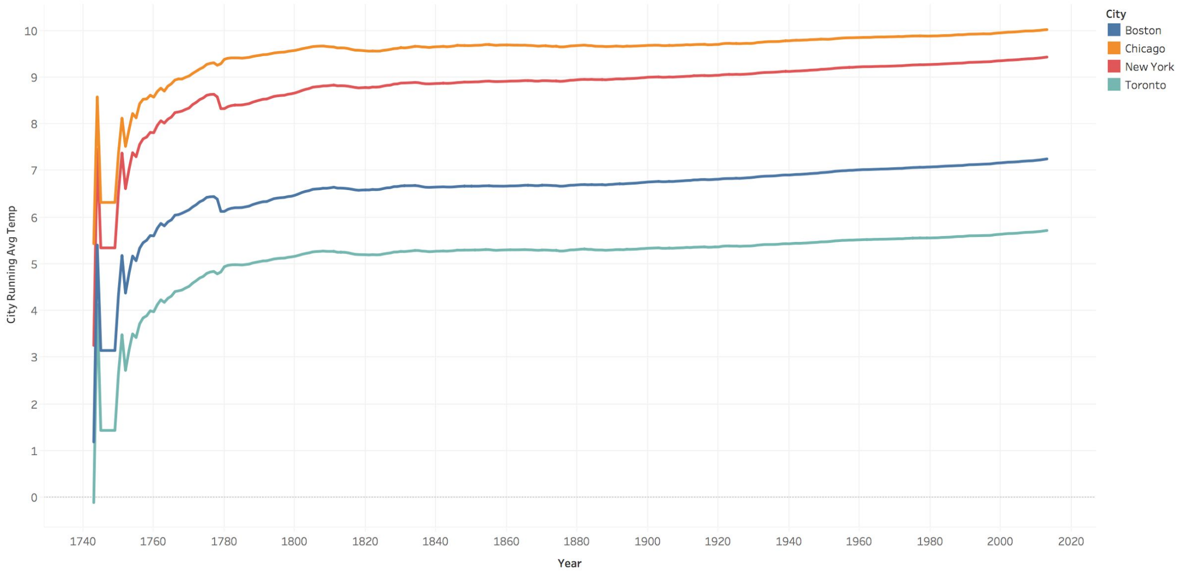

Boston and Global Average Temperatures

The plots of City Running Avg Temp and Global Running Avg Temp for Year. Color shows details about City Running Avg Temp and Global Running Avg Temp. The data is filtered on City and Year (global). The City filter keeps Boston. The Year (global) filter ranges from 1750 to 2015.

3. EDA - Selected Cities

Wrangling	Same as previous.
Manipulation	
Observations	4 Cities' running averages all display data that looks dramatically different prior to 1780. This could be a sudden spike in temperature, or more likely improvement in measurement instrumentation. Can't really tell for sure though.
Notes	It's probably a good idea to filter out data prior to 1780.

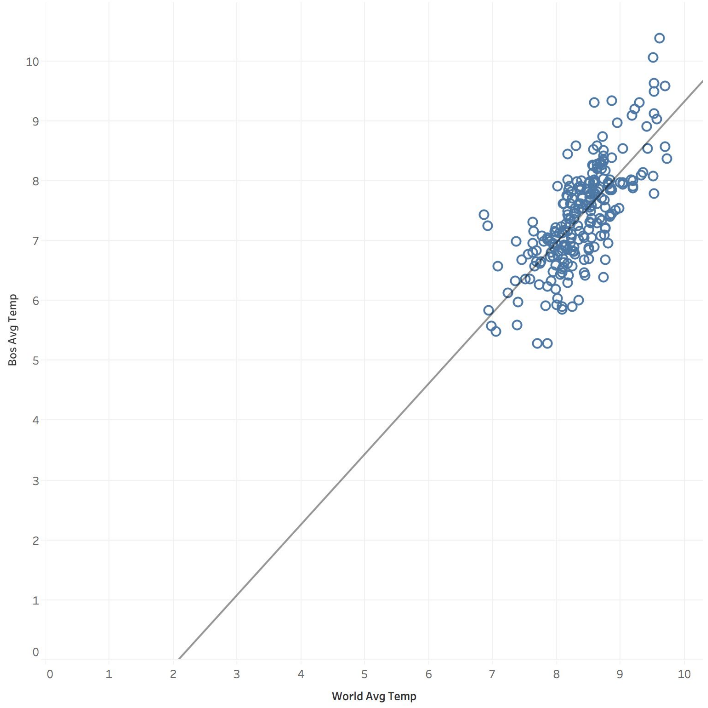
EDA - My Cities



The trend of City Running Avg Temp for Year. Color shows details about City. The view is filtered on City, which keeps Boston, Chicago, New York and Toronto.

4. Correlation

Wrangling	I created a single CSV with Boston and world data only, so I could work with it in Python.
Manipulation	Deleted data outside years 1781 - 2013 so I'd have data points with both temps associated. Used Tableau's native CORR function but also tested that the sq root of the r-squared value matched the output of CORR.
Observations	Data points themselves have a high positive linear correlation.
Notes	



Correlation Scatter Plot

World Avg Temp vs. Bos Avg Temp.

Correlation
0.737438348

Appendix: Description of Trend Models Automatically Generated by Tableau (Scatter)

Boston:

A linear trend model is computed for average of Avg Temp given Year. The model may be significant at p <= 0.05.

Model formula: Measure Names*(Year + intercept)
Number of modeled observations: 263
Number of filtered observations: 1
Model degrees of freedom: 2
Residual degrees of freedom (DF): 261
SSE (sum squared error): 260.57
MSE (mean squared error):
0.998351
R-Squared:
0.256387
Standard error:
0.999175
p-value (significance):
< 0.0001

Global:

A linear trend model is computed for average of Global Avg Temp (global) given Year. The model may be significant at p <= 0.05.

Model formula: Measure Names*(Year + intercept)
Number of modeled observations: 264
Number of filtered observations: 0
Model degrees of freedom: 2
Residual degrees of freedom (DF): 262
SSE (sum squared error): 54.4088
MSE (mean squared error):
0.207667
R-Squared:
0.374684
Standard error:
0.455705
p-value (significance):
< 0.0001