Data Analysis on Historical Weather of Kyiv, Ukraine

Brad Kreider

Abstract

This paper is a summary of data gathering and statistical analysis performed on historic weather data on the city of Kyiv, Ukraine. Our data analysis will be focusing primarily on 2 questions of interest:

1. How have historic average temperatures in Kyiv changed over time?
2. Have historic average high temperatures in Kyiv increased over time?

1. Data Gathering

The historical weather data that is used for this analysis is sourced from Open-Meteo, a free and open source weather API for non-commercial use. I was able to pull the data from the Kyiv region using the Open-Meteo API, my source code is include below:

https://github.com/bradley-kreider/Data-Analysis-Project---Kyiv-Weather-Data

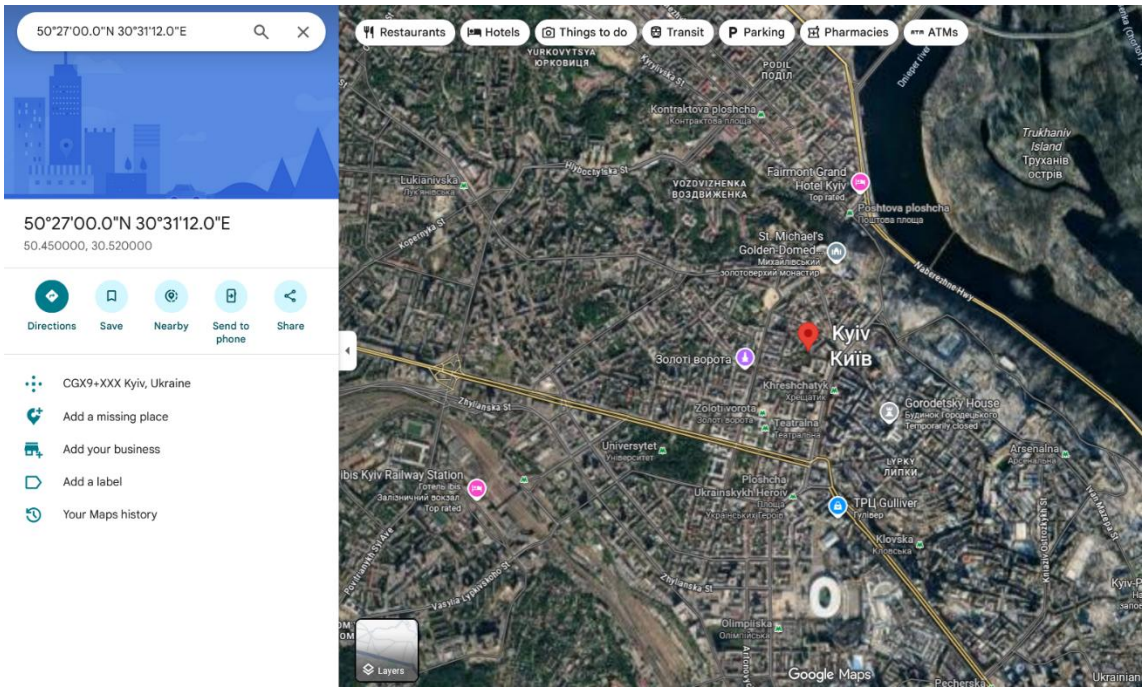
1.1 Data Validity

Open-Meteo is a reputable source for historical weather data. Their extensive database spans over 80 years of highly detailed and accurate weather information, which made them the perfect candidate for this analysis.

In terms of my source code where I obtained the data using the Open-Meteo API, the following code segment demonstrates the parameters which were included when I executed the query:

```
1.    params = {
2.        "latitude": 50.45,
3.        "longitude": 30.52,
4.        "start_date": "1940-01-01",
5.        "end_date": "2025-12-31",
6.        "daily": "temperature_2m_mean,temperature_2m_max,temperature_2m_min",
7.        "timezone": "Europe/Kyiv"
8.    }
```

The latitude and longitude that the data is being extracted from is (50.45, 30.52) which corresponds to the city center of Kyiv, Ukraine.



The start and end dates cover the range from years 1940 – 2025, which is the time frame that we are considering in this analysis.

The query also requests for : “temperature_2m_mean”, “temperature_2m_max”, “temperature_2m_min”

Overall, the data for this analysis has been sourced reliably and is trustworthy and reliable for the purposes of this analysis

1.2 Data Manipulation

The query that was ran to achieve the data from the Open-Meteo API provided us with daily summary statistics for mean, maximum, and minimum values in °C at the 2-meter level for Kyiv Ukraine from January 1, 1940 to December 31, 2025. Thus we have 1 row entry for each day in this range containing each of these 3 temperature readings. The summary of the first 6 rows of the CSV file gathered from the Open-Meteo API is:

Code:

```
%%R
weather <- read.csv('kyiv_temperature_1940_2025.csv')
head(weather)
```

First 6 lines of `'kyiv_temperature_1940_2025.csv'`

```
      time                temperature_2m_mean temperature_2m_max temperature_2m_min
1 1940-01-01                  NA               -11.5               -16.8
2 1940-01-02                -14.4              -12.8               -16.6
3 1940-01-03                -15.5              -11.9               -17.7
```

```
5.  4 1940-01-04               -7.4               -5.1              -11.5
6.  5 1940-01-05               -8.8               -7.0              -12.2
7.  6 1940-01-06               -9.8               -6.7              -12.0
```

To properly look at the historical change in temperature in Kyiv, simply looking across the thousands of individual days will not do. For our analysis, we will need to aggregate our data for each year to achieve an average temperature for each year, as completed below:

```
 1. %%R
 2. # Convert time to Date and extract year
 3. weather$time <- as.Date(weather$time)
 4. weather$year <- as.numeric(format(weather$time, '%Y'))
 5.
 6. # Create annual average data frame, omitting NAs to ensure valid means
 7. annual_weather <- aggregate(temperature_2m_mean ~ year, data = weather, FUN = mean, na.rm =
TRUE)
 8. head(annual_weather)
```

Looking at our `'annual_weather'` data frame, it is now structured as:

```
1.  year           temperature_2m_mean
2. 1 1940            5.260822
3. 2 1941            5.333699
4. 3 1942            5.437808
5. 4 1943            7.296164
6. 5 1944            7.988798
7. 6 1945            6.768493
```

To properly look at the historic mean maximum temperatures in Kyiv, we will once again need to partition our data in half, looking specifically at the maximum temperatures. We will split the data at the year 1982, creating exactly 2 buckets (eras early/late) with exactly 43 years in each bucket:

```
1. %%R
2.
3. annual_max <- aggregate(temperature_2m_max ~ year, data = weather, FUN = mean, na.rm = TRUE)
4.
5. early <- annual_max$temperature_2m_max[annual_max$year <= 1982]
6. late  <- annual_max$temperature_2m_max[annual_max$year >= 1983]
7.
8. annual_max$era <- ifelse(annual_max$year <= 1982, "1940–1982", "1983–2025")
9.
```

Looking at our `annual_max` data frame, it is now structured as:

```
1. year temperature_2m_max       era
2. 1 1940           8.862295 1940–1982
3. 2 1941           8.623836 1940–1982
4. 3 1942           9.182740 1940–1982
5. 4 1943          11.043288 1940–1982
6. 5 1944          11.398361 1940–1982
7. 6 1945          10.238630 1940–1982
```

With our data properly collected and formatted, we are ready to start the analysis.

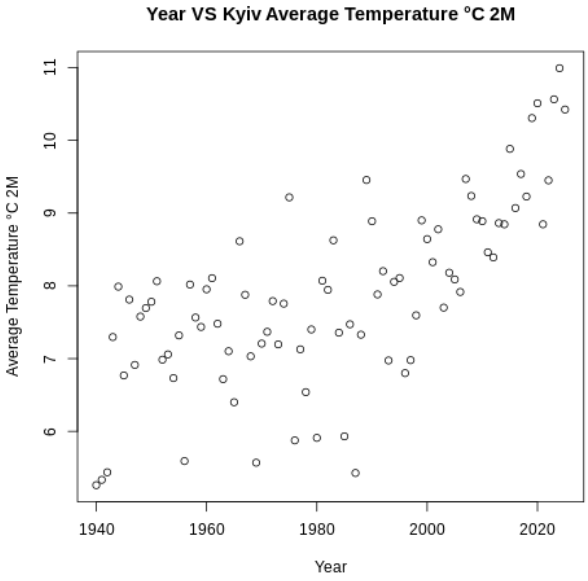
2. Data Summary and Visualization

Summary statistics for average temperature in Kyiv, Ukraine:

```
--- Summary Of Annual Weather Data 2m_mean ---
      year      temperature_2m_mean   predicted
 Min.   :1940   Min.   : 5.261      Min.   :6.396
 1st Qu.:1961   1st Qu.: 7.108      1st Qu.:7.130
 Median :1982   Median : 7.844      Median :7.865
 Mean   :1982   Mean   : 7.865      Mean   :7.865
 3rd Qu.:2004   3rd Qu.: 8.743      3rd Qu.:8.599
 Max.   :2025   Max.   :10.991      Max.   :9.333


Standard deviation - 1.25584

Range – (5.260822, 10.991257)

Spread - 5.730435
```

Looking at the summary statistics, it seems that we may have some lower and upper outliers, as we see actual temperature data somewhat well above and below the predicted values.

We also see a standard deviation of 1.25584 °C for the mean temperatures, indicating that taking the average over the year certainly serves to make the data more consistent with less variability. This is also reflected in the median/mean values of 7.844 °C and 7.865 °C respectively, which also shows that there is not a ton of skew within the data set.



Considering our summary statistics and looking at the graph of year VS average temperature in Kyiv, it appears that the data may follow a positive weak linear correlation. This will be important moving forward as we move towards analysis. We will look to see if the temperature in Kyiv increasing over time is a statistically significant trend.

We also see that we potentially have some lower outliers around the year 1980. This will potentially throw off any sort of linear trend to the data, and will likely be something to take into consideration for our conditions.

3. Data Analysis – Question 1

After sourcing our data, we can proceed forward with our analysis for our first question. All statistical analysis will take place at the .95 confidence level.

3.1 How have historic average temperatures in Kyiv changed over time?

Earlier, we noted that the historic yearly average temperatures in Kyiv as demonstrated graphically appeared to be roughly following a weak positive linear correlation. We will now examine this possibility from a formal statistical perspective.

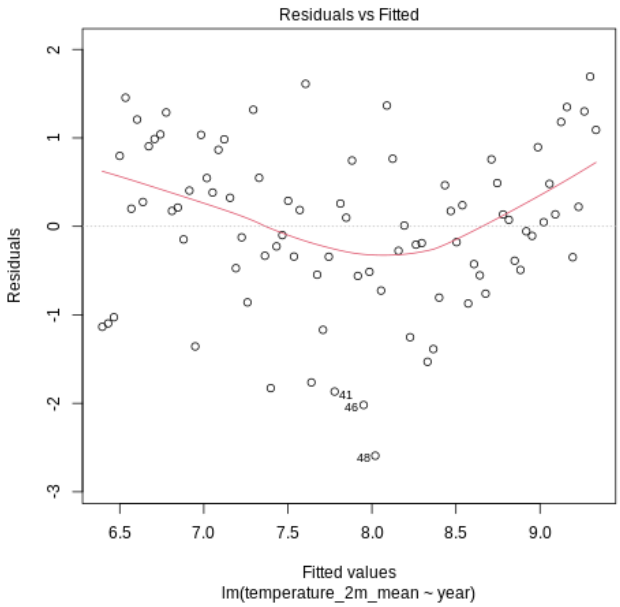
3.2 Linear Regression Model

Considering our data, we will be fitting a linear regression model onto our data for historic yearly average temperatures in Kyiv.

3.2.1 Conditions

Before we can fit the linear model to our data, we must properly check for our conditions for setting. It is important to note that our sample was not taken randomly, so we move into the conditions already knowing that we are in violation of randomness.
The conditions that we must satisfy exist as follows:

1. *Linearity*



Although our scatter plot for the data did appear to show a weak linear positive trend, looking at the graph of our residuals vs fitted values, we clearly see a curve that indicates that the

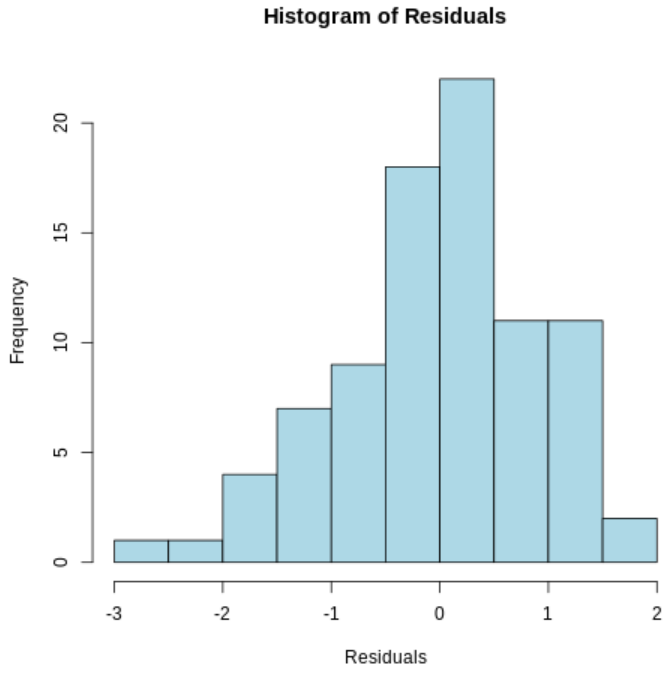
relationship between the mean temperature and year is not linear. We also do see some lower outliers for the years of 1941, 1946, and 1948. Overall, we definitely have some concerns about linearity for this setting, so we will proceed with caution.

Our linearity concerns indicate that a linear model might not be the best fit for this data. We acknowledge this reality and are going to proceed with this model for the purposes of simplification.

2. *Independence*

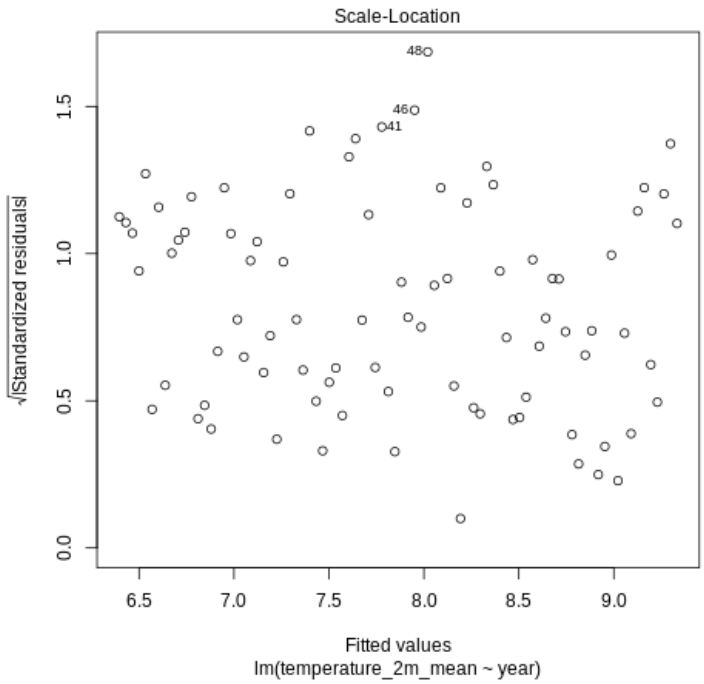
Because we are looking at historical weather data, it is inevitable that we must question independence between our samples for our average yearly temperature. It is entirely possible that the ending of one year could be affecting the weather for the beginning of the next year, although this would likely be in some very small way. Despite this, because we have a large sample size and are taking the yearly average of the average daily temperature for each day, we are minimizing the affects of any of this lack of independence within the study. We will move forward in recognition of this and proceed with caution.

3. *Normality*



Considering our normality condition, we can see that the residuals have a unimodal distribution that is roughly bell shaped and centered at 0. The left hand tail does appear to extend further than the right hand side, which could potentially be a result of the lower outliers that were addressed above. Overall, we do have a few normality concerns here, but we will proceed with caution.

4. Equal Variance



To check equal variance, we will look at the plot of fitted values vs the square root of the standardized residuals. Looking at the plot, we can see that the data is reasonably spread out in a random fashion. Thus, our equal variance condition is satisfied.

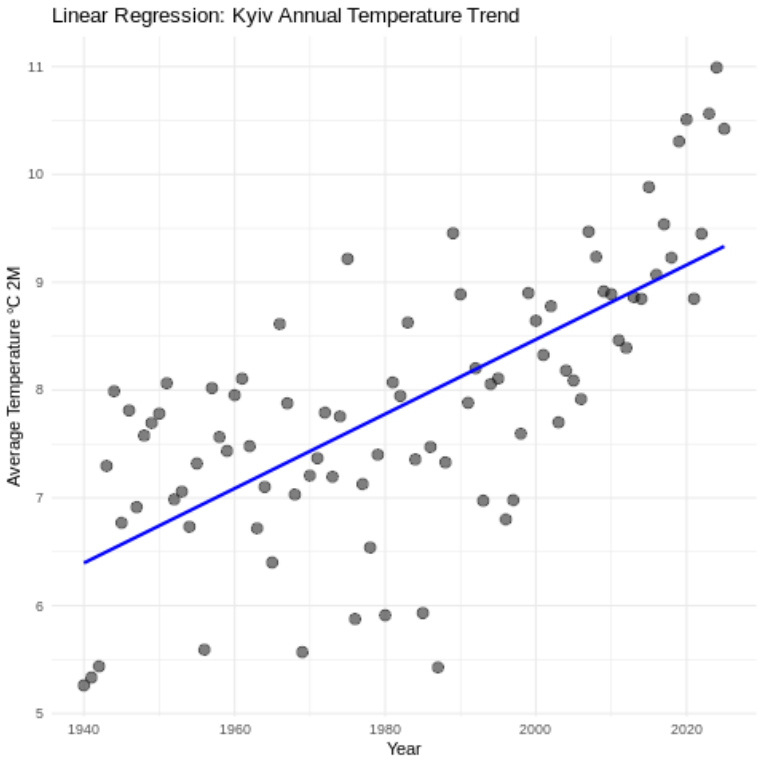
Considering all of the conditions, we definitely have concerns about linearity, independence, and normality, but we are going to move forwards and proceed with caution

3.2.2 Linear Regression Model Fit

We will use R to fit a linear regression model onto our average yearly temperature data set:

```
%%R
library(ggplot2)

linModel <- lm(temperature_2m_mean ~ year, data = annual_weather)
print(summary(linModel))

annual_weather$predicted <- predict(linModel, newdata = annual_weather)
```

This yields a linear model `'linModel'` that is overlayed onto the data below:



3.2.3 Summary Statistics

Here are the following summary statistics from the result of our linear regression model:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.5924 -0.5087  0.0857  0.6952  1.6928

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -60.631447   7.906278  -7.669 2.78e-11 ***
year          0.034550   0.003988   8.664 2.81e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.918 on 84 degrees of freedom
Multiple R-squared:  0.4719,       Adjusted R-squared:  0.4656
F-statistic: 75.07 on 1 and 84 DF,  p-value: 2.814e-13
```

3.2.4 Conclusions

As we can see, our linear regression model that uses year to predict average temperature in Kyiv, Ukraine is: $\hat{y} = 0.034550x - 60.631447$. From this model, we can see that we ultimately do have a positive slope between year and average temperature, indicating that temperature seems to be increasing in Kyiv over time. This is also indicated by our p-value of `2.814e-13` and F-statistic `75.07` which informs us to reject our null hypothesis that there is no linear relationship between year and average temperature.

The correlation coefficient R is approximately `0.6870`, so our positive correlation is of moderate strength.

Our R^2 value of `0.4719` indicates that 47.19% of the variation in the average temperature is explained by the year in our linear regression model. Over half of the variation is left unexplained, thus, coupling with our concerning conditions for this method, it is evident that this simple linear model is not the best fit for this data.

We can certainly take away that the data seems to have a positive slope in terms of temperature increase over time, but with our questionable conditions and low R^2 value, we should not use this model to predict average yearly temperature in Kyiv moving forwards.

4. Data Analysis – Question 2

4.1 Have historic average high temperatures in Kyiv increased over time?

In our first analysis of Kyiv's historical temperature data, we concluded that the Kyiv's mean temperature over the last 80 years has a positive slope. Thus, we will now take a closer look a the average mean high temperatures each year to see if they have increased over time.

4.2 Two Sample t-Test of Means

To examine if the average high temperatures in Kyiv have increased over time, we will divide our dataset into two halves ([1940 – 1982], [1983 – 2025]) and perform a two-sample t-test of means to look for a difference in the two periods.

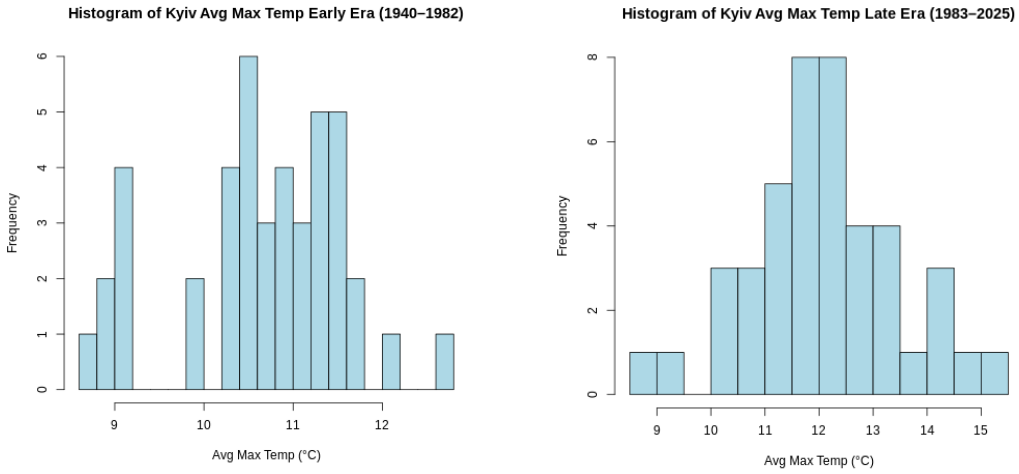
Although we are technically in a large sample setting with `n=43` for each sample, we will run this under a t-test. With this sample size, the Z and t distributions should be almost identical, with the t-test giving us a slightly more conservative perspective. Because we are using R for our analysis, running a t-test will also provide us with an easier workflow as we can use the built in t-test functionality in R.

4.2.1 Conditions

Before we can perform the two-sample t-test, we must examine the conditions required to perform the test. It is important to note that our sample was not taken randomly, so we move into the conditions already knowing that we are in violation of randomness.
The conditions that we must satisfy exist as follows:

1. Independence
   Once again, we are looking at historical weather data for this question. Although we are splitting our historical data into two separate groups, we still have some minor concerns about data from the 'early' section having an impact on data from the 'late' section. If there were to be any lack of independence as a result of this, the results would likely be very minimal, so we will recognize our slight concern with independence and proceed with caution.

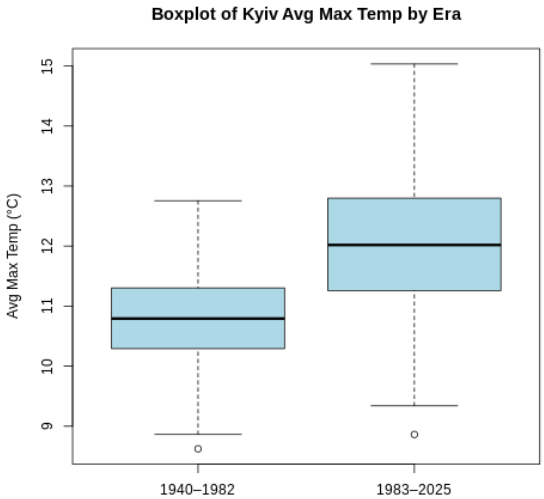
2. Normality
   Considering our normality condition, we will look at a histogram of each era (early, late) in order to view the distribution of each to check for normality:



Looking at our histograms, we have a few concerns. Starting with the early era, our distribution appears to be unimodal, but we definitely see some irregularities, including some upper outliers and high frequency in the lower regions. For the late era, we see a much more normal looking distribution. Still, the late era is not perfectly normal, being bimodal and having some lower outliers. We will take note of this and proceed forwards with caution.

3. Unknown Variance

For our unknown variance condition, we will look at a side by side boxplot of our early and late eras to check for equal variance between the two groups.



Considering our boxplot, we definitely see unequal lengths (unequal variances) between the two groups. The late era is also definitely skewed towards the higher temperature range, spreading between about 9.5 °C and 15 °C. Both groups also appear to have a lower outlier. Overall, we definitely have some concerns with our conditions for this test. We will move forwards with the test with caution about our conclusions.

4.2.2 t-Test

After having checked the conditions, we can proceed to performing our two-sample t-test of means.

```
1. t.test(late, early, alternative = "greater")
```

4.2.3 Summary Statistics

The summary statistics for our two-sample t-test are as follows:

```
 1. Two Sample t-test
 2.
 3. data:  late and early
 4. t = 5.6905, df = 84, p-value = 9.027e-08
 5. alternative hypothesis: true difference in means is greater than 0
 6. 95 percent confidence interval:
 7.  1.009349      Inf
 8. sample estimates:
 9. mean of x mean of y
10.  12.07031  10.64412
```

4.2.4 Conclusions

Looking at our summary statistics for our test, we can see that we have a p-value of `9.027e-08`, or approximately 0, indicating that the we have strong evidence to reject the null hypothesis that the two different eras (early and late) have the same mean maximum temperature. In fact, we can see that the late era's mean maximum temperature of `12.07031` is significantly higher than the early era's mean maximum temperature of `10.64412`.

Thus, addressing our second question for this analysis, historic high average temperatures in Kyiv have certainly increased over time.

5. Conclusions

After having performed data collection and statistical analysis with R, we can now consider what our conclusions state about the broader context of Kyiv's weather.

5.1 General Conclusion

Both our linear regression model of average yearly temperatures and our two sample t-test of means for the average maximum temperature confirmed that the temperature in Kyiv has gotten warmer over the period of interest 1940-2025.

5.2 Limitations

Despite this overarching conclusion, we must reiterate that we cannot extend these generalizations very far. In both of our statistical processes, we noted that some conditions were violated or questionable.

Thus, our conclusions of direction are likely correct (with small p-values) but there is certainly room for error in our numbers and calculations.

As mentioned above, because some of the conditions for our linear regression model were violated, that model is not fit to be used to predict the temperature moving forward. It is simply limited to being a decent estimate of Kyiv's temperature between 1940-2025.

5.3 Further Notes

Where do we go from here? The purpose of this statistical weather analysis was to do some preliminary investigative work to answer some temperature related questions around Kyiv's weather. Although we found statistically significant results, we cannot make any formal conclusions as to why the temperature has been increasing in Kyiv from 1940-2025. We can only speculate on the factors behind the data.

Appendix A: Formal Six-Step Hypothesis Test: Question 1

Regression

Hypothesis: $H_O : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Where β_1 is the slope of the regression line that uses year to predict the mean temperature in Kyiv, Ukraine.

Conditions: Our simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ is correct and the errors come randomly from a normally distributed population with mean 0.

Comments: As discussed above, we do have some concerns with linearity. We also have concerns with independence as we have yearly data that might not be independent year to year. In terms of normality, we have further concerns as our distribution does not seem to be precisely normal, exhibiting some skew. Our data was also not selected randomly in any way. The variance of our data does appear to be random, displaying no particular trend or direction.

Overall, we will proceed with the analysis in caution.

Rejection Region: We will reject H_O if $TS > 3.955$

Test Statistic: $TS = 75.07$

P-value: $p = 9.027e\text{-}08$

Conclusion: We have strong evidence to suggest ($p < 0.001$) that the slope of the regression line that uses year to predict the mean temperature in Kyiv, Ukraine is not 0. In fact, it is positive. Despite this, because of our questionable conditions, we will proceed with cautions and not use this model to predict the mean temperature is Kyiv moving forwards.

Appendix B: Formal Six-Step Hypothesis Test: Question 2

Case 3: 2 Sample hypothesis test for $\mu_1 - \mu_2$, small sample, unknown variances, normal data

Hypothesis: $H_O : \mu_2 - \mu_1 = 0$

$H_A : \mu_2 - \mu_1 > 0$

Where μ_1 is the mean maximum temperature in Kyiv, Ukraine from 1940-1982 and μ_2 is the mean maximum temperature in Kyiv, Ukraine from 1983-2025

Conditions: We have independent, random samples from two normally distributed populations, with variances that are unknown.

Comments: As discussed above, we do have some concerns with independence. Our samples are also not randomly selected and we have some concerns about normality after looking at our distributions. We do have unknown variances.

Overall, we will proceed with the analysis in caution.

Rejection Region: We will reject H_O if TS > 1.663

Test Statistic: TS = 5.6905

P-value: p = 9.027e-08

Conclusion: There is enough evidence to conclude that the difference between the mean maximum temperature in Kyiv, Ukraine from 1940-1982 and the mean maximum temperature in Kyiv, Ukraine from 1983-2025 is not 0. In fact, it is less than 0.

References

Free Open-Source Weather API | Open-Meteo.com. (n.d.). Open-Meteo.com. https://open-meteo.com