**Weather Analysis Project Documentation**

**Overview**

This document outlines the data preparation, analysis, and insights derived from the weather dataset. The project aims to explore the relationships between weather conditions and the advisability of outdoor activities.

**Data Preparation**

**Data Loading**

The dataset was loaded using Pandas from a CSV file named weather_data.csv. The data consists of the following columns:

- outlook: Weather outlook (e.g., overcast, rainy)
- temperature: Temperature condition (e.g., hot, mild, cool)
- humidity: Humidity level (e.g., high, normal)
- windy: Whether it is windy (True/False)
- play: Whether it is advisable to play or not (yes/no)

**Initial Data Inspection**

The initial inspection revealed the following sample data:

| **outlook** | **temperature** | **humidity** | **windy** | **play** |
|---|---|---|---|---|
| overcast | hot | high | False | yes |
| overcast | cool | normal | True | yes |
| overcast | mild | high | True | yes |
| overcast | hot | normal | False | yes |
| rainy | mild | high | False | yes |

**Missing Values Check**

No missing values were found in the dataset:

```
Missing values before cleaning:
outlook        0
temperature    0
humidity       0
windy          0
play           0
dtype: int64
```

**Data Cleaning**

Although no missing values were found, a deprecation warning was encountered regarding the fillna method. The deprecated method was replaced with obj.ffill().

**Final Data Inspection**

The dataset remained unchanged after cleaning:

| **outlook** | **temperature** | **humidity** | **windy** | **play** |
|---|---|---|---|---|
| overcast | hot | high | False | yes |
| overcast | cool | normal | True | yes |
| overcast | mild | high | True | yes |
| overcast | hot | normal | False | yes |
| rainy | mild | high | False | yes |

**Advanced Analysis**

**Handling Outliers**

No numerical columns with significant outliers were found in the dataset. The Z-score approach would be useful if there were numerical columns with possible outlier values.

**Insights Derived**

- **Data Consistency**: The dataset does not have any missing values or obvious outliers.
- **Data Characteristics**:
  - The play column contains uniform values, indicating limited or biased data.
  - The outlook and temperature columns contain consistent values with no apparent variation.
- **Visualization**: A plot of temperature over time or other relevant visualizations could provide additional insights if the dataset had more varied data and a date column.

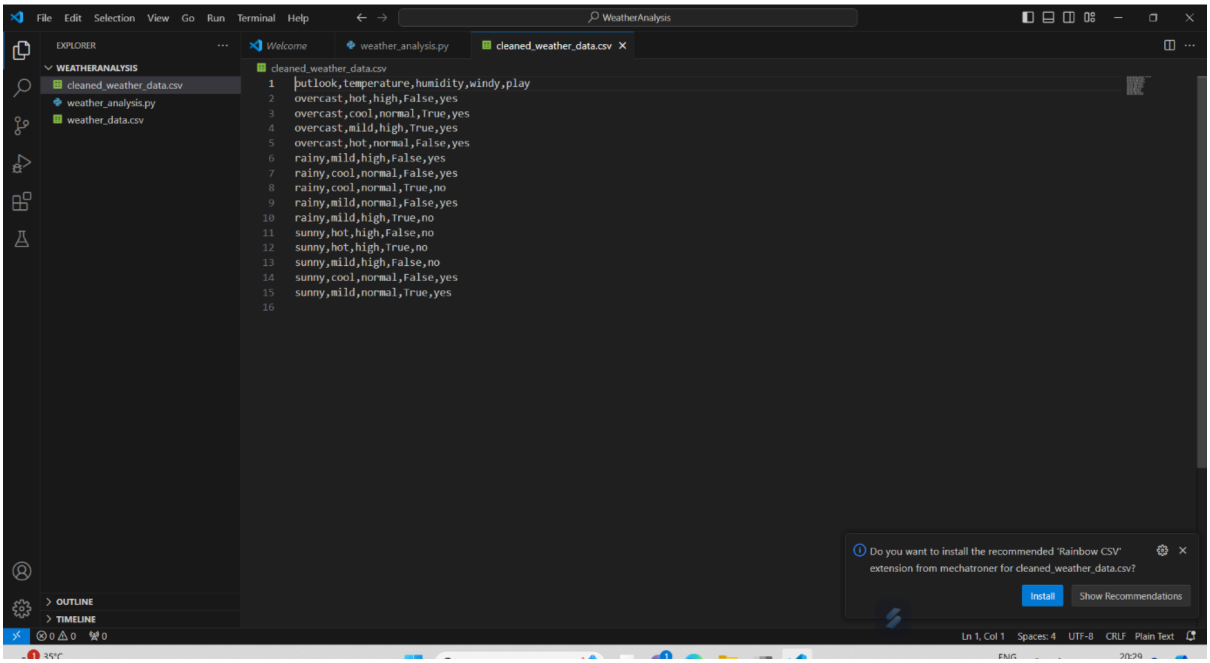
**Future Improvements**

- **Expand Dataset**: Include a wider range of observations to gain more insights.
- **Include More Features**: Add additional features or columns to enhance the analysis and allow for more robust conclusions.
- **Advanced Analytics**: Apply machine learning techniques to predict or classify data based on weather conditions.

**Conclusion**

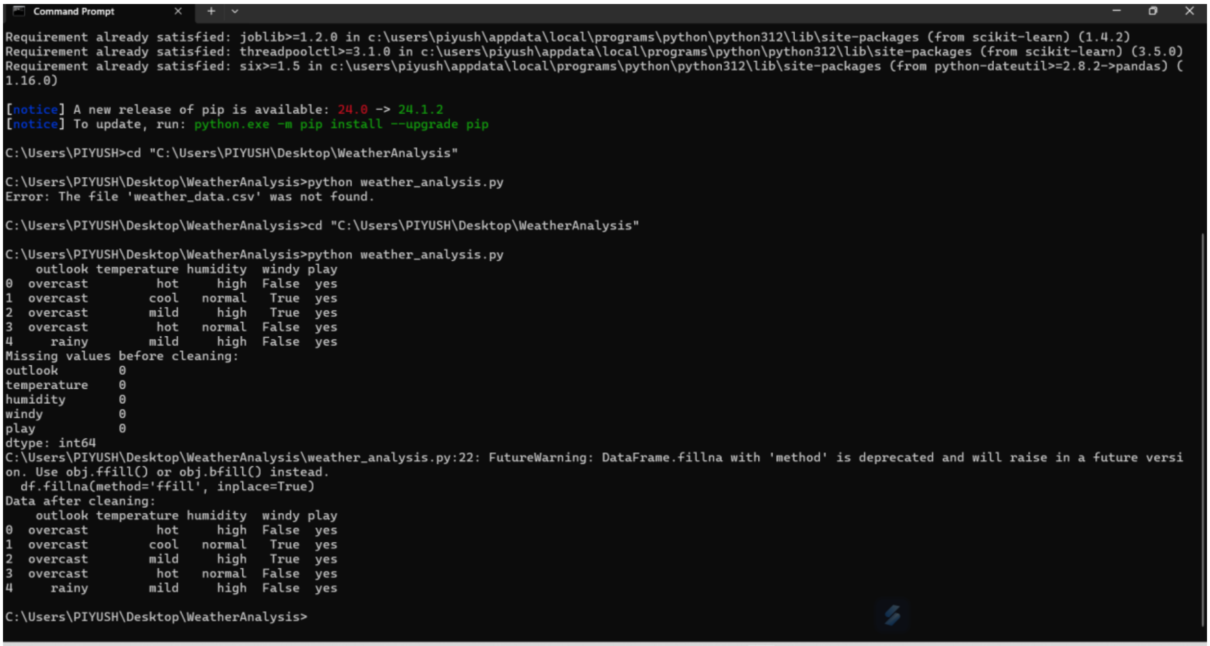
The dataset was successfully loaded, and preliminary data cleaning confirmed the absence of missing values. Basic inspections revealed consistent data, but lack of variability hindered deeper analysis. Future work could focus on expanding the dataset and applying more complex analytical techniques.

SCREENSHOTS-

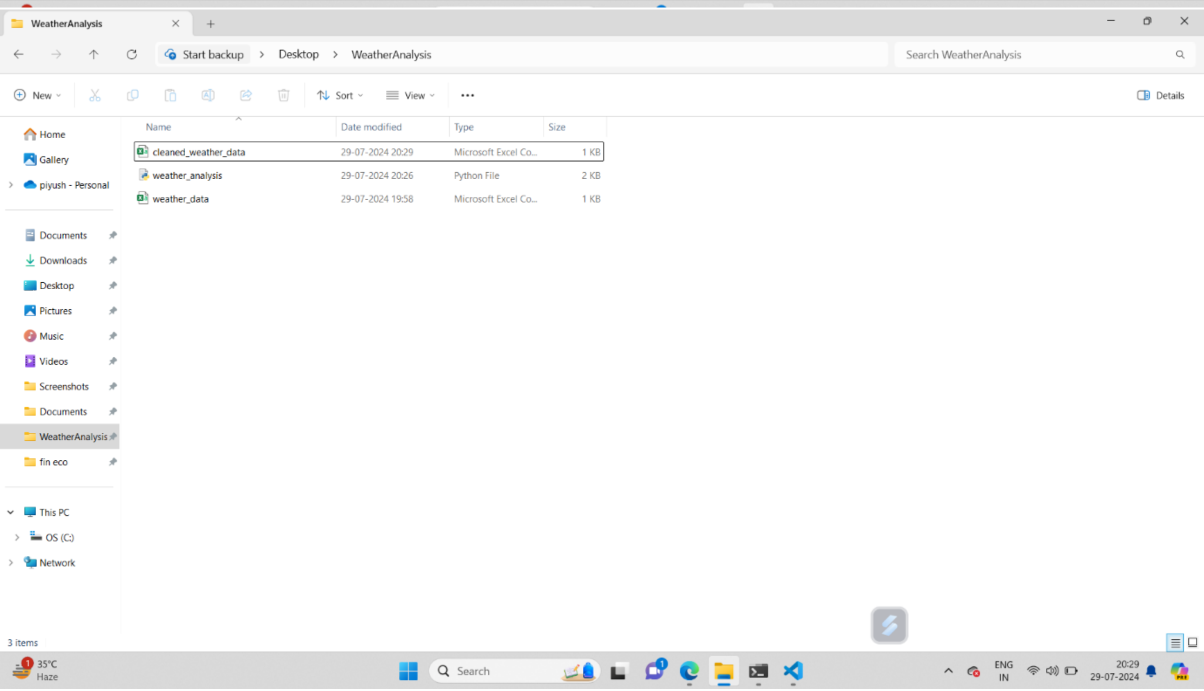


```
outlook,temperature,humidity,windy,play
overcast,hot,high,False,yes
overcast,cool,normal,True,yes
overcast,mild,high,True,yes
overcast,hot,normal,False,yes
rainy,mild,high,False,yes
rainy,cool,normal,False,yes
rainy,cool,normal,True,no
rainy,mild,normal,False,yes
rainy,mild,high,True,no
sunny,hot,high,False,no
sunny,hot,high,True,no
sunny,mild,high,False,no
sunny,cool,normal,False,yes
sunny,mild,normal,True,yes
```

Do you want to install the recommended 'Rainbow CSV' extension from mechatroner for cleaned_weather_data.csv?

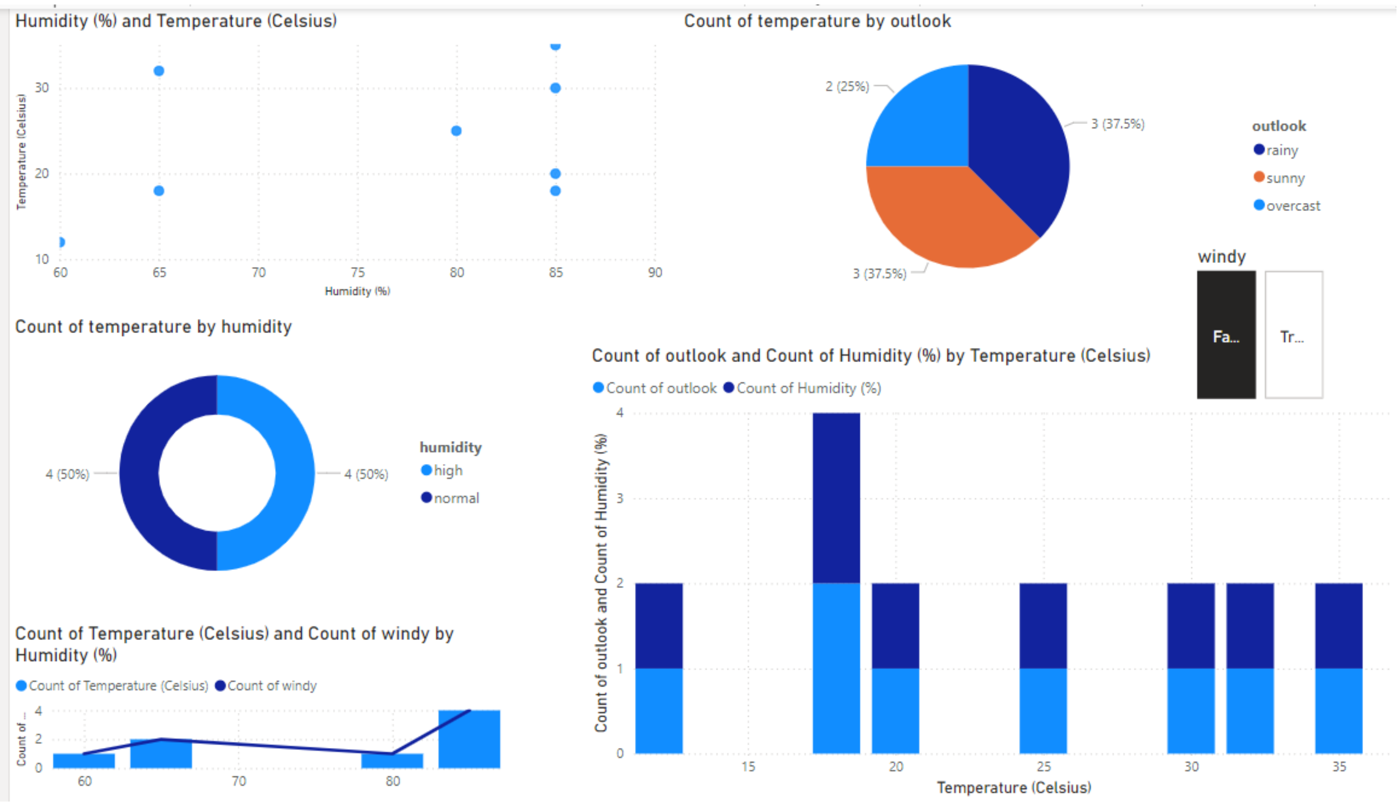


```
Requirement already satisfied: joblib>=1.2.0 in c:\users\piyush\appdata\local\programs\python\python312\lib\site-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\piyush\appdata\local\programs\python\python312\lib\site-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: six>=1.5 in c:\users\piyush\appdata\local\programs\python\python312\lib\site-packages (from python-dateutil>=2.8.2->pandas) (
1.16.0)

[notice] A new release of pip is available: 24.0 -> 24.1.2
[notice] To update, run: python.exe -m pip install --upgrade pip

C:\Users\PIYUSH>cd "C:\Users\PIYUSH\Desktop\WeatherAnalysis"

C:\Users\PIYUSH\Desktop\WeatherAnalysis>python weather_analysis.py
Error: The file 'weather_data.csv' was not found.

C:\Users\PIYUSH\Desktop\WeatherAnalysis>cd "C:\Users\PIYUSH\Desktop\WeatherAnalysis"

C:\Users\PIYUSH\Desktop\WeatherAnalysis>python weather_analysis.py
    outlook temperature humidity  windy play
0  overcast         hot     high  False  yes
1  overcast        cool   normal   True  yes
2  overcast        mild     high   True  yes
3  overcast         hot   normal  False  yes
4     rainy        mild     high  False  yes
Missing values before cleaning:
outlook        0
temperature    0
humidity       0
windy          0
play           0
dtype: int64
C:\Users\PIYUSH\Desktop\WeatherAnalysis\weather_analysis.py:22: FutureWarning: DataFrame.fillna with 'method' is deprecated and will raise in a future versi
on. Use obj.ffill() or obj.bfill() instead.
  df.fillna(method='ffill', inplace=True)
Data after cleaning:
    outlook temperature humidity  windy play
0  overcast         hot     high  False  yes
1  overcast        cool   normal   True  yes
2  overcast        mild     high   True  yes
3  overcast         hot   normal  False  yes
4     rainy        mild     high  False  yes

C:\Users\PIYUSH\Desktop\WeatherAnalysis>
```



| Name | Date modified | Type | Size |
|---|---|---|---|
| cleaned_weather_data | 29-07-2024 20:29 | Microsoft Excel Co... | 1 KB |
| weather_analysis | 29-07-2024 20:26 | Python File | 2 KB |
| weather_data | 29-07-2024 19:58 | Microsoft Excel Co... | 1 KB |

Power bi


Humidity (%) and Temperature (Celsius)
Temperature (Celsius)
30
20
10
60
65
70
75
80
85
90
Humidity (%)
Count of temperature by outlook
2 (25%)
3 (37.5%)
3 (37.5%)
outlook
● rainy
● sunny
● overcast
windy
Fa...
Tr...
Count of temperature by humidity
4 (50%)
4 (50%)
humidity
● high
● normal
Count of outlook and Count of Humidity (%) by Temperature (Celsius)
● Count of outlook ● Count of Humidity (%)
Count of outlook and Count of Humidity (%)
4
3
2
1
0
15
20
25
30
35
Temperature (Celsius)
Count of Temperature (Celsius) and Count of windy by Humidity (%)
● Count of Temperature (Celsius) ● Count of windy
Count of ...
4
2
0
60
70
80

### Visualizations Overview

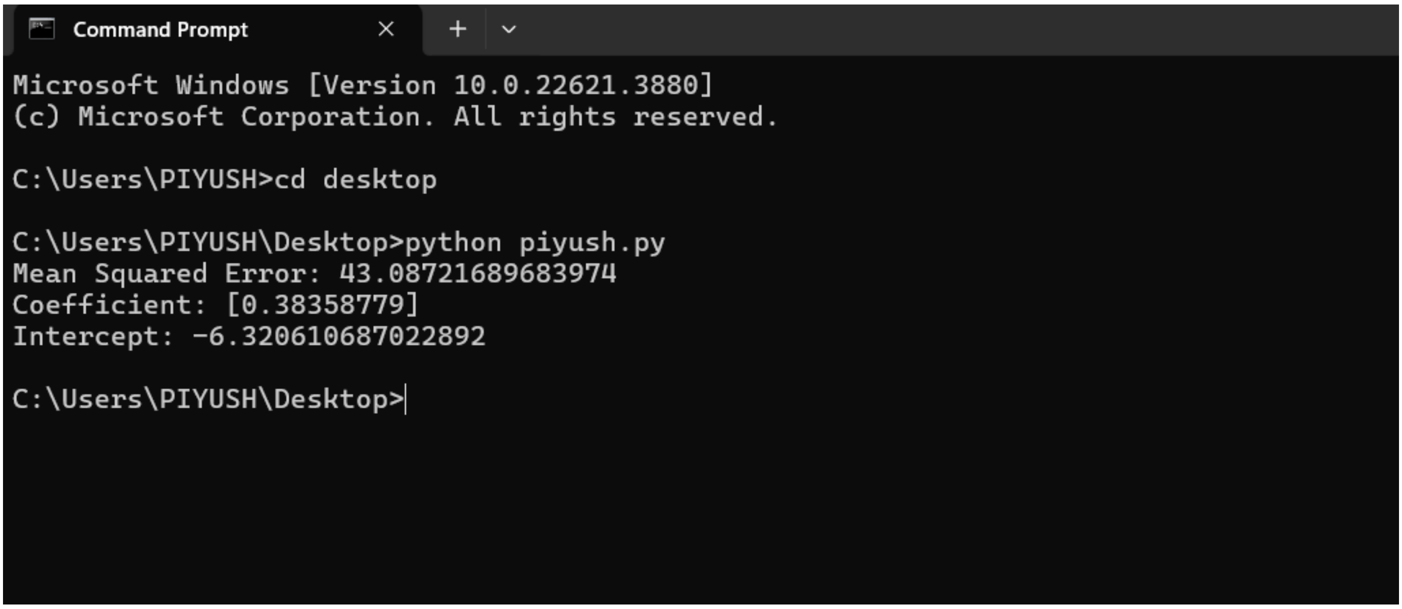
The dashboard presents a basic overview of the relationship between humidity, temperature, outlook, and windy conditions. The visualizations used are:

- **Scatter Plot:** Visualizes the relationship between humidity and temperature.
- **Pie Chart:** Shows the distribution of temperature across different outlooks.
- **Pie Chart:** Displays the distribution of humidity levels.
- **Stacked Column Chart:** Compares the count of outlook and humidity by temperature.
- **Line Chart:** Attempts to show the relationship between temperature and windy conditions, but the visualization is unclear.

### Data Insights

Based on the provided visualizations, we can infer the following:

1. **Positive Correlation:** There seems to be a positive correlation between humidity and temperature, as indicated by the scatter plot.
2. **Temperature Distribution:** The pie chart shows a relatively even distribution of temperature across different outlooks.
3. **Humidity Distribution:** The pie chart indicates a similar proportion of high and normal humidity levels.
4. **Outlook and Humidity by Temperature:** The stacked column chart provides a basic overview of how outlook and humidity vary with temperature. However, the visualization is complex and could be improved for better clarity.
5. **Windy and Temperature:** The line chart is not effective in conveying the relationship between temperature and windy conditions. A different visualization might be more suitable.



```
Microsoft Windows [Version 10.0.22621.3880]
(c) Microsoft Corporation. All rights reserved.

C:\Users\PIYUSH>cd desktop

C:\Users\PIYUSH\Desktop>python piyush.py
Mean Squared Error: 43.08721689683974
Coefficient: [0.38358779]
Intercept: -6.320610687022892

C:\Users\PIYUSH\Desktop>
```

 Mean Squared Error: 43.08721689683974 indicates the calculated mean squared error of the regression model. This metric measures the average squared difference between the predicted and actual values. A lower value generally indicates a better fit of the model to the data.

Coefficient: [0.38358779] represents the coefficient of the humidity variable in the linear regression equation. This value indicates the change in temperature for a one-unit increase in humidity.

Intercept: -6.320610687022892 is the intercept of the regression line. It represents the predicted temperature when humidity is zero.

**Overall:** The analysis provides a preliminary understanding of the relationship between humidity and temperature. A linear regression model was used to predict temperature based on humidity. The model's performance, as measured by MSE, needs further evaluation in comparison to other models or benchmarks. The coefficient for humidity indicates a positive relationship between the two variables.

**Limitations and Next Steps**

- The analysis is based on a limited dataset and a simple linear regression model. More complex models or additional features might improve predictive accuracy.
- The interpretation of the intercept should be considered carefully in the context of the data.
- Further analysis could include exploring other factors influencing temperature, evaluating different regression models, and assessing the model's performance on a larger dataset.

**In conclusion,** the initial analysis provides a foundation for understanding the relationship between humidity and temperature. However, further investigation is required to draw more definitive conclusions and improve the predictive capabilities of the model.

……………..end of report…………………….