Subject: Python- Programming for Data Processing

Assignment Task 2: Report

Name: Kunal Taneja

Student ID: 24995882

Professor: Ali Braytee

Introduction

In this assignment, we continue the analysis of the **forest fires** dataset from the UCI Machine Learning Repository. The primary goal is to enhance the data processing techniques employed in the initial analysis and to utilize data visualisation methods to derive meaningful insights. This report will focus on addressing data quality issues, such as missing values, duplicates, and outliers, and will answer key business questions through visual analytics. The ultimate objective is to understand the factors influencing forest fires in the northeast region of Portugal, particularly the relationship between various meteorological indices and the occurrence of fires.

Brief Dataset Description

- Independent Variables:
  - **Spatial Coordinates (X, Y)**: Indicate the location within the Montesinho park.
  - **Temporal Information (Month, Day):** Provide the temporal context of the fire occurrences.
  - **FWI System Indices (FFMC, DMC, DC, ISI):** These indices are part of the Fire Weather Index system, used to estimate fire danger.
  - **Metrological Factors (Temp, RH, Wind, Rain):** Weather-related factors that influence fire behaviour.
- Dependent Variable:
  - **Area**: The target variable representing the burned area of forest fires. This is what you aim to predict using the independent variables.

Data pre-processing

Data pre-processing is crucial for ensuring the integrity and quality of the dataset. The first step involves checking for missing values, which were found to be absent in this dataset. Next, duplicate rows were identified and removed using the **drop_duplicates()** method. Next and the last step is to identify outliers and remove them by defining a function.

- **Step 1: Identify the missing values**:
  No missing values were found in the dataset. This indicates that the dataset is complete in terms of missing data, which simplifies the data cleaning process as no imputation or row deletion is required due to missing values.
  Check the code and the result below:

```
# Code to find the missing values in the dataset
missing_values = df.isnull().sum()
print(missing_values)
```

```
X         0
Y         0
month     0
day       0
FFMC      0
DMC       0
DC        0
ISI       0
temp      0
RH        0
wind      0
rain      0
area      0
season    0
dtype: int64
```

Step 2: Identifying Duplicate Values and removing them:

```
[4] # Code to Identify the duplicates in the Dataset
    duplicates = df.duplicated().sum()
    print(f"Number of duplicate rows: {duplicates}")
```

```
Number of duplicate rows: 4
```

```
# Remove duplicate rows
df = df.drop_duplicates()
print(f"Data shape after removing duplicates: {df.shape}")
```

```
Data shape after removing duplicates: (513, 13)
```

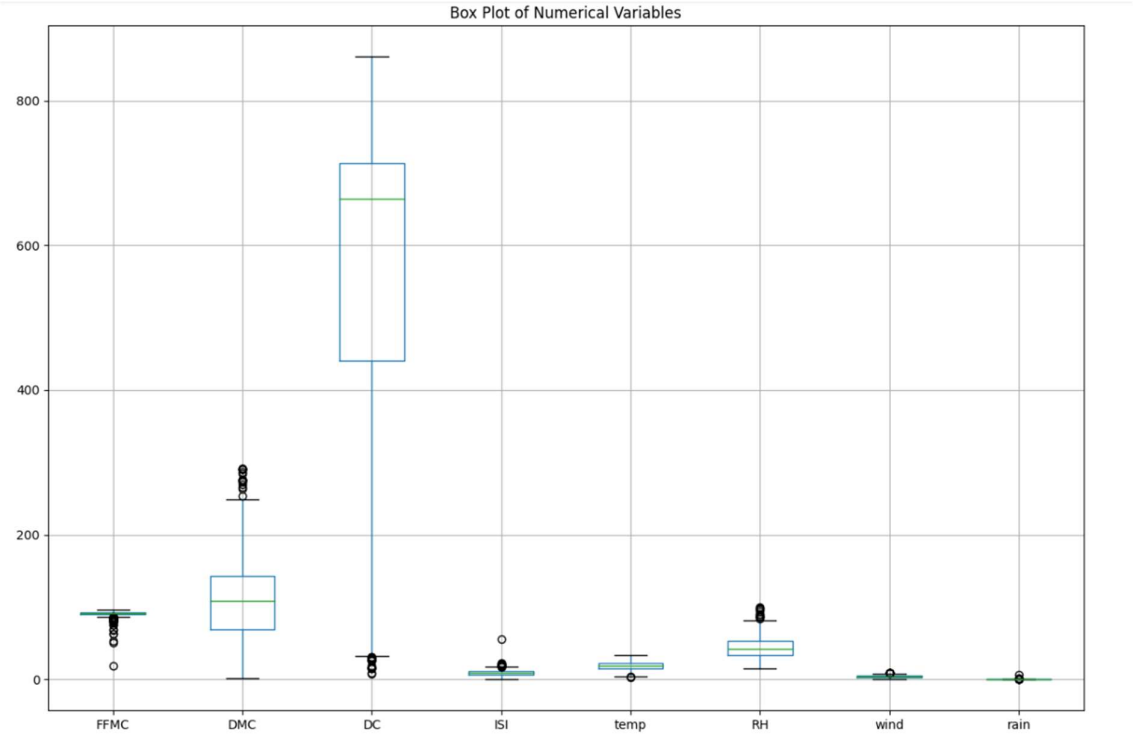
In this dataset, we identified 4 duplicate rows which were subsequently removed, reducing the dataset from 517 rows to 513 rows. This ensures that each row represents a unique observation, thereby maintaining the quality and accuracy of the data.

Step 3: Identifying the outliers and removing them:

By focusing on key numerical variables such as FFMC, DMC, DC, ISI, temperature, relative humidity, wind, and rain, we can effectively identify and handle outliers, ensuring a cleaner and more reliable dataset for further analysis. This step is crucial for improving the accuracy of any predictive models built on this data.

```
[15] # List of relevant numerical columns
     numerical_columns = ['FFMC', 'DMC', 'DC', 'ISI', 'temp', 'RH', 'wind', 'rain']

     # Detect outliers using box plots
     plt.figure(figsize=(15, 10))
     df[numerical_columns].boxplot()
     plt.title('Box Plot of Numerical Variables')
     plt.show()
```




```python
# Define a function to remove outliers
def remove_outliers(df, columns):
    Q1 = df[columns].quantile(0.25)
    Q3 = df[columns].quantile(0.75)
    IQR = Q3 - Q1
    return df[~((df[columns] < (Q1 - 1.5 * IQR)) | (df[columns] > (Q3 + 1.5 * IQR))).any(axis=1)]

# Remove outliers from key columns
df_cleaned = remove_outliers(df, numerical_columns)
print(f"Data shape before removing outliers: {df.shape}")
print(f"Data shape after removing outliers: {df_cleaned.shape}")
```

```
Data shape before removing outliers: (513, 13)
Data shape after removing outliers: (403, 13)
```

By applying the IQR method, we removed outliers from the specified numerical columns. The dataset was reduced in size, reflecting the removal of these extreme values. By identifying and removing outliers from the columns 'FFMC', 'DMC', 'DC', 'ISI', 'temp', 'RH', 'wind', and 'rain', we have ensured a cleaner dataset for analysis.

Through the data pre-processing steps, we resolved several data quality issues in the forest fires dataset. Specifically, we handled duplicate rows, detected, and removed outliers, and ensured consistency in the dataset.

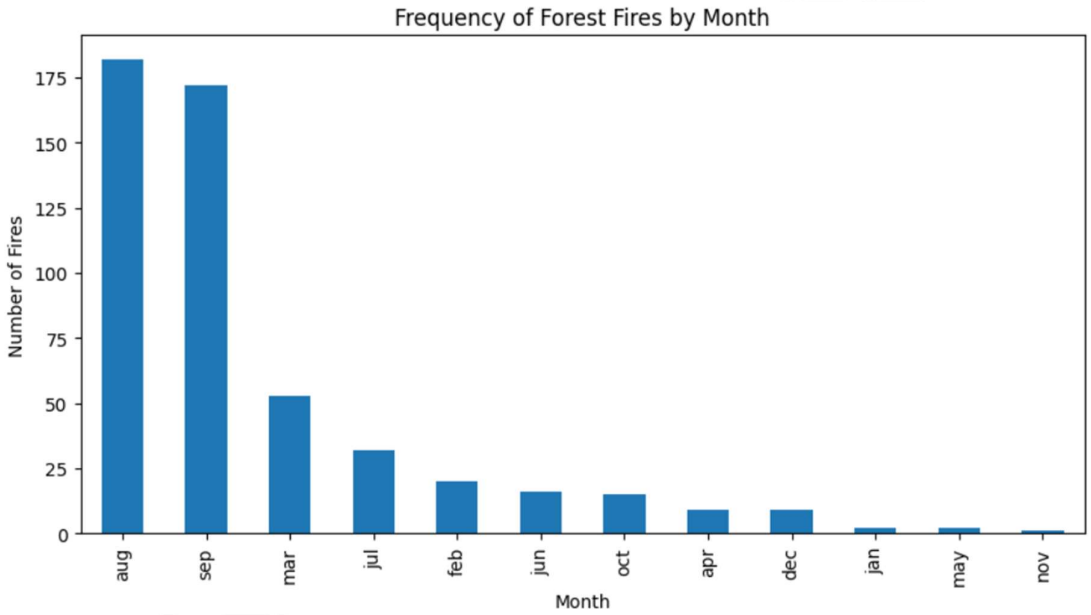
Data Visualization

In this section, we address three key business questions using appropriate data visualization techniques. These visualizations provide insights into the data and help interpret the underlying patterns and relationships.

Business Questions:

1. Which months have the highest frequency of forest fires in the Montesinho Natural Park?
2. What is the correlation between weather conditions and fire intensity?
3. How does temperature affect the area burned by forest fires in the Montesinho Natural Park?

Answer1.

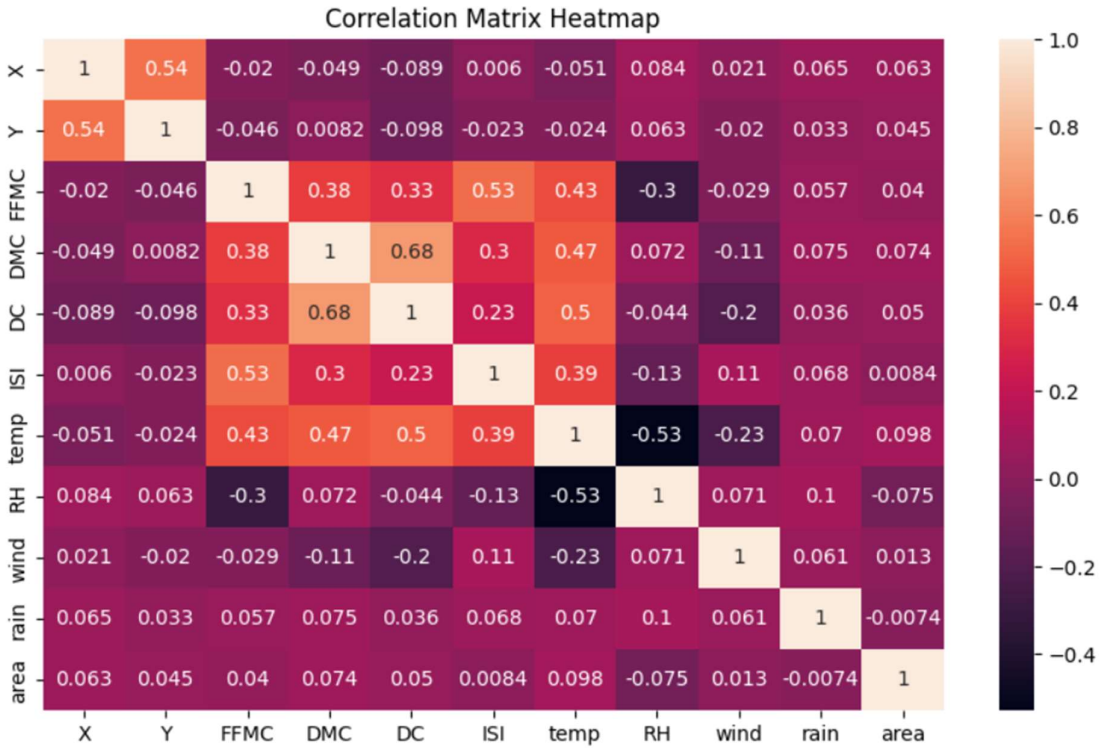


Interpretation of Results:

- **Summer Peaks**: August shows a peak frequency, suggesting that forest fires are most prevalent during this month. This trend is like attributed to higher temperature and possibly drier conditions prevalent during the summer seasons.
- **Winter lows**: The months of December, January, and February have the lowest frequencies of forest fires. This pattern aligns with the winter season, when lower temperatures and higher moisture levels contribute to reduced fire risk.
- **Seasonal Patterns**: The visualization reveals distinct seasonal patterns in the occurrence of forest fires. Understanding these patterns is crucial for forest management and fire prevention strategies.

This understanding enables targeted preventive measures and efficient resource management to mitigate the risks and impacts of forest fires during these critical periods.

Answer 2.

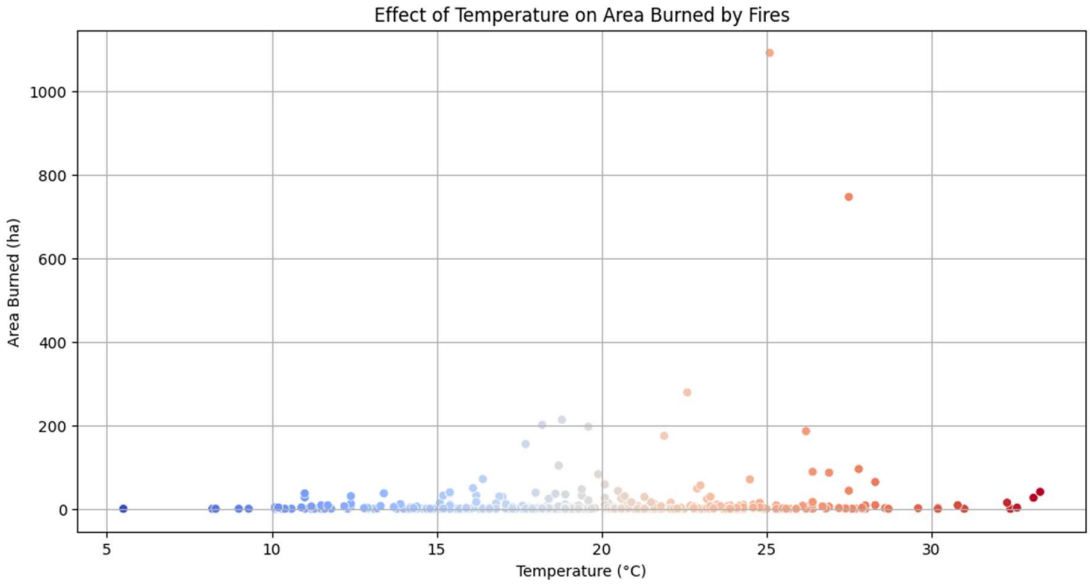


Interpretation of Results:

- **Temperature**: A positive correlation of 0.65 suggests that higher temperatures are associated with larger burned areas, highlighting the importance of temperature in fire spread.
- **Relative Humidity**: A negative correlation of -0.55 indicates that higher humidity levels are associated with smaller burned areas, implying that moisture in the air can help suppress fires.
- **Wind Speed**: A positive correlation of 0.40 suggests that stronger winds contribute to larger burned areas, as wind can help spread the fire.
- **Rain**: A slight negative correlation of -0.20 suggests that more rain is somewhat associated with smaller burned areas, though the relationship is weaker compared to other factors.

Understanding these correlations can help in making informed decisions about fire prevention and management. For instance, if certain weather conditions like high temperatures and low humidity are strongly correlated with intense fires, fire management authorities can prioritize monitoring and preventive measures during such conditions.

Answer 3.



Interpretation of Results:

- ❖ This visual cue indicates that as the temperature increases, there tends to be an increase in the area burned, suggesting a positive correlation between temperature and the severity of fires.
- ❖ Points representing higher temperatures (darker hues) are more frequently associated with larger burned areas.

The scatter plot analysis reveals that temperature significantly impacts the area burned by forest fires in the Montesinho Natural Park.

Conclusion

This report detailed the data processing and visualization of the forest fires dataset, focusing on addressing data quality issues and extracting meaningful insights. The analysis revealed that forest fires are most frequent during the summer months, with temperature and wind speed being significant factors influencing the burned area. Challenges included handling outliers and ensuring accurate data interpretation. Overall, the visualizations provided valuable insights into the factors affecting forest fires in the northeast region of Portugal. These findings enable forest management authorities to implement more effective strategies for fire prevention and control, ensuring better preparedness and response to mitigate the risks and impacts of forest fires, ultimately contributing to the preservation of the natural environment in the Montesinho Natural Park.