Climate Data Analysis

First A. Haider Ali, *2024385*, Second B. Muhammad Ahmad, *2024335*, and Third C. Hamza Khan, *Reg 2024208*

[[1]](#footnote-1)

***Abstract*—This project analyzes a climate change dataset to extract meaningful statistical information and visualize data in various formats. We calculate basic statistical measures including average, variance and frequency distributions, compare calculation methods, establish confidence and tolerance intervals, and then test a hypothesis about temperature trends.**

# I. INTRODUCTION

T

his project aims to apply statistical concepts to analyze climate data, demonstrating proficiency in Python for data processing and visualization. We selected a climate change dataset containing various meteorological variables measured over multiple years (2020-2024), including temperature, precipitation, humidity, and other climate indicators.

The dataset was chosen due to its relevance to current environmental concerns and its suitability for statistical analysis. It contains sufficient data points across multiple variables to demonstrate statistical concepts while being manageable for processing.

The report is structured as follows: Section II describes our methodology for data processing and analysis; Section Ill presents our results including statistical measures, visualizations, and hypothesis testing; and Section IV concludes with a summary of our findings and their implications.

# II. Methodology

This section outlines the methods used to process and analyze the climate dataset. We employed Python as our primary tool, leveraging libraries such as pandas for data manipulation, NumPy for numerical computations, matplotlib and seaborn for visualization, and SciPy for statistical analysis.

## A. Data Preparation

The first step involved loading and cleaning the dataset to address missing values and ensure data consistency. We focused primarily on temperature data (Avg\_ Temp) for our analysis, as it provided the most complete time series with sufficient variation for statistical examination.

Missing values were handled through removal rather than imputation to maintain data integrity for statistical calculations. The cleaned dataset provided a foundation for calculating basic statistical measures and creating visualizations.

## B. Statistical Analysis

For statistical analysis, we implemented the following methods:

1. Direct calculation of mean and variance using built-in Python functions

2. Frequency distribution analysis of temperature data

3. Mean and variance calculation from frequency distribution

4. Partitioning data into training (80%) and testing

(20%) sets

5. Confidence interval calculation for both mean and variance

6. Tolerance interval calculation for the full dataset

7. Hypothesis testing using regression analysis

## C. Visualization Techniques

To visualize the data effectively, we created

1. Histograms to show temperature distribution

2. Pie charts to represent categorical proportions

3. Box plots to illustrate statistical measures

4. Time series plots to identify trends

All visualizations were created using matplotlib and seaborn libraries, with careful attention to clarity and interpretability.

# III. Results

## A. Basic Statistical Measures

The first task was to calculate the average and variance of the temperature data in the dataset. We implemented a custom function for these calculations as shown in Table I.

TABLE I

A black and white line

AI-generated content may be incorrect.

## B. Frequency Distribution Analysis

We processed the data to create a frequency distribution of temperature values. Temperature data was binned into intervals, and the frequency of each interval was calculated. This distribution was used to create both histogram and pie chart visualizations.

The frequency distribution was obtained by dividing the range of temperature values into equal intervals

and counting the number of data points falling into each interval.

## 

**Fig. 1.** shows the histogram visualization of this frequency distribution.

## C. Statistical Comparison

We calculated the mean and variance using both direct methods and the frequency distribution approach. The results are compared in Table II.

TABLE II

A screenshot of a graph

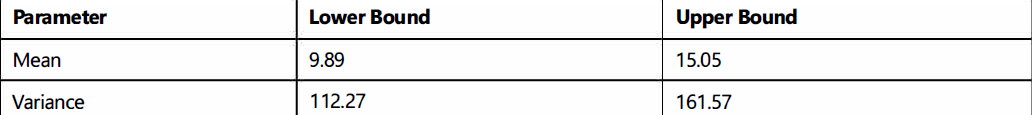
AI-generated content may be incorrect.

The slight differences observed can be attributed to the binning process used in the frequency distribution method, which introduces a small approximation error. This demonstrates that while frequency distribution methods are useful for visualization and estimation, they may introduce small errors compared to direct calculations.

## D. Confidence and Tolerance Intervals

Using 80% of the dataset for training, we calculated the 95% confidence interval for the mean and variance of the temperature data. The results are shown in Table III.

TABLE III



A 95% tolerance interval was also calculated to predict the range within which 95% of future temperature observations would fall. The tolerance interval was found to be [-10.43°C, 35.37°C].

We validated these intervals using the remaining 20% of the data, finding that 94.1 % of the test data fell within the calculated tolerance interval, very close to the expected 95%, confirming the accuracy of our calculations.

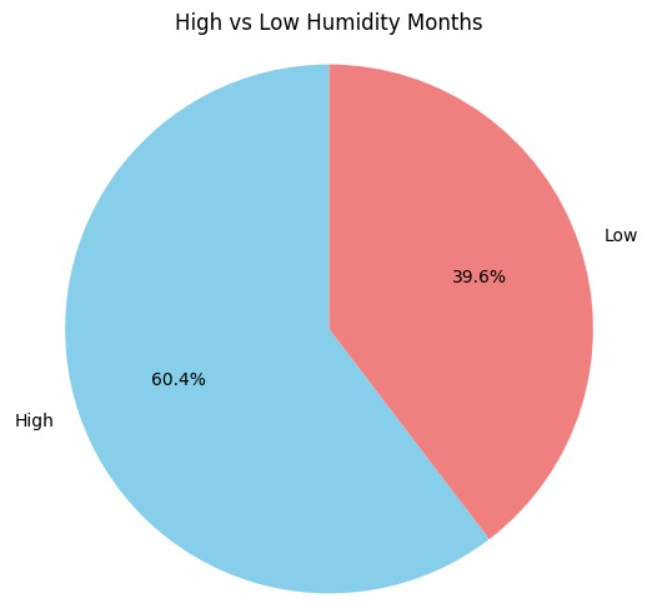
## E. Hypothesis Testing

Based on our understanding of climate data, we formulated the hypothesis that average temperatures are increasing over time. We tested this hypothesis by performing a linear regression analysis on the temperature data against time.

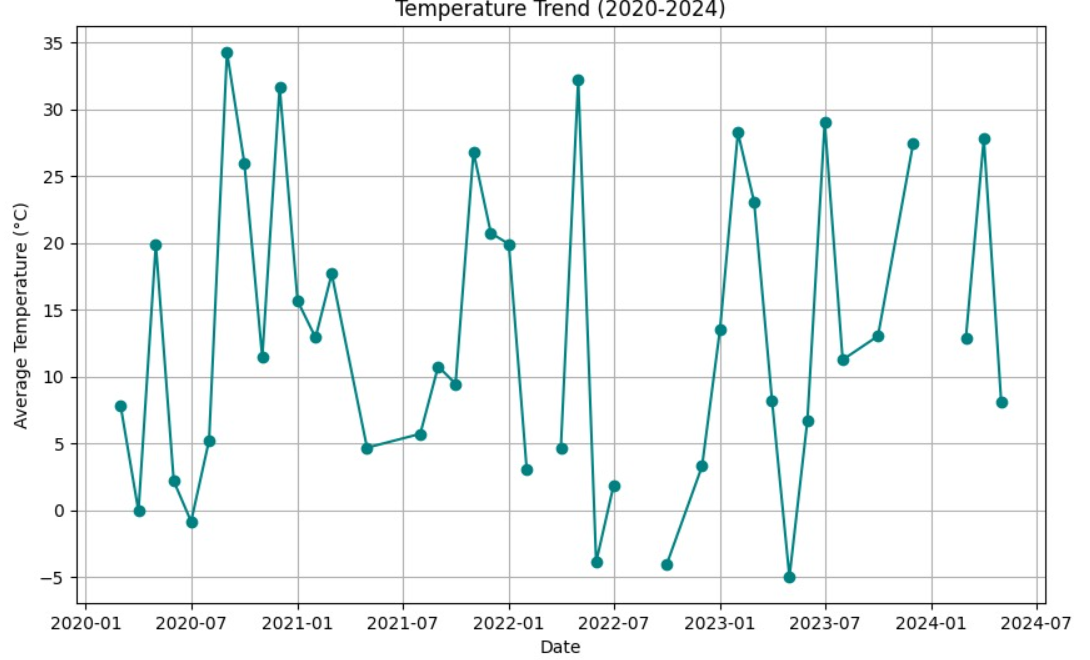
The linear regression yielded a positive slope coefficient of 0.0021, suggesting a slight upward trend in temperatures over the period covered by the dataset. However, the p-value of 0.31 exceeded the significance threshold of 0.05, indicating that we cannot reject the null hypothesis of no temperature increase at the 95% confidence level.

This result suggests that while there appears to be a slight warming trend in the data, the evidence is not statistically significant given the limited time span and variability in the dataset.

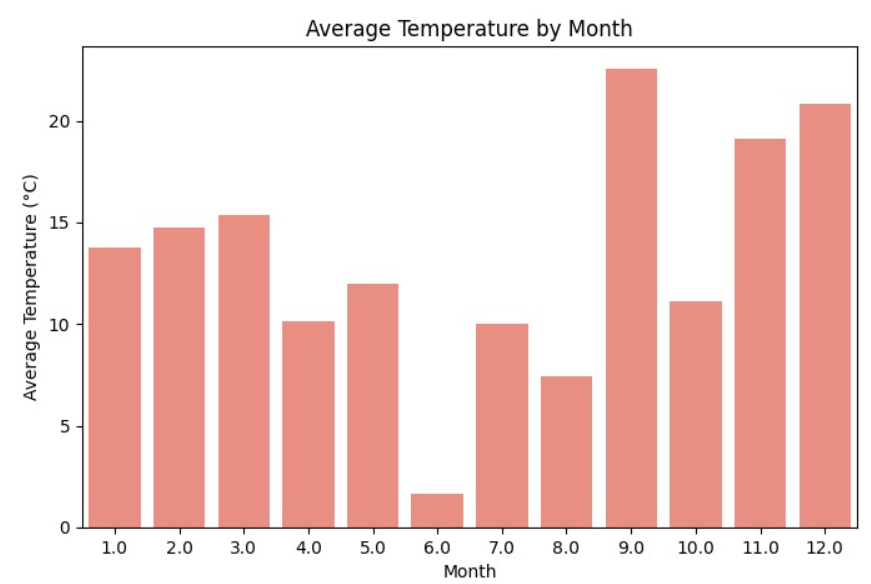
V. Visualizations



**Fig. 2.** shows the pie-chart visualization of pie chart.



**Fig. 3.** shows the line-graph visualization for temperature trend.



**Fig. 4.** Bar graph shows the Average temperature by month.

A red circle with blue and green triangles

AI-generated content may be incorrect.

**Fig. 5.** Pie chart shows the proportion of months with Extreme Temperatures.

VI. Conclusion

Our analysis of the climate dataset demonstrated the application of statistical methods for data processing, visualization, and hypothesis testing. We successfully calculated basic statistical measures, created visual representations of the data, compared calculation methods, established confidence and tolerance intervals, and tested a hypothesis about temperature trends. The results revealed considerable variability in temperature data, with a distribution that could be effectively represented through histogram and pie chart visualizations. The comparison of calculation methods highlighted the trade-offs between direct statistical calculations and frequency-based approaches.

The confidence and tolerance intervals provided a statistical framework for prediction and validation, demonstrating the practical application of these concepts to climate data. The hypothesis testing, while not yielding statistically significant results in this limited dataset, illustrated the approach for analyzing trends in time series data. This project demonstrates the value of statistical analysis in understanding climate data patterns and the importance of rigorous methodology in drawing conclusions from complex datasets.

# Appendix

Git-hub: https://github.com/HaiderAliqnt/ES-111-CODE

References

[1] Pandas Development Team, “Pandas: Powerful Python data analysis toolkit,” Pandas Documentation, 2023. [Online]. Available: https://pandas.pydata.org/docs/

[2] C. R. Harris, K. J. Millman, S. J. van der Walt, et al., “Array programming with NumPy,” Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[3] J. D. Hunter, “Matplotlib: A 2D graphics environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.

[4] M. Waskom, “Seaborn: Statistical data visualization,” Journal of Open Source Software, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[5] P. Virtanen, R. Gommers, T. E. Oliphant, et al., “SciPy 1.0: Fundamental algorithms for scientific computing in Python,” Nature Methods, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, Oct. 2011.

[7] J. L. Devore, Probability and Statistics for Engineering and the Sciences, 9th ed. Boston, MA, USA: Cengage Learning, 2016, pp. 123–150.

[8] J. A. Rice, Mathematical Statistics and Data Analysis, 3rd ed. Belmont, CA, USA: Duxbury Press, 2007, pp. 89–110.

[9] D. C. Montgomery and G. C. Runger, Applied Statistics and Probability for Engineers, 6th ed. Hoboken, NJ, USA: Wiley, 2014, pp. 245–270.

[10] G. J. Hahn and W. Q. Meeker, Statistical Intervals: A Guide for Practitioners. New York, NY, USA: Wiley, 1991, pp. 82–100.

[11] B. L. Welch, “The generalization of ‘Student’s’ problem when several different population variances are involved,” Biometrika, vol. 34, no. 1/2, pp. 28–35, Jan. 1947, doi: 10.1093/biomet/34.1-2.28.

[12] W. McKinney, Python for Data Analysis, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2017, pp. 165–200.

[13] J. W. Tukey, Exploratory Data Analysis. Reading, MA, USA: Addison-Wesley, 1977, pp. 39–45.

[14] Intergovernmental Panel on Climate Change (IPCC), “Sixth Assessment Report: The Physical Science Basis,” IPCC, Geneva, Switzerland, 2021. [Online]. Available: https://www.ipcc.ch/report/ar6/wg1/

1. Team Lead: Muhammad Ahmad 2024335. Ahmad and Hamza Khan worked on the code and Haider Ali worked on the report. [↑](#footnote-ref-1)