Statistical Analysis of Meteorological Data to Investigate the Impact of Climate Change

1. Introduction

Climate change stands out as a primary challenge of our generation. In the news, we hear about glaciers melting, polar bears losing their habitat, and CO_2 levels rising. An ordinary person can not relate to those. Can we do a local analysis to find signs of climate change using data analysis in Python?

The National Oceanic and Atmospheric Administration has a website climate.gov which shows other signs of climate change such as Japanese Cherry Blossoms flowering earlier, increasing rate of receding glaciers, and the bleaching of coral reefs. The same site shares temperature data for the continent which shows that the average land and ocean temperature increased approximately 0.5°C for the last 20 years (NOAA Climate.org). It is hard for humans to comprehend this 0.5°C change over 20 years, because of the daily variation in temperature.

Further, this is not entirely accurate because climate change has a lot of variation from region to region. Certain areas might be more impacted by climate change than others. The goal of this paper was to explore meteorological data for the past 70 to 80 years to statistically demonstrate the impact of climate change in a handful of United States cities. In this paper, we looked at different ways to study meteorological data from different regions. We looked at plots of average temperature on a particular day over a century, modeled temperature over a year and compared it to historical data, and finally looked at extreme temperature events.

2. Methods

2.1 Study Systems

To find weather data for cities in the United States, we used the National Centers for Environmental Information (NCEI). The most consistent and reliable data in the NCEI's stations came from large airports and universities. These datasets contain information on wind, precipitation, daylight duration, snowfall, maximum and minimum temperatures, and other daily climate characteristics. Most of these had 25,000 recorded days.

The data in CSV files were uploaded and analyzed by Pandas data frameworks. These data frames allowed quick manipulation of the data. Then, the individual data rows were removed if they contained null values. Also, rows containing multiple "9"'s in a row were considered invalid according to the NCEI's documentation and were removed accordingly.

For the analysis presented in this paper, the data used was from the John Glenn Columbus International Airport (1948 to 2022), Miami International Airport (1948 to 2022), Baltimore/Washington International Thurgood Marshall Airport (1939 to 2022), and Pittsburgh International Airport (1945 to 2022). The data contained columns that contained the date and minimum and maximum temperatures for each day. A new column was created that was the average of the minimum and maximum temperatures

2.2 Daily average temperature for a particular day over multiple years

This analysis is focused on studying data for one particular day over the entire recorded time for a city. The best-fit line and r-squared is determined using NumPy and plotted using Matplotlib.

2.3 Comparing historical average temperature fit to individual years

One can visualize average temperature over a year as an inverted parabola. To create a baseline for the first 30 years, we used the least squares method employing a 30-year time frame which provided a substantial dataset for statistical analysis. Goodness of fit to quadratic equation for the 30-years was calculated as r-squared using NumPy. Next, goodness of fit of data for subsequent individual years to the baseline equation were determined. The best fit line and corresponding r-squared is plotted for every single year. Significant deviation in temperature would show up as a lower r-squared value.

This goodness of fit was plotted as a function of year using Matplotlib. Using this plot, a linear regression was performed on the r-squared.

2.4 Comparing number of statistically extreme daily high temperature events per year

The data was first grouped by the day of the year and a mean and standard deviation was calculated for each day. The following equation was used to calculate the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

Where σ is the standard deviation, x_i is a value, μ is the average value, and N is the number of values. For every year and every day, the number of occurrences where a day's high temperature exceeded 2 or 3 times the standard deviation for that day was counted. To calculate the occurrences the following equations were used.

$$Tobs > Tavg + 2\sigma$$
$$Tobs > Tavg + 3\sigma$$

Where T_{obs} is the observed temperature, T_{avg} is the average temperature for the day, and σ is the standard deviation for the day. Two graphs for each city were created for 2 sigma and 3 sigma excursions, respectively.

3. Results and Discussion

3.1 Daily average temperature for a particular day over multiple years

Fig 1 shows the average temperature on January 1 at the Columbus Airport from 1948 to 2022. Note that temperatures over the years range from -2°C to 24°C. The slope of this graph tells us that every 10 years there is around a 0.5°C increase in average temperature on January 1. Fig 2 shows the average temperature on January 1 at the Miami Airport from 1948 to 2022. Note that temperature over the years ranges from 11°C to 26°C. The slope of the graph shows that every 10 years there is around a 0.6°C increase in average temperature on January 1. Fig 3 shows the average temperature on January 1 at Cornell University in Ithaca, New York. Note that the temperature over the years range from -14°C to 12°C. The slope of this graph indicates that every 10 years there is around a 0.1°C increase in average temperature on January 1. The three cities are showing differences in variation over the years.

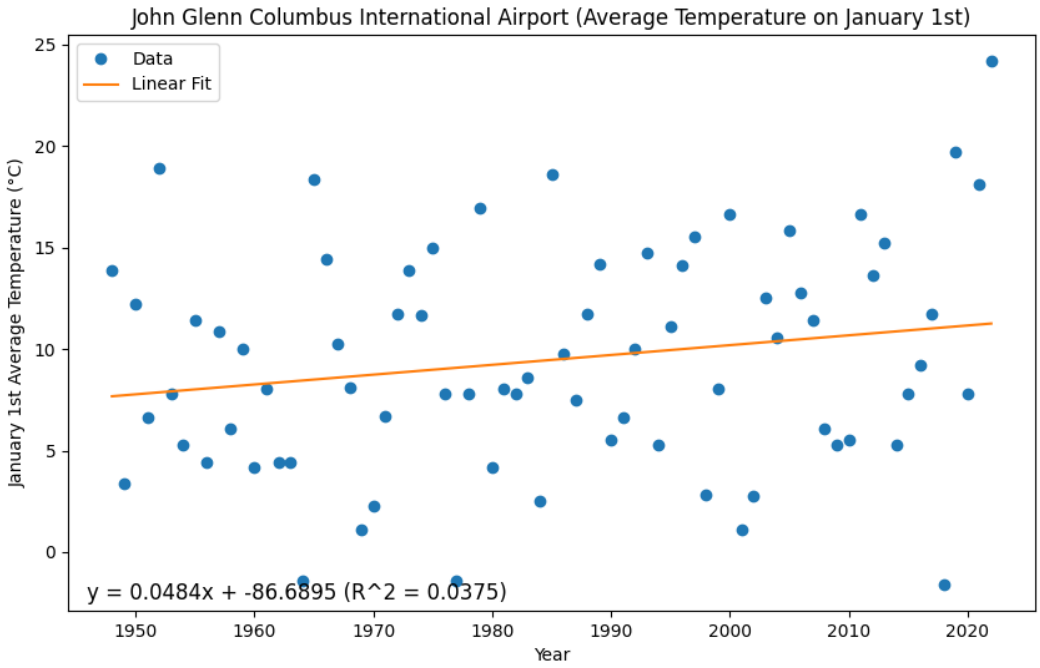

Fig. 1. John Glenn Columbus International Airport shows an upward trend of the average temperature on January 1 but the r-squared value is very low (0.0375). Despite observing a 0.48°C/10 year temperature increase, this temperature increase is not statistically significant.

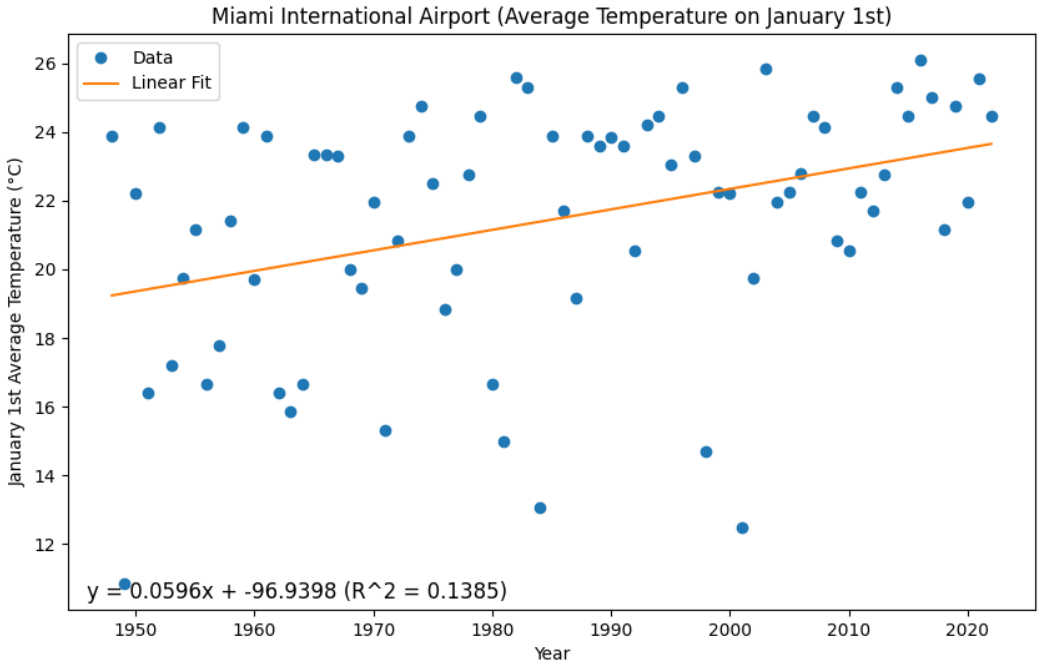

Fig. 2. Miami International Airport shows an upward trend of the average temperature on January 1 but the r-squared value is 0.1385.

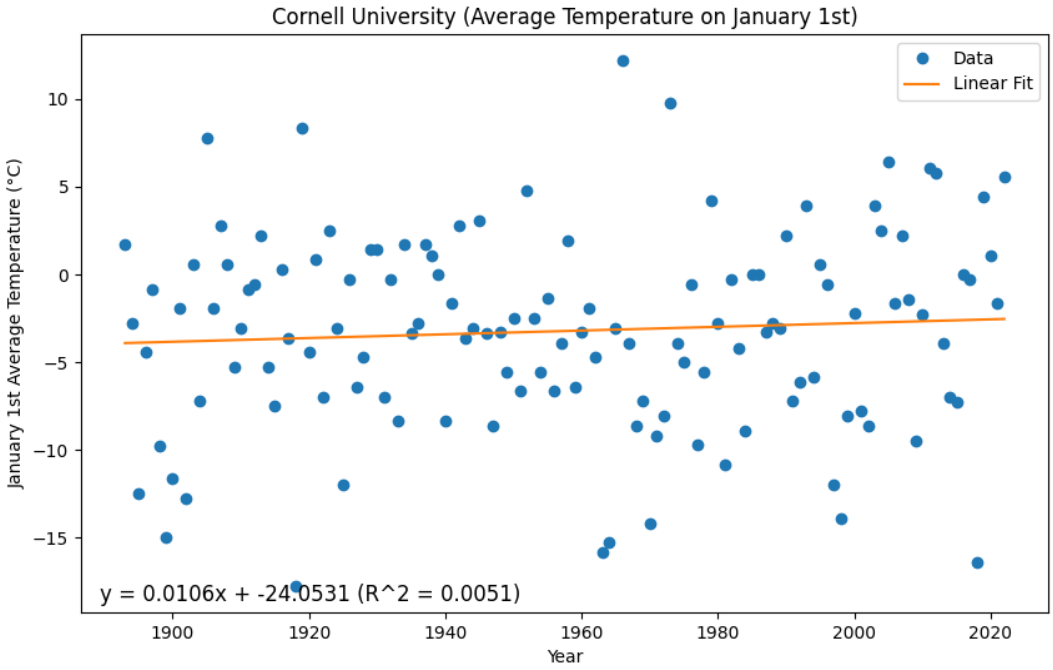

Fig. 3. Cornell University shows an upward trend of the average temperature on January 1 but the r-squared value is 0.0051.

Looking at the graphs, there is almost no correlation between the linear fit and the data. The line may show a trend in temperature but it does not capture the full data. In all three cases, the temperature rise per year is increasing on January 1, but the r-squared value of the graph is quite low and visually most of the data points do not fall on the line. A common person would not observe the small increase predicted by the line unless viewing the fit. Both statistics and our own visual analysis shows insufficient evidence that climate change is occurring.

3.2 Comparing historical average temperature fit to individual years

As described in section 2.3, a best fit quadratic equation based on first 30 years of daily temperature records was determined using least square analysis. Fig 4 and 5 show the daily average temperatures for 1948 and 2020, respectively along with the line based on the first 30 years of baseline. When comparing the data and the baseline quadratic equation, 1948 temperature is better described by the line than the 2020 data. The r-squared value of 1948 is 0.78 while 2020 has an r-squared of 0.61.

In Fig 6 and 7, the r-squared value between each of the year's temperature data and the best-fit curve from the first 30 years is plotted for Columbus Airport and Miami Airport. The r-squared value for each year is plotted. In each case, the r-squared value representing the baseline is seen to be showing worse fit. However, the slope of the line from Columbus Airport is much lower than the slope of the line from Miami Airport indicating that Miami is showing greater differences in weather from 1948 to 2022.

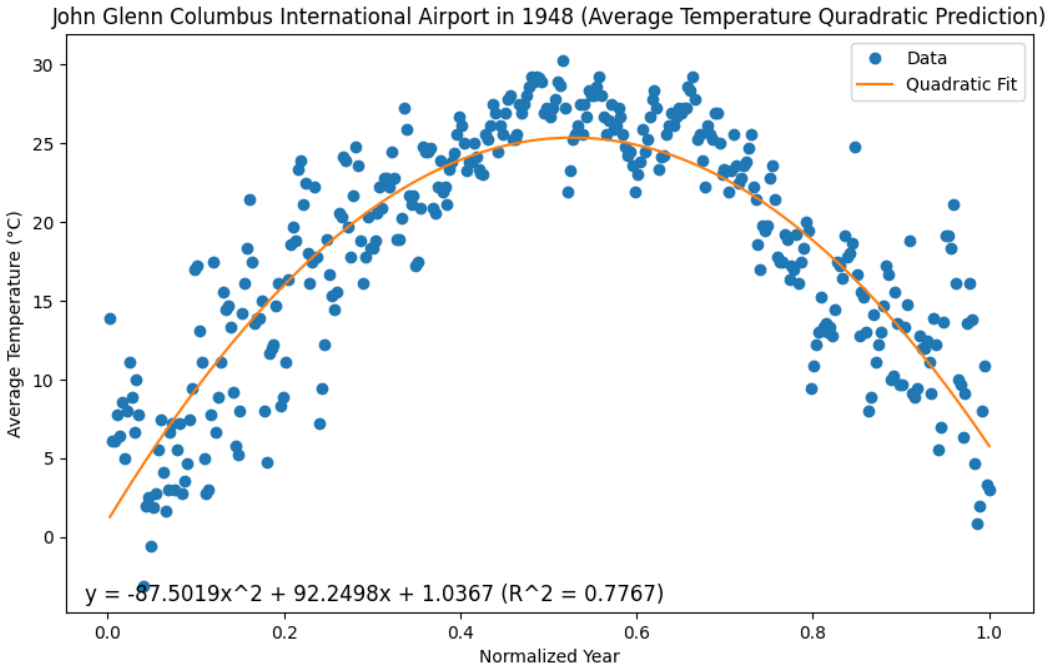

Fig. 4. John Glenn Columbus International Airport daily average temperature in 1948 with quadratic least square fit from the first 30 years. Shows an r-squared value of 0.7767 which is a decent fit for the data. In other words, the quadratic fit fits the data quite well.

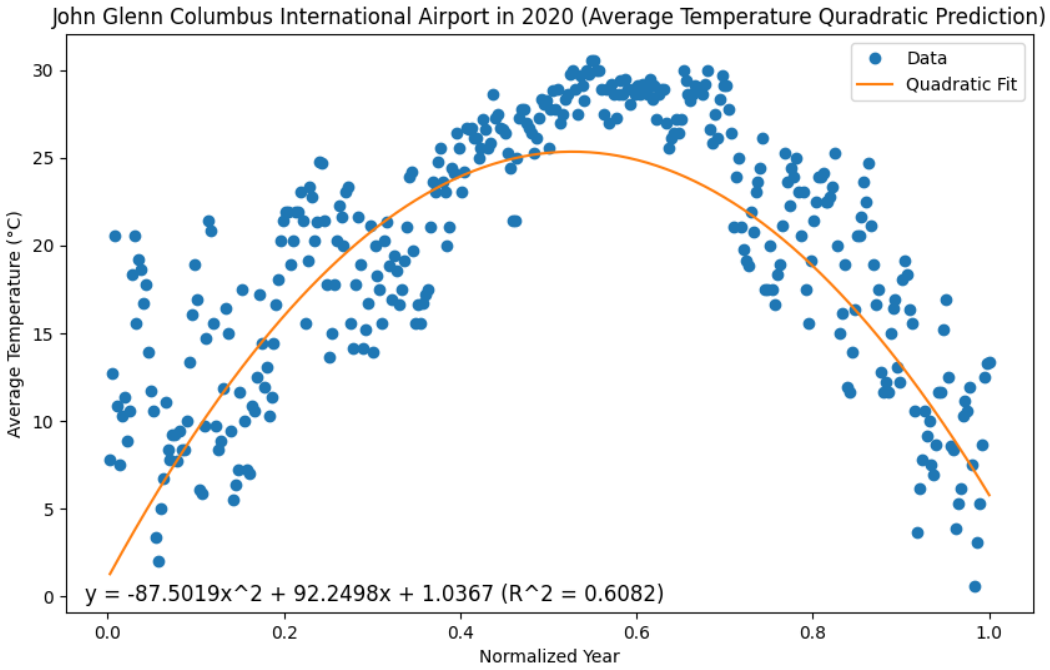

Fig. 5. John Glenn Columbus International Airport daily average temperature in 2020 with quadratic least square fit from the first 30 years. Shows an r-squared value of 0.6082 which is a more inaccurate fit for the data. The quadratic fit doesn't fit well and there are a lot of points higher than the predicted curve.

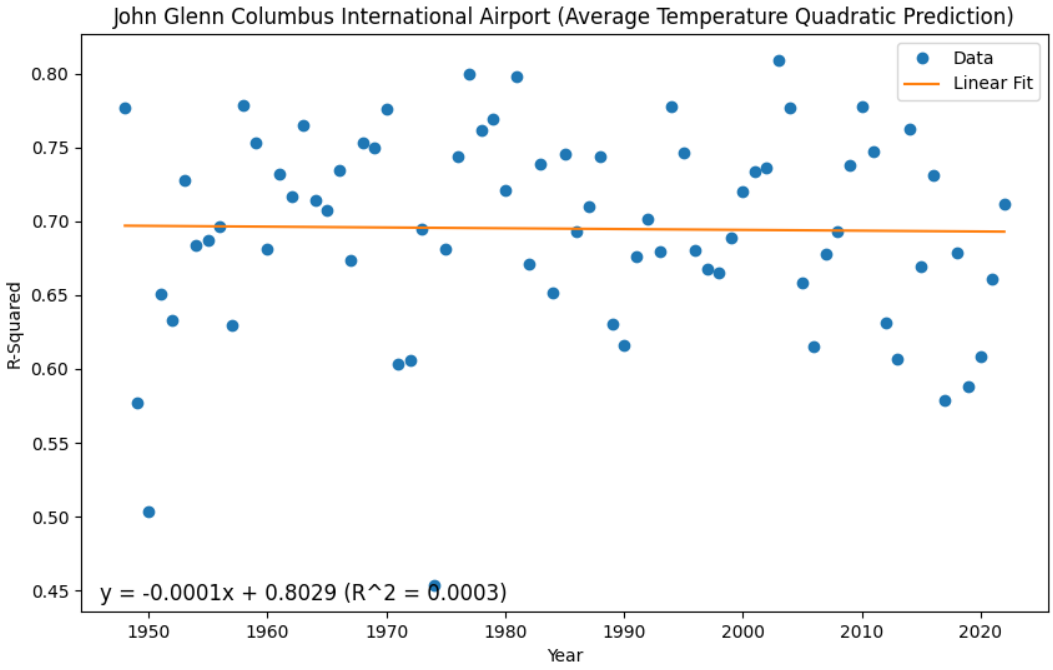

Fig. 6. John Glenn Columbus International Airport shows a negligible decrease in r-squared from the predicted average temperature for the first 30 years. The r-squared value is still low.

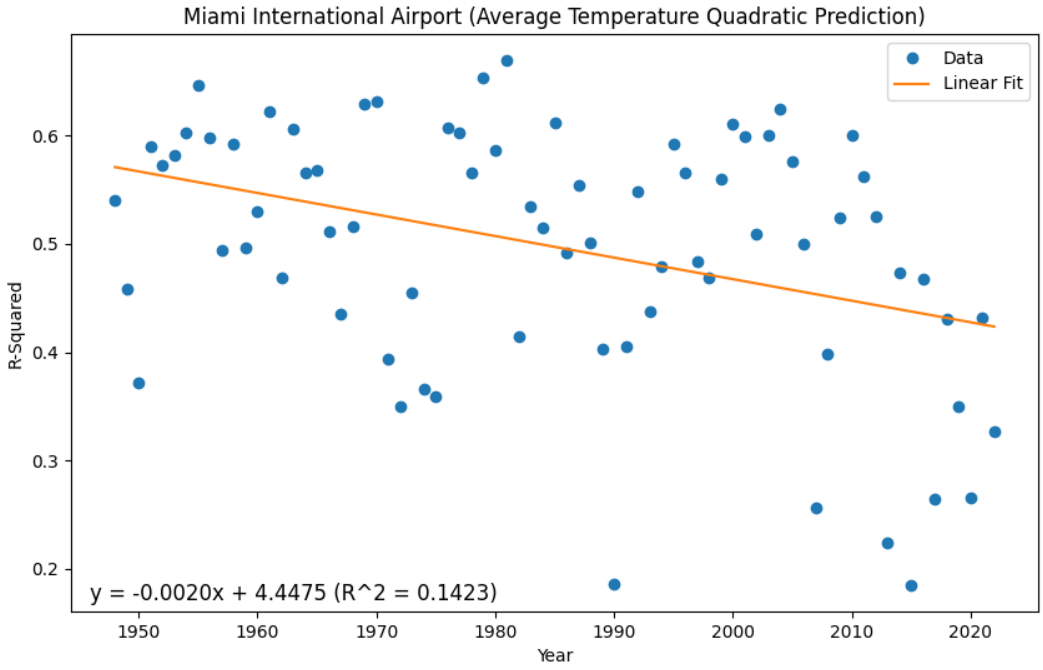

Fig. 7. Miami International Airport shows a steep decrease in r-squared from the predicted average temperature for the first 30 years. The r-squared value is still low.

When the r-squared value decreases, it means that the variability between the predicted and actual values is becoming greater. For both Columbus and Miami, there is a negative trend in the r-squared value over the years. However, in both cases the r-squared value of the graph is quite low as visually most of the data points do not fall on the line. This analysis is not conclusive evidence that climate change is occurring; however, the r-squared for Miami is higher than the prior analysis.

3.3 Comparing number of statistically extreme daily high temperature events per year

In the previous two sections, there has not been conclusive evidence that climate change in occurring. However in a recent publication, Katz and Brown show that the number of extreme events increase as climate change happens. A method to identify these extreme events is by looking at the number of occurrences when a day in a year has an high temperature value that exceeds the day's mean temperature plus 2 or 3 times the standard deviation.

As seen in Fig 8, 9, 10, 11, 14, and 15, there is a clear increase in the frequency of these extreme high temperature events. However, stations such as Cornell University, show barely any correlation in the data. This means that some cities are not experiencing climate change as much as other cities.

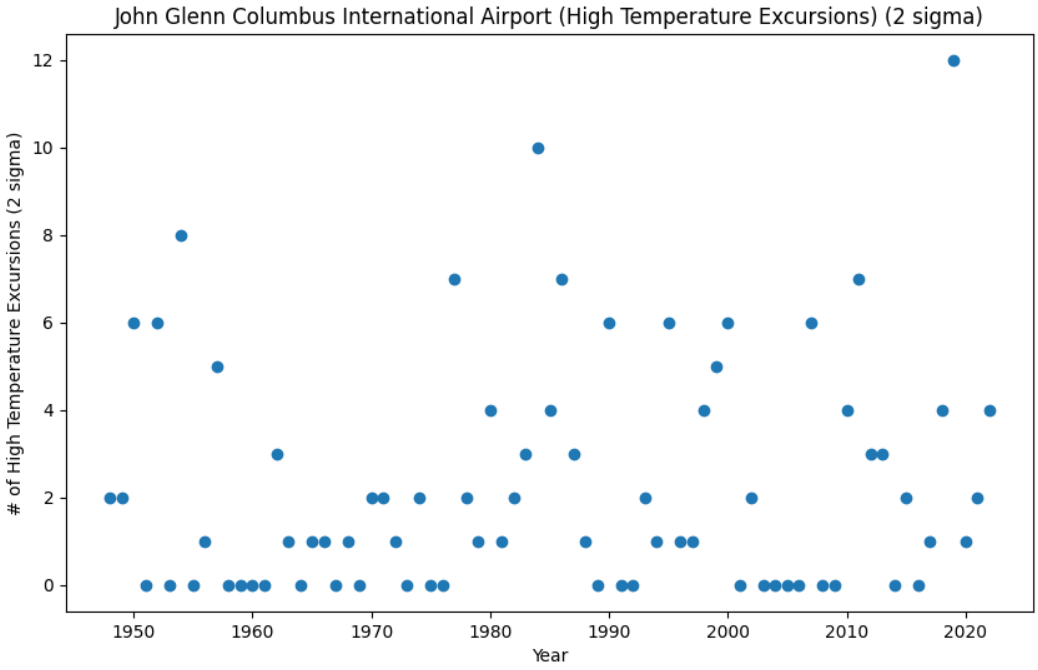

Fig. 8. John Glenn Columbus International Airport shows a positive trend in the frequency of high temperature excursions that are more than 2 standard deviations away from the average.

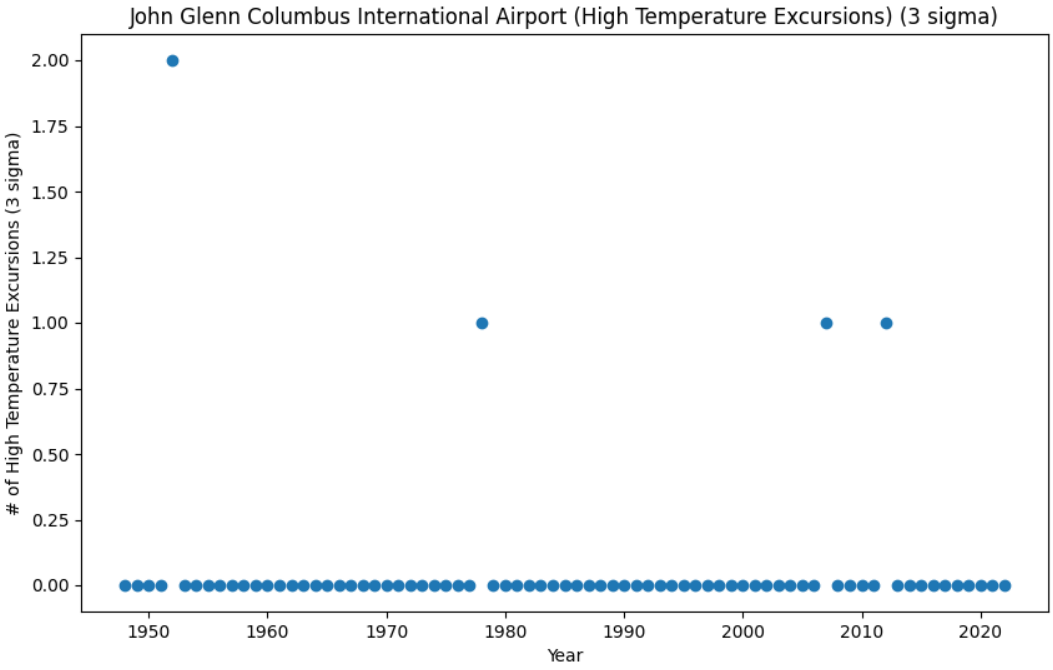

Fig. 9. Miami International Airport shows a positive trend in the frequency of high temperature excursions that are more than 2 standard deviations away from the average.

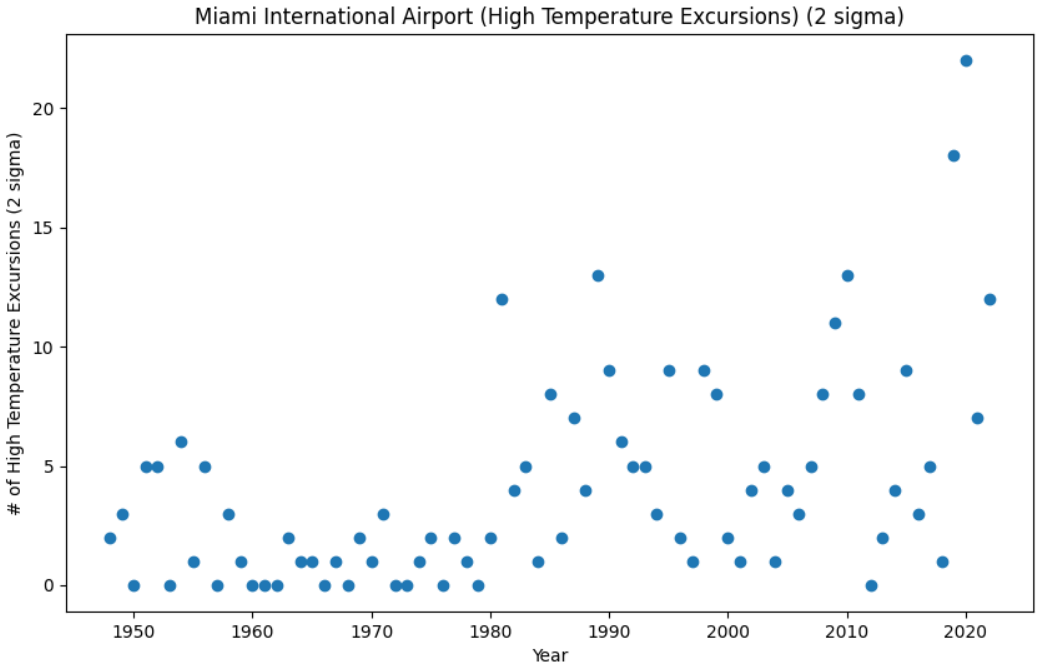

Fig. 10. John Glenn Columbus International Airport shows a positive trend in the frequency of high temperature excursions that are more than 3 standard deviations away from the average.

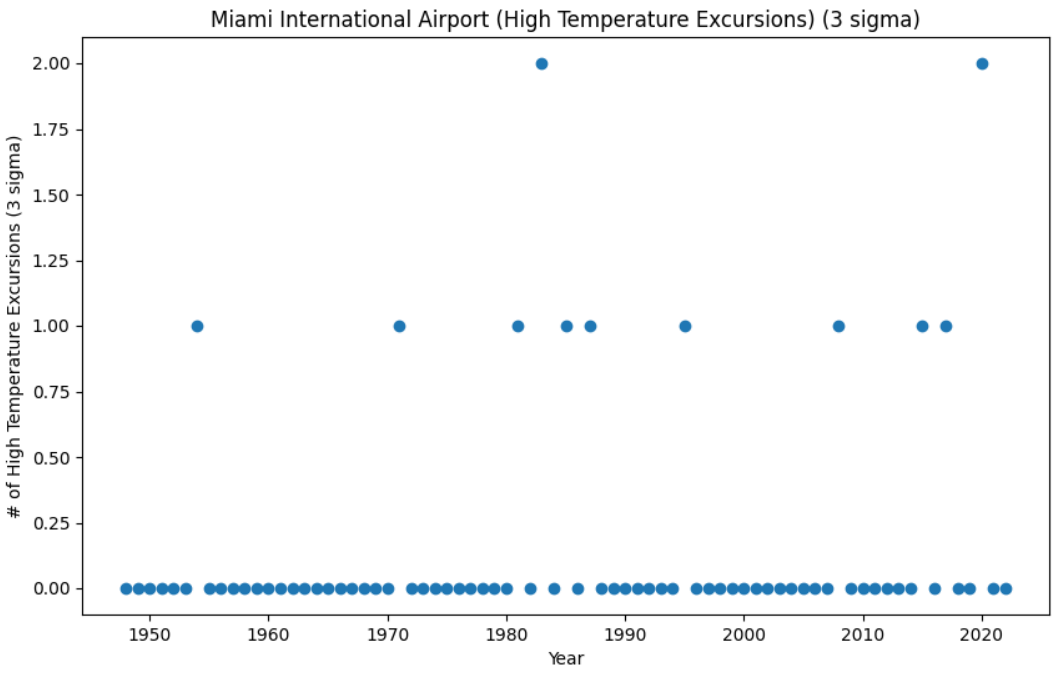

Fig. 11. Miami International Airport shows a positive trend in the frequency of high temperature excursions that are more than 3 standard deviations away from the average.

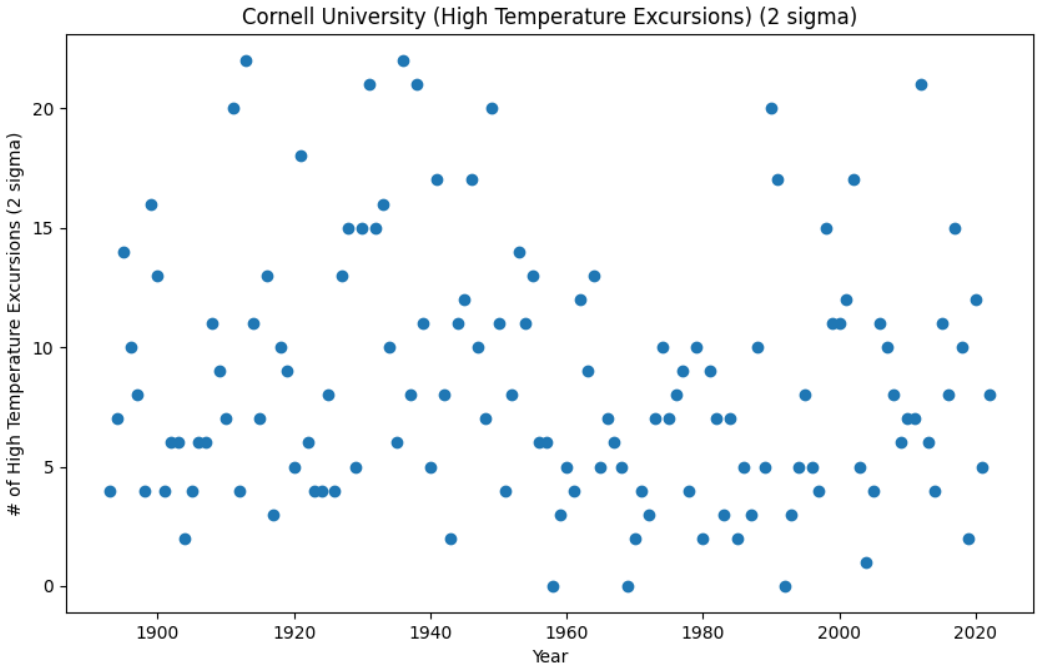

Fig. 12. Cornell University shows no correlation in the frequency of high temperature excursions that are more than 2 standard deviations away from the average.

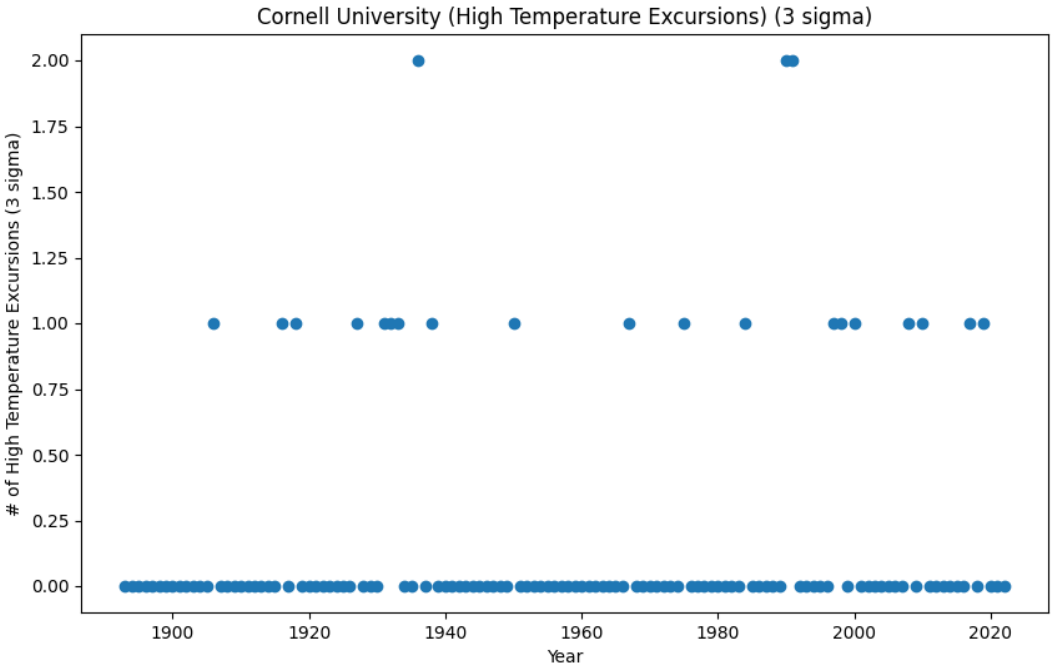

Fig. 13. Cornell University shows no correlation in the frequency of high temperature excursions that are more than 3 standard deviations away from the average.

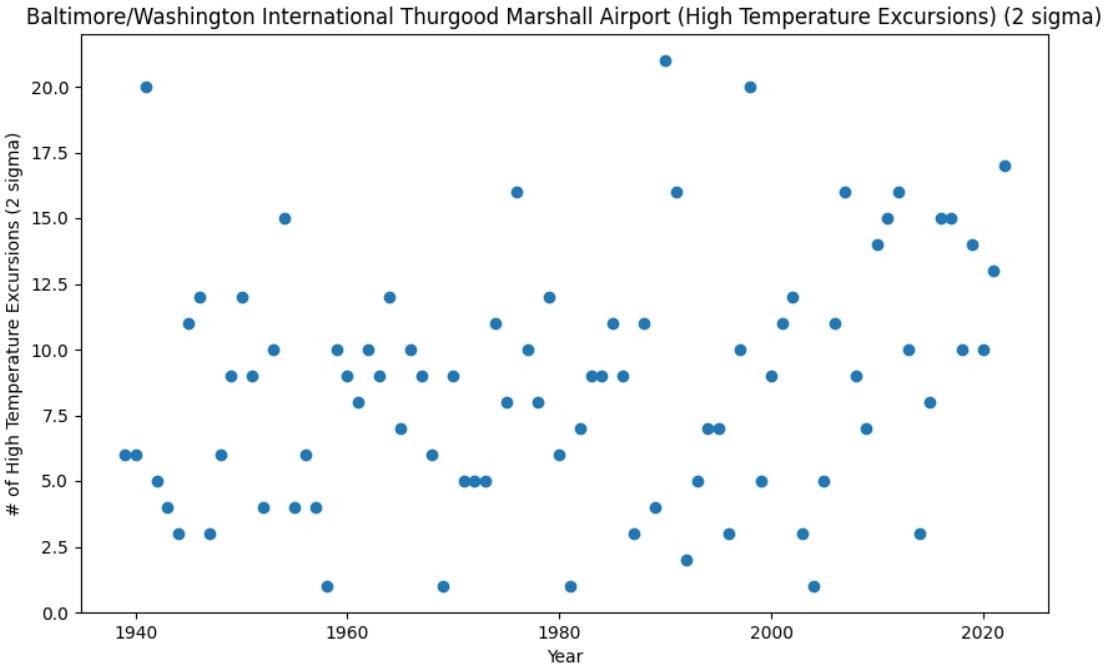

Fig. 14. Baltimore/Washington International Thurgood Marshall Airport shows a positive trend in the frequency of high temperature excursions that are more than 2 standard deviations away from the average.

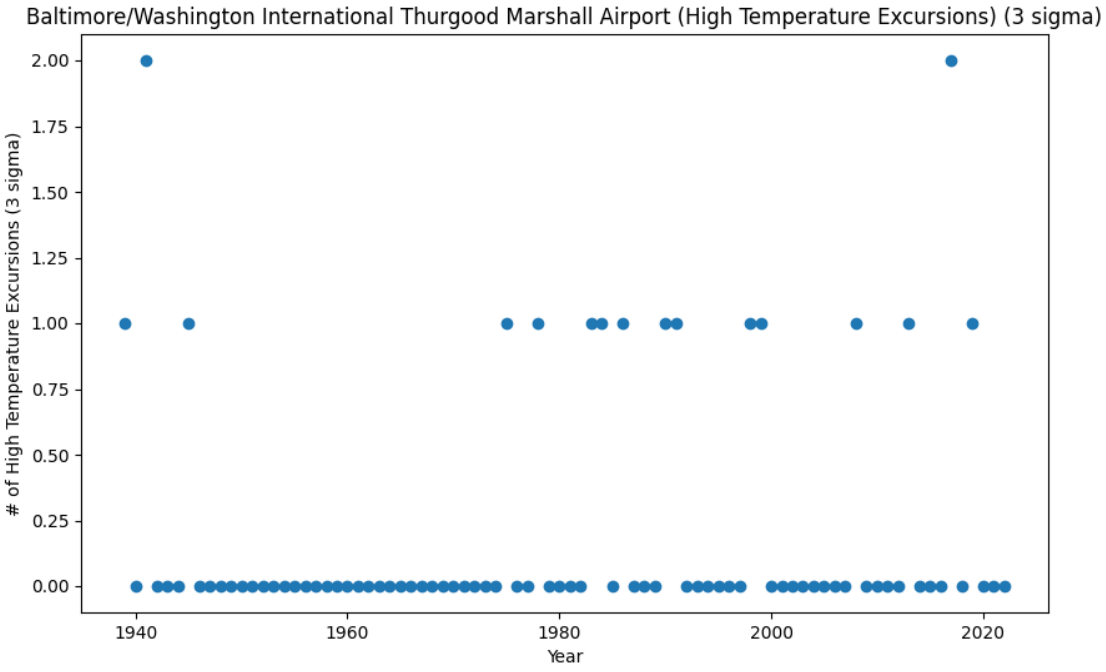

Fig. 15. Baltimore/Washington International Thurgood Marshall Airport shows a positive trend in the frequency of high temperature excursions that are more than 3 standard deviations away from the average.

Although the data varied from city to city, the data for Columbus, Miami, and Baltimore show a positive trend in the frequency of these temperature excursions. However, some cities such as Cornell show less shifts than others. The average temperature is becoming more variable and shows the impact of climate change. This variability is a good indicator of climate change (Katz and Brown). We are finding a similar result to Katz and Brown in Miami and Columbus.

4. Conclusion

Climate change is a complex topic. Despite all the data collected by scientists from ocean heat to species growing earlier, an average person has a difficult time grasping climate change. It is not easy to conclude that climate change is impacting us because there is too much variation in weather from day to day. Looking at a single day every year yields a low r-squared, and is not conclusive of climate change Though creating a regression for the year shows a better r-squared, it is not enough to make a conclusion. The extreme events visually and statistically show that the number of extreme events show more consistent data. These extreme events show variability and sensitivity which is even more important than increases in average temperature (Katz and Brown). The small subset of data shows that some cities will experience more significant changes than others as seen in the graphs.

This variation might be because of a city's proximity to water or the equator, population, or even topography. This definitely requires more research.

5. References

NOAA Climate.gov. (2021). *Global Temperature Anomalies - Graphing Tool*. https://www.climate.gov/maps-data/dataset/global-temperature-anomalies-graphing-tool

Katz, R. W., & Brown, B. G. (1992). Extreme events in a changing climate: Variability is more important than averages. *Climatic Change*, *21*(3), 289–302. https://doi.org/10.1007/bf00139728