FINAL PROJECT REPORT

MULTIVARIATE DATA ANALYSIS (STAT-6388)

June 17, 2025

Submitted By
Solayman Hossain Emon (ID-80744292)

Contents

US Climate Data Analysis (Arizona)

1 Project Overview

The project utilizes a dataset containing weather data collected from multiple stations across Arizona (2014-2024). The data cleaning process involves handling missing values and selecting relevant variables for analysis. Exploratory data analysis (EDA) techniques, including histograms, boxplots, and time series plots, are employed to understand the distribution and trends of these climate variables across different weather stations. Statistical analyses such as t-tests are performed to compare climate variables between distinct periods, possibly representing different seasons or climatic conditions. Furthermore, clustering techniques are applied to identify groups of weather stations with similar climate characteristics. The project aims to provide insights into the variability and patterns of weather conditions across Arizona, potentially informing decision-making processes in areas such as agriculture, urban planning, and disaster preparedness.

2 Data Description

The dataset has been downloaded from the National Centers for Environmental Information (NOAA) [1]. After navigating through the data, we initially found plenty of missing values inside the dataset. Thus, we select the subset of features those have comparatively less missing values. Through data cleaning, we able to extract the data of 49 different stations over Arizona. According to our analysis, the key features would be:

Variable	**Description**
STATION	Weather station identifier
DATE	Observed Date
TAVG	Average temperature
PRCP	Total precipitation
CLDD	Cooling degree days
HTDD	Heating degree days (HTDD)
EMXT	Extreme maximum temperature for the period
DX70	Number days with maximum temperature > 70 F (21.1C)

Table 1: Key variables after the data cleaning

3 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) delved into various aspects of the dataset, encompassing distribution analysis, correlation examination, and trend analysis. In the distribution analysis (Weather Variable), we meticulously studied the distribution patterns of the key weather variables. Moving forward, our correlation analysis explored the relationships between different weather parameters. Furthermore, we conducted trend analysis (Time Series), scrutinizing the temporal patterns and fluctuations within the dataset to discern any discernible trends or seasonality.

3.1 Distribution (Weather Variable)

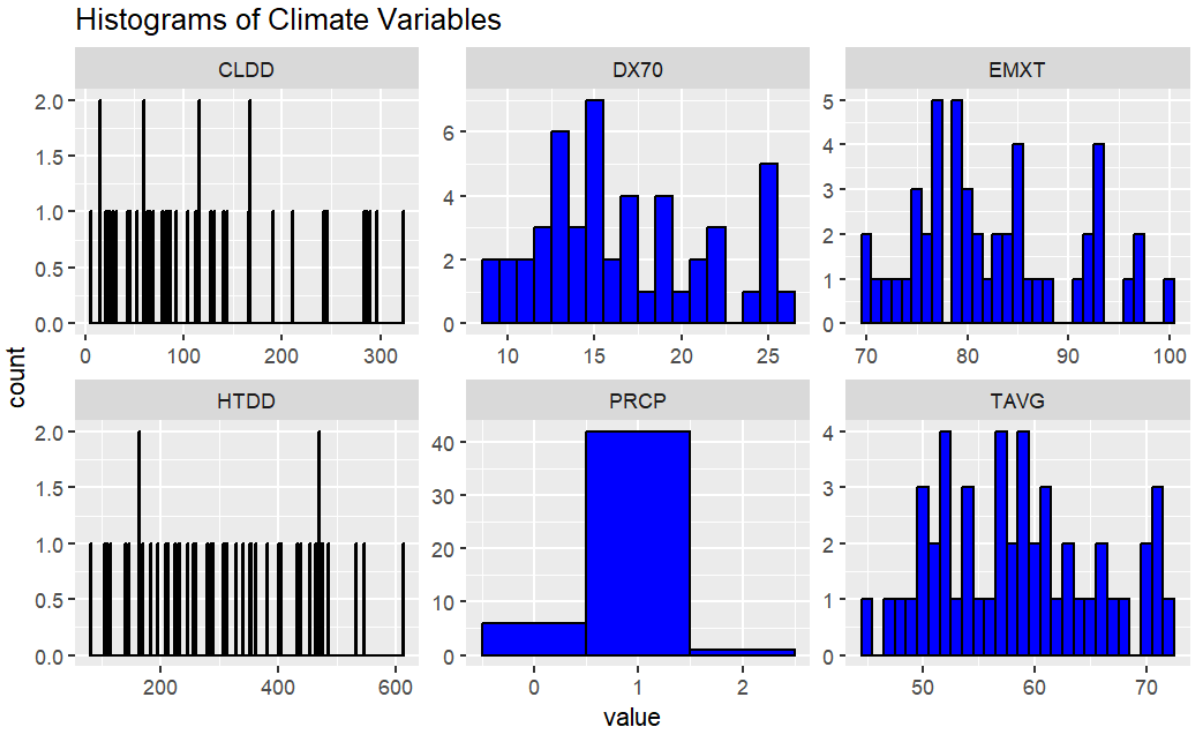


Figure 1: Histograms of climate variables

Histograms serve as powerful tools for visualizing the distribution of values associated with parameter. In the Figure 1, we visualize the distribution of different climate situations throughout the study period. We clearly observed that the features are not normally distributed.The histograms depicting CLDD and HTDD exhibit closely a uniform distribution, characterized by irregular peaks. The other weather variables have the skewed distribution.

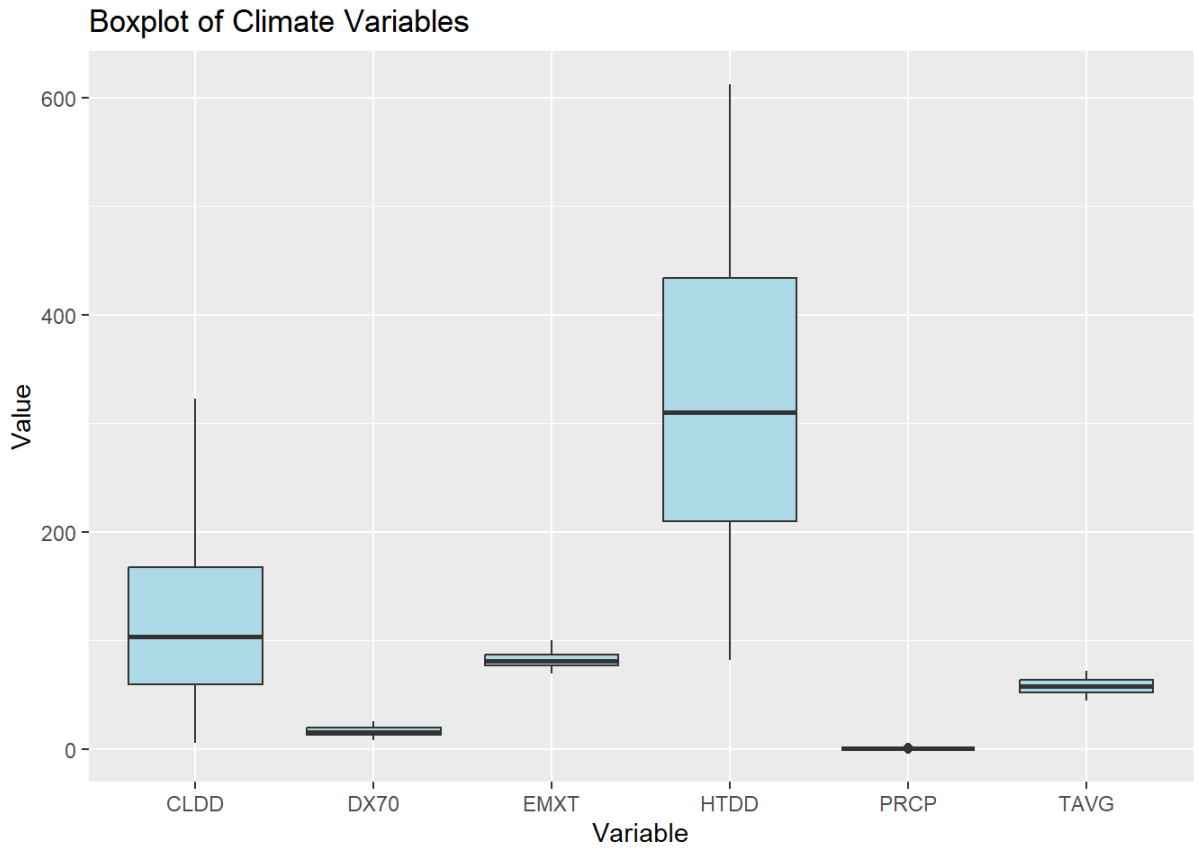


Figure 2: Boxplot of climate variables

The boxplots (Figure 2) provide valuable insights into the climate characteristics represented by different variables. CLDD exhibits a wide range of values, with the majority clustered towards the lower end, indicating fewer instances of extreme cold. DX70 demonstrates less variability, suggesting more consistency in the number of days exceeding 70°F, with the median leaning towards fewer such days. EMXT reveals significant monthly variations in highest temperatures without any outliers, highlighting the consistency of extreme heat events. Conversely, HTDD's narrow boxplot near the axis suggests numerous months with low heating degree days, possibly indicating a milder climate. PRCP's concentrated range suggests infrequent high precipitation, with outliers hinting at rare months with substantial rainfall. Lastly, TAVG's boxplot depicts a symmetric distribution of average temperatures, indicating moderate variability across months. Overall, these boxplots collectively characterize a climate profile with occasional extremes but generally moderate conditions across various meteorological parameters.

3.2 Correlation Analysis

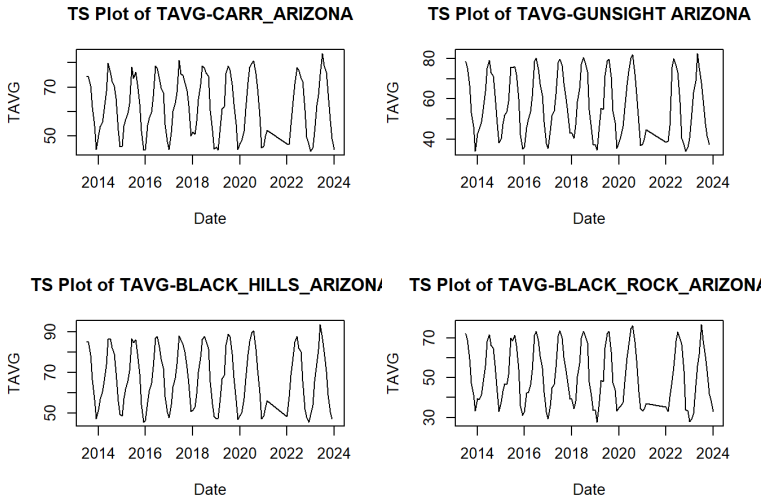
The correlation plot and matrix offer valuable insights into the relationships between different weather parameters. Through these visualizations, we can discern patterns of association, identifying whether certain variables tend to vary together or exhibit opposing trends.

Figure 3: Correlation plot

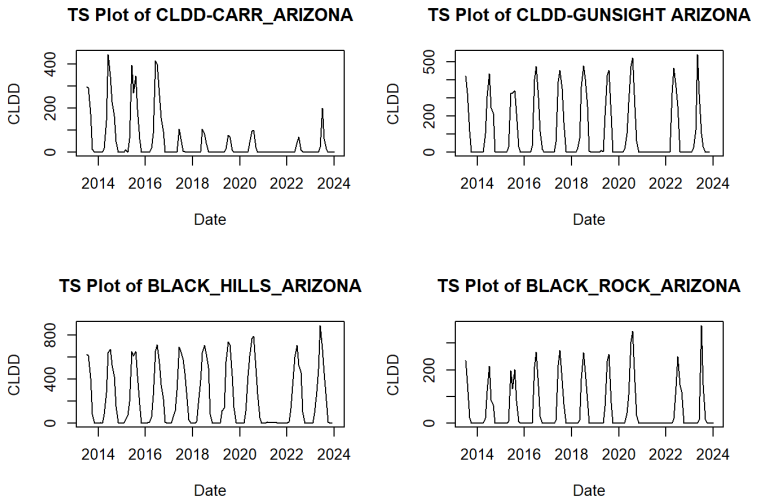
From the correlation matrix, we observed that PRCP (Total Precipitation) has relatively lower linear association with other climate variables, as expected since Arizona has lots of desert. TAVG (Average Temperature) indicating strong linear correlation with the other climate variables. It is also make sense, since temperatures are naturally higher in the state of Arizona.

3.3 Trend Analysis (Time Series)

In our time series analysis, we delved into the temporal dynamics of the weather data, unveiling underlying patterns, trends, and seasonality. From R coding, We randomly selected 4 stations and performed trend analysis for the key features.

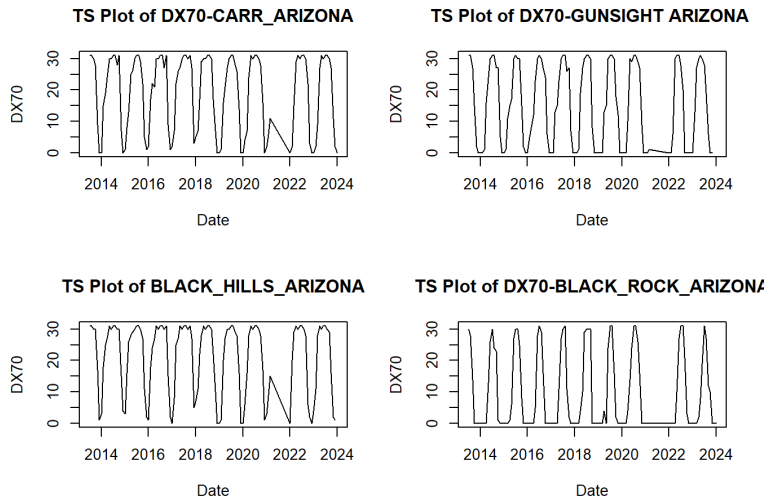


Interpretation: Among the random 4 stations, we observed that the range of the middle of the year has higher temperature, on the other hand start and end of the month has relatively lower temperature. For the year of 2022, there some inconsistence in pattern compared to other periods.

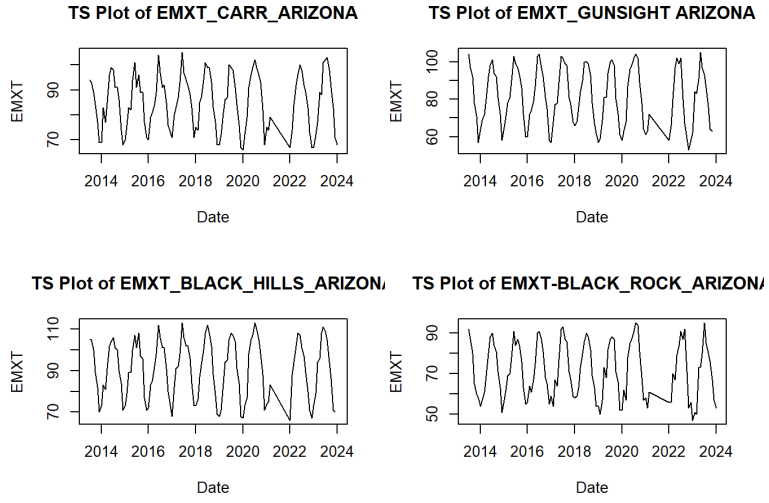


Interpretation: For CLDD (Cooling degree days), accept one station, the trend is pretty

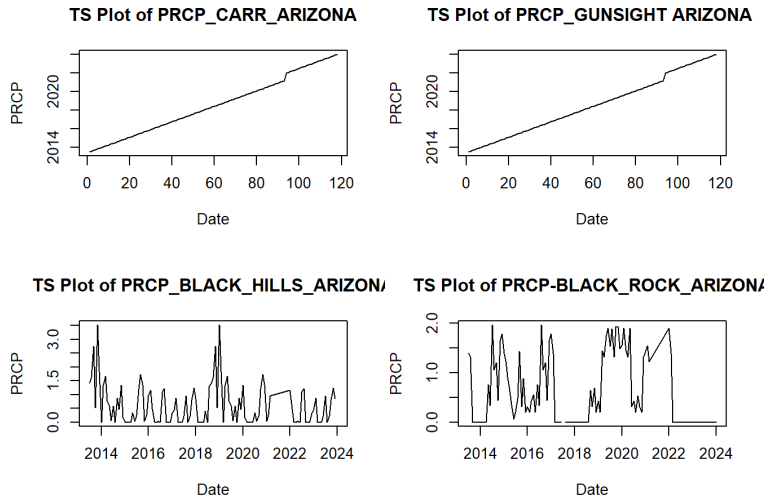
much consistent. However, the peak points sometimes not closely related.



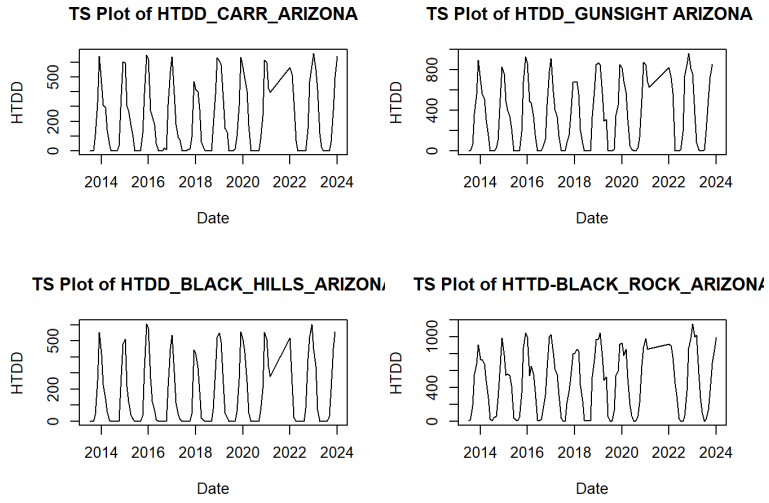
Interpretation: For DX70 (Number days with maximum temperature > 70 F (21.1C)), the trend seems consistent over periods. We observed pretty high peak here, since temperature of Arizona is naturally high.



Interpretation: Regarding the variable Extreme maximum temperature for the period (EMXT), we also observed roughly similar trend with DX70.



Interpretation: For Total precipitation (PRCP), the trend is pretty insconsistent. This also make sense, bacause precipitation incidents are naturally sudden in the Arizona.



Interpretation: For Heating degree days (HTDD), the trend has roughly similar to the EMXT and DX70.

4 Test Statistic

We previously observed (Figure 1) that the data is neither normally distributed nor independent. Thus, we conduct separate independent two-sample t-tests two periods defined by "Apr-Sep" and "Oct-Mar". Let's say, April to September (μ_1) and October to March (μ_2). Therefore, the Hypothesis would be:

$$\mathbf{H}_0 : \mu_1 - \mu_2 = 0$$

$$\mathbf{H}_1 : \mu_1 - \mu_2 \neq 0$$

Mathematically, the t-test itself involves comparing means of two samples and calculating the t-statistic, which is given by:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means,
- s_1^2 and s_2^2 are the sample variances,
- n_1 and n_2 are the sample sizes.

And the degrees of freedom expression:

$$df = n_1 + n_2 - 2$$

4.1 Test Result

Variable	Apr-Sep (μ_1)	Oct-Mar (μ_2)
CLDD	222.71126	11.66041
DX70	26.018119	7.018405
EMXT	92.53007	71.75965
HTDD	88.98458	544.21003
TAVG	69.33628	47.30733
PRCP	0.6227525	0.6322194

Table 2: Mean values by period

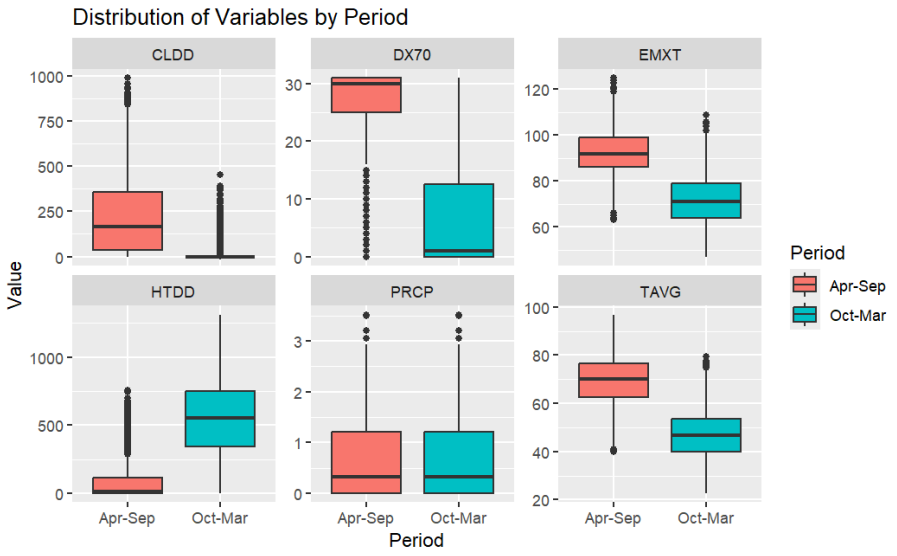


Figure 4: Distribution by period (Apr-Sep & Oct-Mar)

Conclusion: Alternative Hypothesis ($\mathbf{H}_1$) is true. We have enough evidence to conclude that the mean of October to March minus the mean of April to September is not zero.

5 Clustering

K-means clustering: K-means clustering is an unsupervised learning algorithm used for partitioning data into distinct groups based on similarity. When applied to weather station data, it can help identify patterns and similarities among different locations based on their weather characteristics.

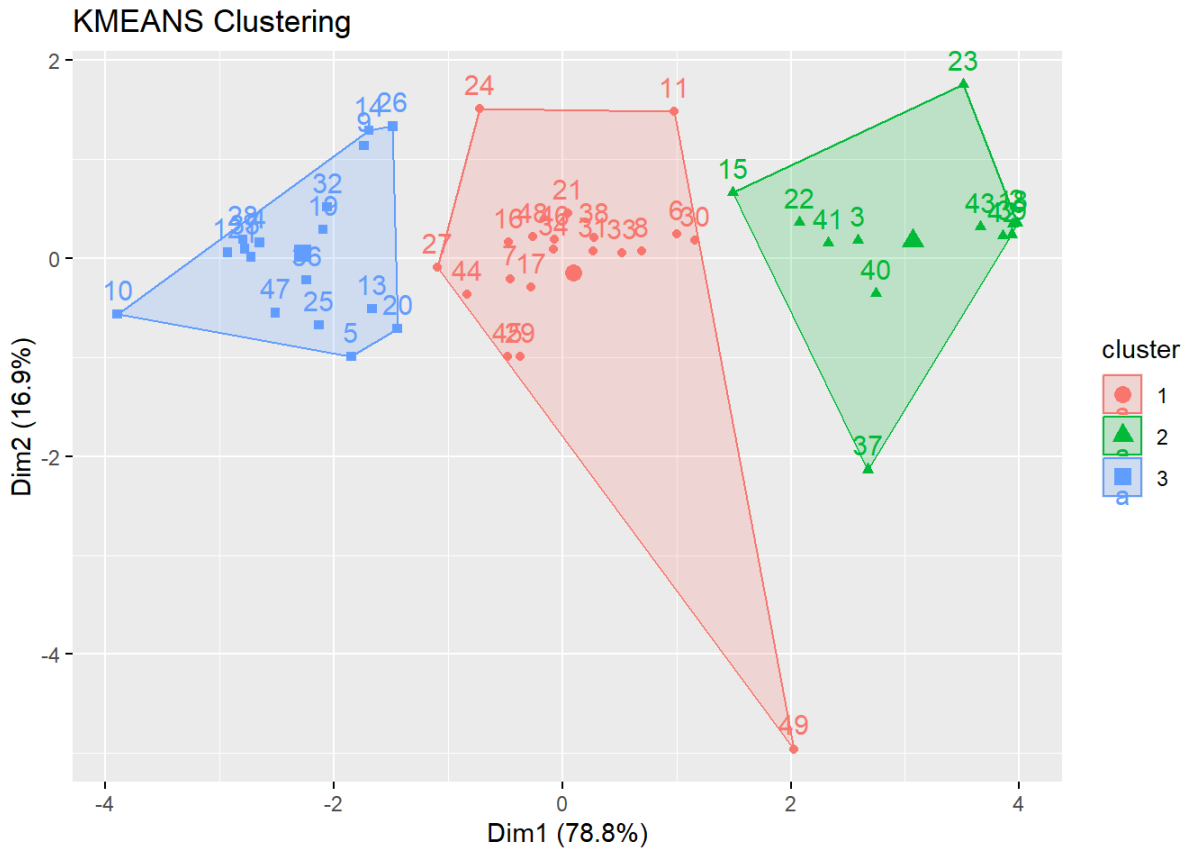
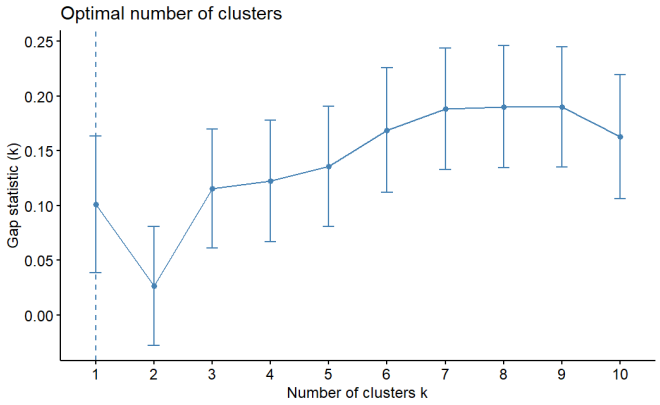


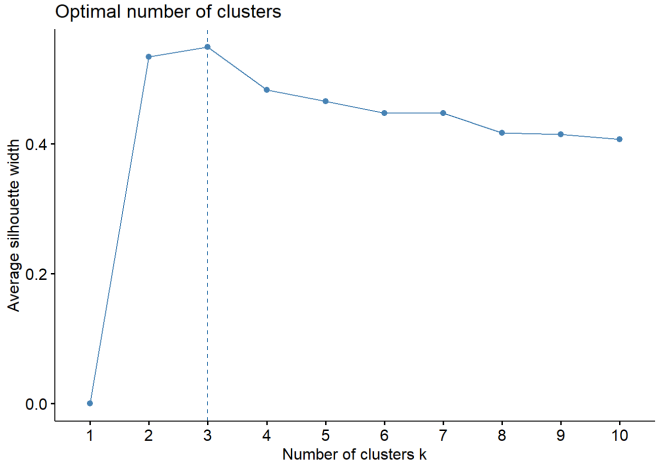
Figure 5: Initial Clustering (K-Means)

In this study, we observed a reasonable separation from the K-Means Clustering algorithm. However, we need to analysis further to find the optimal number of cluster to separate the observations. However, we didn't perform any experiment yet how should we choose the appropriate/optimal number of clusters (k).

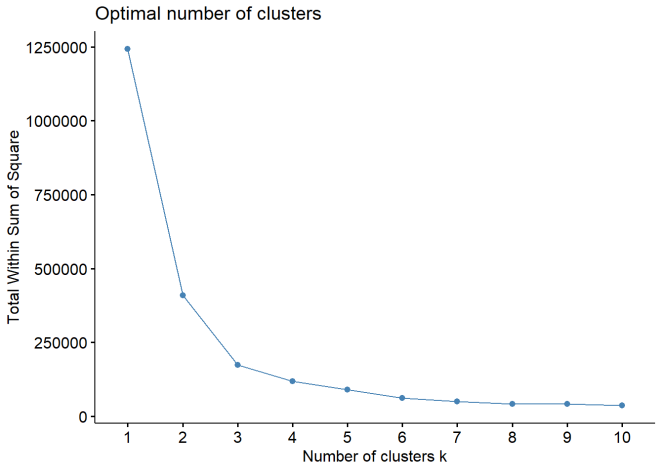
Choosing Appropriate K:



(a) Gap Statistics



(b) Avg Silhouette



(c) Sum of Square

Figure 6: Optimal number of clusters (k)

According to the given plots, $k = 3$ would be a reasonable choice for separating the given observations. From the R output, we also get the **WB** Ratio of **0.162**, which is relatively smaller.

Conclusion: The optimal number of clusters $k = 3$

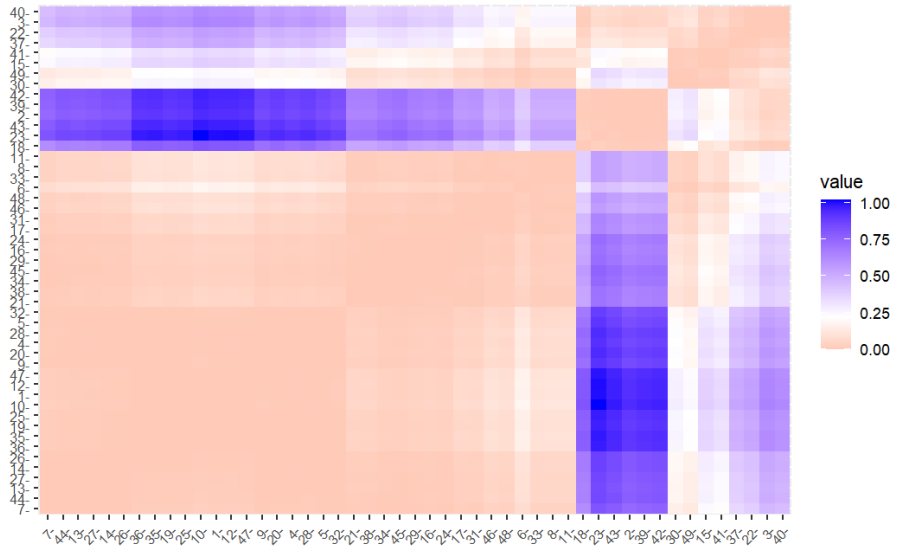


Figure 7: Visualize distances (pearson correlation)

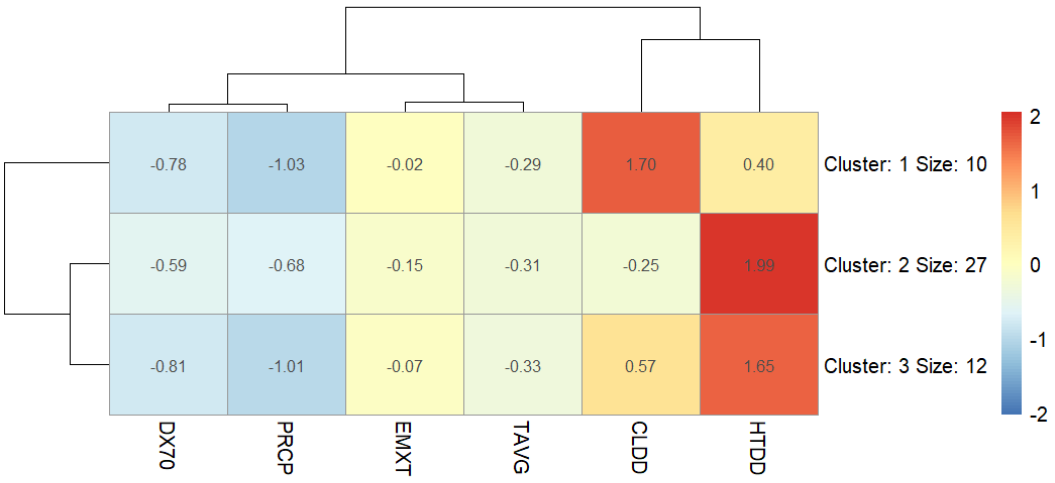


Figure 8: Heat map (clustering result)

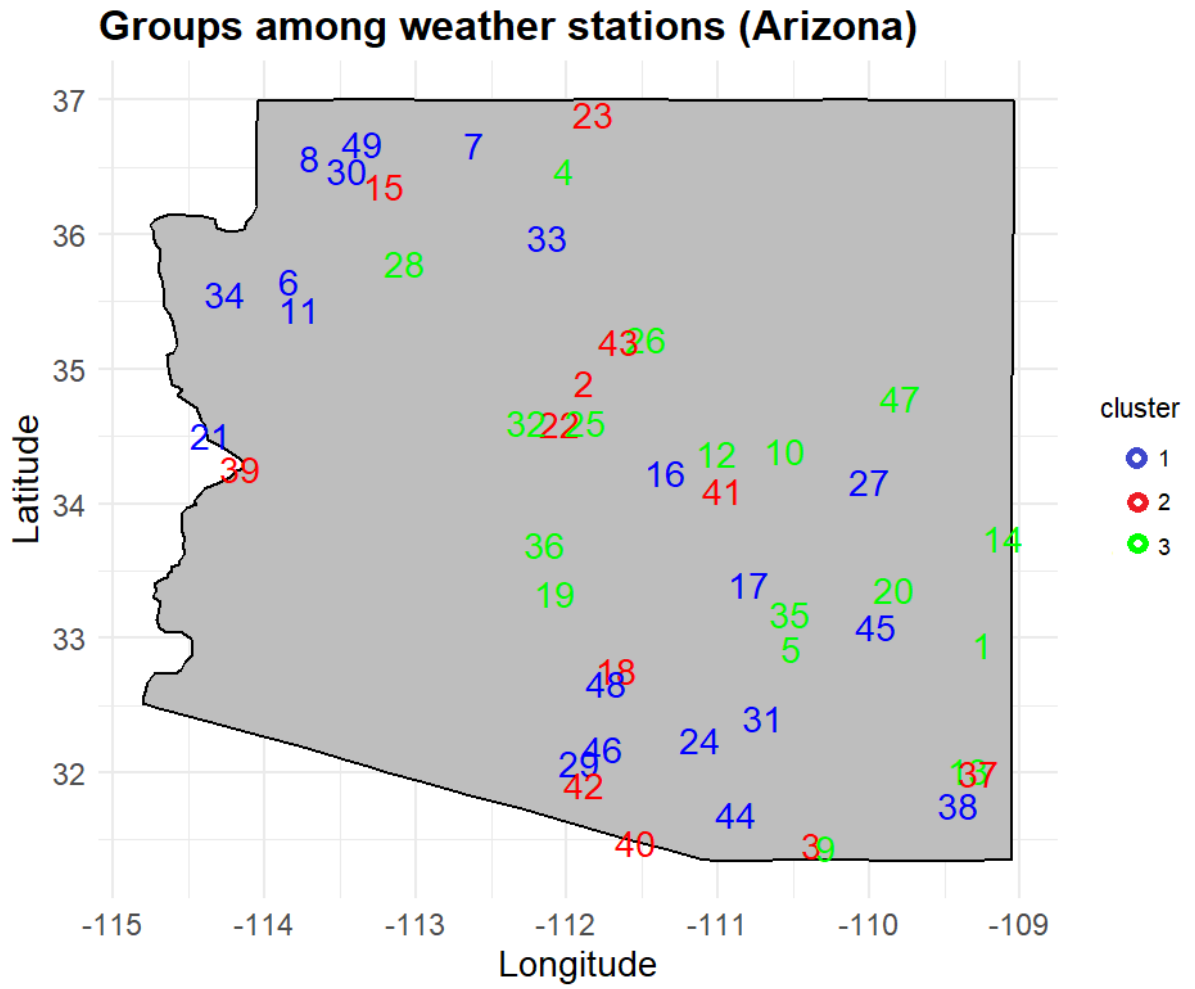


Figure 9: Clusters among the weather stations

6 Conclusion

6.1 Findings

The analysis revealed diverse climate patterns across Arizona, with varying distributions, correlations, and temporal trends among climate variables. Stations exhibited differences in temperature extremes, precipitation levels, and heating/cooling degree days, reflecting the state's diverse geography and climate zones. Seasonal variations were evident, with distinct patterns observed between different periods of the year. Additionally, clustering analysis identified groups of stations with similar climate profiles, suggesting regional similarities in weather patterns.

6.2 Insights on Decision Process

- **Agriculture:** Understanding climate variability can aid in crop selection, planting schedules, and irrigation management.
- **Urban Planning:** Knowledge of climate patterns can inform infrastructure development, such as building design and energy usage planning.
- **Disaster Preparedness:** Insights into extreme weather events can help in developing strategies for mitigating risks associated with floods, droughts, and heatwaves.
- **Tourism:** Understanding seasonal climate variations can inform tourism planning and marketing efforts.

Overall, the insights provided by this report can support informed decision-making processes in various sectors, contributing to the resilience and sustainability of communities across Arizona.

7 References

[1] National Oceanic and Atmospheric Administration (NOAA). National center for environmental information. https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-month.

[2] Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36.2 (2003): 451-461.