Kigali Independent University (ULK)

Department of Computer Science - Data Science

Module : R programming

Group Assignment

Submission date: 29/5/2025

Roll numbers:

202312282

202311474

202311779

202311744

202312358

202310037

202312135

202312242

Temperature Data Analysis Report

1. Introduction

This analysis explores temperature trends using the `city_temperature.csv` dataset. The goal is to understand patterns and relationships between average temperature and other variables such as region, country, city, and time-based factors like year and month. The analysis includes data cleaning, visualization, and statistical tests such as correlation and ANOVA.

2. Data Preprocessing

Data Inspection

- The dataset contains a large number of observations with various columns such as `Region`, `Country`, `State`, `City`, `Year`, `Month`, `Day`, and `AvgTemperature`.
- Initial checks revealed missing values and duplicates.

Handling Missing Values

- Columns with more than 40% missing data were excluded.
- Rows with any remaining `NA` values were removed using `na.omit()`.

Removing Duplicates

- Duplicate rows were detected and removed using `distinct()`.

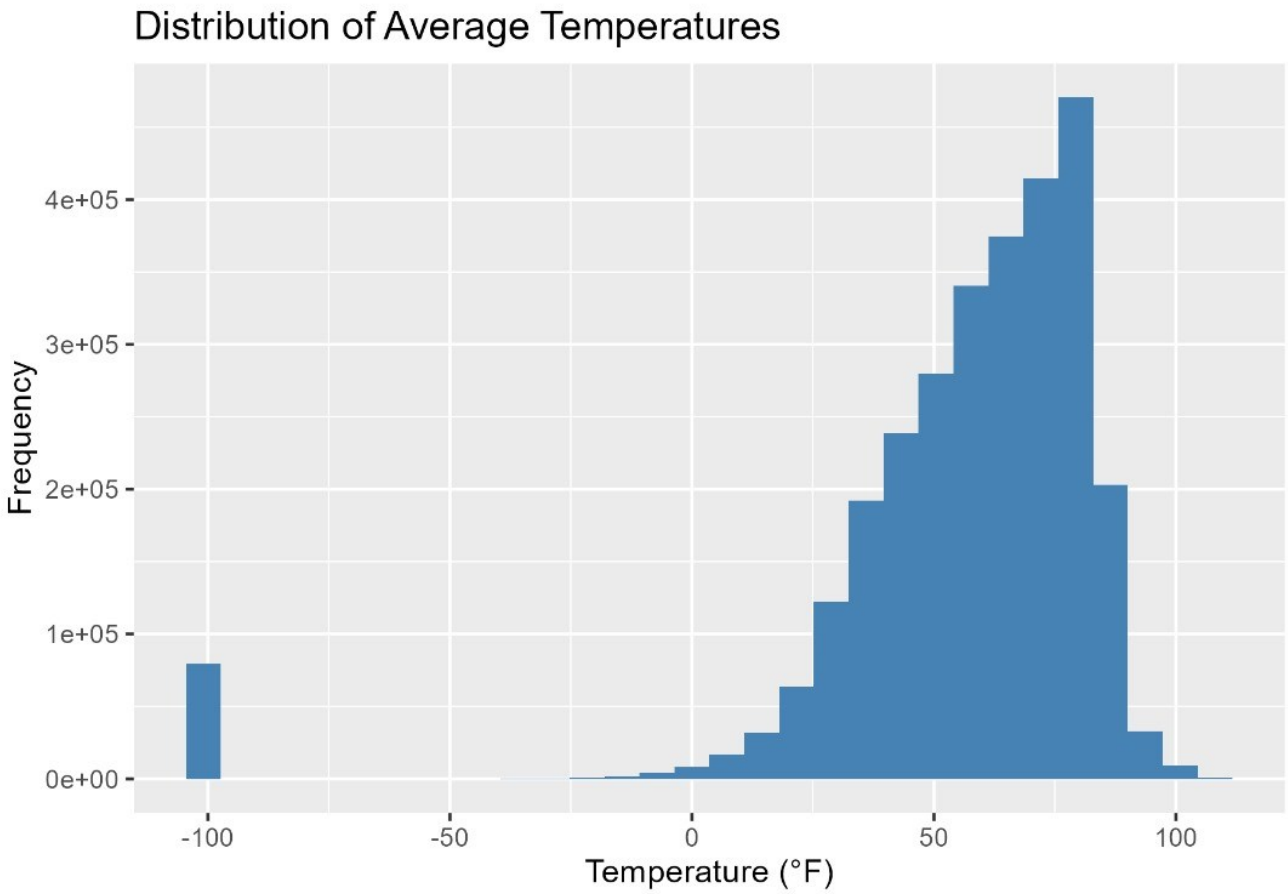
Final Clean Dataset

- The cleaned dataset (`clean2`) contains no missing values or duplicates.
- Outliers with `AvgTemperature < -50` were removed to ensure meaningful visualizations.

3. Visualizations (Include Captions)

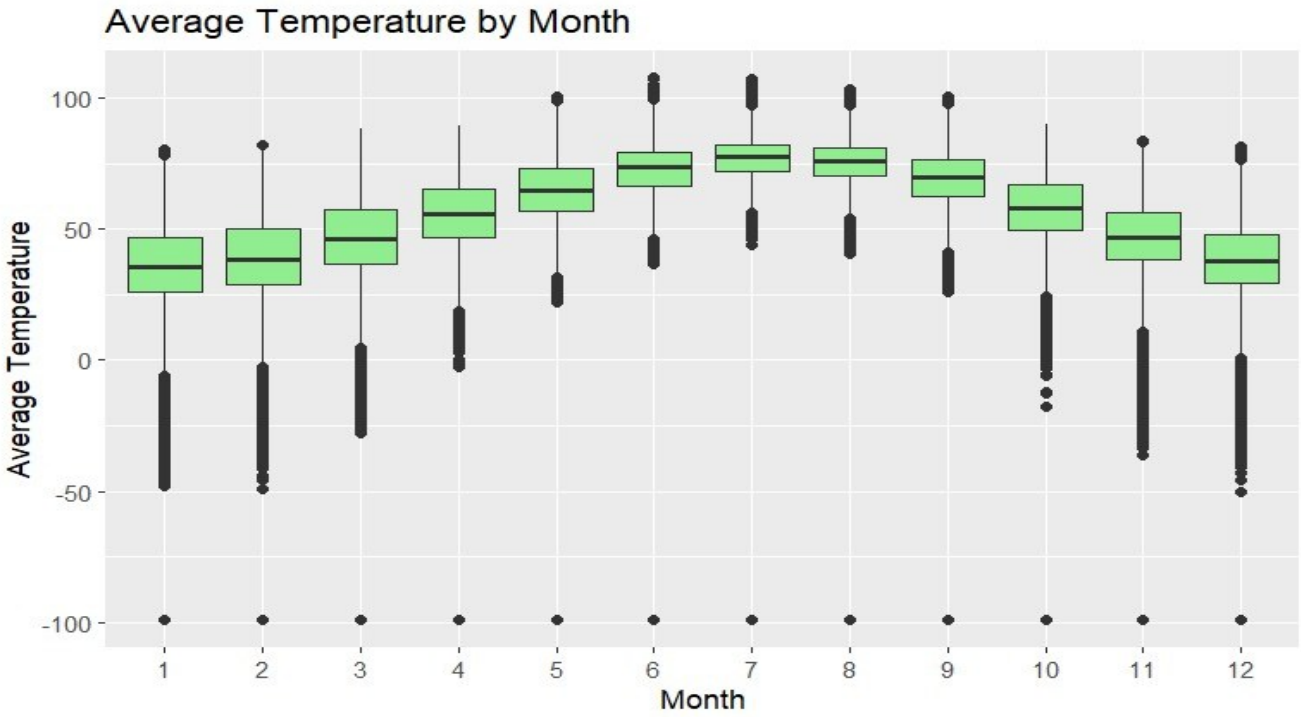
Histogram of Average Temperature

Shows distribution of global temperatures. Most values lie between 40–80°F.



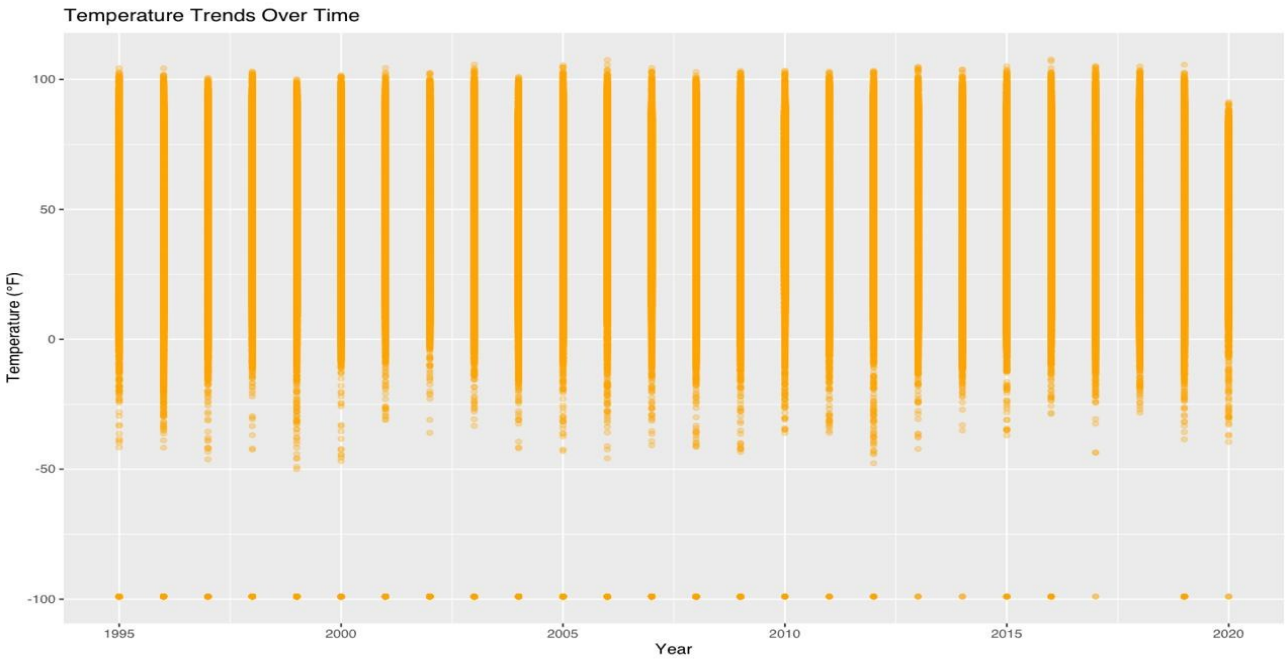
🎁 Boxplot of Average Temperature by Month

Shows monthly variations; warmer in mid-year months, cooler at start and end of the year.



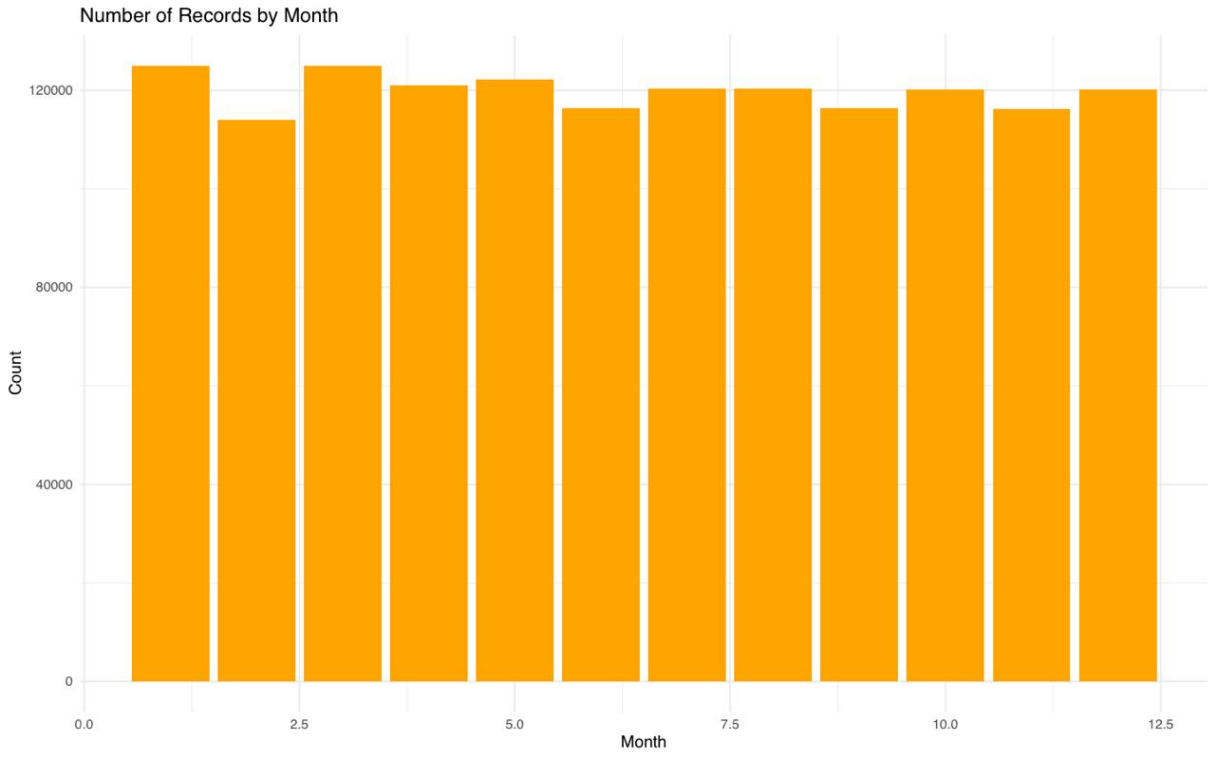
🟤 Scatter Plot (Temperature vs Year)

Reveals long-term trends and variability—potential warming signal.



🌍 Bar Chart (Data by Region)

Uneven representation, with some regions like North America and Asia having more data.



4- Initial Insights Summary (2–4 Paragraphs)

The city_temperature.csv dataset contains over 2.8 million observations of daily average temperatures collected from cities around the world. The key variables include both categorical (Region, Country, State, City) and numerical types (Month, Day, Year, AvgTemperature). These variables provide rich temporal and spatial information that can be highly relevant to environmental and agricultural research. For instance, average temperature trends across years and months are crucial indicators of climate change and seasonal variability—both of which influence crop growth cycles, water availability, and regional planning in agriculture.

Through visualizations such as histograms, boxplots, scatter plots, and correlation heatmaps, several initial patterns were observed. For example, the distribution of average temperatures revealed seasonal fluctuations, with noticeable variation by month. Boxplots grouped by month highlighted higher temperatures in mid-year months and lower in early and late months, consistent with typical climatic cycles. Scatter plots of average temperature over years hinted at a slight warming trend in some regions, although further statistical validation is needed. The correlation matrix showed strong relationships between date-related variables (Month, Day, Year), but average temperature showed relatively weak correlation with these individually—suggesting that additional derived features (e.g., seasons) may better capture meaningful trends.

During data cleaning, one of the main challenges was the presence of missing values, especially in the State column and occasionally in AvgTemperature. These missing entries were handled using na.omit() for simplicity, though a more robust approach could involve imputing missing values or analyzing their distribution. Duplicate records were also found and removed using the unique() function. Outliers in temperature values were visually detected in boxplots (e.g., extremely low or high values), which may result from data entry errors or represent rare extreme weather events—both of which are important to flag when modeling.

Looking ahead, this dataset holds potential for multiple research directions. One could investigate long-term temperature trends in specific regions to quantify climate change impacts. Machine learning tasks such as time series forecasting (e.g., predicting next year's temperature profile) or classification (e.g., identifying regions at risk of extreme temperature events) are also promising. Integrating this temperature data with agricultural yield data could open up predictive models for crop production based on climate conditions. Additionally, clustering cities by temperature patterns might reveal ecological zones or climate similarity regions that could guide agricultural policy and infrastructure planning.

5- Correlation Analysis

The Pearson correlation test was used to check relationships between average temperature and numerical variables:

Variable	Correlation (r)	p-value	Interpretation
Day	~0.0064	< 2.2e-16	Very weak positive correlation, statistically significant.
Year	~-0.044	< 2.2e-16	Very weak negative correlation, statistically significant.

5. ANOVA (Analysis of Variance)

To assess the effect of categorical variables on average temperature:

Variable	Df	F-value	p-value	Interpretation
Region	6	90695	< 2e-16	Significant differences in temperature between regions.
Country	124	8893	< 2e-16	Significant differences in temperature between countries.
State	52	8809	< 2e-16	Significant differences in temperature between states.
Month	11	125358	< 2e-16	Strong seasonal variation in temperature.

6. Conclusion

This analysis revealed clear geographic and temporal patterns in temperature data:

- **Month** was the most significant factor influencing temperature, showing strong seasonal variation.
- **Region, Country, and State** also significantly influenced average temperatures.
- While **`City`** was excluded from the final ANOVA summary, other categorical variables sufficiently
- explained the variation in temperature.

Even though correlation with **`Day`** and **`Year`** was statistically significant, their effect sizes were negligible.