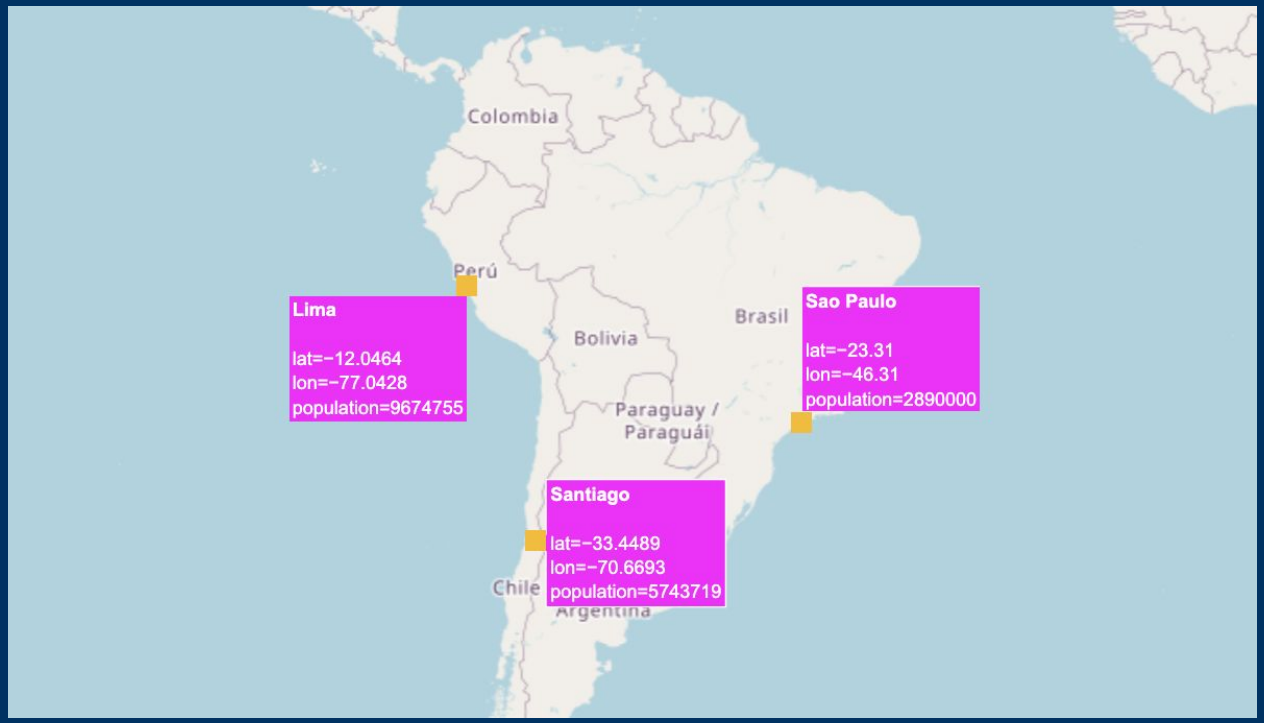
Exploratory Data Analysis of Temperatures in South American Cities Since 1960

John Bang
Stephanie Owyang

South American Cities — A Study Over Time

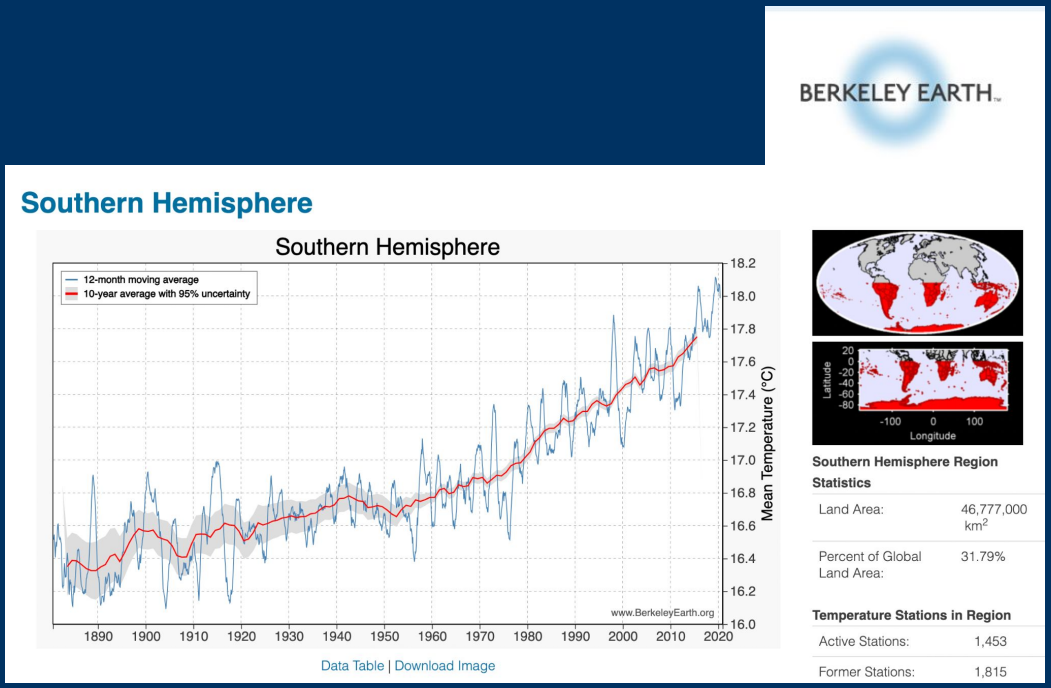


Several South American cities in Brazil, Chile, and Peru have become megacities in the 20th century

Some guiding research was how human activity might be related to temperature change

Berkeley

DATA SET | BERKELEY EARTH



Berkeley Earth

Independent U.S. non-profit organization focused on environmental data science and analysis.

Primary Data Sets

South America

Lima, Peru

Santiago, Chile

São Paulo, Brail

Berkeley

John

DATA SET | BERKELEY EARTH

Year	Month	Monthly Anomaly	Monthly Unc.	Annual Anomaly	Annual Unc.	Five-year Anomaly	Five-year Unc.	10-year Anomaly	10-year Unc.	20-year Anomaly	20-year Unc.
1834	6	-1.499	1.648	-0.636	0.484	-0.741	0.483	-0.828	0.413	NaN	NaN
1834	7	0.089	1.22	-0.6	0.766	-0.745	0.484	-0.837	0.414	NaN	NaN
1834	8	-1.256	1.187	-0.567	0.858	-0.771	0.467	-0.839	0.416	NaN	NaN
1834	9	-0.553	1.245	-0.694	0.903	-0.793	0.466	-0.841	0.418	NaN	NaN
1834	10	-1.47	1.468	-0.763	1.038	-0.807	0.489	-0.852	0.419	NaN	NaN
1834	11	-1.094	3.651	-0.913	1.155	-0.811	0.492	-0.863	0.422	NaN	NaN
1834	12	-0.64	1.289	-0.898	1.073	-0.845	0.468	-0.867	0.427	NaN	NaN
1835	1	-0.618	2.722	-0.919	0.981	-0.859	0.455	-0.877	0.43	NaN	NaN
1835	2	0.117	1.618	-0.879	0.856	-0.86	0.449	-0.889	0.429	NaN	NaN

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.

John

Berkeley

DATA CLEANING | DOLOR

```
34  % The current region is characterized by:
35  %%   Name: Peru
36  %%   Latitude Range: -18.34 to -0.03
37  %%   Longitude Range: -81.34 to -68.68
38  %%   Area: 1299433.22 km^2
39  %%   Percent of global land area: 0.883 %
40  %%   Approximate number of temperature stations: 77
41  %%   Approximate number of monthly obeservations: 32259
42  %
43  % Note that all results reported here are derived from the full field
44  % analysis and will in general include information from many additional
45  % stations that border the current region and not just those that lie
46  % within this region.  In general, the temperature anomaly field has
47  % significant correlations extending over greater than 1000 km, which
48  % allows even distant stations to provide some insight at times when
49  % local coverage may be lacking.
50  %
51  %%  Estimated Jan 1951-Dec 1980 absolute temperature (C): 19.97 +/- 0.20
52  %
53  % Estimated Jan 1951-Dec 1980 monthly absolute temperature (C):
54  %       Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
55  %%     20.61  20.74  20.70  20.22  19.72  18.90  18.56  19.24  19.91  20.13  20.46  20.50
56  %%  +/-   0.21   0.22   0.22   0.21   0.21   0.21   0.22   0.21   0.21   0.23   0.23   0.21
57  %
58  % For each month, we report the estimated land-surface anomaly for that
59  % month and its uncertainty.  We also report the corresponding values for
60  % year, five-year, ten-year, and twenty-year moving averages CENTERED about
61  % that month (rounding down if the center is in between months).  For example,
62  % the annual average from January to December 1950 is reported at June 1950.
63  %
64  % Values are reported as missing (i.e. NaN) when station coverage within
65  % the region becomes too low, even though a limited number of observations may
66  % still have been made.  Time averages over intervals with some missing data will
67  % be reported as long as at least 75% of the necessary values are available.
68  %
69  %                 Monthly          Annual          Five-year        Ten-year        Twenty-year
70  % Year, Month,  Anomaly, Unc.,  Anomaly, Unc.,  Anomaly, Unc.,  Anomaly, Unc.,  Anomaly, Unc.
71
72    1892     1    -1.434  1.183       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
73    1892     2    -1.265  1.034       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
74    1892     3    -1.199  0.588       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
75    1892     4    -0.971  0.508       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
76    1892     5    -0.594  0.681       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
77    1892     6     0.462  2.126       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
78    1892     7       NaN    NaN       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
79    1892     8       NaN    NaN       NaN   NaN       NaN   NaN       NaN   NaN       NaN   NaN
```

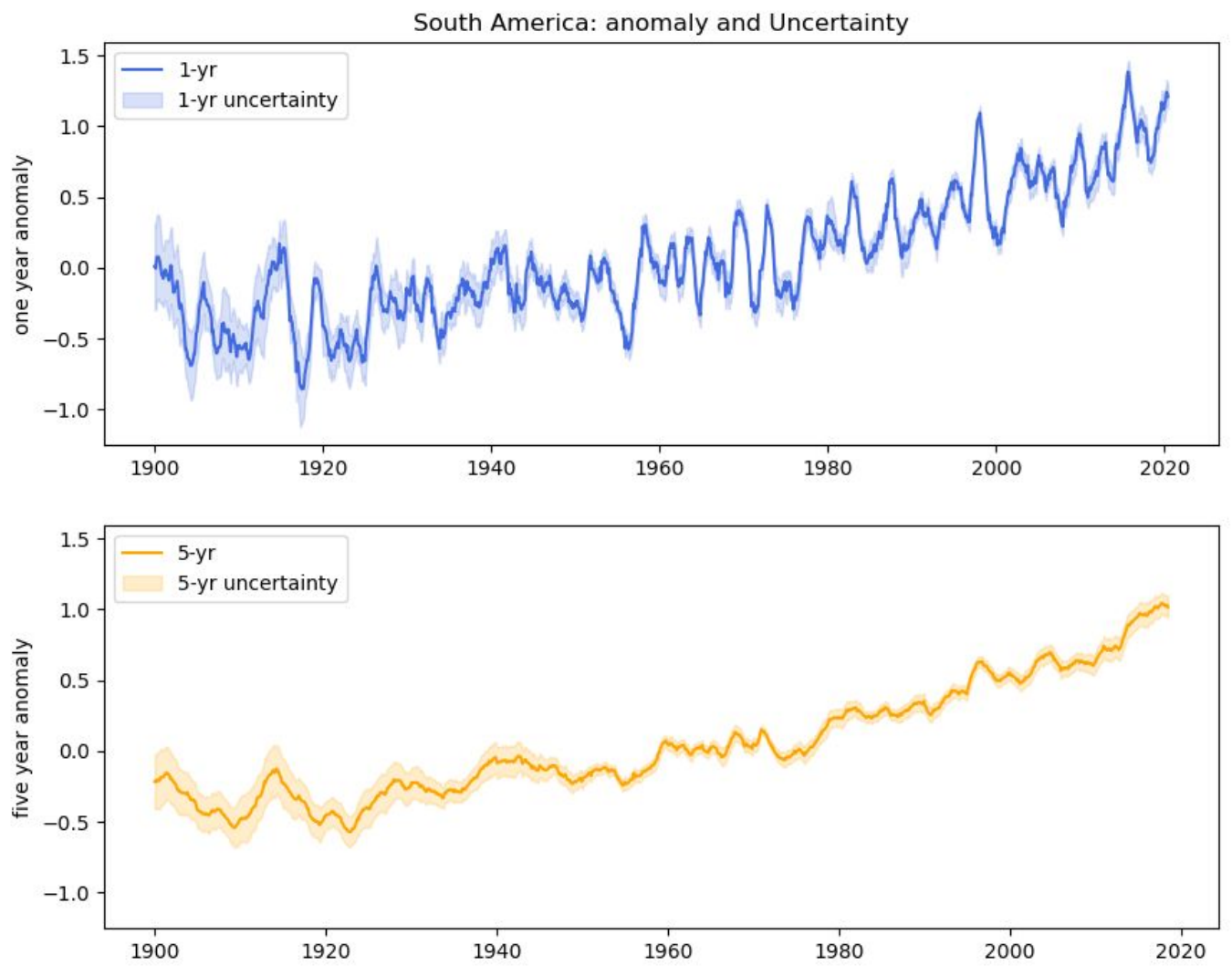
Data Cleaning

Commented Header description
Two line header
Strange spacing
Separator
String parsing
NaN’s

Stephanie

Berkeley

DATA SET | BERKELEY EARTH



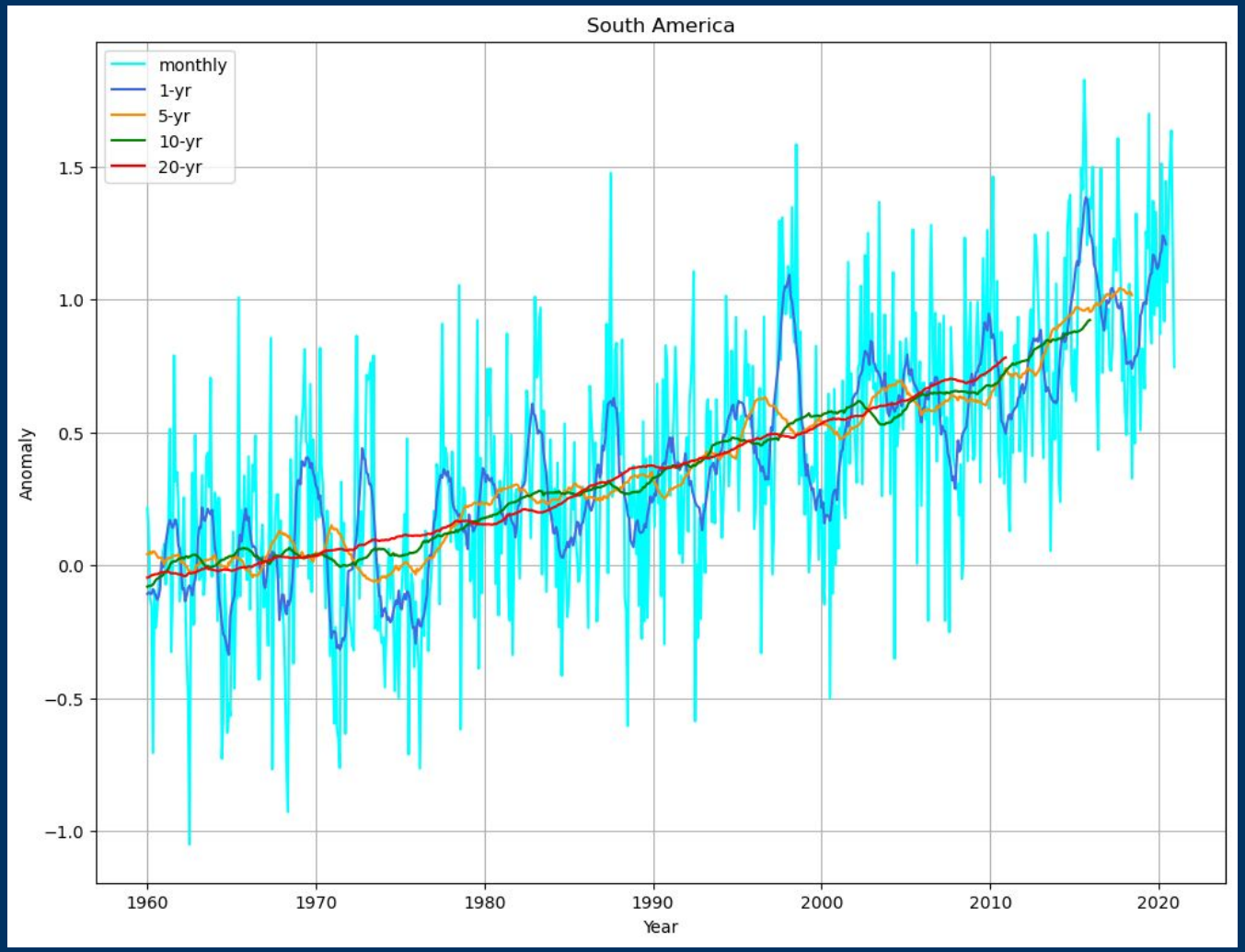
What data are we focusing on?

Intersection of data between our datasets

Stephanie

Berkeley

DATA SET | BERKELEY EARTH



What does the initial data look like?

We opted for 5-year anomalies because the data is less noisy.

Stephanie

Berkeley

FOCUSED QUESTIONS

1. How does each city compare to the rest of South America?
2. Which cities have higher temperatures 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?

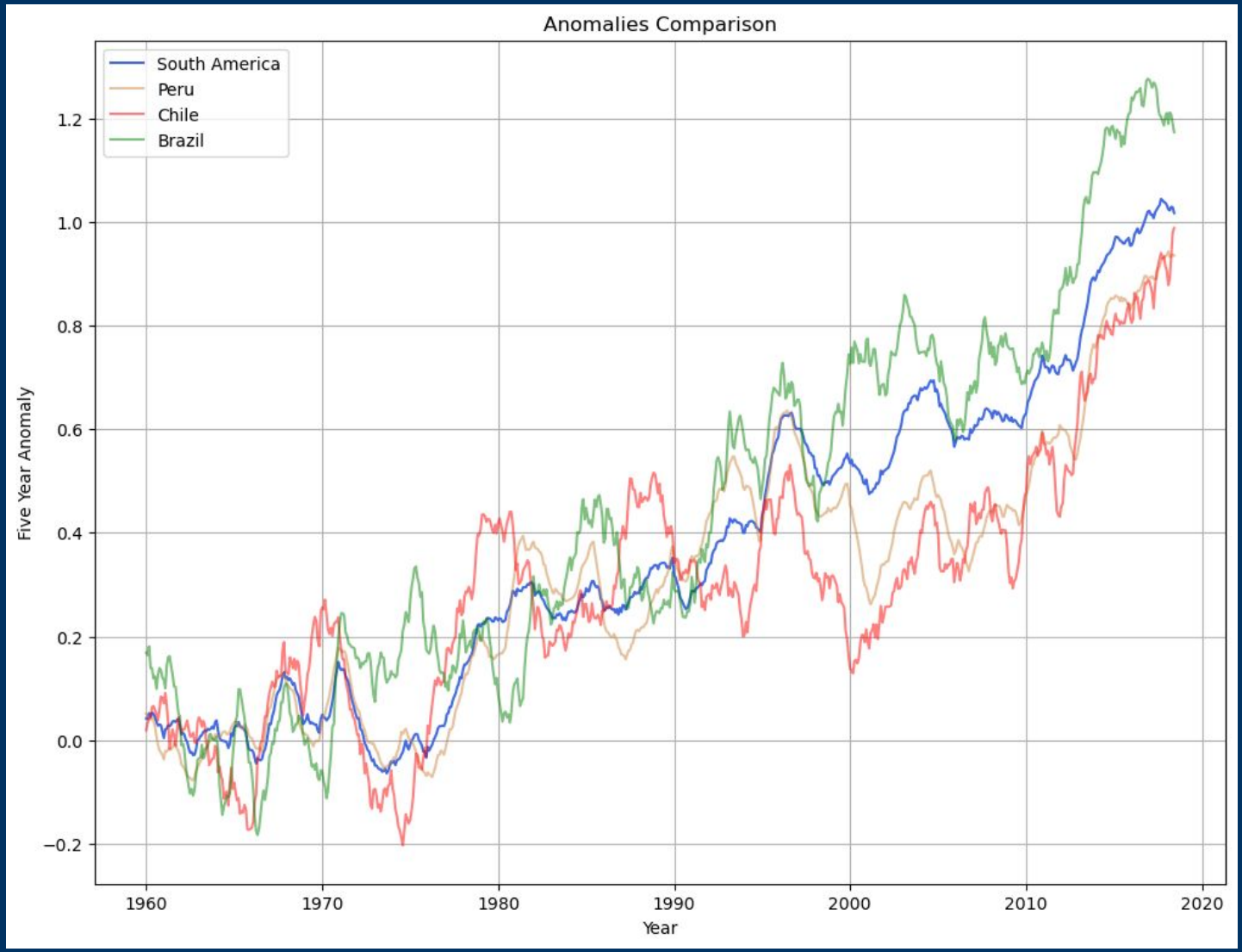
Stephanie

Berkeley

QUESTIONS

How does each city compare to the rest of South America?

Which cities have higher temperatures over time?

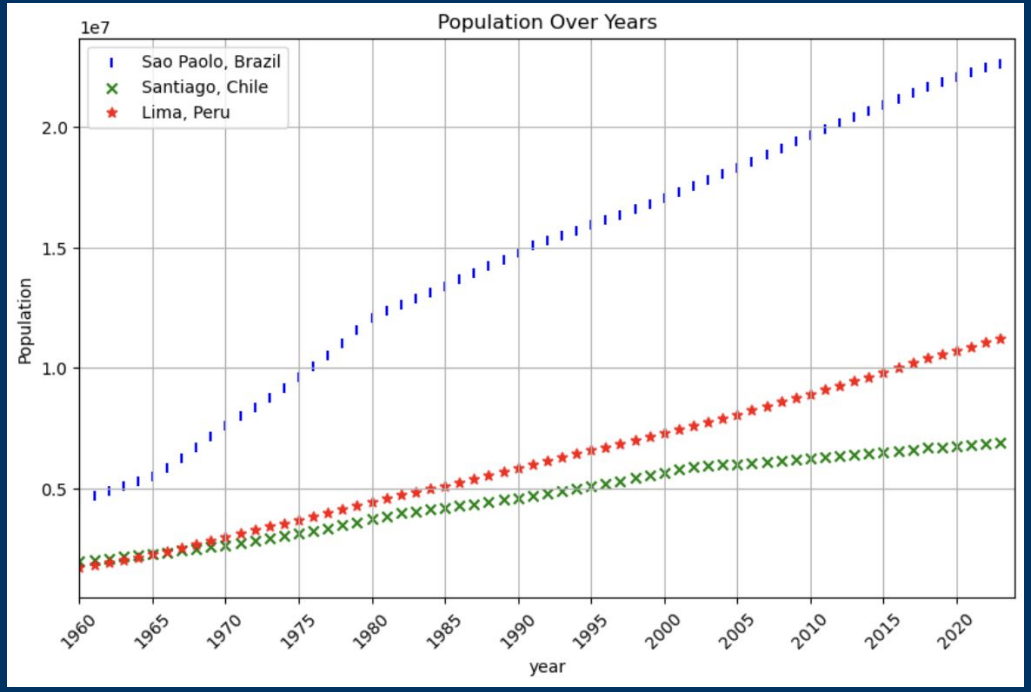


Berkeley

Stephanie

QUESTIONS

Do the temperature trends correlate to population trends?



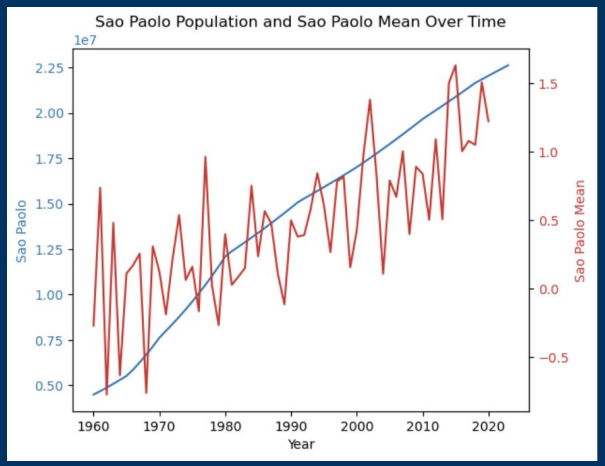
	year	Sao Paolo, Brazil	Santiago, Chile	Lima, Peru	date_formatted
0	1960	4493182.0	1979927.0	1755920.0	1960-01-01
1	1961	4681086.0	2047034.0	1845658.0	1961-01-01
2	1962	4878624.0	2106056.0	1946579.0	1962-01-01
3	1963	5084497.0	2166780.0	2053037.0	1963-01-01
4	1964	5299360.0	2229343.0	2165476.0	1964-01-01
...	...	...	...	...	...
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

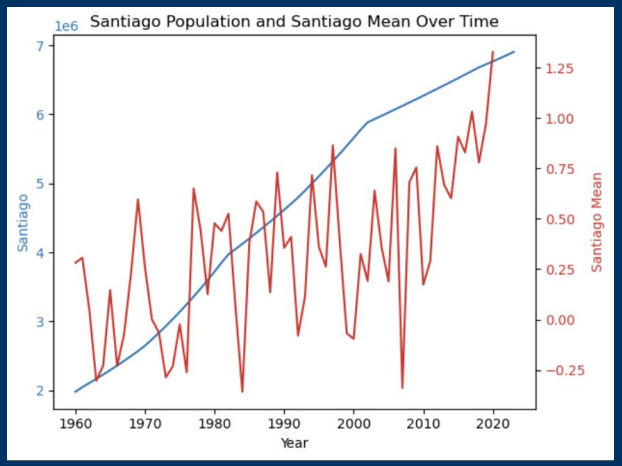
Berkeley

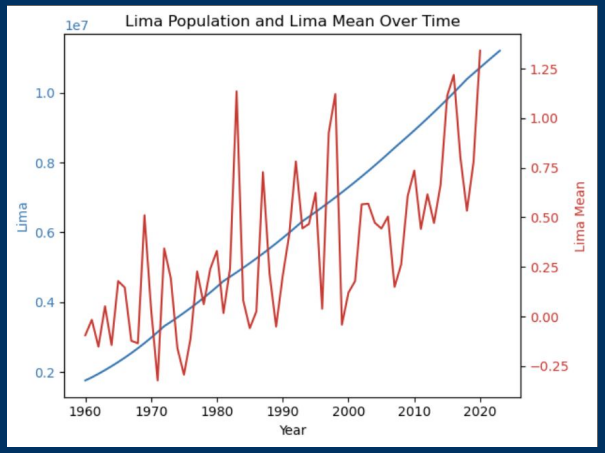
John

QUESTIONS

Do the temperature trends correlate to population trends?





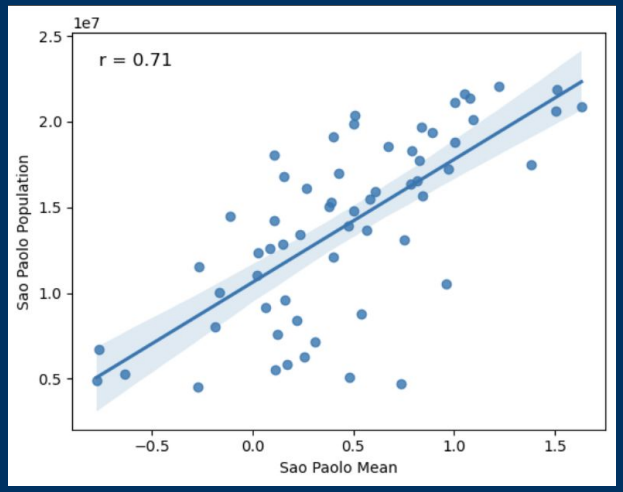


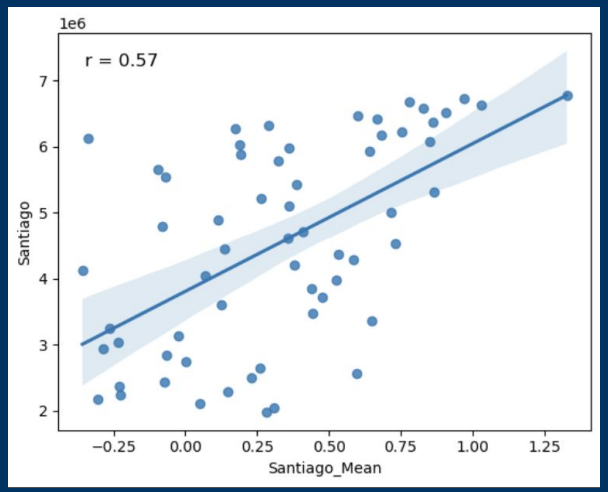
John

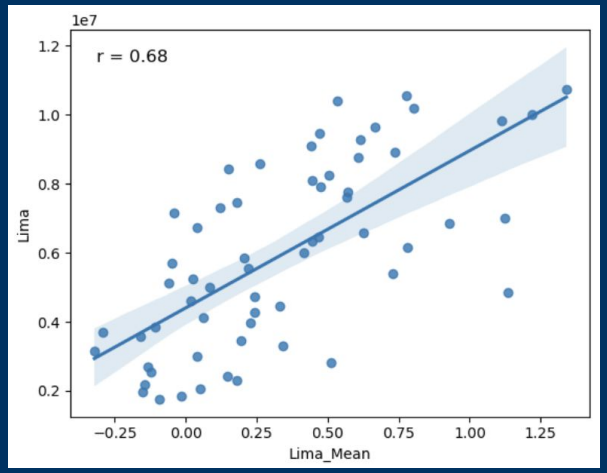
Berkeley

QUESTIONS

Do the temperature trends correlate to population trends? Pearson correlation values



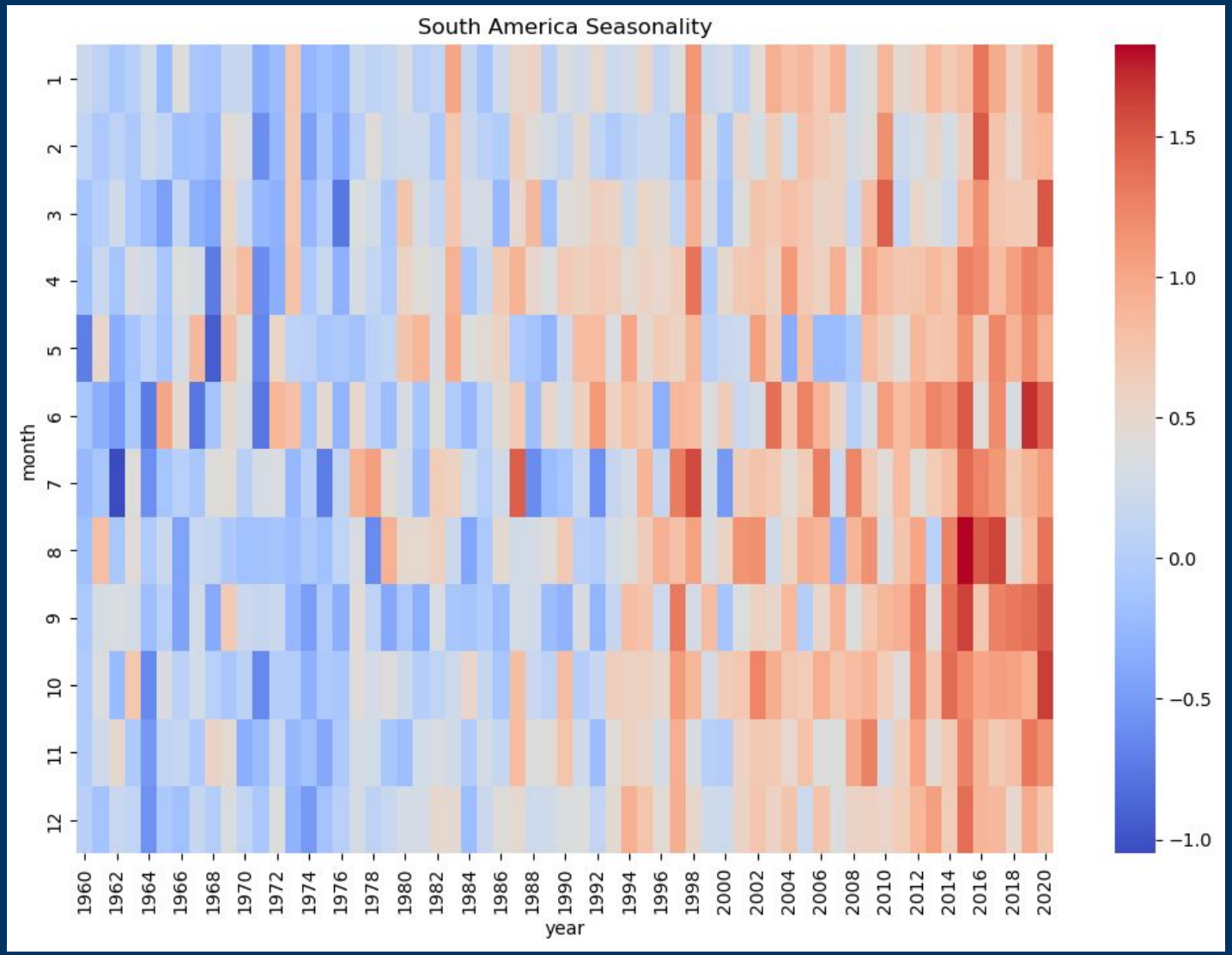




John

Berkeley

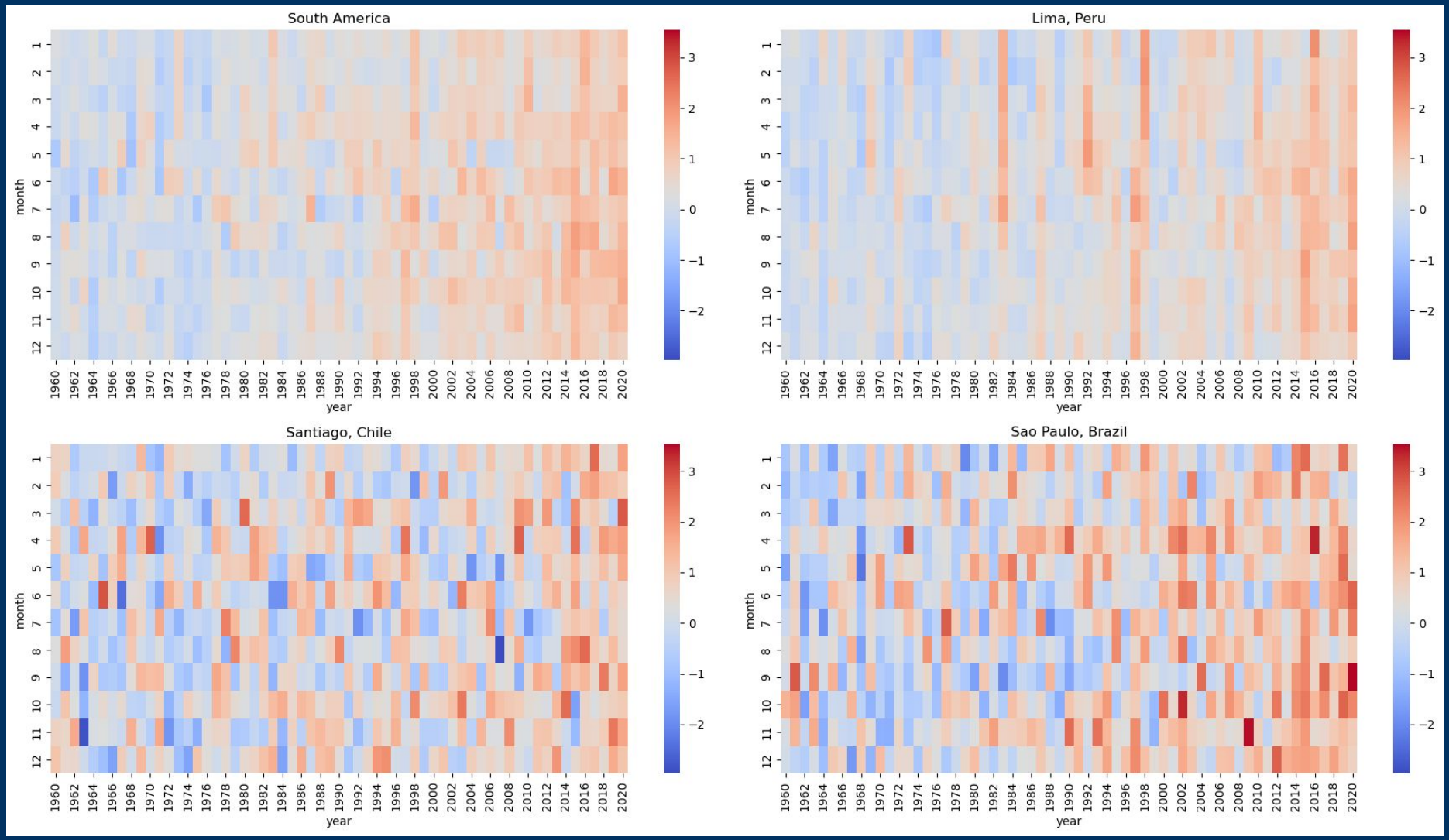
QUESTIONS | Is there a particular season where the anomaly and increased or decreased faster than others?



Berkeley

Stephanie

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



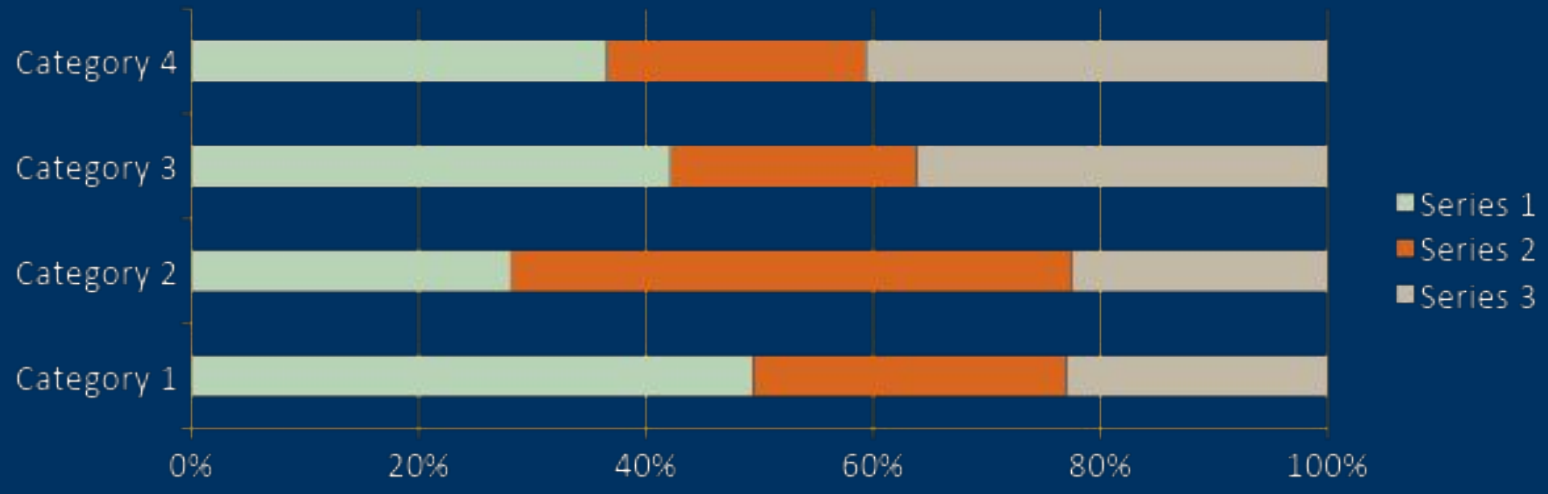
Berkeley

Stephanie

THANK YOU!

Berkeley

Open Sans 42 pt



Berkeley

Header
Open Sans 18 pt

Open Sans 14 pt

Nemo enim ipsam volptatem quia voluptas sit aspernatur aur adit amet eius modi tempora incidunt ut labore et dolore magnam

Lorem	Dolor	It	Enim	Color
43	60000	5600	1700	34000
35	55000	89056	4359	0349
435	3245	23243	25567	123
56	7	45	09867	456
345	768	0980	43138	2389
Fuga	Modi	Tempora	Nihil	Incidunt
Cor	Et vel	Wquis	Autem	Nulla

Berkeley

Header Open Sans 22 pt

Open Sans 18 pt

Berkeley

Header Open Sans 42 pt

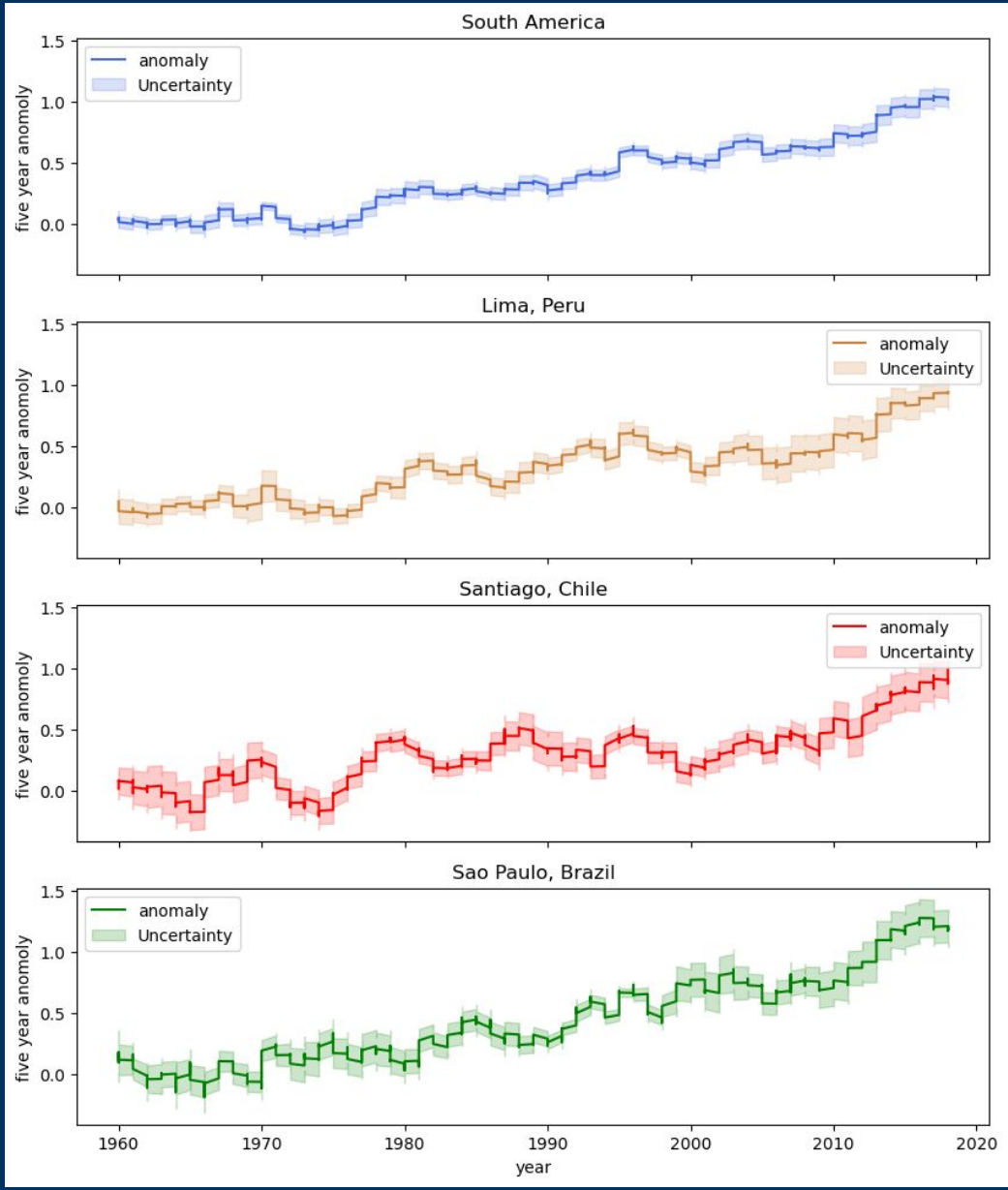
- Body copy, Open Sans 18 pts
  - Add your secondary bullet point here
    - Add your tertiary bullet point here

Berkeley

QUESTIONS

How does each city compare to the rest of South America?

Which cities have higher temperatures over time?



Berkeley