University of Calgary

STAT 327: Statistics for the Physical and Environmental Sciences

Final Project

Global Anomaly Trends: A Comparative Study of Hemispheric Temperatures

Authors:
Hala Abdelbaki, Zoya Malik

Instructor:
Claudia Mahler

April 15, 2024

TABLE OF CONTENTS

SECTION 1

Introduction

1.1 Abstract

This report investigates historical temperature anomalies to predict future conditions and to analyze differences between the Northern and Southern Hemispheres. Using regression analysis and hypothesis testing, we aim to understand long-term climate patterns and provide insights into regional climatic variations.

1.2 Introduction

In an era marked by growing concerns over climate change and its far-reaching implications, the utilization of advanced data analysis techniques becomes crucial in understanding and addressing environmental challenges. Leveraging the wealth of publicly available climate data from esteemed sources like NASA and other leading organizations, the focal point of this project is to elucidate important aspects of climate patterns and the variances observed across different regions.

The project is directed by two fundamental questions:

1. *How does the historical data of temperature anomalies predict future anomalies for specific seasons or months in the Northern and Southern Hemispheres?*

2. *Are there any significant differences in temperature anomalies between the Northern and Southern Hemisphere?*

For clarity throughout the report, we will use certain technical terms pertinent to the field of climate science. These terms include “ temperature anomalies” which refer to the deviations from a baseline temperature, and “historical climate data,” which denotes previously recorded data used to establish trends and predict future conditions.

To address the research questions, the following statistical methods were employed:

- Regression Analysis: Used to predict future temperature anomalies based on historical data.
- Two-Sample t-test/Mann-Whitney U test: Chosen based on the data distribution, these tests compare temperature anomalies between the two hemispheres.

Assumptions such as linearity were considered and tested in the regression model. The choice of test (t-test or Mann-Whitney U test) was dependent on the normality and variance of the temperature data.

SECTION 2

Sample Data Collection Techniques and EDA

2.1 Sample and Sampling Method

For this research project, we acquired publicly available climate data from data.world, a platform hosting various datasets, including those sourced from reputable organizations like NASA. The dataset in question originated from the NASA Goddard Institute for Space Studies and was utilized in an article titled "What's Really Warming the World?" published by Bloomberg.com on June 24, 2015, authored by Eric Roston and Blacki Migliozzi.

The sampling method for this dataset involves convenience sampling, as the data is freely accessible online and does not require specific permission or approvals for acquisition.

2.1.1 Variables

Temperature Anomalies: The primary variable, sourced from NASA's climate research, represents deviations from the long-term average temperature for specific seasons or months. These anomalies are crucial for understanding climate trends and identifying patterns of global warming or cooling.

Geographic Location: Geographic coordinates (latitude and longitude) are included in the dataset, enabling spatial analysis and comparisons between different regions worldwide. This variable facilitates the examination of temperature variations across diverse geographical settings.

Time: Time-related variables, such as year, month, and season, are provided in the dataset, and are crucial in tracking climate patterns over periods and understanding how they evolve.

2.2 Exploratory Data Analysis (EDA)

Upon obtaining the dataset, we conducted exploratory data analysis (EDA) to gain insights into its structure, content, and characteristics. Our EDA process involved the following steps:

Data Verification: We verified the integrity and authenticity of the dataset by cross-referencing it with its original source at the NASA Goddard Institute for Space Studies and validating its relevance to the research questions at hand.

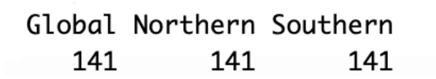
Data Preprocessing: To process the data so it is ready for analysis and visualizations, we imported the data into OpenRefine for cleaning purposes. At first glance, we noticed missing cell values for the year 1880 for all the Hemisphere variables. Though we did decide to keep the rows for the year 1880, because deleting rows can lead to biases in the analysis, especially if the missingness is not random. It could potentially distort the results and lead to incorrect conclusions, and for time series analysis, maintaining a continuous timeline is important even if some data points are missing. Removing rows could disrupt the sequence of data, affecting analyses that depend on temporal continuity. This was also especially because the other seasons had complete data.

We also noticed that after downloading the dataset, the “Year” variable was not in a date format, and the Global Temperature Anomalies were not in a numerical value, they were in a text data type. To create visualizations out of the variables, it was crucial to change these into numeric and date data types, so we did this in OpenRefine by “transforming” the cells.

Luckily for us, this dataset was acquired from a publicly available data source website data.world, where most datasets seemed to already be free of any corrupt or incorrectly formatted/duplicated data. This would not have been the case if we were to obtain our own data. We did not remove any outlier values either, as this could be indicative of any heat spikes or significant events in history and would create a bias or forced trends.

Descriptive Statistics:

Frequency table for the Hemisphere variable:



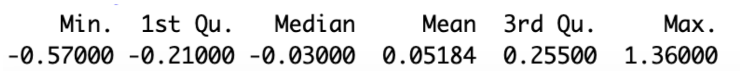
```
Global Northern Southern
   141      141      141
```

Summary Statistics :

Mean: 0.05184397 Standard deviation: 0.3667811

Median: -0.03 Variance: 0.1345283

Summary, Range, and IQR:



```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-0.57000 -0.21000 -0.03000  0.05184  0.25500  1.36000 
```

Range = $x_{max} - x_{min} = 1.36 - (- 0.57) = 1.93$

IQR = $Q_3 - Q_1 = 0.255 - (- 0.21) = 0.465$

- Mean (0.05184397): On average, the temperature anomalies in the Northern Hemisphere are about 0.052 degrees Celsius above the baseline. This suggests a tendency for warmer-than-average conditions.
- Median (-0.03): Half of the temperature anomalies are below -0.03 degrees Celsius, and the other half are above. Since the median is slightly below zero, it indicates that the center of the data distribution is marginally cooler than the baseline.
- Standard Deviation (0.3667811): The temperature anomalies vary from the mean by about 0.367 degrees Celsius on average. This value indicates that there is variability in the temperature anomalies, with some years experiencing more significant differences from the average than others.
- Variance (0.1345283): The variance, being the standard deviation squared, further indicates the degree of spread in the temperature anomalies. A variance of about 0.135 suggests that while there is variability, the majority of temperature anomalies are relatively close to the mean.
- Minimum (-0.5700): The lowest recorded temperature anomaly is -0.57 degrees Celsius, meaning the coolest anomaly compared to the baseline was substantially lower than average.

- Maximum (1.3600): Conversely, the highest temperature anomaly is 1.36 degrees Celsius, indicating that the warmest anomaly was significantly higher than the baseline.
- Range (1.93): The range is a measure of the total spread of the data, calculated as Maximum - Minimum. A range of 1.93 degrees Celsius indicates that there is a substantial difference between the coldest and warmest anomalies recorded.
- Interquartile Range (IQR) (0.465): The IQR is the range of the middle 50% of the data and is less affected by outliers. It's calculated as the Third Quartile - First Quartile. An IQR of 0.465 degrees Celsius indicates that the central majority of the anomalies are within a relatively narrow band of temperature differences, suggesting less variability in the bulk of the data compared to the full range.

Data Visualization: To visualize our dataset, we made graphs on Tableau and RStudio. This really helped us to draw hypotheses and inferences on our dataset, and visualize it overall. This first graph was made in Tableau. We knew we wanted a graph that drew a temperature comparison between the Hemispheres, though it seemed difficult given our variables in the dataset. We came up with the solution to "Pivot" our data to create a single measure rather than the 12 month variables to create pivoted field values with the corresponding temperature anomalies. We then created a calculated field for a yearly average from the 12 months to aggregate the measures. Then building our visualization by segregation by hemisphere was possible.

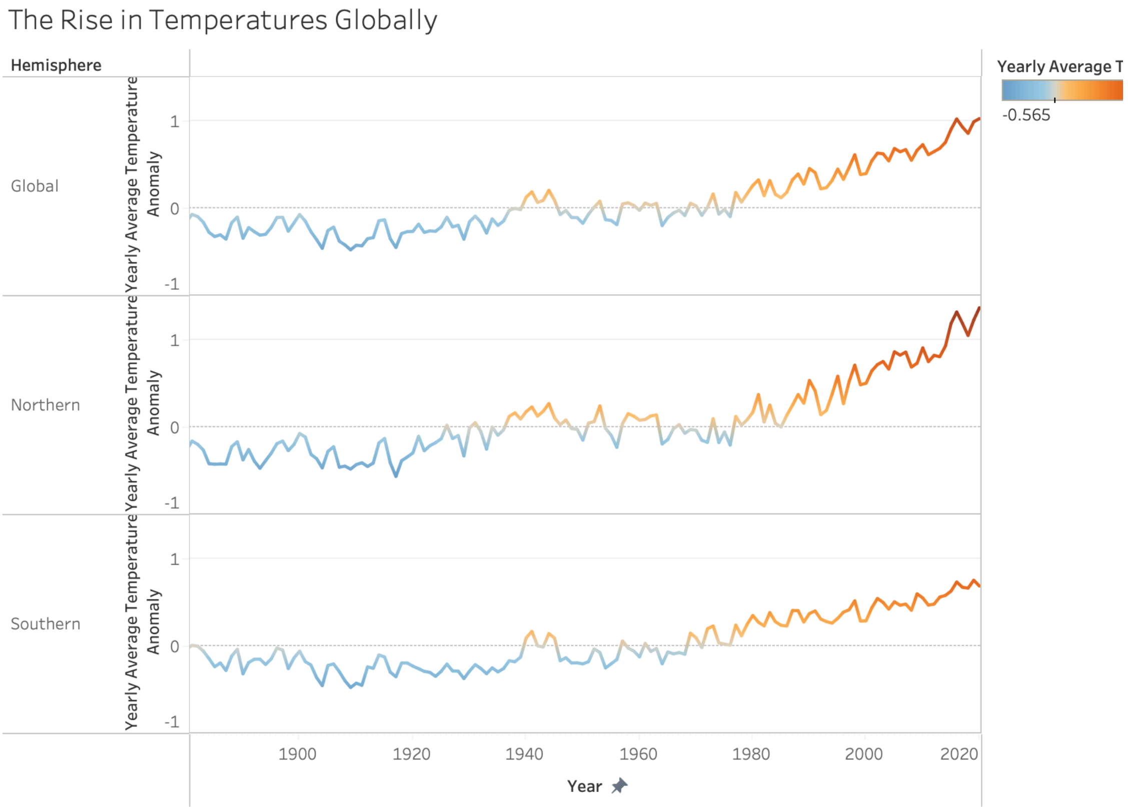


figure 1
This graph was created using Tableau. The global and regional temperature anomalies over the years show that the temperature has risen over the last century. The temperature has risen higher in the Northern Hemisphere than in the Southern Hemisphere.

It is also cool to note, the temperature anomaly spike in the years around roughly 1940 *could* be related to the events of World War II

Our next graphs were constructed in RStudio. Now with the visualizations for the month variables in mind, we saw that it was difficult to read and infer anything from this graph. Though it is still possible to see that the Northern Hemisphere had greater temperature spikes in all months. The following is this graph:

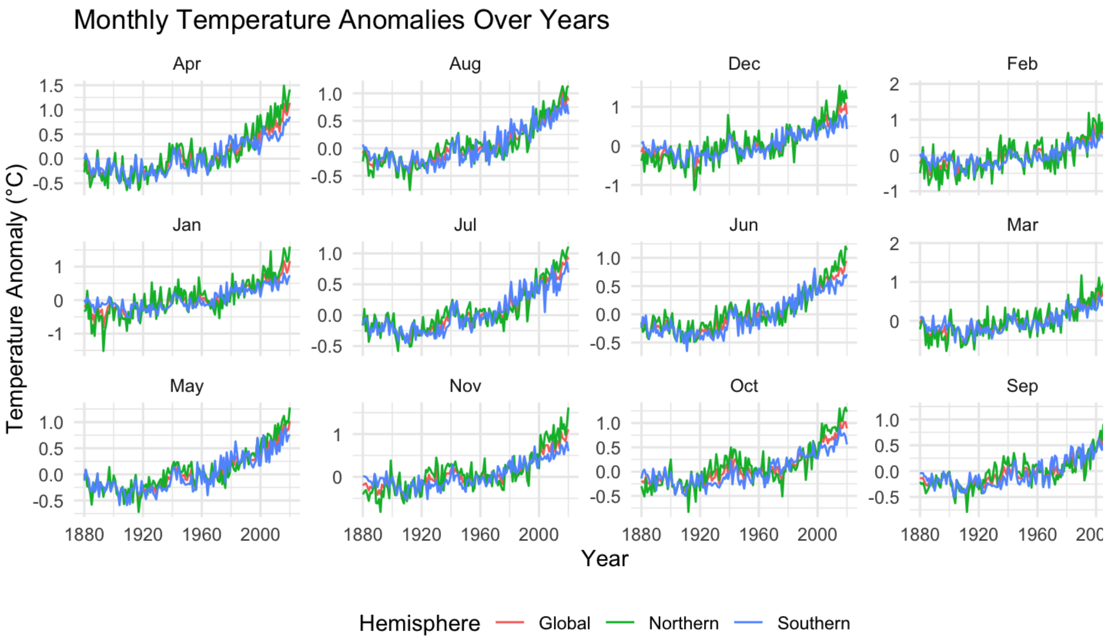


figure 2

We also created visualizations for the seasons. This graph is a bit easier to read than the previous monthly visual, and it is still segregated by Hemisphere. The key difference in this graph is that it showcases seasonal data, rather than the monthly data. This is especially useful for when we draw conclusions with the seasons in mind. The variables used are:

- DJF: December of the previous year, January, February (Northern Winter/ Southern Summer)
- MAM: March, April, May (Northern Spring / Southern Autumn)
- JJA: June, July, August (Northern Summer / Southern Winter)
- SON: September, October, November (Northern Autumn / Southern Spring)

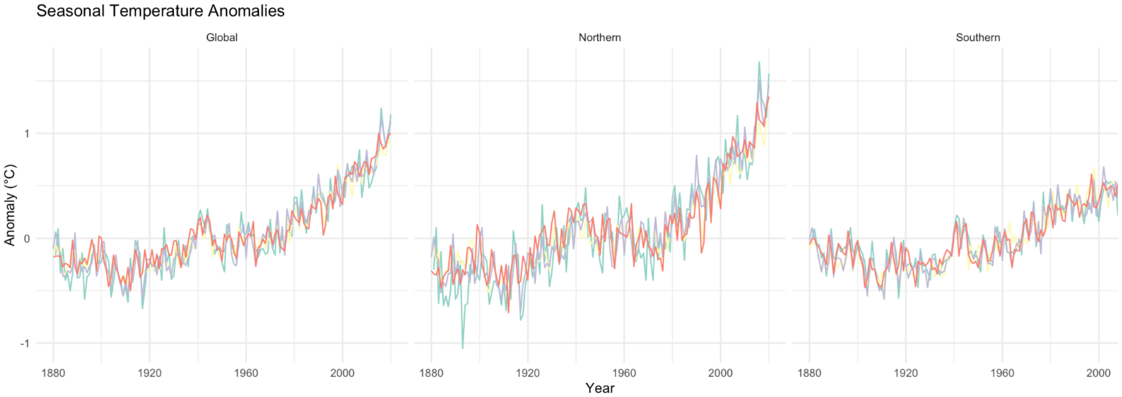


figure 3

We also thought a box-plot graph would be a great visualization for these variables, as box-plots are great for depicting the distribution of data across different categories or over time. This helps us compare the central tendency and variability of temperature anomalies across different categories (e.g., hemispheres, seasons) at a glance. This is especially helpful when wanting to identify the differences between groups. The following is this graph:

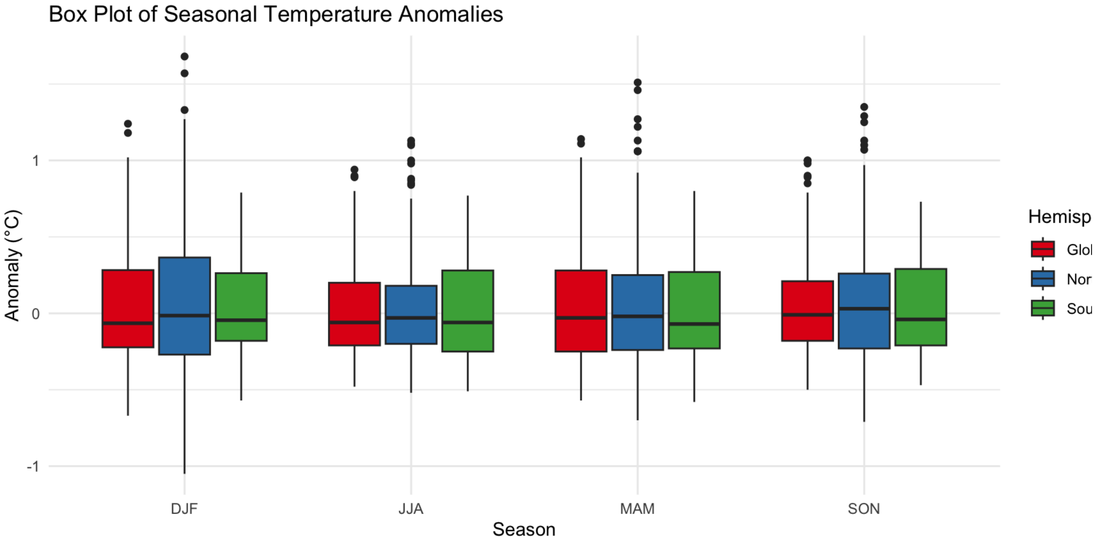
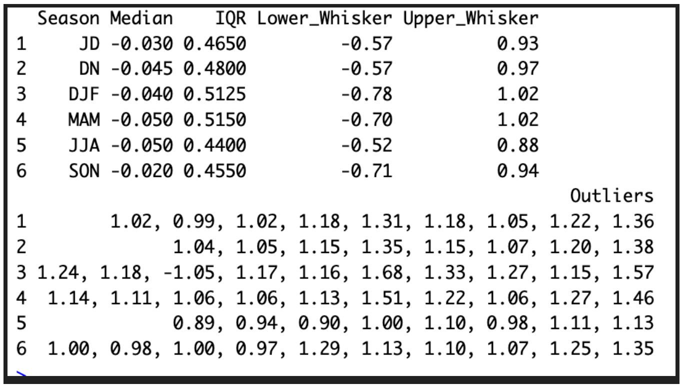


figure 4

In this graph, Hemisphere (Categorical Variable) represents the data category, with each color representing a level of the categorical variable. The x-axis represents the seasons, which are categorical in nature. The y-axis represents the temperature anomalies, which are continuous numeric variables. An anomaly is the deviation of temperature for a given period (in this case, seasonal) from a long-term average or baseline. A positive anomaly indicates that the observed temperature was warmer than the baseline period average. A negative anomaly indicates cooler than the baseline period average.

- Median: The line within each box shows the median temperature anomaly for each category (Global, Northern, Southern) and season. The median is the value that divides the dataset into two equal halves, with 50% of the values falling below and 50% above.
- Interquartile Range (IQR): The box itself spans from the 25th percentile (the lower edge of the box) to the 75th percentile (the upper edge), encompassing the middle 50% of the data. It provides a sense of the variability of the anomalies around the median.
- Whiskers: The whiskers extend from the IQR to the furthest points that are within 1.5 times the IQR from the upper and lower quartiles. They provide a visual indication of the range of typical data points, excluding outliers.
- Outliers: Dots or points that lie beyond the whiskers are outliers. They represent temperature anomalies that are significantly higher or lower than most of the data and may indicate unusual climate events.



```
  Season Median    IQR Lower_Whisker Upper_Whisker
1     JD -0.030 0.4650         -0.57          0.93
2     DN -0.045 0.4800         -0.57          0.97
3    DJF -0.040 0.5125         -0.78          1.02
4    MAM -0.050 0.5150         -0.70          1.02
5    JJA -0.050 0.4400         -0.52          0.88
6    SON -0.020 0.4550         -0.71          0.94
                                                          Outliers
1       1.02, 0.99, 1.02, 1.18, 1.31, 1.18, 1.05, 1.22, 1.36
2             1.04, 1.05, 1.15, 1.35, 1.15, 1.07, 1.20, 1.38
3 1.24, 1.18, -1.05, 1.17, 1.16, 1.68, 1.33, 1.27, 1.15, 1.57
4  1.14, 1.11, 1.06, 1.06, 1.13, 1.51, 1.22, 1.06, 1.27, 1.46
5             0.89, 0.94, 0.90, 1.00, 1.10, 0.98, 1.11, 1.13
6  1.00, 0.98, 1.00, 0.97, 1.29, 1.13, 1.10, 1.07, 1.25, 1.35
>
```

We calculated these exact values. R Code for generating this output is attached in a separate document

SECTION 3

Analysis

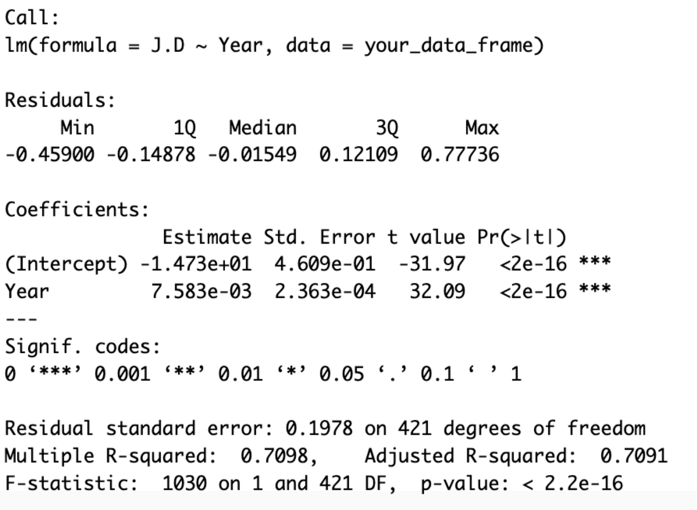
3.1 Research Question 1

How does the historical data of temperature anomalies predict future anomalies for specific seasons or months in the Northern and Southern Hemispheres?

Null Hypothesis (H_0): The slope of the regression line is zero ($\beta = 0$). (No relationship between the independent and dependent variables)

Alternative Hypothesis (H_A): The slope of the regression line is not zero ($\beta \neq 0$). (No relationship exists between the independent and dependent variables)

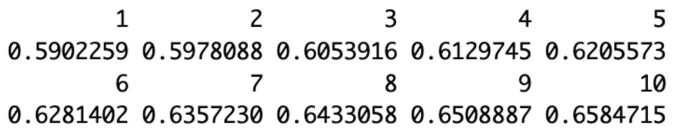
Test Statistic and p-value obtained using R:



```
Call:
lm(formula = J.D ~ Year, data = your_data_frame)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.45900 -0.14878 -0.01549  0.12109  0.77736 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.473e+01  4.609e-01  -31.97   <2e-16 ***
Year         7.583e-03  2.363e-04   32.09   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1978 on 421 degrees of freedom
Multiple R-squared:  0.7098,	Adjusted R-squared:  0.7091 
F-statistic:  1030 on 1 and 421 DF,  p-value: < 2.2e-16
```

Assuming $\alpha = 0.05$, and using the p-value $= 2 \cdot 10^{-16}$, $2 \cdot 10^{-16} < 0.05$. Therefore we reject the null hypothesis, indicating there is a significant relationship between the years and temperature anomalies.

Then to predict future values for future years, we use the predict() function in R (specified on a separate document). We were then given the following output[1]



```
        1         2         3         4         5 
0.5902259 0.5978088 0.6053916 0.6129745 0.6205573 
        6         7         8         9        10 
0.6281402 0.6357230 0.6433058 0.6508887 0.6584715 
```

The numbers 1 - 10 are identifiers for each prediction corresponding to each year from 2021 to 2030. The numbers below it are the predicted temperature anomalies for each respective year. The values represent the anomaly above a baseline, indicating how much warmer or cooler the global temperature is expected to be compared to the historical average.

2.2 Research Question 2

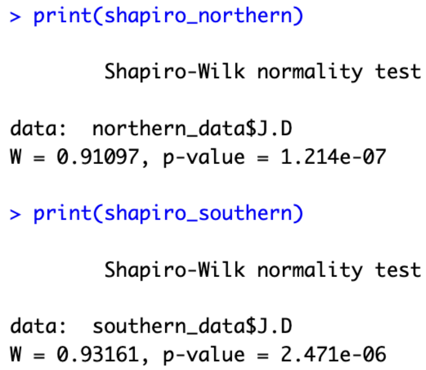
Are there any significant differences in temperature anomalies between the Northern and Southern Hemisphere?

Null Hypothesis (H_0): There is no significant difference in the median temperature anomalies between the Northern Hemisphere and the Southern Hemisphere. This hypothesis assumes that any observed differences in the medians of the two samples are due to random sampling variation.

Alternative Hypothesis (H_1): There is a significant difference in the median temperature anomalies between the Northern Hemisphere and the Southern Hemisphere. This suggests that the differences observed are not just due to chance, and that one hemisphere might exhibit higher or lower temperature anomalies than the other.

Shapiro-Wilk test to statistically assess the normality of these distributions:

[1] *Note: For this step of our testing, we used a separate source for help. The website is noted in the references section*



```
> print(shapiro_northern)

        Shapiro-Wilk normality test

data:  northern_data$J.D
W = 0.91097, p-value = 1.214e-07

> print(shapiro_southern)

        Shapiro-Wilk normality test

data:  southern_data$J.D
W = 0.93161, p-value = 2.471e-06
```

For the Northern Hemisphere, the p-value is

$1.214 \cdot 10^{-7}$

For the Southern Hemisphere, the p-value is

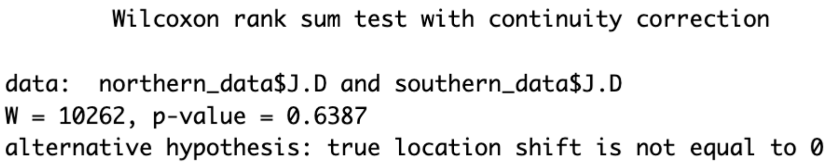
$2.471 \cdot 10^{-6}$

Both p-values are much less than the commonly used significance level of 0.05, indicating that the temperature anomalies for both hemispheres are not normally distributed.

Given the results of the Shapiro-Wilk test indicating non-normal distributions for both the Northern and Southern Hemisphere temperature anomalies, using a parametric test like the Two-Sample t-test would not be appropriate. This is because the fundamental assumption of normality required for the t-test is violated, which could lead to inaccurate results and conclusions.

Therefore, the Mann-Whitney U test, a non-parametric test, is chosen as it does not require the assumption of normality. This test will compare the ranks of the data rather than their means, making it suitable for analyzing the differences between the temperature anomalies of the two hemispheres under the observed conditions of non-normality.

As our next step, we use the wilcox.test function which provides a way to compare two independent samples to see if they come from the same distribution. Which in this case, translates to checking if there is a statistically significant difference between the temperature anomalies in the Northern and Southern Hemispheres. The following is a result of this test:



```
        Wilcoxon rank sum test with continuity correction

data:  northern_data$J.D and southern_data$J.D
W = 10262, p-value = 0.6387
alternative hypothesis: true location shift is not equal to 0
```

Our result is
Test statistic = 10262
P-value = 0.6387

Since the p-value is greater than the conventional threshold of 0.05, there is not enough evidence to reject the null hypothesis. The test does not provide sufficient evidence to conclude that there is a significant difference in the temperature anomalies between the Northern and Southern Hemispheres.

SECTION 4

Conclusion

The primary goal of our project was to harness historical climate data to predict future temperature anomalies and to discern any significant differences in these anomalies between the Northern and Southern Hemispheres. Through our careful analysis, we were able to draw several key conclusions that both confirmed and challenged our initial hypotheses.

First and foremost, our analysis revealed no substantial statistical difference in temperature anomalies between the two hemispheres. This outcome was contrary to initial expectations, which were influenced by commonly held perceptions of the Northern Hemisphere experiencing greater environmental variability. Despite these expectations, the application of the Mann-Whitney U test demonstrated that the variations observed were not significant enough to establish a distinct pattern between the hemispheres over the studied period. This finding suggests a shared experience between the hemispheres in response to global climatic shifts, highlighting the universal reach of climate change.

The implications of these findings are substantial. They imply that climate change, as reflected in temperature anomalies, is not confined to a single region but is a pervasive global phenomenon. This universal pattern reinforces the need for collective global action in response to climate change, as the impacts are not isolated by geography.

In reflecting upon our research process, we recognize the potential for improvement in several areas. A more granular approach to data, including finer geographic detail and a broader range of climatic variables, could offer deeper insights. Additionally, the inclusion of more recent data could provide an updated perspective on current trends. Our data preprocessing decisions, particularly around the treatment of missing values and the transformation of data types, were necessary and considered, yet they also highlight the importance of meticulous data management in climate research.

Throughout this project, we navigated challenges that are typical in large datasets analysis. One of the most significant was ensuring data integrity after transformation and managing missing data without introducing bias. Our commitment to preserving the dataset’s continuity was a critical component of our analysis, ensuring that our conclusions were drawn from a complete and uninterrupted historical timeline.

To conclude, our investigation found no statistically significant difference in temperature anomalies between the Northern and Southern Hemispheres, emphasizing the globally uniform impact of climate change. The lack of significant difference underscores the need for a united front in addressing and mitigating the effects of climate change. Future research could build on our findings by incorporating more detailed spatial analysis, including a wider array of climate indicators, or exploring the socio-economic consequences of these climate patterns. Our work contributes to the broader dialogue on climate change and serves as a stepping stone for future explorations into this critical global issue.

References

CN, P. (2022, September 30). *How to use the predict() function in R programming*. DigitalOcean. https://www.digitalocean.com/community/tutorials/predict-function-in-r

Graphs in R. RCODER. (n.d.). https://r-coder.com/r-graphs/

Holtz, Y. (n.d.). *Line chart*. the R Graph Gallery. https://r-graph-gallery.com/line-plot.html

Quantitative skills for biology. 10 Making graphs in R. (n.d.). https://ahurford.github.io/quant-guide-all-courses/graph.html