|  |  |  |
| --- | --- | --- |
| **Project title: Weather condition in Tehran 2023-2024** | | |
| **CMSC6950 Fall 2025** | **Mona Hejazi** | **201996350** |

**Introduction**

**Project Aim**

This project investigates a one-year weather dataset for Tehran, with a primary focus on identifying extreme values in temperature and precipitation. Through statistical analysis, the study aims to uncover insights into Tehran’s climatic trends and extreme weather events over the observed period.

**Background**

Weather data serves as a crucial resource for understanding climatic patterns, forecasting, and decision-making in various sectors. By analyzing meteorological parameters such as temperature, precipitation, and wind patterns, we can identify anomalies, explore correlations, and assess their potential impacts. This project explores the statistical distribution of key weather variables for Tehran, delving into both summary statistics and outlier detection.

**Dataset Overview:**

The dataset contains daily weather records for Tehran from **October 1, 2023**, to **September 30, 2024**. It comprises 366 entries, representing each day of the year, and includes 11 meteorological variables:

- `date`: Date of the weather record

- `tavg`: Average temperature (°C)

- `tmin`: Minimum temperature (°C)

- `tmax`: Maximum temperature (°C)

- `prcp`: Precipitation (mm)

- `snow`: Snowfall (cm) — with limited data

- `wdir`: Wind direction (degrees)

- `wspd`: Wind speed (m/s)

- `wpgt`: Wind gust (m/s) — with missing data

- `pres`: Atmospheric pressure (hPa)

- `tsun`: Sunshine duration (hours) — missing data

**Objectives**

The project’s primary objectives include:

1) Analyzing daily averages, minimum and maximum temperatures, and precipitation trends across the year.

2) Identifying and visualizing extreme values in temperature and precipitation using statistical methods.

3) Exploring correlations between key variables to understand weather dynamics in Tehran.

**Hypothesis**

1) The average temperature, wind speed, and precipitation do not follow a normal distribution.

2) There is no linear correlation between average temperature and precipitation.

3) A linear correlation exists between wind speed and precipitation.

4) Removing outliers enhances the reliability of statistical analysis.

**Source :** https://meteostat.net/en/place/ir/tehran?s=40754&t=2023-10-01/2024-09-30

**Data Processing and Analysis of Tehran's Weather Data**

**Overview of Data Inspection and Handling Missing Values**

The project commenced with an in-depth inspection of the dataset to understand its structure and identify potential issues such as missing data. Several columns, including snow, wpgt, and tsun, exhibited substantial amounts of missing values. Given their limited data availability, these columns were excluded from further analysis to maintain the integrity and relevance of the findings.

**Summary Statistics**

Key variables—`tavg` (average temperature), `tmin`(minimum temperature), `tmax` (maximum temperature), and `prcp` (precipitation)—were analyzed to compute summary statistics. These included metrics such as the mean, standard deviation, minimum, maximum, and interquartile range (25th, 50th, and 75th percentiles). The summary statistics served as a foundational step to understand data distributions, detect anomalies, and gain insights into the central tendencies and variability within the dataset.

**Analyzing Extreme Temperature Values in Tehran**

**Workflow Overview**

To examine Tehran's weather trends over one year, I developed a Python-based analytical pipeline. The workflow was initiated by importing and preprocessing the dataset from an Excel file. The date column was converted to a datetime format and set as the index to enable time-series operations. This setup facilitated a structured approach to analyzing temporal trends in weather data.

**Identifying and Highlighting Extreme Values**

A custom function was implemented to extract extreme temperature values, pinpointing:

1) The highest maximum temperature along with its corresponding date.

2) The lowest minimum temperature and its occurrence date.

3) These findings were crucial for understanding significant deviations in Tehran’s weather patterns.

**Visualization of Temperature Trends**

To communicate insights effectively, a comprehensive visualization was created:

**Daily Trends:** The plot featured daily minimum, maximum, and average temperatures, with shaded areas representing the range between minimum and maximum values for enhanced clarity.

**Extreme Highlights:** Markers and annotations were used to emphasize extreme temperature points, ensuring they stood out within the broader trends.

**Mean Temperature:** A horizontal line denoting the annual mean temperature was added to provide contextual reference against daily values.

The plot was further customized with:

1) Clear axis labels and a descriptive title.

2) An informative caption to contextualize the visualization within the project’s objectives.

3) Enhanced aesthetics, ensuring alignment with professional documentation standards.

**Output and Documentation**

This systematic approach not only highlights extreme weather events but also underscores broader temperature patterns in Tehran, offering a valuable perspective for stakeholders interested in climate trends.

A graph showing the temperature of the year

Description automatically generated

**Temperature Analysis: Deviations from Historical Mean by ±1.5 Standard Deviations**

**Objective and Methodology**

This analysis focuses on identifying temperature trends that deviate significantly from the historical mean, using a threshold of ±1.5 standard deviations (SD). The aim was to classify daily temperatures into three categories:

- Above the Mean by +1.5 SD (significantly warmer days).

- Below the Mean by -1.5 SD (significantly cooler days).

- Within ±1.5 SD (normal range).

**Data Preparation**

**Data Source:** The dataset was imported from an Excel file.

**Preprocessing:** The date column was converted to a datetime format for proper indexing and time-series analysis.

**Statistical Computations:** The script computed the historical mean and standard deviation for daily average temperatures (tavg). These values were used to define thresholds for categorizing the data.

**Categorization and Monthly Trends**

Each day was classified into one of the three categories based on its temperature value:

- Days exceeding the upper threshold (+1.5 SD) were labeled as "Above."

- Days below the lower threshold (-1.5 SD) were labeled as "Below."

- Remaining days were categorized as "Normal."

The script further analyzed these trends by month:

**Monthly Counts:** The number of days in each category was tallied for every month.

**Majority Trends:** For each month, the predominant category was identified, indicating whether the month was predominantly warmer, cooler, or within normal temperature ranges.

**Visualization and Outputs**

**Scatter Plot:** A scatter plot was generated with color-coded markers to represent the temperature categories:

-Red dots for "Above" temperatures.

-Blue dots for "Below" temperatures.

-Green dots for "Normal" temperatures.

**Summary Table:** A detailed table summarizing monthly trends was created, showing the count of days in each category and the majority trend for each month.

**Key Findings**

Based on the analysis and results presented in Table 2:

**Warmest Month:** `July` recorded the highest number of days with temperatures significantly above the threshold, marking it as the warmest month of the year.

**Coldest Month:** `February` had the highest number of days with temperatures significantly below the threshold, making it the coldest month of the year.

This analysis provides a detailed understanding of temperature deviations from the historical mean, highlighting seasonal patterns and significant anomalies. The insights gained can support climate monitoring, urban planning, and resource management strategies in Tehran.

A graph showing the temperature of the day

Description automatically generatedA table with numbers and symbols

Description automatically generated

**Seasonal Temperature Analysis**

To divide the dataset into the four seasons (Winter, Spring, Summer, Fall), I will use the following general guidelines for seasonal divisions in Tehran:

- **Winter**: December 21 to March 20

- **Spring**: March 21 to June 20

- **Summer**: June 21 to September 20

- **Fall**: September 21 to December 20

This analysis focuses on average temperature data collected for Tehran from December 2023 to December 2024. To begin, I loaded the dataset from an Excel file and converted the date column into a datetime format for easier manipulation. Seasonal date ranges for Winter, Spring, Summer, and Fall were then defined, and the dataset was segmented accordingly. For each season, I extracted the average temperature data to prepare for visualization.

To better understand the distribution of temperature data across seasons, I generated two boxplots: one before and one after removing outliers. Outliers were identified using two methods: the Z-Score (±1.5 standard deviations) and the Interquartile Range (IQR). In the first plot, outliers identified by the Z-Score method were marked in gold, while those identified by the IQR method were highlighted in orange. Each plot was annotated with the mean and standard deviation for the temperature data of each season.

Following the identification of outliers, I removed them from the dataset and recalculated the statistics. The second boxplot displayed the data after outlier removal. Additionally, I calculated and displayed the percentage of data removed for each season as well as overall.

To enhance the clarity of the visualizations, I added a caption explaining the plots, included a simple legend to differentiate between the types of outliers, and incorporated gridlines for better readability. This analysis provided a clear view of seasonal temperature trends and demonstrated the influence of outliers on the data distribution.

A screenshot of a graph

Description automatically generated

The information of outlier removal data:

| **Season** | **Number of Outliers Removed** | **Percentage of Data Removed** |
| --- | --- | --- |
| Winter | 17 | 18.68% |
| Spring | 18 | 19.57% |
| Summer | 11 | 11.96% |
| Fall | 0 | 0.00% |
| **Total** | **46** | **16.14%** |

**Seasonal Temperature Heatmap for Tehran**

In this analysis, I processed a weather dataset to create a seasonal temperature heatmap for Tehran in the year 2024. First, I loaded the dataset from an Excel file and converted the 'date' column to a datetime format for easier manipulation. I then defined the date ranges for each season: winter, spring, summer, and fall. Using these seasonal ranges, I segmented the dataset accordingly. Next, I resampled the data to daily averages for each season and ensured consistency in the dataset length by filling any missing values using forward-fill and backward-fill methods. I combined the seasonal data into a single matrix, where each season was represented as a column. I then visualized this matrix as a heatmap, with color representing temperature, ranging from blue (cold) to red (warm). Labels for the axes and a color bar were added for clarity, and the plot was saved as a PDF. The resulting heatmap provides a clear visual representation of the average daily temperatures across the four seasons in Tehran.

A close-up of a temperature chart

Description automatically generated

**Statistical Analysis of Weather Data**

In this analysis, I loaded and preprocessed weather data from an Excel file, focusing on temperature (`tavg`), wind speed (`wspd`), and precipitation (`prcp`). The 'date' column was converted to datetime format, and it was set as the index for easier time-based analysis. I then performed a statistical analysis of the distributions of these variables by plotting histograms and kernel density estimates (KDE) for each column. For each variable, I calculated and reported the skewness, kurtosis, and performed the Shapiro-Wilk normality test to assess whether the data followed a normal distribution. The results of these statistical analyses were compiled into a table, which was saved as a CSV file. The plots, including both histograms and KDEs, were saved in a PDF file with a caption, providing a comprehensive visual summary of the distribution trends. The final output includes both the figure and the statistical analysis table, which were saved in the specified output folder for further review and documentation.

A diagram of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

**Weather Data Analysis: Temperature and Precipitation Trends in Tehran**

In this analysis, I processed and visualized weather data for Tehran by focusing on the trends in temperature (minimum, maximum, and average) and precipitation over time.  First, I loaded and prepared the data by converting the 'date' column to a datetime format and setting it as the index. I then created a combined plot displaying temperature trends in light colors and precipitation with dark purple bars. The plot highlights extreme values for both temperature and precipitation, including the highest and lowest recorded values with annotations. I also calculated the correlation between temperature and precipitation and displayed it on the plot. Furthermore, I identified and printed the months that had less precipitation than the overall average. The plot is saved as a PDF file with a caption describing its contents, and the results, including months with low precipitation, are printed for further analysis. Finally, I displayed the plot, showing the temperature and precipitation trends over time with the necessary annotations and labels for clarity.

A graph with numbers and lines

Description automatically generated with medium confidence

The driest months of the year, with precipitation levels below the mean, are October, December, June, July, and August.

**Visualizing Temperature and Precipitation Trends**

In this analysis, I aimed to explore the temperature and precipitation trends over time using various time series plots. To achieve this, I utilized several visualization techniques to examine the data comprehensively. Below are the steps I followed, and the corresponding plots I created:

**Line Plot (Time Series Plot)**

To begin, I created a line plot to visualize the temperature and precipitation trends over time. This plot displays continuous data for both temperature and precipitation, where the x-axis represents the date, and the y-axis shows the values for temperature and precipitation. This plot effectively shows how both variables evolve over the entire period of the dataset.

In this analysis, I processed weather data from Tehran to calculate and visualize seasonal statistics related to temperature, precipitation, and wind speed. The dataset, stored in an Excel file, was loaded and the 'date' column was converted into a datetime format for easier manipulation. I defined the seasonal date ranges for winter, spring, summer, and fall, then segmented the data accordingly. For each season, I calculated key statistics such as the minimum, maximum, mean, median, and standard deviation for temperature, precipitation, and wind speed. These statistics were organized into a table for clearer presentation. Additionally, I created a grouped bar chart to visually compare the average values of temperature, precipitation, and wind speed across the four seasons. The bar chart was annotated with precise values to improve readability. To supplement the visual representation, I added a detailed table beneath the figure summarizing the calculated statistics. Finally, I saved the figure and table as a PDF document for further use.

A graph of different colored bars

Description automatically generated with medium confidence

**Correlation heatmap**

In this analysis, I created two types of visualizations to explore the relationships between various weather variables in the dataset. First, I developed a correlation heatmap to examine how numerical variables such as temperature, precipitation, wind speed, and atmospheric pressure are interrelated. The heatmap was generated by selecting relevant numerical columns, excluding variables such as snow, wind gust, and solar radiation. I computed the correlation matrix for these selected variables and visualized the results using a heatmap, with annotated correlation coefficients and a color scale. Additionally, I included a legend to describe each variable for better clarity. The heatmap was saved as a PDF in the specified output folder.

**Pairplot**

Next, I generated a pairplot to visually analyze the pairwise relationships between the same weather variables. The pairplot included kernel density estimations (KDE) on the diagonal to show the distributions of each variable and scatterplots for their pairwise relationships. Similar to the heatmap, I added a legend with descriptions of each variable. This pairplot was also saved as a PDF for further use. Both visualizations provide insights into the correlations and distributions of weather data, aiding in understanding the patterns and interactions between the key weather variables.

A graph showing weather varities

Description automatically generated with medium confidenceA group of blue dots

Description automatically generated

Figure 8: The correlation heatmap and pairplot illustrate relationships among selected weather variables in Tehran.

Here’s a table representing the correlation scale based on the descriptions provided:

A black and white box with black text

Description automatically generated

This table provides a quick reference for understanding the strength and direction of the correlation between two variables.

This correlation heatmap shows the pairwise correlation coefficients between four weather variables: average temperature (tavg), precipitation (prcp), wind speed (wspd), and atmospheric pressure (pres). The values range from -1 to 1, where red indicates a strong positive correlation and blue indicates a strong negative correlation. Temperature (tavg) has a moderate negative correlation with pressure (pres) (-0.76) and a very weak negative correlation with precipitation (prcp) (-0.16). Wind speed (wspd) shows a weak positive correlation with temperature (tavg) (0.12) and a weak negative correlation with pressure (-0.31). Precipitation (prcp) has very weak correlations with other variables, with the highest being a small negative correlation with temperature (-0.16). This heatmap indicates that temperature and pressure have a stronger inverse relationship, while other variable pairs show minimal or no significant correlations.

In the pairplot, the histograms on the diagonal show the distribution of each variable, with temperature and pressure being more evenly distributed, while precipitation and wind speed are highly skewed. The scatter plots off the diagonal reveal weak or no strong correlations between the variables. Temperature shows a slight negative trend with wind speed and pressure, but there is no clear relationship between precipitation and any other variable. Overall, the plot suggests limited correlations between the variables in the dataset.

**Discussion**

The analysis of Tehran's weather data over one year has revealed significant insights into its climatic patterns and extreme weather events. The study employed rigorous statistical methods, including summary statistics, correlation analysis, and outlier detection, alongside visualization techniques to highlight trends and anomalies in temperature and precipitation.

Key findings demonstrate that Tehran's weather exhibits distinct seasonal characteristics. For example, July was identified as the warmest month, while February emerged as the coldest. This seasonal variation was further supported by the analysis of temperature trends, which showed pronounced deviations during extreme months. The correlation analysis highlighted limited linear relationships between average temperature and precipitation, aligning with the hypothesis that these variables are not strongly correlated. Additionally, the study identified that removing outliers enhanced the reliability of the data, as evidenced by the cleaner distributions post-removal.

Despite the robust methodology, the analysis faced limitations. Significant missing data in variables such as wind gust and sunshine duration restricted a more comprehensive investigation of Tehran's climatic dynamics. Future studies could address this by integrating multiple data sources to fill gaps and expand the scope of analysis.

**Conclusion**

This project underscores the utility of statistical and visualization techniques in climate studies. By analyzing Tehran's weather patterns, it has provided valuable insights into extreme events and seasonal trends that are critical for urban planning and resource management. The identification of outliers and extreme values contributes to understanding climate anomalies, potentially aiding decision-makers in preparing for and mitigating the effects of extreme weather events.

Overall, the findings serve as a foundation for further research into Tehran's weather and climate. Expanding the temporal range and integrating additional meteorological variables could enhance the understanding of the city's long-term climatic trends, supporting broader climate resilience strategies.