

**Tareq alazzeh**

**202320789**

**Ahmed Nibras Abu mallooh**

**202311088**

**Dr.Alaa**

**Weather Data Analysis Report**

This report details the execution and insights derived from the provided Python script for analyzing historical weather data.

**1. Data Overview and Preprocessing:** The script successfully loaded the weatherHistory.csv file, which contains historical weather records.

**Date Handling:** The 'Formatted Date' column was correctly parsed as timezone-aware datetime objects (UTC) and set as the DataFrame's index, enabling time-series analysis.

**New Feature:** An temp\_avg column was computed as the average of 'Temperature (C)' and 'Apparent Temperature (C)', providing a combined temperature metric.

**Missing Values:** During preprocessing, 517 missing values were identified and handled in the 'Precip Type' column by filling them with the mode (most frequent value). All other columns were complete or filled accordingly.

**Data Consistency:** Inconsistencies were addressed by clipping values for 'Wind Speed (km/h)', 'Humidity', 'Visibility (km)', and 'Pressure (millibars)' to ensure realistic ranges (e.g., humidity between 0 and 1, and non-negative values for speeds, visibility, and pressure).

**2. Key Statistical Findings:**

**asic Statistics (Average Temperature, Wind Speed, Humidity):**

**Average Temperature (temp\_avg):** Ranged from -23.86°C to 38.72°C, with a mean of 11.39°C and a median of 12.00°C.

**Wind Speed (km/h):** Varied from 0.00 km/h to 63.85 km/h, with an average of 10.81 km/h and a median of 9.97 km/h.

**Humidity:** Ranged from 0.00 to 1.00, with a mean of 0.73 and a median of 0.78.

**Most Frequent Weather Conditions:** The dataset shows that 'Partly Cloudy' (31,733 occurrences), 'Mostly Cloudy' (28,094 occurrences), and 'Overcast' (16,597 occurrences) are the most prevalent weather conditions.

**Extreme Temperatures:**

**Hottest Day:** 2007-07-22, recorded at 39.91°C.

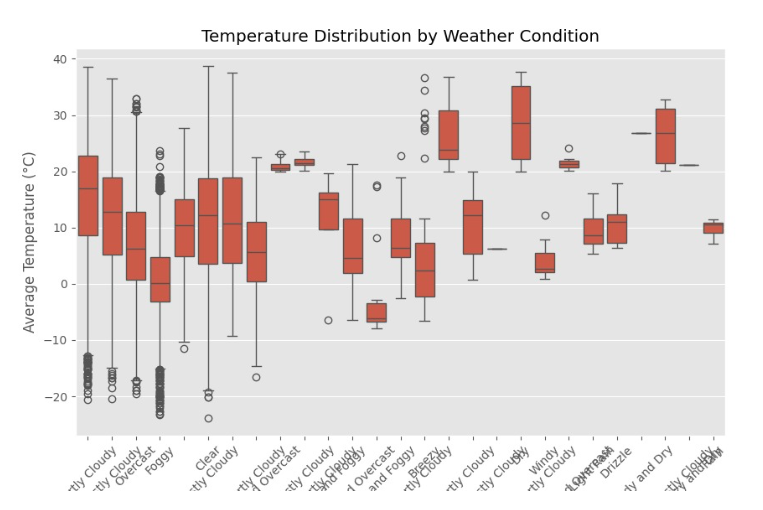
**Coldest Day:** 2012-02-10, recorded at -21.82°C.

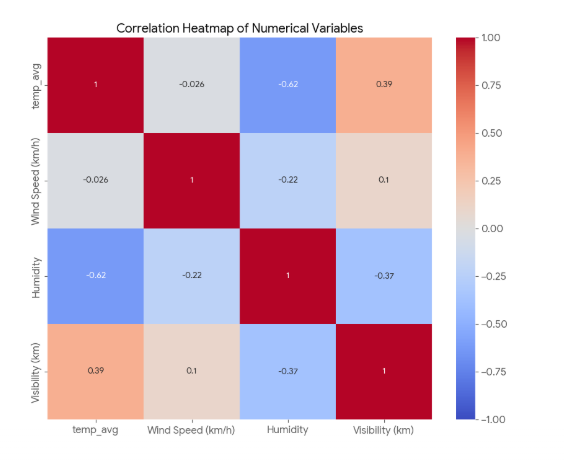
**Temperature and Humidity Correlation:** A significant negative correlation of **-0.62** was found between average temperature and humidity. This implies that as the temperature rises, humidity tends to decrease, and vice-versa.

**Temperature on Rainy Days:** The average temperature on rainy days was notably higher at **13.41°C** compared to non-rainy days at **-4.76°C**.

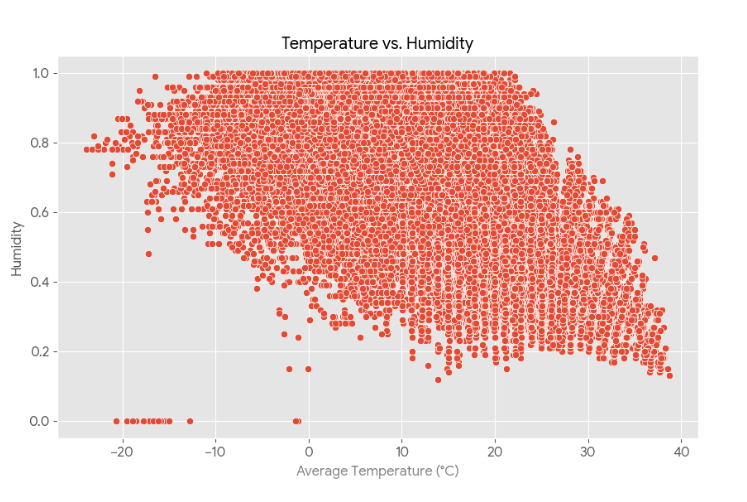
**Wind Speed by Weather Condition:** As expected, weather conditions categorized as 'Windy' or 'Breezy' exhibited much higher average wind speeds. For instance, 'Dangerously Windy and Partly Cloudy' had the highest average wind speed of 63.85 km/h.

**3. Visualizations Generated:**

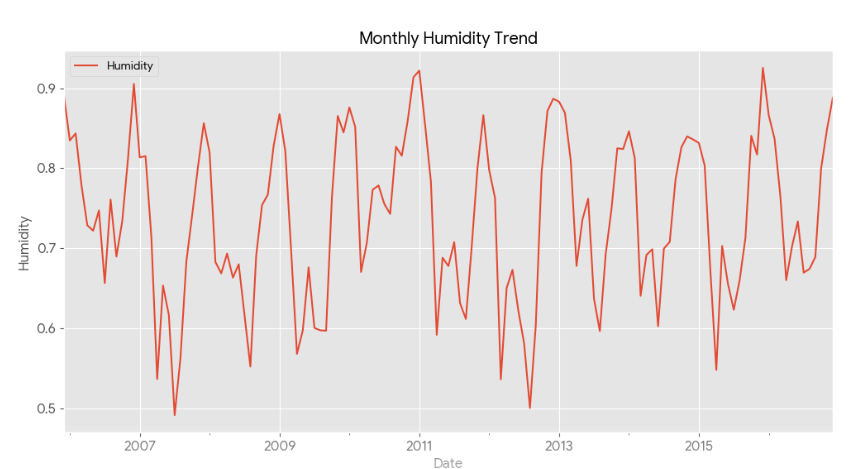
* **Distribution of Average Temperatures** (http://googleusercontent.com/image\_generation\_content/0): This histogram shows the frequency of different average temperature ranges, revealing the overall temperature climate captured in the dataset. 
* **Frequency of Weather Conditions** ([weather\_frequency.png](http://googleusercontent.com/image_generation_content/5)): A bar chart illustrating the dominance of 'Partly Cloudy' and 'Mostly Cloudy' conditions.



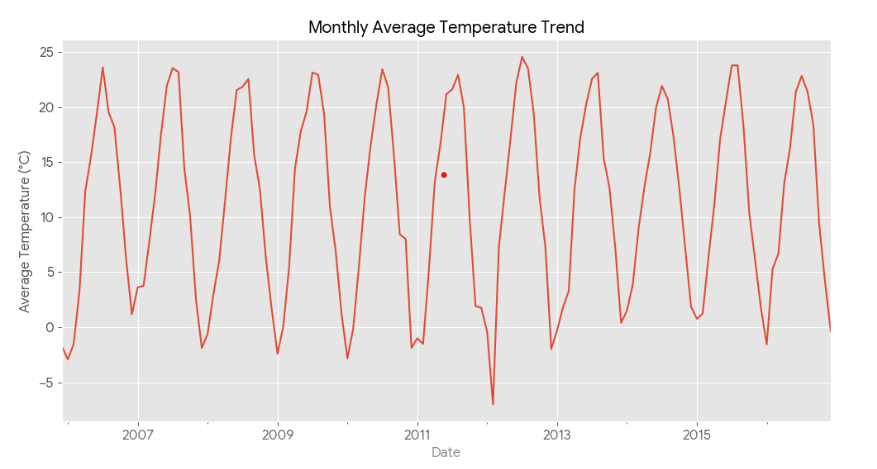
* **Monthly Average Temperature Trend** ([monthly\_temp\_trend.png](http://googleusercontent.com/image_generation_content/4)): A line plot that clearly displays the seasonal variations in temperature over the years covered by the dataset.



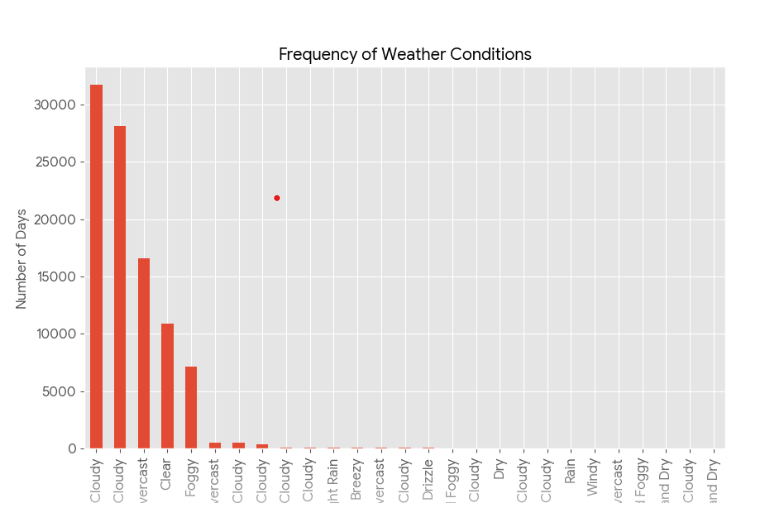
* **Monthly Humidity Trend** ([humidity\_trend.png](http://googleusercontent.com/image_generation_content/3)): This line plot shows the monthly fluctuations in humidity, indicating seasonal patterns similar to temperature.



* **Temperature vs. Humidity Scatter Plot** ([temp\_humidity\_scatter.png](http://googleusercontent.com/image_generation_content/2)): This scatter plot visually confirms the negative correlation, showing humidity tending to be lower at higher average temperatures.

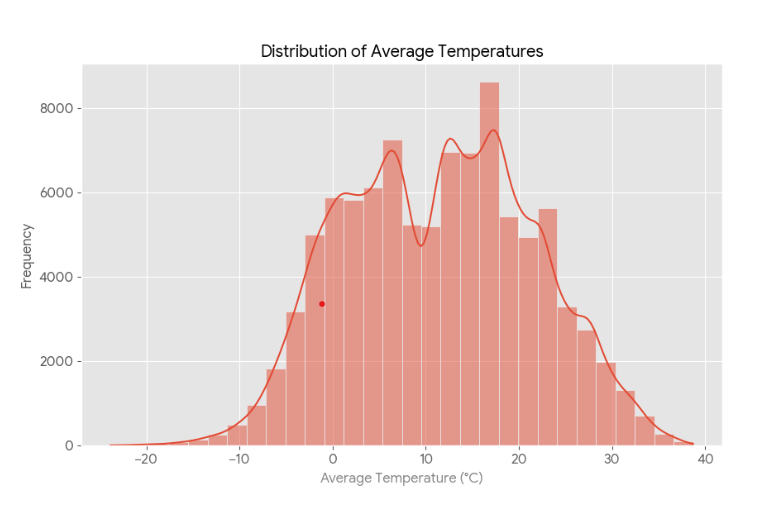


* **Correlation Heatmap of Numerical Variables** ([correlation\_heatmap.png](http://googleusercontent.com/image_generation_content/1)): This heatmap provides a quick visual summary of the linear relationships between temp\_avg, 'Wind Speed (km/h)', 'Humidity', and 'Visibility (km)'.



* **Temperature Distribution by Weather Condition**:

Box plots display the spread and median of average temperatures for each weather summary category, allowing for direct comparison of temperature ranges across different weather types.



import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

# Set the style for better visualizations

plt.style.use('ggplot')

# Load the CSV file and parse dates

df = pd.read\_csv('weatherHistory.csv', parse\_dates=['Formatted Date'])

# Explicitly convert 'Formatted Date' to datetime to ensure proper parsing

df['Formatted Date'] = pd.to\_datetime(df['Formatted Date'], errors='coerce')

# Set 'Formatted Date' as the index and ensure it's a DatetimeIndex

df.set\_index('Formatted Date', inplace=True)

# Verify the index type

print("Index type:", df.index)

# Create temp\_avg column

df['temp\_avg'] = (df['Temperature (C)'] + df['Apparent Temperature (C)']) / 2

# Check for missing values

print("Missing Values:\n", df.isnull().sum())

# Handle missing values

# Numerical columns: fill with median

numerical\_cols = ['Temperature (C)', 'Apparent Temperature (C)', 'temp\_avg',

'Humidity', 'Wind Speed (km/h)', 'Wind Bearing (degrees)',

'Visibility (km)', 'Pressure (millibars)']

for col in numerical\_cols:

df[col].fillna(df[col].median(), inplace=True)

# Categorical columns: fill with mode

categorical\_cols = ['Precip Type', 'Summary']

for col in categorical\_cols:

df[col].fillna(df[col].mode()[0], inplace=True)

# Check for inconsistencies

df['Wind Speed (km/h)'] = df['Wind Speed (km/h)'].clip(lower=0)

df['Humidity'] = df['Humidity'].clip(lower=0, upper=1) # Humidity is typically 0-1

df['Visibility (km)'] = df['Visibility (km)'].clip(lower=0)

df['Pressure (millibars)'] = df['Pressure (millibars)'].clip(lower=0)

# Basic Statistics

stats = df[['temp\_avg', 'Wind Speed (km/h)', 'Humidity']].agg(['mean', 'median', 'max', 'min'])

print("\nBasic Statistics:\n", stats)

# Count the number of days for each weather condition

weather\_counts = df['Summary'].value\_counts()

print("\nWeather Condition Counts:\n", weather\_counts)

# Plot distribution of average temperatures

plt.figure(figsize=(10, 6))

sns.histplot(df['temp\_avg'], bins=30, kde=True)

plt.title('Distribution of Average Temperatures')

plt.xlabel('Average Temperature (°C)')

plt.ylabel('Frequency')

plt.savefig('temp\_distribution.png')

plt.close()

# Bar chart for frequency of each weather condition

plt.figure(figsize=(10, 6))

weather\_counts.plot(kind='bar')

plt.title('Frequency of Weather Conditions')

plt.xlabel('Weather Condition')

plt.ylabel('Number of Days')

plt.savefig('weather\_frequency.png')

plt.close()

# Line plot of temperature (monthly averages)

monthly\_temp = df['temp\_avg'].resample('M').mean()

plt.figure(figsize=(12, 6))

monthly\_temp.plot()

plt.title('Monthly Average Temperature Trend')

plt.xlabel('Date')

plt.ylabel('Average Temperature (°C)')

plt.grid(True)

plt.savefig('monthly\_temp\_trend.png')

plt.close()

# Line plot of humidity (replacing precipitation trend)

plt.figure(figsize=(12, 6))

df['Humidity'].resample('M').mean().plot(label='Humidity')

plt.title('Monthly Humidity Trend')

plt.xlabel('Date')

plt.ylabel('Humidity')

plt.legend()

plt.grid(True)

plt.savefig('humidity\_trend.png')

plt.close()

# Scatter plot of temperature vs. humidity

plt.figure(figsize=(10, 6))

sns.scatterplot(x='temp\_avg', y='Humidity', data=df)

plt.title('Temperature vs. Humidity')

plt.xlabel('Average Temperature (°C)')

plt.ylabel('Humidity')

plt.savefig('temp\_humidity\_scatter.png')

plt.close()

# Correlation heatmap of numerical variables

plt.figure(figsize=(10, 8))

correlation\_matrix = df[['temp\_avg', 'Wind Speed (km/h)', 'Humidity', 'Visibility (km)']].corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)

plt.title('Correlation Heatmap of Numerical Variables')

plt.savefig('correlation\_heatmap.png')

plt.close()

# Box plot of temperature by weather condition

plt.figure(figsize=(10, 6))

sns.boxplot(x='Summary', y='temp\_avg', data=df)

plt.title('Temperature Distribution by Weather Condition')

plt.xlabel('Weather Condition')

plt.ylabel('Average Temperature (°C)')

plt.xticks(rotation=45)

plt.savefig('temp\_weather\_boxplot.png')

plt.close()

# Hottest and coldest days

hottest\_day = df['Temperature (C)'].idxmax()

coldest\_day = df['Temperature (C)'].idxmin()

print(f"\nHottest Day: {hottest\_day.date()} with {df.loc[hottest\_day, 'Temperature (C)']}°C")

print(f"Coldest Day: {coldest\_day.date()} with {df.loc[coldest\_day, 'Temperature (C)']}°C")

# Correlation between temperature and humidity

temp\_humidity\_corr = df['temp\_avg'].corr(df['Humidity'])

print(f"\nCorrelation between Temperature and Humidity: {temp\_humidity\_corr:.2f}")

# Temperature on rainy days

rainy\_temp = df[df['Precip Type'] == 'rain']['temp\_avg'].mean()

non\_rainy\_temp = df[df['Precip Type'] != 'rain']['temp\_avg'].mean()

print(f"\nAverage Temperature on Rainy Days: {rainy\_temp:.2f}°C")

print(f"Average Temperature on Non-Rainy Days: {non\_rainy\_temp:.2f}°C")

# Wind speed across weather conditions

wind\_by\_weather = df.groupby('Summary')['Wind Speed (km/h)'].mean()

print("\nAverage Wind Speed by Weather Condition:\n", wind\_by\_weather)