Weather Data Analysis Report

Title Page

Problem Statement: Analyze the historical weather data using Python libraries (Pandas, Matplotlib, Seaborn) and provide insights through data visualization.

Prepared by: Aryan Singh

Roll No: 202401100400054

Introduction

The objective of this analysis is to explore historical weather data to derive meaningful insights. The dataset includes variables like temperature, apparent temperature, humidity, wind speed, visibility, and precipitation type. The analysis aims to identify patterns, correlations, and trends using data visualization techniques.

Methodology

1. Data Loading: The dataset was loaded using Pandas.

2. Data Preparation: Extracted date components (Year, Month, Day) and handled datetime parsing.

3. Visualization Tools: Utilized Matplotlib and Seaborn for creating graphs and plots.

4. Analysis Performed:

- Temperature Distribution
- Average Temperature by Month
- Correlation Heatmap
- Temperature vs. Apparent Temperature
- Precipitation Type Distribution
- Wind Speed Distribution
- Humidity vs. Visibility
- Pressure Trend Over Time

Code

 # importing Libraries

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

data = pd.read_csv('/content/weatherHistory.csv')

data.head()
```

	Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Loud Cover	Pressure (millibars)	Daily Summary
0	2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251.0	15.8263	0.0	1015.13	Partly cloudy throughout the day.
1	2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259.0	15.8263	0.0	1015.63	Partly cloudy throughout the day.
2	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204.0	14.9569	0.0	1015.94	Partly cloudy throughout the day.
3	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269.0	15.8263	0.0	1016.41	Partly cloudy throughout the day.
4	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259.0	15.8263	0.0	1016.51	Partly cloudy throughout the day.

 # Data Cleaning (Handle missing values, incorrect data types, etc.)

 # Remove rows with missing temperature values

```
data.dropna(subset=['Temperature (C)'], inplace=True)
```

 # Convert data types if necessary

 # Convert 'Date' column to datetime objects

```
data['Formatted Date'] = pd.to_datetime(data['Formatted Date'], utc = True)

data['Year'] = data['Formatted Date'].dt.year

data['Month'] = data['Formatted Date'].dt.month

data['Day'] = data['Formatted Date'].dt.day

data.head()
```

	Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Loud Cover	Pressure (millibars)	Daily Summary
0	2006-03-31 22:00:00+00:00	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251.0	15.8263	0.0	1015.13	Partly cloudy throughout the day.
1	2006-03-31 23:00:00+00:00	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259.0	15.8263	0.0	1015.63	Partly cloudy throughout the day.
2	2006-04-01 00:00:00+00:00	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204.0	14.9569	0.0	1015.94	Partly cloudy throughout the day.
3	2006-04-01 01:00:00+00:00	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269.0	15.8263	0.0	1016.41	Partly cloudy throughout the day.
4	2006-04-01 02:00:00+00:00	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259.0	15.8263	0.0	1016.51	Partly cloudy throughout the day.

 # Exploratory Data Analysis and Visualization

 # 1. Temperature Trends

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

plt.plot(data['Formatted Date'], data['Temperature (C)'], marker='o', linestyle='-')

plt.xlabel("Date")

plt.ylabel("Temperature (°C)")

plt.title("Temperature Trends Over Time")

plt.grid(True)

plt.show()
```

 # sets the plot style

```
sns.set_style('whitegrid')
```

 # 2. Average Temperature by Month

```
monthly_temp = data.groupby('Month')['Temperature (C)'].mean().reset_index()


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

sns.lineplot(x='Month', y='Temperature (C)', data=monthly_temp, marker='o')

plt.title('Average Temperature by Month', fontsize=16)

plt.xlabel('Month', fontsize=14)

plt.ylabel('Average Temperature (C)', fontsize=14)
```

```
plt.xticks(range(1, 13))
plt.show()

# 3. Correlation Heatmap of Weather Variables
plt.figure(figsize=(10, 8))
correlation_matrix = data.corr(numeric_only=True)
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap of Weather Variables', fontsize=16)
plt.show()

# 4. Temperature vs. Apparent Temperature
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Temperature (C)', y='Apparent Temperature (C)', hue='Humidity',
data=data, palette='viridis', alpha=0.6)
plt.title('Temperature vs. Apparent Temperature', fontsize=16)
plt.xlabel('Temperature (C)', fontsize=14)
plt.ylabel('Apparent Temperature (C)', fontsize=14)
plt.show()

# 5. Precipitation Type Distribution
plt.figure(figsize=(8, 6))
precip_counts = data['Precip Type'].value_counts()
sns.barplot(x=precip_counts.index, y=precip_counts.values, palette='pastel')
plt.title('Precipitation Type Distribution', fontsize=16)
plt.xlabel('Precipitation Type', fontsize=14)
plt.ylabel('Count', fontsize=14)
```

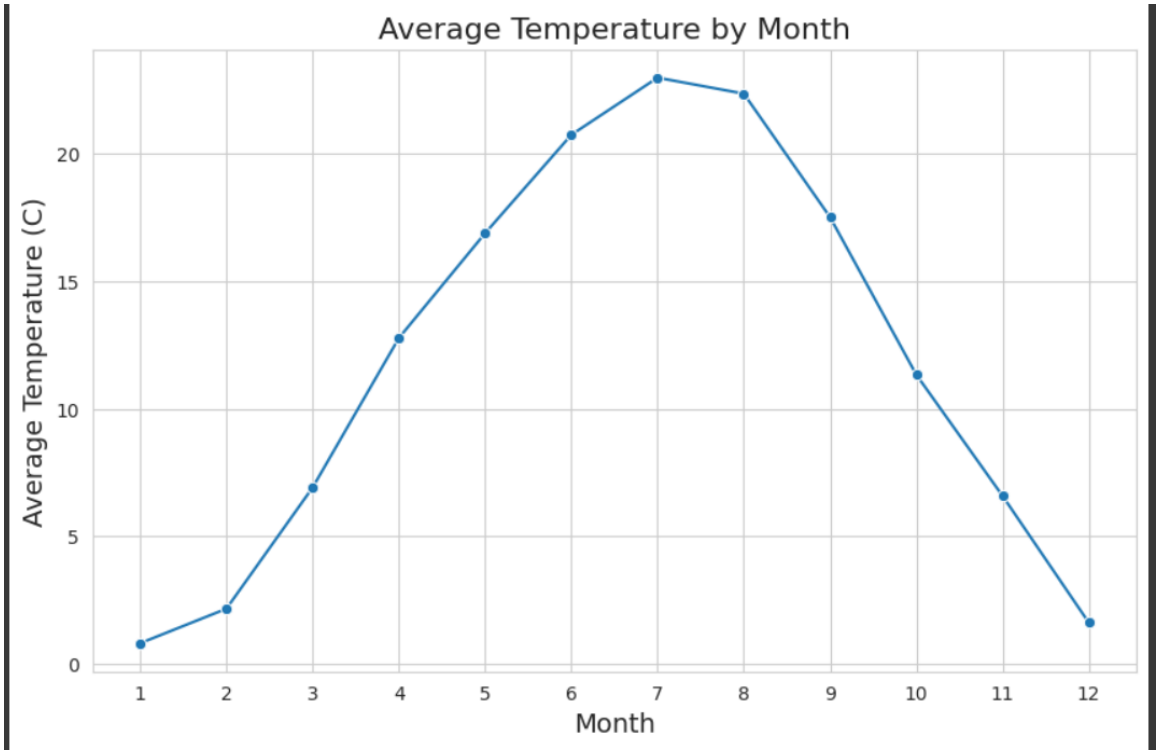
```
plt.show()

# 6. Wind Speed Distribution
plt.figure(figsize=(10, 6))
sns.histplot(data['Wind Speed (km/h)'], bins=50, kde=True, color='lightgreen')
plt.title('Wind Speed Distribution', fontsize=16)
plt.xlabel('Wind Speed (km/h)', fontsize=14)
plt.ylabel('Frequency', fontsize=14)
plt.show()

# 7. Humidity vs. Visibility
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Humidity', y='Visibility (km)', hue='Temperature (C)', data=data,
palette='cool', alpha=0.6)
plt.title('Humidity vs. Visibility', fontsize=16)
plt.xlabel('Humidity', fontsize=14)
plt.ylabel('Visibility (km)', fontsize=14)
plt.show()

# 8. Pressure Trend Over Time
plt.figure(figsize=(10, 6))
sns.lineplot(x='Formatted Date', y='Pressure (millibars)', data=data, color='purple')
plt.title('Pressure Trend Over Time', fontsize=16)
plt.xlabel('Date', fontsize=14)
plt.ylabel('Pressure (millibars)', fontsize=14)
plt.xticks(rotation=45)
```

plt.show()

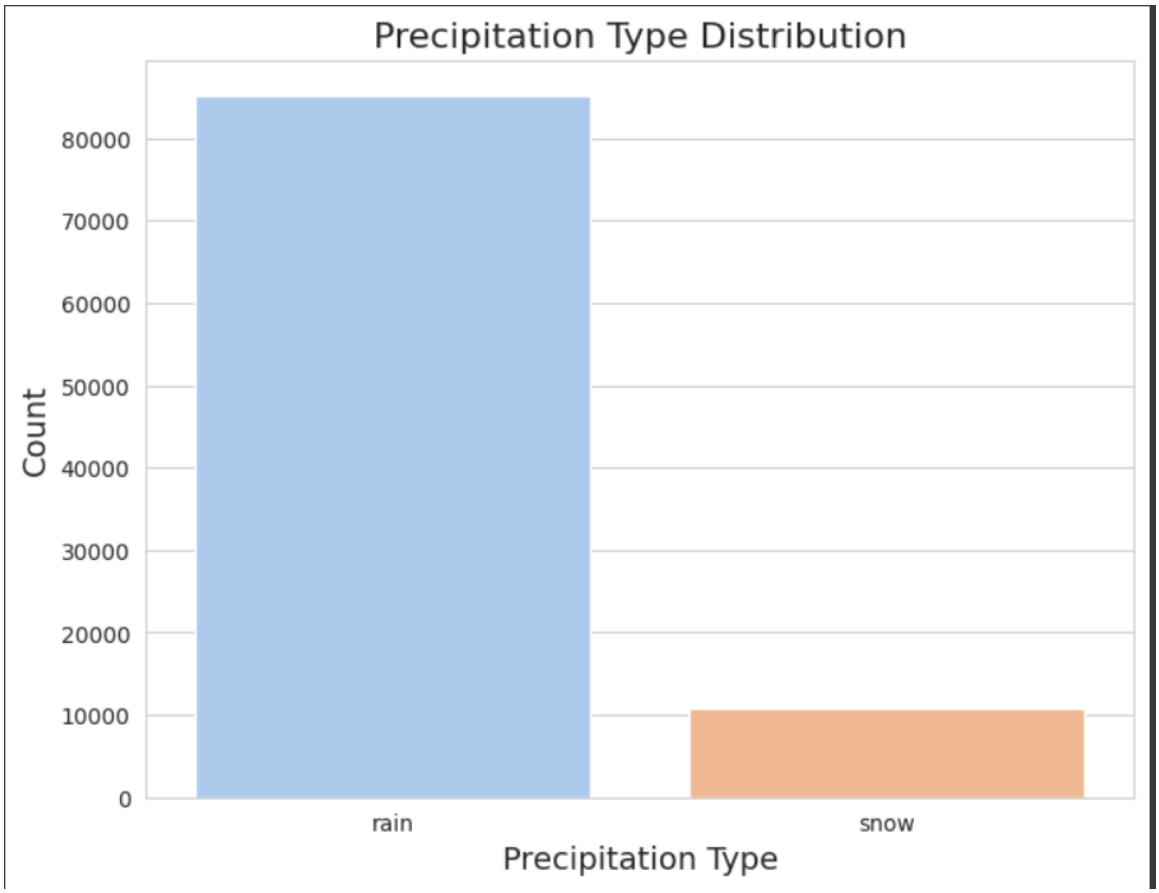
Output/Result

The analysis revealed:

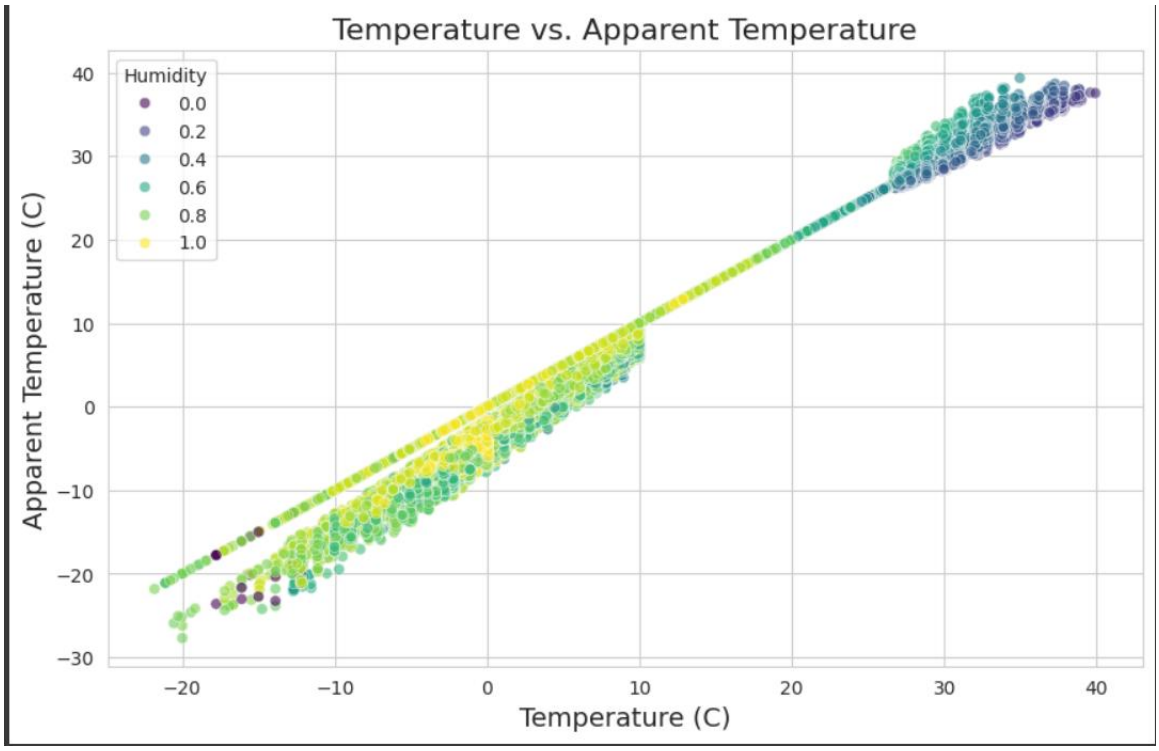
- Temperature Trends: Clear seasonal patterns in temperature.



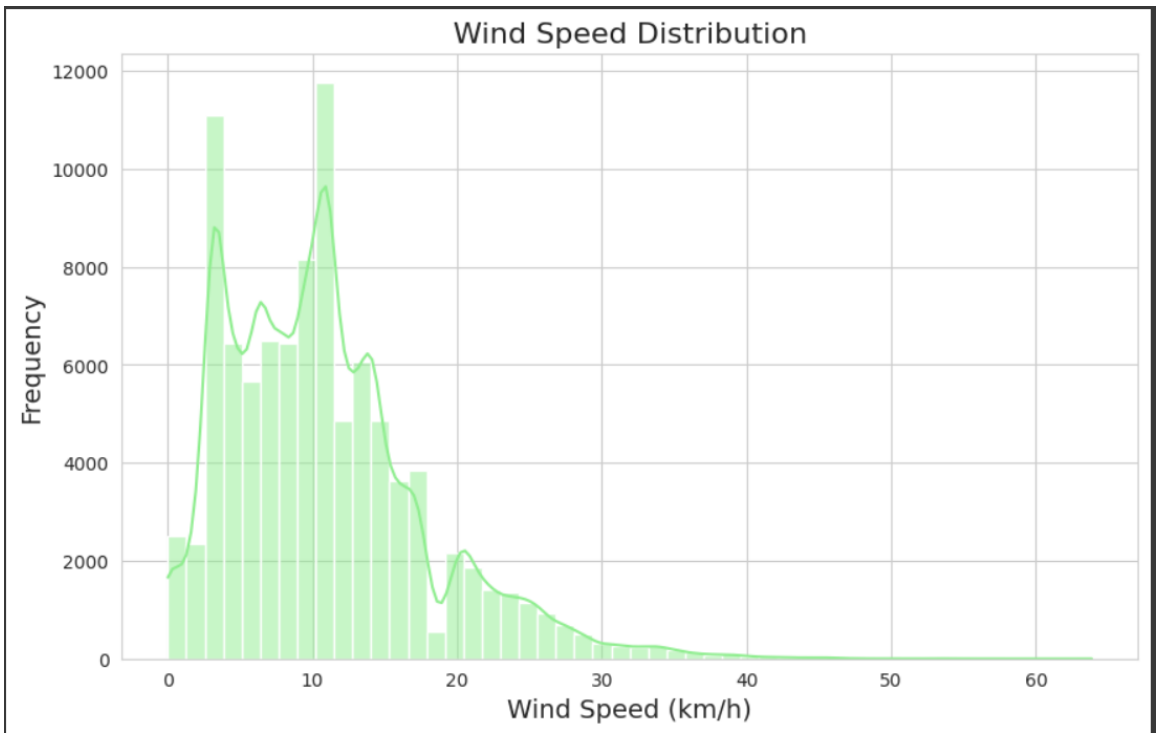
- Precipitation Types: 'Rain' was the most common precipitation type.



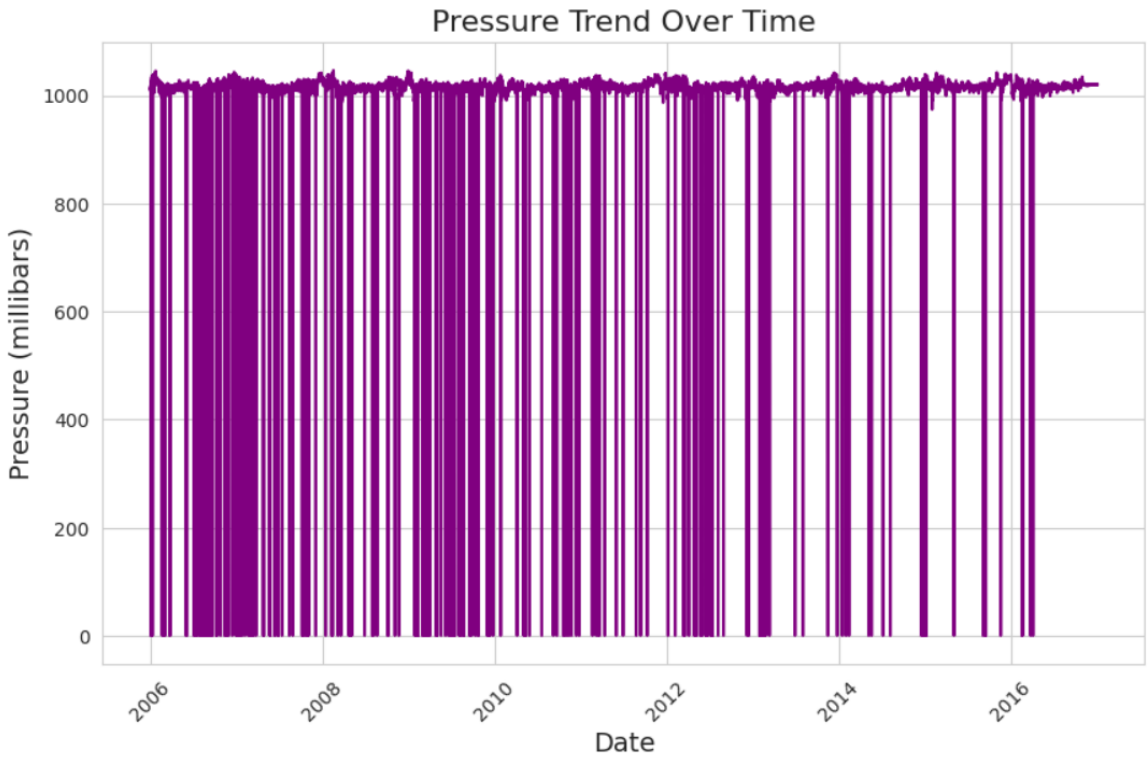
- Correlation Insights: High correlation between actual and apparent temperature.



- Wind Speed & Humidity: Provided distribution and relational insights.



- Pressure Trends: Displayed temporal variations in atmospheric pressure.



References/Credits

- Dataset: Provided as weatherHistory.csv

- Libraries: Pandas, Matplotlib, Seaborn

Submission

Files Uploaded to GitHub:

- .ipynb

- PDF Report

- README file with project details