Exploratory Data Analysis of Temperatures in South American Cities Since 1960

John Bang, Stephanie Owyang
Github: https://github.com/gowrath/Project2_SOwyang_JBang

Introduction

Existing literature from a comprehensive research book, Global Urban Heat Island Mitigation, explains how the growth of megacities produces an urban heat island effect, wherein surface temperatures are higher in cities due to the density of human structures.

Exploratory Questions

1. How does each city compare to the rest of South America?
2. Which cities have higher temperature anomalies over time?
3. Do the temperature trends correlate to population trends?
4. Is there a particular season where the anomaly increased or decreased faster than others?

Source Data

Our initial analysis stemmed from the opening of the Berkeley Earth dataset of global temperatures. It provides high-resolution land and ocean time series data and gridded temperature data (Berkeley, 2024). The dataset has data stretching from 1850, which we did a preliminary analysis of. We finalized on studying the years past 1960 due to a secondary dataset we used, which we will detail in a subsequent section.

We analyzed a subset of the data, focusing on temperatures in South American cities since 1900. We chose cities instead of countries because of the granularity of the data; in addition, the three cities chosen—Lima, Santiago, and Sao Paulo—are the three megacities in their respective countries (Peru, Chile, and Brazil). Again, our approach was guided by a background assumption that dense concentrations of pavement and buildings may exacerbate heat waves.

The time series data set had columns for monthly, annual, five-year, to twenty-year anomalies (fluctuations, or moving averages) and uncertainty, which is a plus/minus range of the anomaly. While comprehensive, most of the anomalies were missing for five-year, ten-year, and twenty year moving averages. So we settled on using the monthly and annual anomalies, because of the limitations of our dataset.

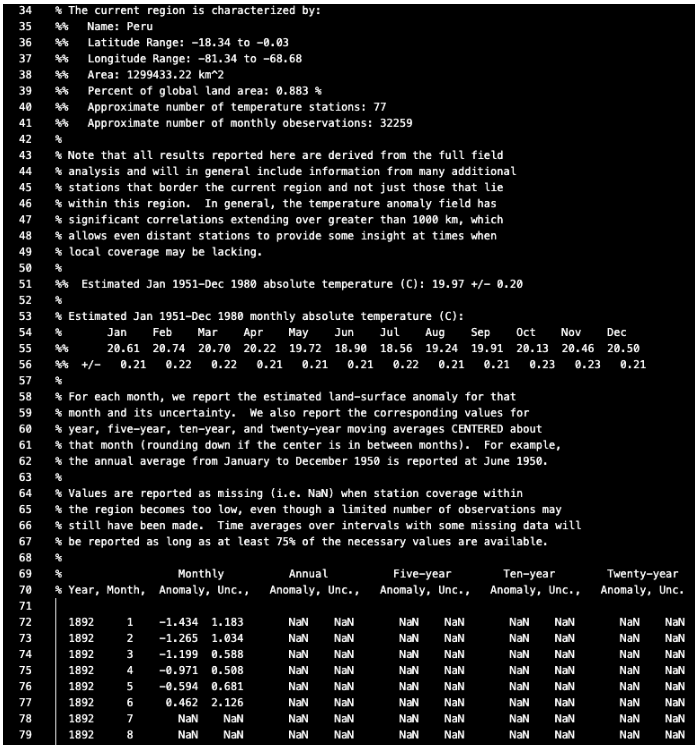
Data Structure

There are 1980 observations and 12 columns. There are 12 variables including:

- **year:** The year the data was collected or calculated
- **month:** The month the data was collected or calulated
- **monthly_anom:** This is the monthly anomaly, where the anomalies are reported relative to the historical average.
- **Monthly_unc:** This is the monthly uncertainty which represents the 95% confidence interval for statistical noise and spatial undersampling effect.
- **one_yr_anom**: Moving average centered about that month's anomaly over one year
- **one_yr_unc**: Moving average centered about that month's uncertainty over one year
- **five_yr_anom**: Moving average centered about that month's anomaly over five years
- **five_yr_unc**: Moving average centered about that month's uncertainty over five years
- **ten_yr_anom**: Moving average centered about that month's anomaly over ten years
- **ten_yr_unc**: Moving average centered about that month's uncertainty over ten years
- **twenty_yr_anom**: Moving average centered about that month's anomaly over twenty years
- **twenty_yr_unc**: Moving average centered about that month's uncertainty over twenty years

Out of these variables, we focused on year, month one_yr_anom, one_yr_unc, five_yr_anom, and five_yr_unc. We looked at all variables to compare the trends, although we knew there was more data in the smaller year moving averages.

Data Cleaning and Sanity Checks



```
% The current region is characterized by:
%%   Name: Peru
%%   Latitude Range: -18.34 to -0.03
%%   Longitude Range: -81.34 to -68.68
%%   Area: 1299433.22 km^2
%%   Percent of global land area: 0.883 %
%%   Approximate number of temperature stations: 77
%%   Approximate number of monthly obeservations: 32259
%
% Note that all results reported here are derived from the full field
% analysis and will in general include information from many additional
% stations that border the current region and not just those that lie
% within this region.  In general, the temperature anomaly field has
% significant correlations extending over greater than 1000 km, which
% allows even distant stations to provide some insight at times when
% local coverage may be lacking.
%
%%  Estimated Jan 1951-Dec 1980 absolute temperature (C): 19.97 +/- 0.20
%
% Estimated Jan 1951-Dec 1980 monthly absolute temperature (C):
%       Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
%%      20.61  20.74  20.70  20.22  19.72  18.90  18.56  19.24  19.91  20.13  20.46  20.50
%%  +/-   0.21   0.22   0.22   0.21   0.21   0.21   0.22   0.21   0.21   0.23   0.23   0.21
%
% For each month, we report the estimated land-surface anomaly for that
% month and its uncertainty.  We also report the corresponding values for
% year, five-year, ten-year, and twenty-year moving averages CENTERED about
% that month (rounding down if the center is in between months).  For example,
% the annual average from January to December 1950 is reported at June 1950.
%
% Values are reported as missing (i.e. NaN) when station coverage within
% the region becomes too low, even though a limited number of observations may
% still have been made.  Time averages over intervals with some missing data will
% be reported as long as at least 75% of the necessary values are available.
%
%                 Monthly          Annual          Five-year        Ten-year        Twenty-year
% Year, Month,  Anomaly, Unc.,   Anomaly, Unc.,   Anomaly, Unc.,   Anomaly, Unc.,   Anomaly, Unc.

  1892     1    -1.434  1.183       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
  1892     2    -1.265  1.034       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
  1892     3    -1.199  0.588       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
  1892     4    -0.971  0.508       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
  1892     5    -0.594  0.681       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
  1892     6     0.462  2.126       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
  1892     7       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
  1892     8       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN       NaN    NaN
```

The data sets were originally text files, one for each geographic region: the continent of South America, and the cities of Lima (Peru), Santiago (Chile) and Sao Paulo (Brazil). Each file had a comment block at the start of the file with varying length before the data was presented. While the comment blocks gave us a lot of background context and useful details about the area of study, we needed to filter and clean a bit before we were able to start manipulating and analyzing the data.

To handle the leading comment, we were able to use the skip_rows argument in pandas.read_csv() to get rid of the header. Unfortunately, this also meant skipping the column names, but we just defined them similarly when making our dataframe.

In order to read in the text file, we needed to specify a separator, and found that between values in each row, the spacing was inconsistent. Just the separator argument was not enough, so we had to individually parse each row. Each row was one string, and we used .split() to get each individual value.

The Sao Paulo data forced us to change the way we read in the file, because there was an encoding error. When looking at the file, we could not find any non-ASCII characters that may cause an issue. Adding the latin-1 encoding fixed the issue here.

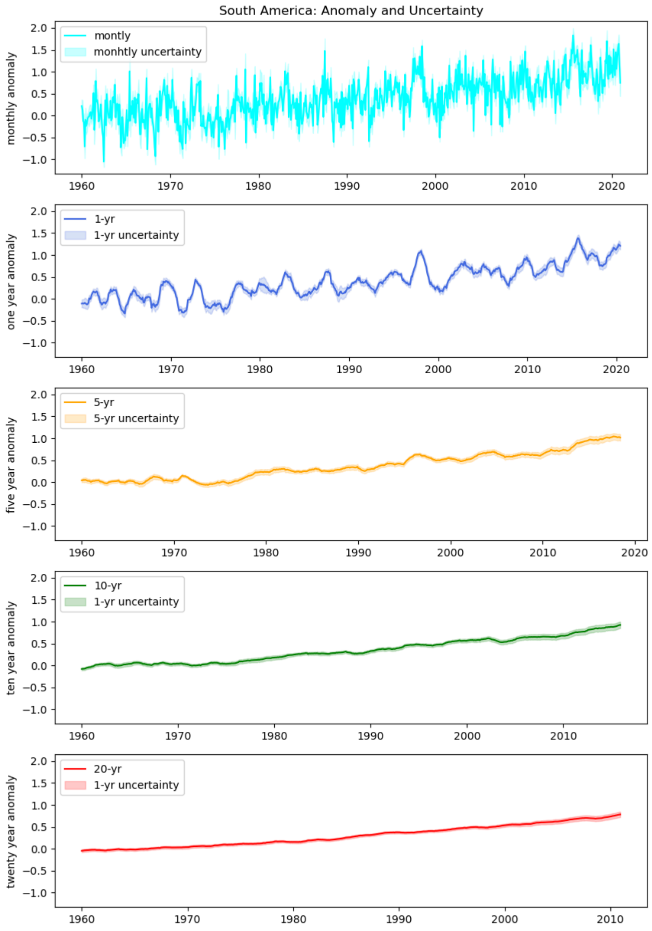
All of the values were imported as strings, so we had to cast all of the anomaly and uncertainty data as floats. We also used the year and month columns to make a datetime timestamp.

Sanity Checks

After we imported each text file, we checked the .shape of each dataframe and saw that the number of rows matched the number of data rows in the text files. Likewise, we made sure the number of columns matched the number of columns we defined.

We could see that there were a lot of NaN's in all of the files, since the data went back to the 1800's and included 10 and 20 year moving averages. A field was also specified as NaN if there were not enough weather stations to accurately calculate an anomaly.

As a sanity check we graphed the data for each region since the files are relatively small. Here we looked for outliers or missing data. Here is an example of the South America data:



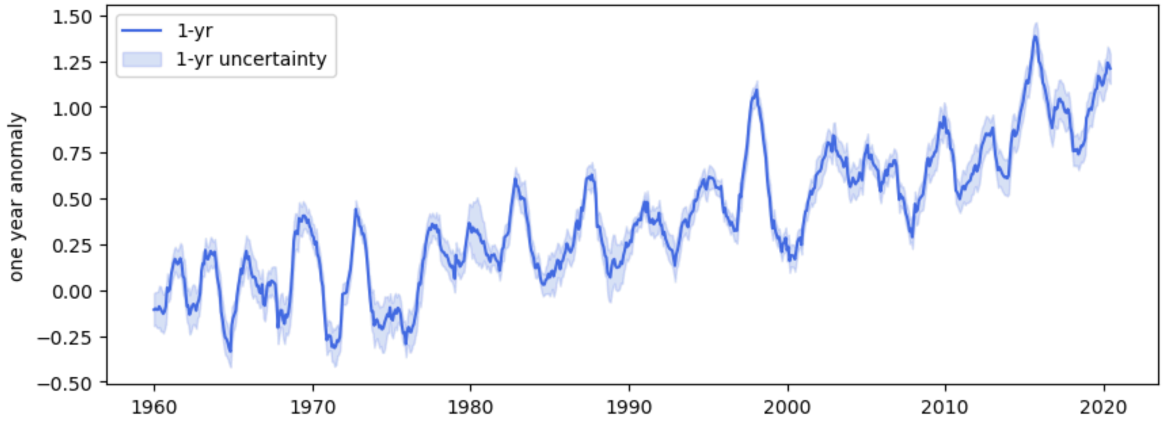
Assumptions

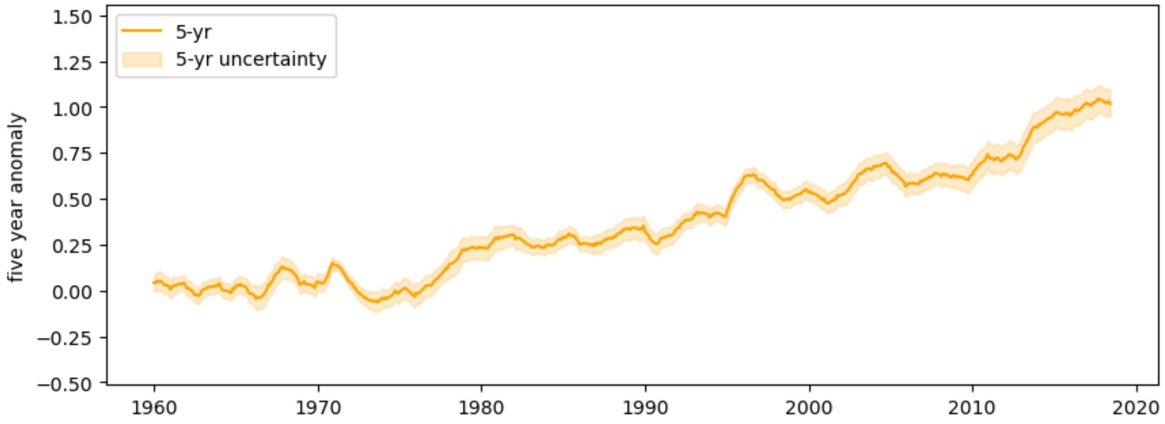
We had some initial and ongoing assumptions about the data. There are not enough data for 10-20 year anomalies because of temperature stations not coming online for 2020 (that would be data in the future, so that's not available). That data could be meaningful, but is omitted for our analysis. In addition, with the paucity of data before 1960 for the 10-year and 20-year anomalies, we dropped those columns so a sense of long-term horizon is missing.

Also an assumption is that anomalies were accurately calculated, or even the baseline was correct in the first place. Uncertainty values also denote that the researchers are aware of the variability in temperatures. We had to use certain time frames for our analysis for a smoother curve and to make better sense of the data.

Initial Exploration

At first, we expected to find insights on the temperature—for annual anomalies to increase over the last 100 years. We also were interested in the uncertainty values because we still did not have a clear picture of what that represented.



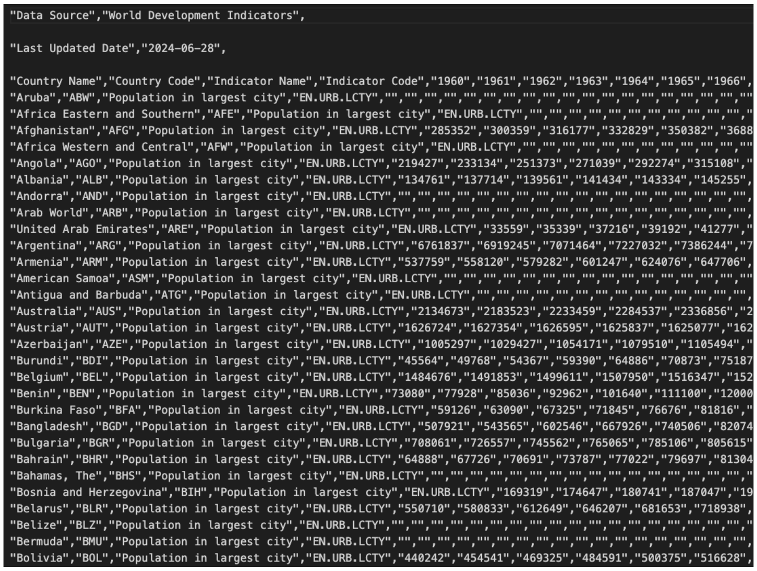


In general, you can see from the charts (created using matplotlib) that anomalies are going up over time in the last 100 years. Shown above are the annual average and five year average. The annual average fluctuates quite a bit from year to year which is expected, but looking at the five year average, we can clearly see an upward trend of positive temperature anomalies after the 1980's.

The uncertainty is much higher at the beginning of the dataset prior to 1950. In both the annual and five year anomaly charts, the shaded region is very pronounced, and decreases sharply in the 1950's and 1960's. On the Berkely Earth website, we found that the number of active weather stations drastically increased in the 1950's and 1960's, likely contributing to the decrease in uncertainty. Since they could measure the temperature of more locations, they could report the anomaly more confidently.

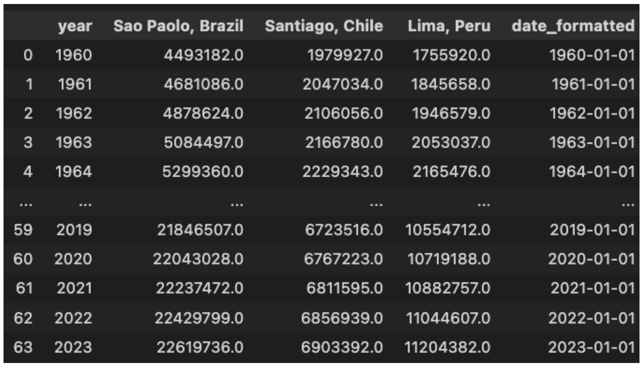
Supplemental Dataset on Population

Then, we added another dataset to further inform our EDA. The supplemental database we used was a population dataset of the cities. We wanted to map that to temperature data. To answer the question on whether the temperature trends correlate to population trends, we obtained population data from the World Bank. In order to work with this data, we had to skip rows, filter out other countries and cities, transpose the columns, and format the dates for analysis.



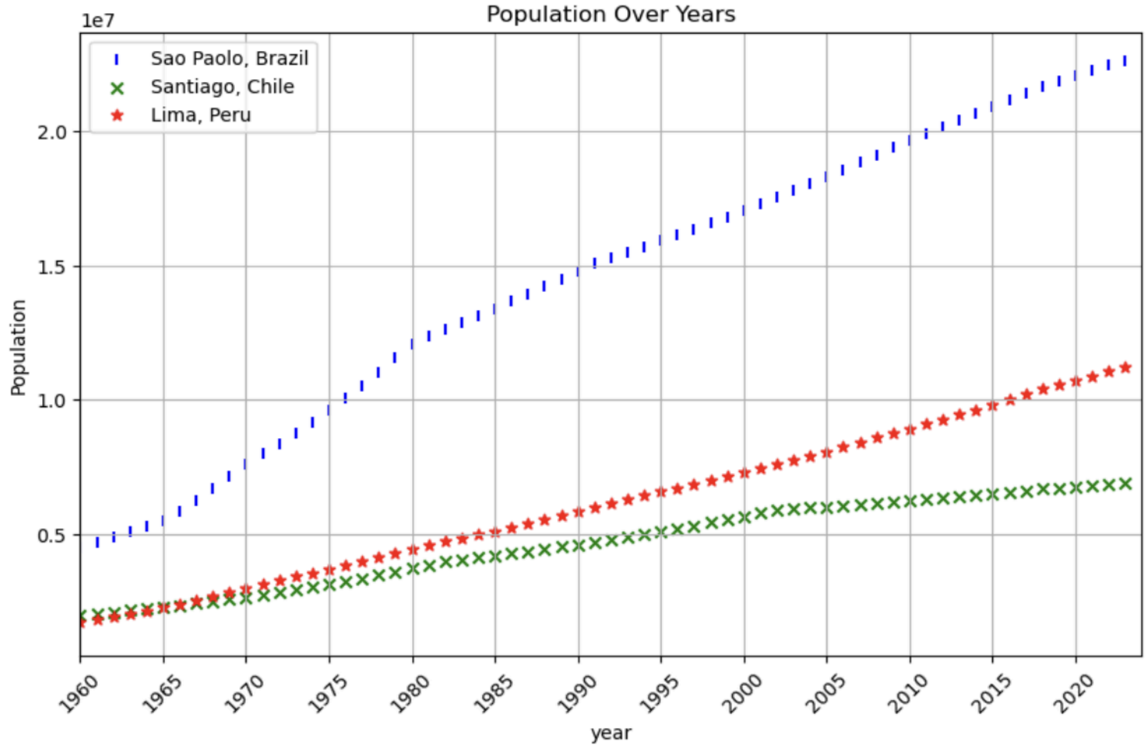
```
"Data Source","World Development Indicators",

"Last Updated Date","2024-06-28",

"Country Name","Country Code","Indicator Name","Indicator Code","1960","1961","1962","1963","1964","1965","1966",
"Aruba","ABW","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","","","","",""
"Africa Eastern and Southern","AFE","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","
"Afghanistan","AFG","Population in largest city","EN.URB.LCTY","285352","300359","316177","332829","350382","3688
"Africa Western and Central","AFW","Population in largest city","EN.URB.LCTY","","","","","","","","","","","",""
"Angola","AGO","Population in largest city","EN.URB.LCTY","219427","233134","251373","271039","292274","315108","
"Albania","ALB","Population in largest city","EN.URB.LCTY","134761","137714","139561","141434","143334","145255",
"Andorra","AND","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","","","","",
"Arab World","ARB","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","","","",
"United Arab Emirates","ARE","Population in largest city","EN.URB.LCTY","33559","35339","37216","39192","41277","
"Argentina","ARG","Population in largest city","EN.URB.LCTY","6761837","6919245","7071464","7227032","7386244","7
"Armenia","ARM","Population in largest city","EN.URB.LCTY","537759","558120","579282","601247","624076","647706",
"American Samoa","ASM","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","","",""
"Antigua and Barbuda","ATG","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","",
"Australia","AUS","Population in largest city","EN.URB.LCTY","2134673","2183523","2233459","2284537","2336856","2
"Austria","AUT","Population in largest city","EN.URB.LCTY","1626724","1627354","1626595","1625837","1625077","162
"Azerbaijan","AZE","Population in largest city","EN.URB.LCTY","1005297","1029427","1054171","1079510","1105494","
"Burundi","BDI","Population in largest city","EN.URB.LCTY","45564","49768","54367","59390","64886","70873","75187
"Belgium","BEL","Population in largest city","EN.URB.LCTY","1484676","1491853","1499611","1507950","1516347","152
"Benin","BEN","Population in largest city","EN.URB.LCTY","73080","77928","85036","92962","101640","111100","12000
"Burkina Faso","BFA","Population in largest city","EN.URB.LCTY","59126","63090","67325","71845","76676","81816","
"Bangladesh","BGD","Population in largest city","EN.URB.LCTY","507921","543565","602546","667926","740506","82074
"Bulgaria","BGR","Population in largest city","EN.URB.LCTY","708061","726557","745562","765065","785106","805615"
"Bahrain","BHR","Population in largest city","EN.URB.LCTY","64888","67726","70691","73787","77022","79697","81304
"Bahamas, The","BHS","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","","","","
"Bosnia and Herzegovina","BIH","Population in largest city","EN.URB.LCTY","169319","174647","180741","187047","19
"Belarus","BLR","Population in largest city","EN.URB.LCTY","550710","580833","612649","646207","681653","718938",
"Belize","BLZ","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","","","","","
"Bermuda","BMU","Population in largest city","EN.URB.LCTY","","","","","","","","","","","","","","","","","","",
"Bolivia","BOL","Population in largest city","EN.URB.LCTY","440242","454541","469325","484591","500375","516628",
```

When we cleaned and reset the index, we got this dataframe for further analysis.



	year	Sao Paolo, Brazil	Santiago, Chile	Lima, Peru	date_formatted
0	1960	4493182.0	1979927.0	1755920.0	1960-01-01
1	1961	4681086.0	2047034.0	1845658.0	1961-01-01
2	1962	4878624.0	2106056.0	1946579.0	1962-01-01
3	1963	5084497.0	2166780.0	2053037.0	1963-01-01
4	1964	5299360.0	2229343.0	2165476.0	1964-01-01
...	...	...	...	...	...
59	2019	21846507.0	6723516.0	10554712.0	2019-01-01
60	2020	22043028.0	6767223.0	10719188.0	2020-01-01
61	2021	22237472.0	6811595.0	10882757.0	2021-01-01
62	2022	22429799.0	6856939.0	11044607.0	2022-01-01
63	2023	22619736.0	6903392.0	11204382.0	2023-01-01

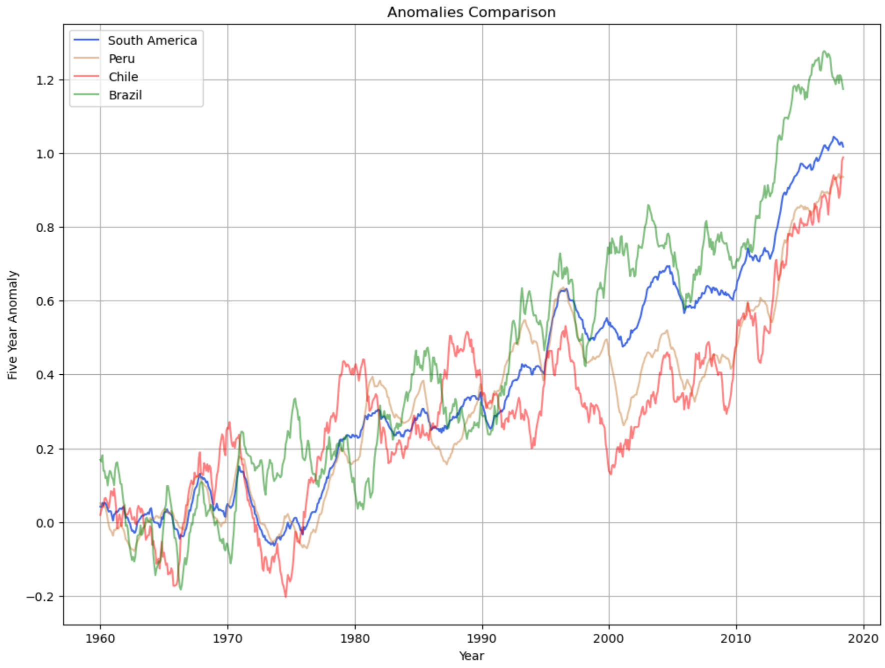
We then plotted the population data using matplotlib:



Results

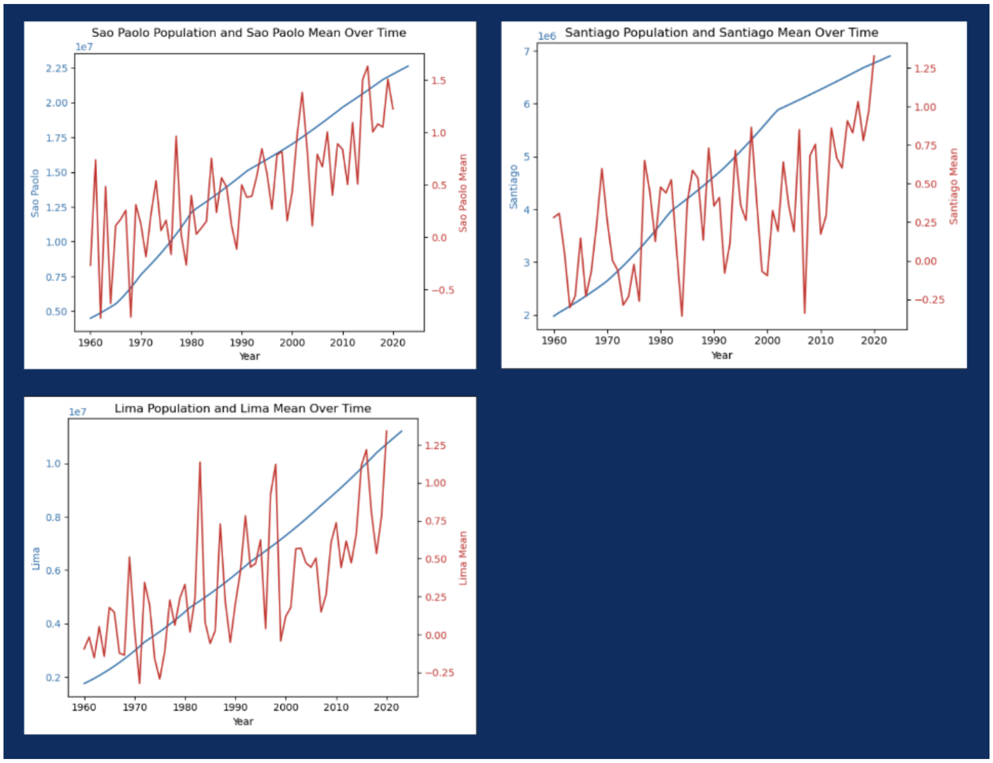
How does each city compare to the rest of South America? Which cities have higher temperature anomalies over time?

Comparing the cities' 5-year anomalies to the South America's 5-year anomaly we see an upward trend for all of them. Chile and Peru follows the South America trends through the 1990's. In 2010, we see that all 3 countries have an increased rate of increasing anomalies, beginning in 2000. Brazil has the highest anomalies of the 3 countries, and is nearly 0.2 degrees higher.

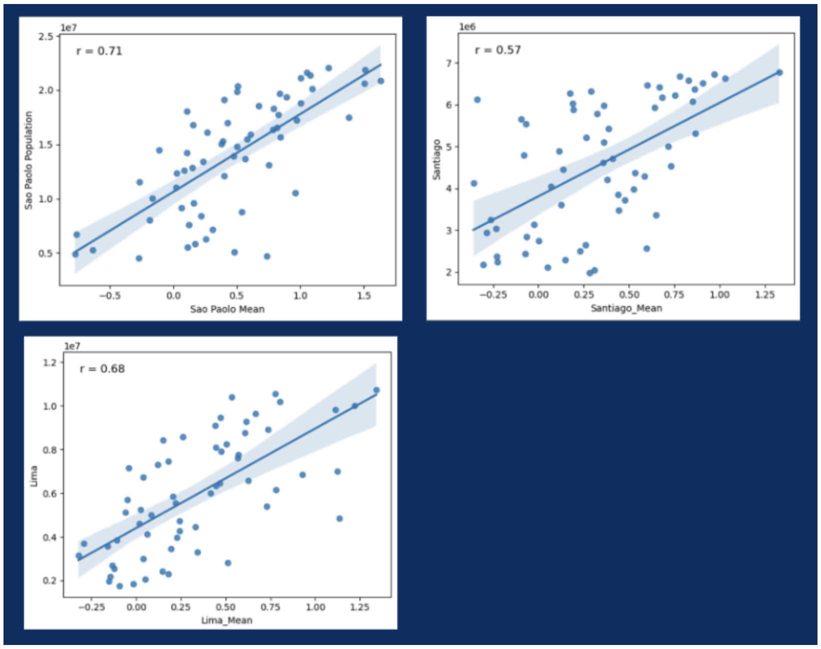


Do the temperature trends correlate to population trends?

To compare the temperature anomalies, to the countries populations, we overlaid population onto the anomaly data, and saw if a correlation could be found.



We also found the Pearson correlation as r.

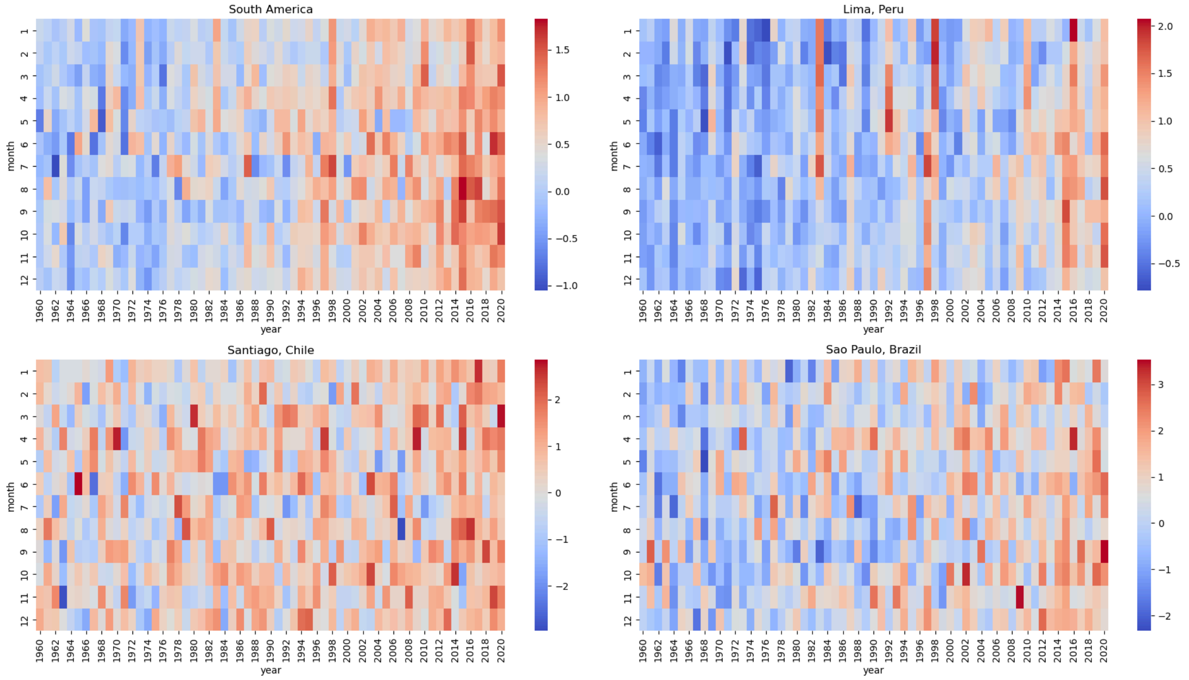


The Pearson R correlation we found was between 0.57 and 0.71. This measures whether temperature vs population are linearly correlated with a perfect positive correlation of 1 or -1. This is not the same as an r-squared value, which a separate analysis would calculate.

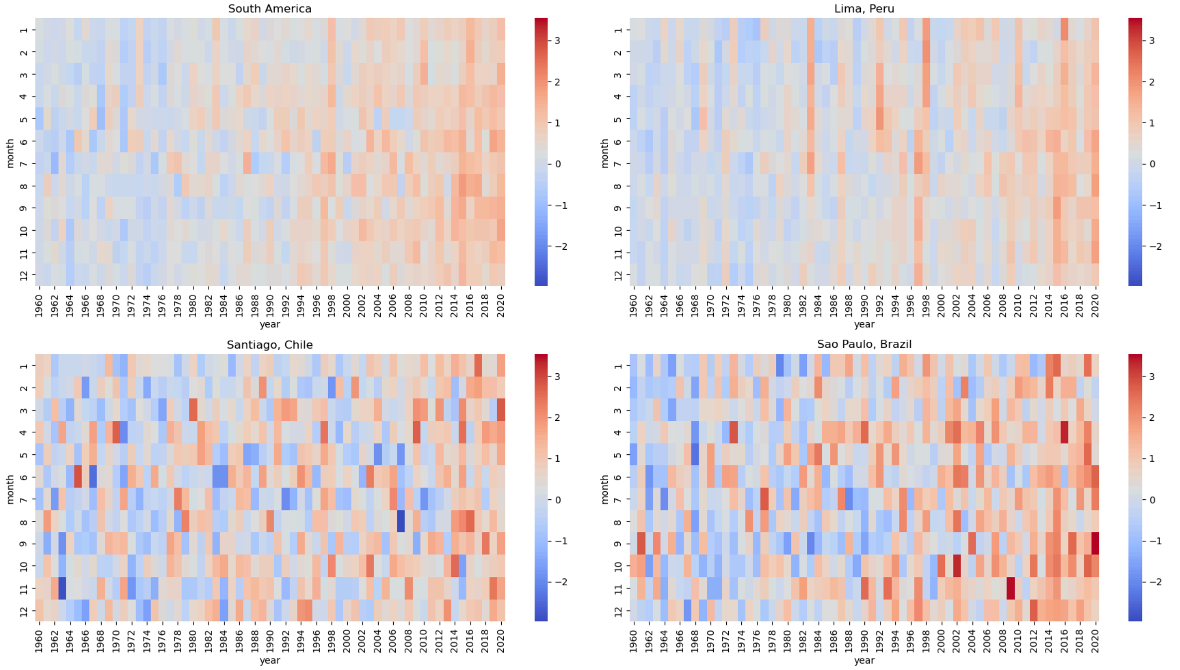
There is evidence to suggest that there is a correlation between population increase and temperature anomalies that is closer to linear than not. However, the two variables are not totally related linearly, which indicates there are factors not included in our analysis. In general, we saw that there is evidence of a linear correlation between temperature increase (anomalies going higher in value) and population growth.

Is there a particular season where the anomaly increased or decreased faster than others?

For seasonality, we looked at the monthly anomalies between 1960 and 2020. Looking at the South American data we can see the anomalies started to increase faster between the months of May and August, which for the Southern Hemisphere is the winter. We see that these months start turning red, and deeper shades of red, indicating a higher increase in temperature anomalies. Looking at the individual countries, this does not seem to hold for the more Southern cities of Santiago and Sao Paulo. Lima is warming more slowly than the South America region as a whole, but we can see that it is also starting to warm in the middle of the year during the winter months starting in 2006.



To compare the countries, we scaled the heat map scale so they yield similar conclusions. South America is in the top left corner, and we see a much less drastic change than the previous slide. Chile and Brazil have much higher extremes. Looking at each country, it does not seem that winter months see the highest increase in temperature anomalies for all countries. However, we do see that beginning in the 2000s, all countries have increasing temperature anomalies independent of seasonality.



Conclusion

Our exploratory data analysis yielded some interesting results, where we can see there is an increase in positive temperature anomalies in South America. The line plots showed a clear upward trend in all time scales such as the monthly and one-year moving average. Exploring the temperature anomalies grouped by month and season showed how extreme some of the southern cities' temperature anomalies can swing year to year. The line plot of the monthly temperature anomalies was very noisy, but when viewed as a heat map, each region showed significant warming in the last few decades.

We see one of the contributing factors could be increasing population, but there are other factors that we did not capture. Some limitations are better population data pre-1960, and the prevalence of climate patterns such as El Nino that would affect anomalies, especially if climate change continues to worsen over time. In addition, the addition of more temperature stations and more accurate readings might shift the values up or down, but it is not clear if that would be significant to alter our findings. In addition, we believe that more cities should be studied, and especially how such anomalies impact local communities. We look forward to further research in this area.