DAND Project 1: Weather Trend Analysis

Summary
This project is to study the global weather trend as well as a local city's weather trend. Data will be extracted from the given database of Udacity using SQL, and loaded to Kaggle jupyter notebook for further exploratory analysis with Python data manipulation and visualization package.

Tools
SQL, Python (pandas, matplotlib, seaborn), Udacity, Kaggle Notebooks

Part I - Data Extraction and Initial Visualization

Step 1: Data Extract

I used SQL to extract relevant data from table global_data and city_data.
I live in Vancouver, which is not on the city_list. So I chose Victoria as the local city and extracted the temperature data for Victoria from the city_data table. I also picked data for US city Dallas as I used to live there and the weather in DFW area is drastically different than Vancouver/Victoria. I then left join the Victoria data and Dallas data to global_data table and merged the tables based on field "year". Finally I filtered the data and selected data between year 1847 and year 2013, as this is the only time period that both Victoria and Dallas has data continuously. I then downloaded the data and saved it to a .csv file called "combo_data.csv". SQL code to extract the data:

- Check what cities are available within country of Canada:
```
SELECT city FROM city_list
WHERE country IN ('Canada')
```

- Check if Dallas is on the list:
```
SELECT * FROM city_list
WHERE city IN ('Dallas')
```

- Extract data for Victoria and Dallas, and merge them with global data:
```
SELECT global_data.year,
        global_data.avg_temp AS global_temp,
        vic.avg_temp AS victoria_temp,
        dal.avg_temp AS dallas_temp
FROM global_data
LEFT JOIN (
    SELECT year, avg_temp FROM city_data
    WHERE city = 'Victoria'
) AS vic
ON vic.year = global_data.year
LEFT JOIN (
    SELECT year, avg_temp FROM city_data
```

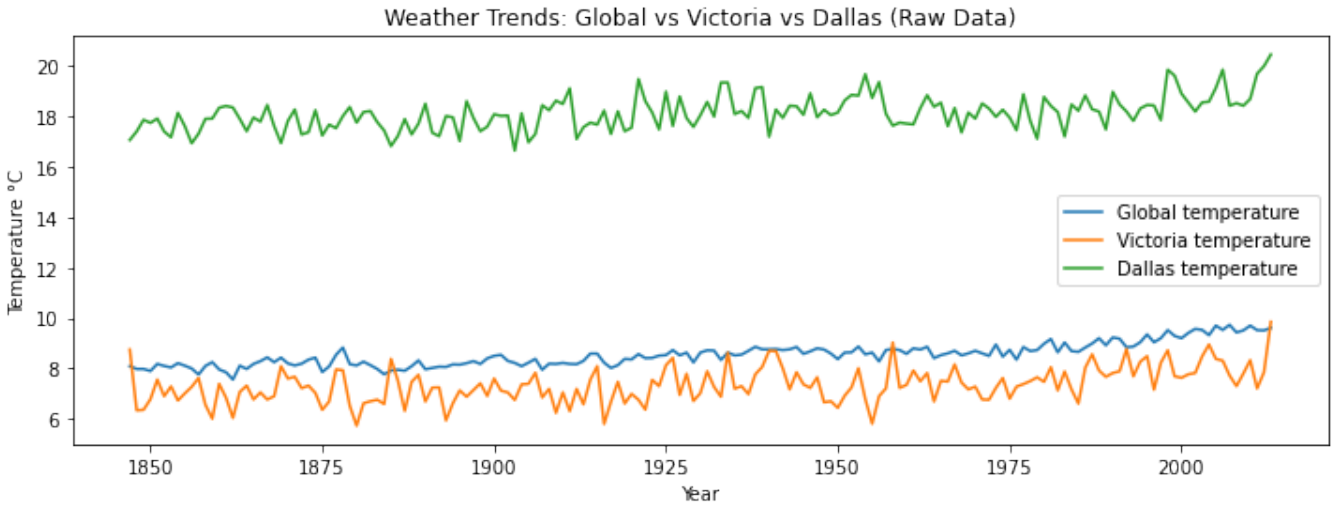
```
        WHERE city = 'Dallas'
    ) AS dal
    ON dal.year = global_data.year
    WHERE global_data.year BETWEEN 1847 AND 2013
    ORDER BY global_data.year
```

Step 2: Load Data to DataFrame

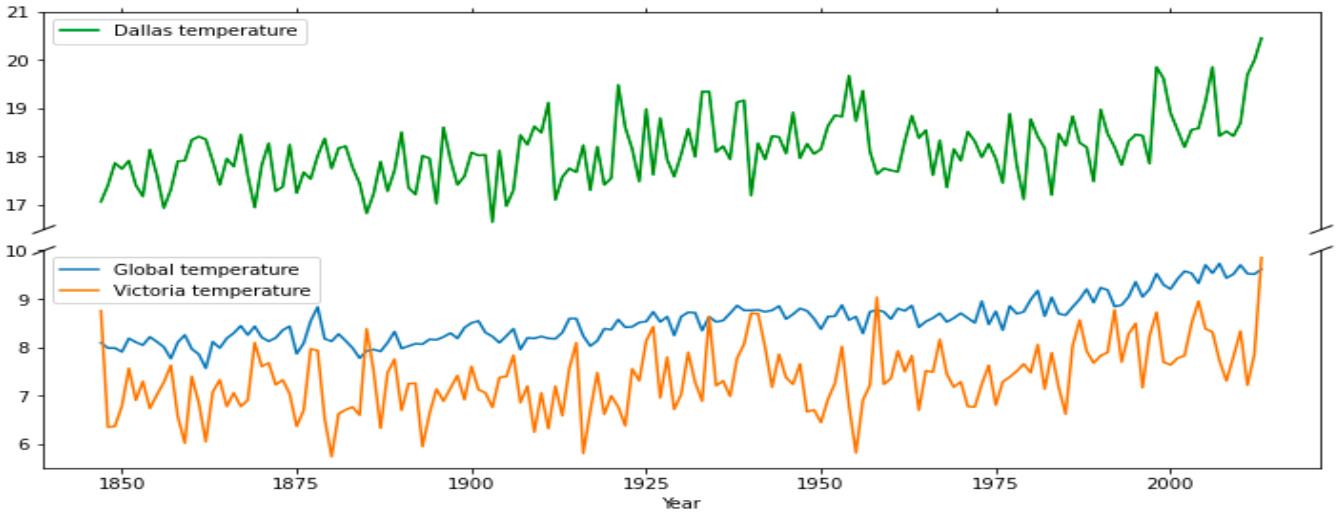
I used the Kaggle jupyter notebook to load the .csv file to Pandas DataFrame named “combo”.

Step 3: Visualization and Exploratory Analysis

I used visualization package matplotlib and seaborn perform the visualization and exploratory analysis. I first visualized the raw data.



The Dallas temperature data is drastically higher on average than global and Victoria data, and visually it’s difficult to interpret the data. So I applied broken axis technique and eliminated the unnecessary blank area on the chart. And below is the updated chart of the raw data:



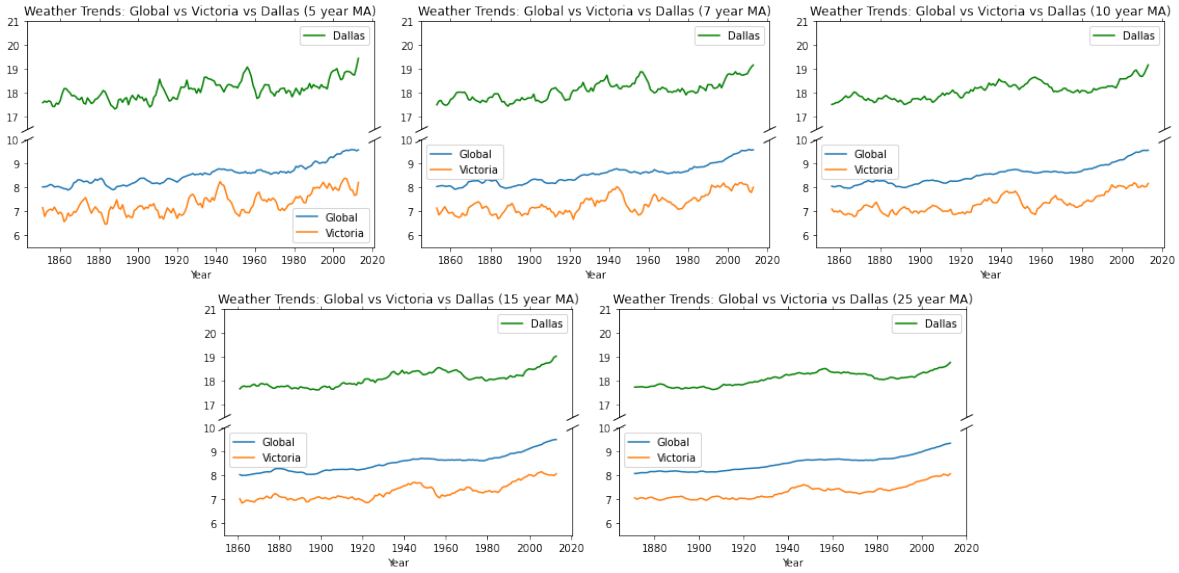
Now we are able to take a closer look. Victoria is colder than the global average, while Dallas is drastically hotter than the global average. And the raw data for Victoria and Dallas shows more oscillatory fluctuations than the global data. So we will need data smoothing to filter out the noise and identify trends in order to have a better understanding of the similarity and/or difference between the datasets. However the moving average window cannot be too long. Otherwise the volatility feature of the Victoria and Dallas temperature data will be completely eliminated. With that said, a moving average window of 30 or 50 years will not be considered in this case.

Step 4: Visualization of Moving Average Statistics

I used simple moving average (SMA) as the data smoothing method. In order to find a good moving average window size, I defined a function called plot_moving_average. This function takes the length of rolling window and the data frame as the inputs, applies rolling average calculation over the entire data frame, and then uses seaborn to line plot the manipulated dataset. Then I called the function five times, each time with a different length of rolling window as the input on the same data frame.

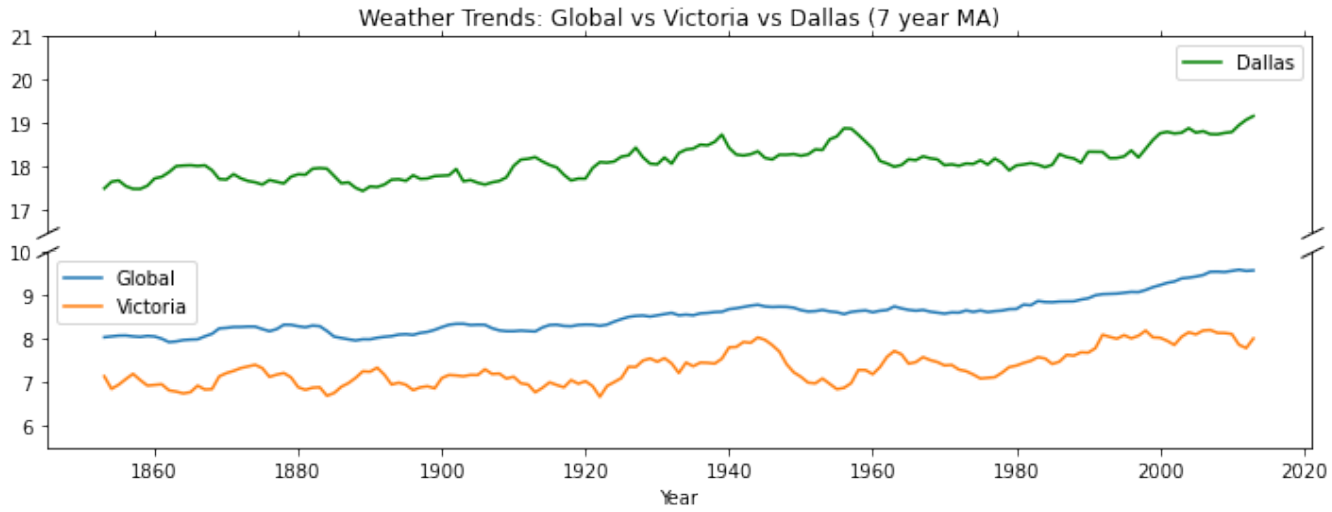
Part II - Data Smoothing and Exploratory Analysis

The following are the line charts for 5, 7, 10, 15, 25 years moving average (MA) temperature trends in Victoria, Dallas and the world.



Step 1: Choose Rolling Window Size

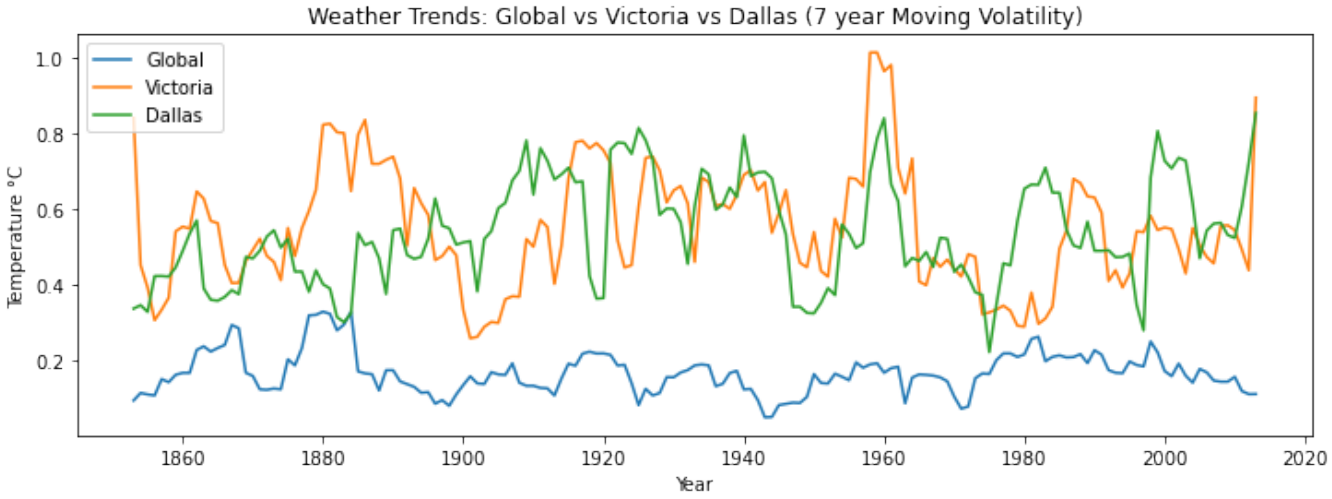
The line chart of 15 years and 25 years MA are relatively too flat to tell any useful information other than that the overall temperature trend is upwards for both global weather and Victoria's local weather and that there was obvious shift around certain periods like 1920s and 1950s.
Either one of the 5 years, 7 years and 10 years MA charts looks plausible. I will take the middle and use 7 years MA chart for further analysis.



Step 2: Further Visual Exploration

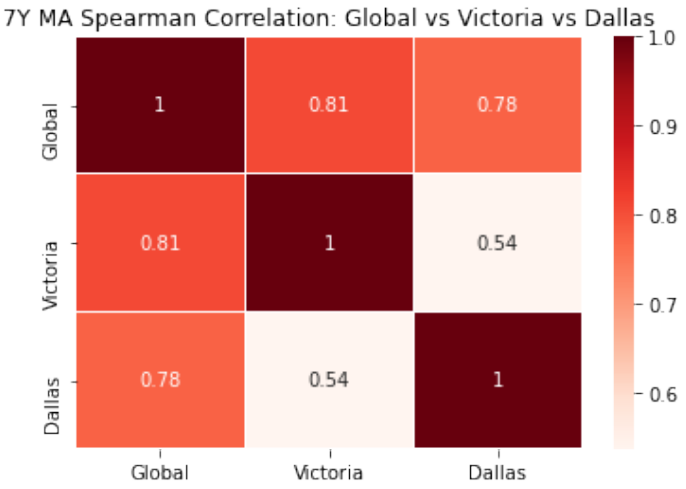
Moving Volatility of Each Dataset

Overall Victoria data shows more fluctuations than Dallas data, and both are relatively highly volatile, especially around 1920, 1960 and 2000. For global data, the volatility has been decreasing since 1980 although the average temperature has been showing a monotonic-ish upwards shift over the same period. Additionally around 1880, both global and local Victoria experienced high fluctuation while Dallas data showed relatively low volatility.



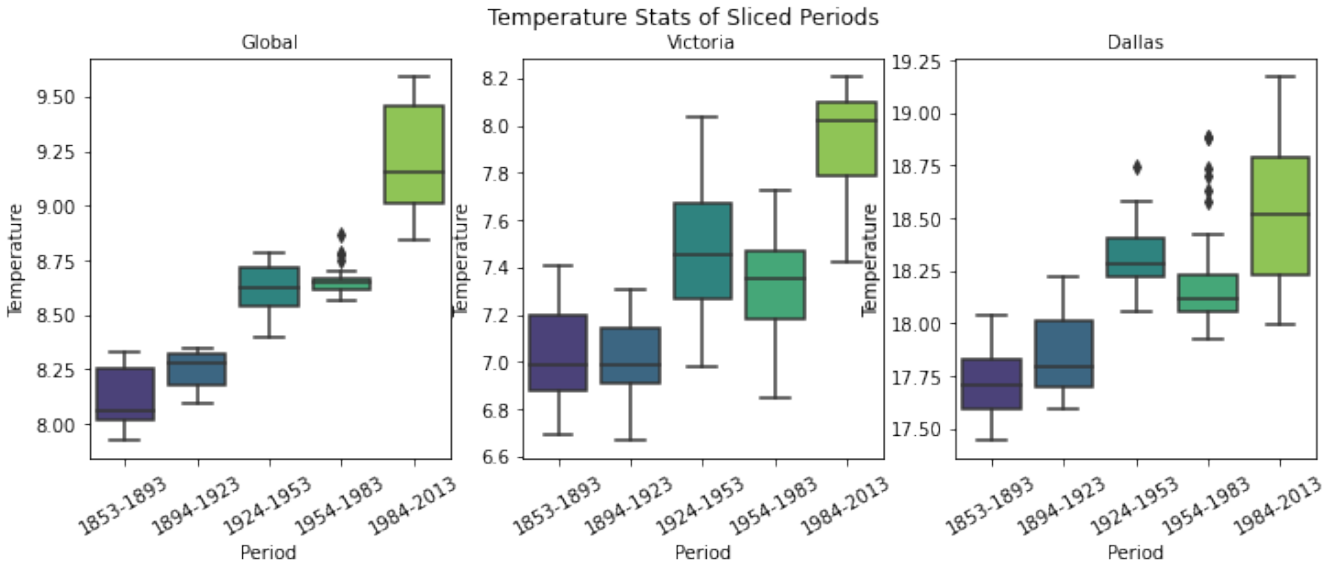
Correlation between Each Dataset

Since I am not sure if a linear relationship exists between the temperature of local cities versus global, I will be using Spearman's Correlation matrix (instead of Pearson's) to assess if the relationship is stronger or weaker across the distribution of the variables.
Correlation between global MA and each city's MA is relatively strong (0.78 and 0.81), while the correlation between Victoria and Dallas is weak (0.54).



Closer Look at Sliced Data

Since the total number of MA observations is 161, I sliced the whole periods of data into 5 sub-periods, each period with a 30-year interval from 1894 to 2013, and 41-year interval for sub-period 1853-1893. Below is a box plot of key statistics of temperature in each location during each sub-period. It shows that for each location, the median has increased overtime, out of which sub-period 1984-2013 has the sharpest increase compared to previous sub-periods. Also with statistical outliers taken into consideration, the gap between highest and lowest temperature within each period has also widened from 1853 to 2013. Except that for Victoria the gap has been consistently wider since 1924. Obviously, Victoria is more volatile than global and Dallas.



Below is the numeric description of key statistics corresponding to the above box plot.

- For global temperature, the 30-year interval average increased from 8.12 to 9.21, the standard deviation increased from 0.13 to 0.27, and the gap between highest/lowest temperature first increased from 0.40 to 0.75. Additionally I added a column “median chng” to calculate the change of median temperature between each sub-period and its previous sub-period. For global, the median change increased from 0.22 to 0.51, meaning the temperature increased at a faster pace overtime.

However during sub-period 1954-1983, the median temperature only increased 0.03 degree compared to previous period, which might be a sign of effective climate control during that period.

- For Victoria temperature, the interval average increased from 7.03 to 7.93, the standard deviation increased from 0.20 to 0.30, and then dropped to 0.23. And the gap between highest/lowest temperature first increased from 0.71 to 1.06, and then decreased to 0.78. For Victoria, the median change increased from 0.00 to 0.67, but interestingly during sub-period 1954-1983 it turned negative, meaning during that period the temperature actually dropped, which agrees with the observation of global trend during the same period.
- For Dallas temperature, the interval average increased from 17.73 to 18.54, the standard deviation increased from 0.17 to 0.33, and the gap between highest/lowest temperature first increased from 0.59 to 1.18. About median change, the situation for Dallas is similar to Victoria. The median change increased from 0.09 to 0.40, but during sub-period 1954-1983 it turned negative as well.

So to summarize:

- On average, temperature around the globe has increased overtime at a faster pace.
- Temperature around the globe shows more oscillatory fluctuation overtime, with higher standard deviation and widened gap between highest/lowest values.
- Overall the median temperate change around the globe has increased overtime. But it showed a drop during period 1954-1983, with both Victoria and Dallas hit a negative point.

Global Temperature:

Period	count	mean	std	min	25%	50%	75%	max	diff	median chng
1853-1893	41.0	8.117596	0.126492	7.925714	8.020000	8.062857	8.252857	8.325714	0.400000	NaN
1894-1923	30.0	8.249476	0.082943	8.091429	8.178214	8.280000	8.323929	8.347143	0.255714	0.217143
1924-1953	30.0	8.620762	0.099105	8.397143	8.541429	8.619286	8.712500	8.784286	0.387143	0.339286
1954-1983	30.0	8.657048	0.064481	8.568571	8.616786	8.646429	8.665357	8.870000	0.301429	0.027143
1984-2013	30.0	9.205476	0.266588	8.842857	9.007857	9.155000	9.457143	9.588571	0.745714	0.508571

Victoria Temperature:

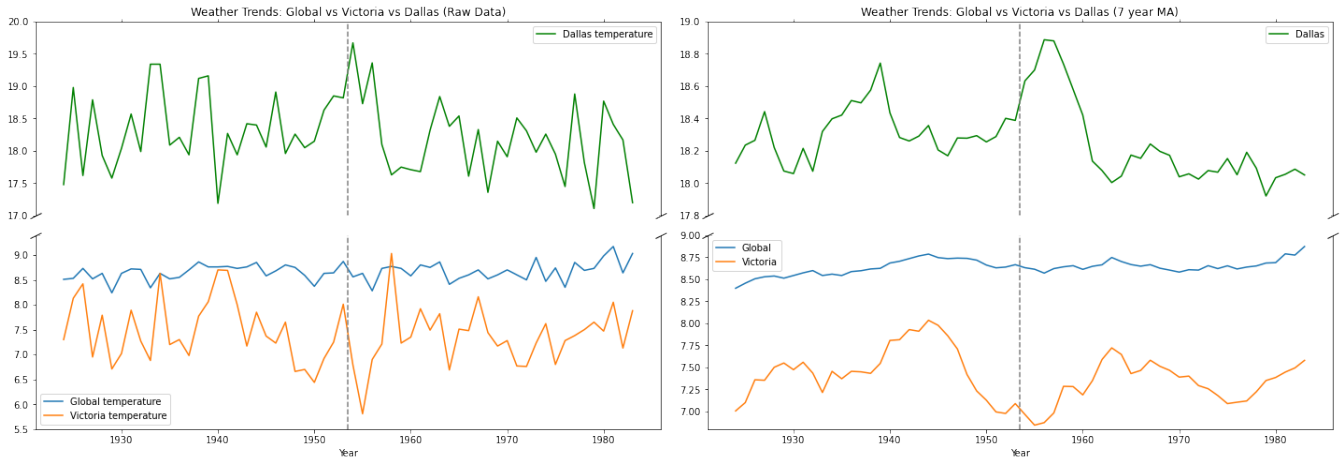
Period	count	mean	std	min	25%	50%	75%	max	diff	median chng
1853-1893	41.0	7.034390	0.198387	6.692857	6.878571	6.984286	7.197143	7.404286	0.711429	NaN
1894-1923	30.0	7.007333	0.147005	6.668571	6.912143	6.987143	7.138929	7.301429	0.632857	0.002857
1924-1953	30.0	7.471714	0.303948	6.978571	7.263214	7.452857	7.671071	8.034286	1.055714	0.465714
1954-1983	30.0	7.317286	0.226328	6.845714	7.183214	7.350714	7.466786	7.721429	0.875714	-0.102143
1984-2013	30.0	7.932810	0.228120	7.422857	7.786429	8.020000	8.100714	8.204286	0.781429	0.669286

Dallas Temperature:

Period	count	mean	std	min	25%	50%	75%	max	diff	median chng
1853-1893	41.0	17.733066	0.173770	17.447143	17.594286	17.707143	17.832857	18.038571	0.591429	NaN
1894-1923	30.0	17.853714	0.191084	17.594286	17.703214	17.793571	18.016429	18.224286	0.630000	0.086429
1924-1953	30.0	18.312476	0.152572	18.060000	18.225000	18.285714	18.400714	18.742857	0.682857	0.492143
1954-1983	30.0	18.231857	0.276314	17.921429	18.053571	18.115000	18.231429	18.887143	0.965714	-0.170714
1984-2013	30.0	18.536429	0.330997	17.992857	18.229643	18.515714	18.786786	19.168571	1.175714	0.400714

Further Look at Significant Changepoint

Since median temperature change showed an opposite trend during sub-period 1954-1983 against other sub-periods, I took a further look at this sub-period. I also added the data from its previous sub-period 1924-1953 just for reference. I added a dashed line in the charts to segregate the two sub-periods, with year 1953 and 1954 on each side of the line.



I first tried to do some research about Dallas. Per Wikipedia (link here), between 1949 and 1957 it was the "1950s Texas drought", during which "the state received 30 to 50% less rain than normal, while temperatures rose above average". Additionally, "Texans experienced the second-, third-, and eighth-driest single years ever in the state - 1956, 1954, and 1951, respectively". This explains the two peaks at year 1954 and 1956 in the upper left chart, and the peaks at point 1956 and 1957 as it calculates 7-year MA.

In 1957, "the state created the Texas Water Development Board, which set into motion a number of water-conservation plans". By 1970, "the number of Texas reservoirs more than doubled, and by 1980, more than 126 major reservoirs had been constructed". Additionally 69 dams were built between 1957 and 1970. This explains why the 7-year MA in Dallas has been decreasing since the peak in 1957. As more water-conservation constructions than ever are in place since 1957, I think it further explains why median temperature for sub-period 1954-1983 (after the drought) is lower than sub-period 1924-1953 (before the drought).

Victoria does not show a strong downward trend as Dallas over sub-period 1954-1983. But its 7-year MA showed a significant peak from late 1930s to late 1940s. According to BC government (link here), the summer precipitation level around Vancouver Island between late 1930s through early 1940s on average is 25% lower than the long-term mean. I think this may contribute to the average high temperature in Victoria during 1924-1953, thus explains the negative value of median change in sub-period 1954-1983.

The global 7-year MA trend is not as volatile as Victoria or Dallas, though it still showed a slight downward trend from mid 1940s to early 1960s. According to NASA research in 2007 (link here), it is likely that "the relatively sudden, massive output of aerosols from industries and power plants contributed to the global cooling trend from 1940-1970".

Part III - Summary of Observation

Observation 1: Both global and local temperature has been increasing overtime. Multiple researches have attributed this to anthropogenic influence (the Canada's climate report as an example).

Observation 2: Temperature around the globe shows more oscillatory fluctuation overtime, with higher standard deviation and widened gap between highest/lowest values.

Observation 3: The median values of both global and local moving-average temperature increased at a faster pace overtime on average.
Observation 4: The median values of global moving-average temperature very small increase over period 1954-1983 (less than 0.03 degree compared to 0.22/0.34/0.50 from other periods). Researchers suspected it was due to "the relatively sudden, massive output of aerosols from industries and power plants".
Observation 5: The median values of Dallas and Victoria moving-average temperature showed negative change over period 1954-1983. For Dallas local, it's due to administrative effort to "set into motion a number of water-conservation plans". For Victoria local, it is possible due to the un-usual low precipitation level prior to that period.
Observation 6: Although Victoria is famously known for its mild climate and all year round greenness, it's surprise to see the temperature data showed a relatively high 7-year MA volatility. It might be because Victoria has so-called micro-climates depending on each area's proximity to water, hills, and mountains, and how protected they are.

Reference

- 1950s Texas drought [https://en.wikipedia.org/wiki/1950s_Texas_drought]
- Long-term Daily Air Temperature and Precipitation Record for the Lake Cowichan Area [https://www.for.gov.bc.ca/hfd/pubs/docs/Tr/TR112.pdf]
- Earth's Temperature Tracker [https://earthobservatory.nasa.gov/features/GISSTemperature/giss_temperature4.php]
- Canada's Changing Climate Report [https://www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/energy/Climate-change/pdf/CCCR_FULLREPORT-EN-FINAL.pdf]