Wenxin Liu

MATH 138

Final Project

An analysis on the temperature increase in the past four Summers at Hofstra

Introduction

As the attention of the world shifts to the ongoing pandemic, we may not forget the global warming that used to be a hot topic people talk about. People used to discuss how polar bears and penguins are losing their habitats because of global warming; another important and relatable proof that people use to argue is that summers are getting hotter each year, which is proved by several news sources. Based on a study that “used 22 climate models” on summer temperatures, an article from CNN concluded that the end of the 21st century would be ”so extreme that they will not have been experienced previously.” While, 58% of the world updates their highest monthly temperature average each year. The cause of which is associated with the rising global greenhouse gas emissions (Christensen). Another article from the New York Times argues that the summer temperature in every decade from 1980s to 2015 keeps getting higher. This means that “most summers are now either hot or extremely hot compared with the mid-20th century” (Popovich & Pearce). Thus, this study aims to verify this increasing temperature trend locally based on the weather station at Hofstra University in the past four years through an ANOVA test and Tests of the Statistical Hypothesis about two means.

Data collection & cleansing

The data was downloaded from WeatherSTEM Portal for Nassau County, specifically the local weather station at Hofstra Soccer Stadium located at the North campus along Hofstra Northern Blvd. The weather station is shown in the figure below.

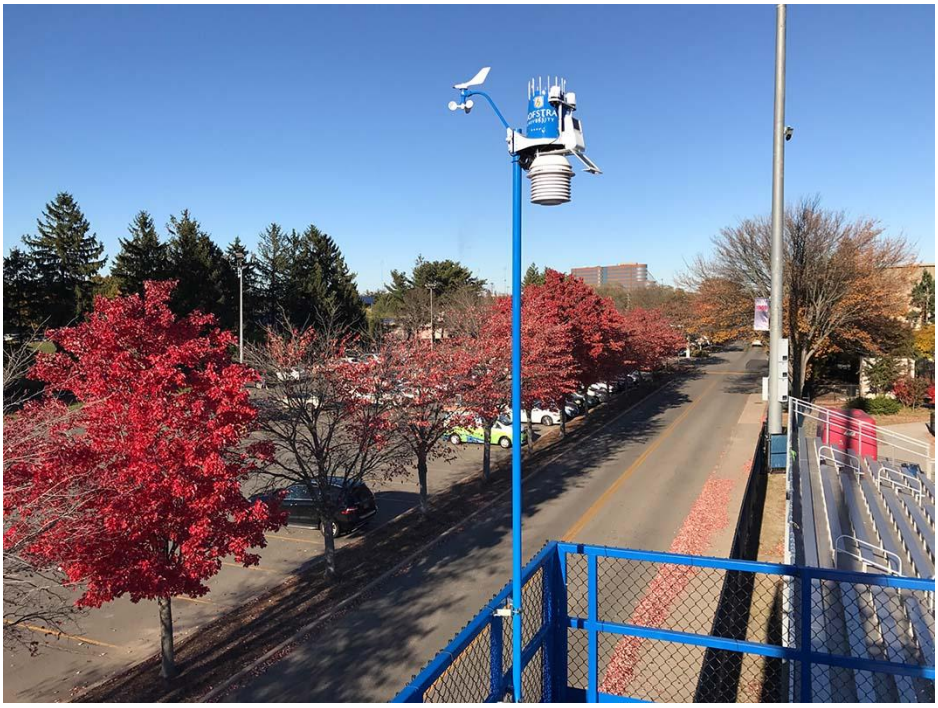


Figure 1: Hofstra Soccer Stadium weather station from nassau-ny.weatherstem.com

From the website, only a temperature readings column, together with a timestamp column, was downloaded as the initial raw data. The Thermometer column shows the temperature in degree Fahrenheit recorded at the instance of time in the corresponding timestamp column, as the figure below shows.

```
Record ID,Timestamp,Thermometer
4339408,2020-06-01 00:00:29,56.9
4339411,2020-06-01 00:01:29,56.8
4339417,2020-06-01 00:02:29,56.8
4339415,2020-06-01 00:03:30,56.7
4339419,2020-06-01 00:04:30,56.7
4339425,2020-06-01 00:05:30,56.7
4339424,2020-06-01 00:06:30,56.6
4339428,2020-06-01 00:07:30,56.6
4339433,2020-06-01 00:08:30,56.6
4339434,2020-06-01 00:09:31,56.5
4339435,2020-06-01 00:10:31,56.5
4339441,2020-06-01 00:11:31,56.4
```

Figure 2: Raw data downloaded from nassau-ny.weatherstem.com

However, there is only one timestamp column to signify the time, date, month, and year the particular reading was recorded from; no separate columns were given for attributes of time, date, month, and year. The author needs to process and strip down the time information into more precise columns to run ANOVA test on. This was done through a while loop in R to loop through each row of data to strip and append the required information separately onto the newly created target dataframe. Other than the need to strip down time information, data cleansing process on the temperature data was also required for weather data. Because weather data was collected at instances of time, there is no guarantee that each reading is exactly one minute apart from each other, which is also the reason why the timestamp is accurate to second. Another concern is that not every minute has a recording for it, while there are more than one recording in some minutes. Therefore, uniformizing the raw data is crucial to guarantee that the minutes with more than one temperature readings do not weigh more than those who do not when the average temperature is computed. Thus, the raw data was first converted into data of 5-minute intervals through the while loop mentioned earlier when stirping the timestamp as well. The data of 5-minute intervals guarantees that each minute of multiples of 5 has a corresponding temperature that is averaged based on whatever was available during the 5 minute interval from the raw data, which comes out like the figure below shows.

```
"","DATETIME","YR","MON","DAT","TIM","TEMP"
"1",202006010000,2020,6,1,0,56.9
"2",202006010005,2020,6,1,5,56.74
"3",202006010010,2020,6,1,10,56.56
"4",202006010015,2020,6,1,15,56.38
"5",202006010020,2020,6,1,20,56.2
"6",202006010025,2020,6,1,25,56.04
"7",202006010030,2020,6,1,30,55.88
"8",202006010035,2020,6,1,35,55.72
"9",202006010040,2020,6,1,40,55.6
"10",202006010045,2020,6,1,45,55.42
```

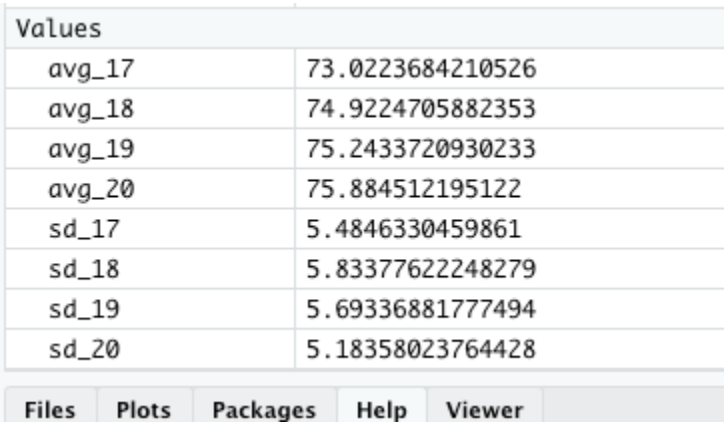
Figure 3: Processed data of 5-minute intervals

Notice that we still possess a column “DATETIME” to be the primary key where each row has a different value for this attribute. After the data has been cleansed, the daily average for each day can now be calculated through simply taking the average of the 120 data points (12

intervals/hour * 24 hours) on each day. The daily average temperature is what we work on and use in the subsequent statistical analysis.

Statistical Analysis

This project utilized ANOVA test and tests of statistical hypotheses for two means to examine the relationship between the temperatures of summer in 2017 through 2020 period. Before any tests were performed, the student first examined the yearly average for the entire three months to confirm that the tests are worth conducting. The yearly temperature average for the entire three months is computed by taking the average of the daily temperatures, where the values are as below.



Values

avg_17	73.0223684210526
avg_18	74.9224705882353
avg_19	75.2433720930233
avg_20	75.884512195122
sd_17	5.4846330459861
sd_18	5.83377622248279
sd_19	5.69336881777494
sd_20	5.18358023764428

Files Plots Packages Help Viewer

Figure 4: Averages and standard deviation for yearly average in June - Aug

An increasing trend in the yearly averages can be observed by looking at the pure averages, which confirms that the tests are worth conducting. Also, we observe that the standard deviation of the four datasets are no farther than twice of any standard deviation in these datasets, which suffices the initial requirement of common variances to conduct an ANOVA test on.

An ANOVA test was performed initially to see whether the four data sets have the same mean in order to later decide whether pairwise hypothesis testings on two means are worth conducting. The other requirement to conduct an ANOVA test was that each of the datasets comes from a normal distribution; this is examined through looking at the qqnorm graphical representation of the four datasets as follows.

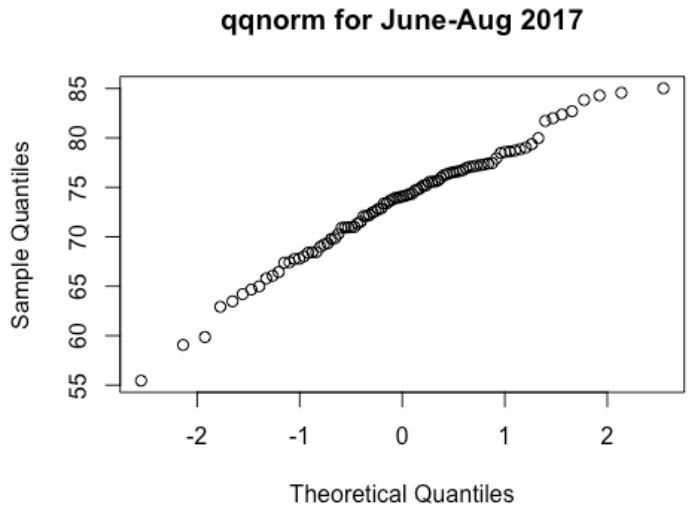


Figure 5: qqnorm for June - Aug 2017

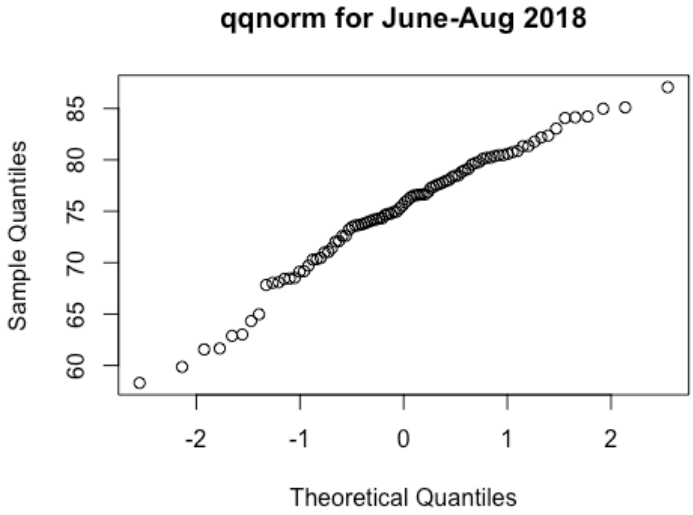


Figure 6: qqnorm for June - Aug 2018

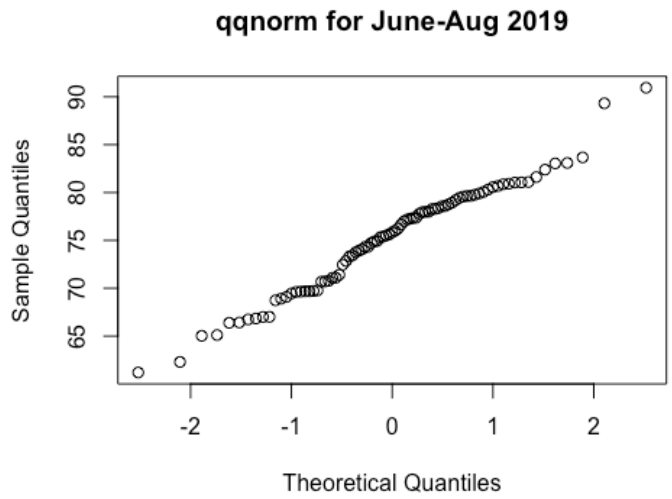


Figure 7: qqnorm for June - Aug 2019

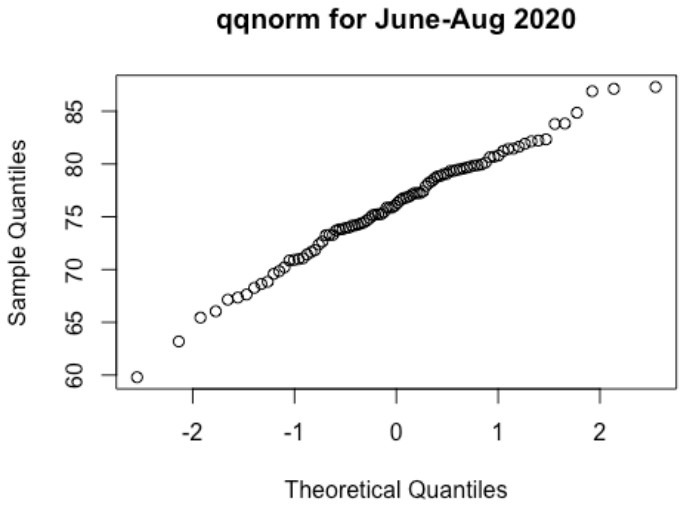


Figure 8: qqnorm for June - Aug 2020

We are able to observe that each of the four qqnorm graphs looks like a straight line, meaning that they are normal enough for our purpose to conduct ANOVA test on in this case. After we have examined the two requirements for ANOVA test, we can now conduct it on our dataset with the column for hourly temperature as the data points and the column for year number as the column of category. The null hypothesis (H_0) is that the four means of the four datasets are equal, while the alternative hypothesis (H_1) is that at least one of the four means is different from the other. The figure below shows how this ANOVA test is conducted on R.

```
> June.Aug17.20 <- read.csv("~/R/June-Aug17-20.txt")
>   View(June.Aug17.20)
> tada = aov(June.Aug17.20$da_TEMP~June.Aug17.20$da_YR)
> summary(tada)
                     Df Sum Sq Mean Sq F value  Pr(>F)
June.Aug17.20$da_YR   1    284  283.62   8.813 0.00319 **
Residuals           359  11554   32.18
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 9: ANOVA test

We see that the p-value of our dataset is computed to be 0.00319 by R. This is lower than any alphas on the tables by the end of the book, which means that with any reasonable alphas the data would not support the null hypothesis (H_0) that the means are the same. Thus, three pairwise hypothesis testing for two means shall be considered to conduct in order to examine whether the increasing trend can be observed in every year; out of the four years, we conduct hypothesis testing on only means of two consecutive years, resulting in 3 tests in total. Also, since we have three months of daily average data for each year and the sample size in each year is bigger than 30 while the actual variance is unknown, case #3 of the hypothesis testing for two means applies; thus, a normal distribution is utilized to fit the test statistic. The setup for the hypothesis testing is as follows.

X: later year dataset	Y: previous year dataset
n: sample size of X $\overline{X}$: sample mean of X s_x: sample standard deviation of X	m: sample size of Y $\overline{Y}$: sample mean of Y s_y: sample standard deviation of Y

Null hypothesis (H_0): $\mu_X = \mu_Y$

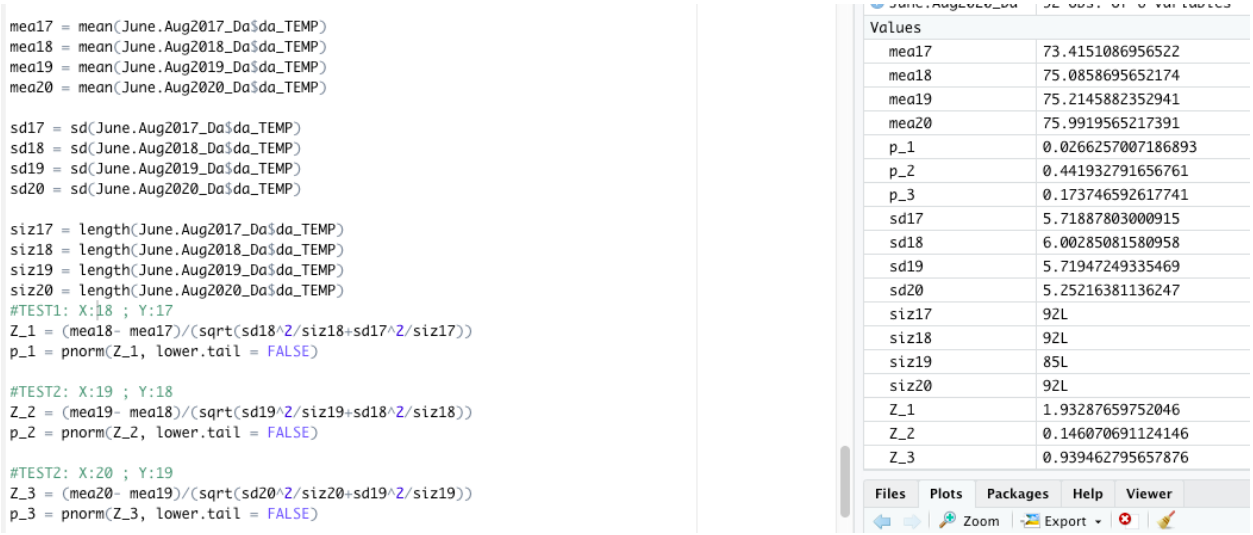
Alternative hypothesis (H_1): $\mu_X > \mu_Y$

Test statistics: $$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Then, the three tests are then carried with the following setups where:

TEST1: X: 2018; Y: 2017
TEST2: X: 2019; Y: 2018
TEST3: X: 2020; Y: 2019

The actual computations are carried out on R to minimize the possibility of human error involvement. The test statistics were computed for each of the three tests separately; then, the p-values of the test statistics were looked up using the command “pnorm” with the option “lower.tail” set to “FALSE” because the alternative hypothesis is that the average of X is bigger than the average Y resulting in a positive test statistic for the cutoff; therefore, upper tailed p-value is desired in this case. A screenshot of the calculation done in R is shown as the figure below.



```
mea17 = mean(June.Aug2017_Da$da_TEMP)
mea18 = mean(June.Aug2018_Da$da_TEMP)
mea19 = mean(June.Aug2019_Da$da_TEMP)
mea20 = mean(June.Aug2020_Da$da_TEMP)

sd17 = sd(June.Aug2017_Da$da_TEMP)
sd18 = sd(June.Aug2018_Da$da_TEMP)
sd19 = sd(June.Aug2019_Da$da_TEMP)
sd20 = sd(June.Aug2020_Da$da_TEMP)

siz17 = length(June.Aug2017_Da$da_TEMP)
siz18 = length(June.Aug2018_Da$da_TEMP)
siz19 = length(June.Aug2019_Da$da_TEMP)
siz20 = length(June.Aug2020_Da$da_TEMP)
#TEST1: X:18 ; Y:17
Z_1 = (mea18- mea17)/(sqrt(sd18^2/siz18+sd17^2/siz17))
p_1 = pnorm(Z_1, lower.tail = FALSE)

#TEST2: X:19 ; Y:18
Z_2 = (mea19- mea18)/(sqrt(sd19^2/siz19+sd18^2/siz18))
p_2 = pnorm(Z_2, lower.tail = FALSE)

#TEST2: X:20 ; Y:19
Z_3 = (mea20- mea19)/(sqrt(sd20^2/siz20+sd19^2/siz19))
p_3 = pnorm(Z_3, lower.tail = FALSE)
```

Values	
mea17	73.4151086956522
mea18	75.0858695652174
mea19	75.2145882352941
mea20	75.9919565217391
p_1	0.0266257007186893
p_2	0.441932791656761
p_3	0.173746592617741
sd17	5.71887803000915
sd18	6.00285081580958
sd19	5.71947249335469
sd20	5.25216381136247
siz17	92L
siz18	92L
siz19	85L
siz20	92L
Z_1	1.93287659752046
Z_2	0.146070691124146
Z_3	0.939462795657876

Files | Plots | Packages | Help | Viewer

Zoom | Export

Figure 10: Pairwise hypothesis testings for two means

We are able to see from the environment window on R that the statistics and the p-values of the three tests are as follows.

	Test statistic	p-value
TEST1	1.9329	0.02663
TEST2	0.1461	0.4419
TEST3	0.9395	0.1737

An appropriate alpha for these three tests can be 0.05, where only TEST1 would reject the null hypothesis that the two means are equal. While, with alpha= 0.05, the data in TEST2 and TEST3 support the null hypothesis.

Results

Although the ANOVA suggests that the means of the four datasets are not the same, we do not perceive a strictly increasing trend in all three pairwise hypothesis testings. A statistically significant p-value (0.02663) is only observed in TEST1 such that the null hypothesis that two means are equal is not supported by the data, where we can conclude that there was an increase in the summer temperature average in 2018 compared to the temperature in summer

2017. But the data in TEST2 and TEST3 support the null hypothesis that the two means are the same, several reasons may be accounted for as the next paragraph shows. This means that ANOVA tests may not be the best to look at the dataset with order involved in it, because ANOVA tests treat all dataset equally and compare all of the dataset at the same time.

This may be a result from the fact that environmental agreements among developing countries were signed when the issue of global warming grabbed more attention from the public before 2018; people are more conscious of the ongoing issue of global warming. Also, it may be inferred that due to the coronavirus pandemic which began in early 2020, more people stayed home and less cars and manmade gas emissions were on roads. However, the p-value of test 3, whose alternative hypothesis was that summer temperature in 2020 is higher than the summer temperature in 2019, was only 0.1737 which is way lower than the p-value of test 2 (0.4419). The most prevalent view should be that people are more conscious of the problems brought by global warming as more news is spread. The author wishes that, with more reports analyzing the effect of climate change and global warming, more audiences can realize the necessity of protecting the environment and save the adorable polar bears and penguins living far away from us yet suffering from what humans did.

Generally, this study served as a good basis in examining the increasing temperature trend during the recent four years locally. However, only four years of data was utilized in this study, which is mainly because the weather data collected at the Hofstra weather station can only be retrieved up to the year 2017. However, the downside with more years of data is that more pairwise hypothesis testing will be expected to be carried out, which can be cumbersome. However, the hypothesis testing may be automated on R given enough motivations.

Works cited:

Christensen, Jen. “Future Summers Will 'Smash' Temperature Records Every Year.” *CNN*, Cable News Network, 17 June 2019, www.cnn.com/2019/06/17/health/climate-change-summer-temperatures-study.

Popovich, Nadja, and Adam Pearce. “It's Not Your Imagination. Summers Are Getting Hotter.” *The New York Times*, The New York Times, 28 July 2017, www.nytimes.com/interactive/2017/07/28/climate/more-frequent-extreme-summer-heat.html.

Keneally, Meghan, and Sandell, Clayton. “These 5 Statistics Show Why We're Experiencing Historically Hot Weather.” *ABC News*, ABC News Network, abcnews.go.com/US/statistics-show-experiencing-historically-hot-weather/story?id=64438226.

https://nassau-ny.weatherstem.com/hofstrasoccer

Dear Dr. Franklin,

In this project, I tried to examine the increasing temperature trend in the past four summers locally at Hofstra with an ANOVA test and three hypothesis tests for two means.

I concentrated most of my efforts on cleansing the data, because sometimes even small misthinking can lead to big mistakes and mess up the output file with incorrect data. Also, it was quite difficult to decide what statistical techniques to use to achieve my goals. The goal can usually be harder to implement given the tools on my hand. It takes time to think through the procedures of what tests I can use to prove my point.

Here's what I think that the strongest parts of my project are. Despite the ANOVA test saying that the means are not the same with an extremely low p-value, only one out of the three pairwise hypothesis tests for two means of two consecutive years rejected the null hypothesis that the two means are the same. The aspects of how different tests for mean deal with this set of data is examined.

I'm especially proud of me being able to cleanse and normalize the raw data from scratch and did not make seemingly obvious errors as the results look reasonable to me.

What I struggled with most was that during the data conversion, I carelessly made mistakes all the time and only noticed them when I have all four files of data converted. Small logical errors can lead to big mistakes in the output when constructing the pseudo code. Also, it was observed that 5 days in the August of 2019 went totally missing from the raw data file I downloaded, presumably because of some technical interruptions of the weather station physically. This resulted in that the year 2019 has 7 less samples compared to the standard 92 days (30+31+31) because I removed the two days with only hours of data to avoid having an unreasonably low daily temperature in the dataset. The bright side was that the two statistical analysis methods that I intend to conduct both allow data sets of different sizes. And I took averages of all days in each summer, so as long as the number of days in the summer did not vary much, the results should not be affected much.

As you read this analysis, please keep in mind that adorable polar bears and penguins are wishing to be saved by humans to slow down the ongoing global warming by protecting the environment.

Sincerely,
Wenxin