Software Packages Project

Student: Matei Ana Maria Valentina

Group: 1104

Introduction

The project has as starting point the dataset of the environmental conditions relevant to forest fires in a national park in Portugal. The variables collected are:

The month and day in which the fires happened.

The FFMC represents the moisture content of litter and other cured fine fuels in a forest stand

The Duff Moisture Code (DMC) represents fuel moisture of decomposed organic material underneath the litter

The Drought code is an indicator of the moisture content in deep compact organic layers. This code represents a fuel layer at approximately 10-20 cm deep. The Drought code fuels have a very slow drying rate, with a time lag of 52 days

The Initial Spread Index (ISI) is a numeric rating of the expected rate of fire spread. It is based on wind speed and FFMC.

The temperature in Celsius registered in the day of the fire.

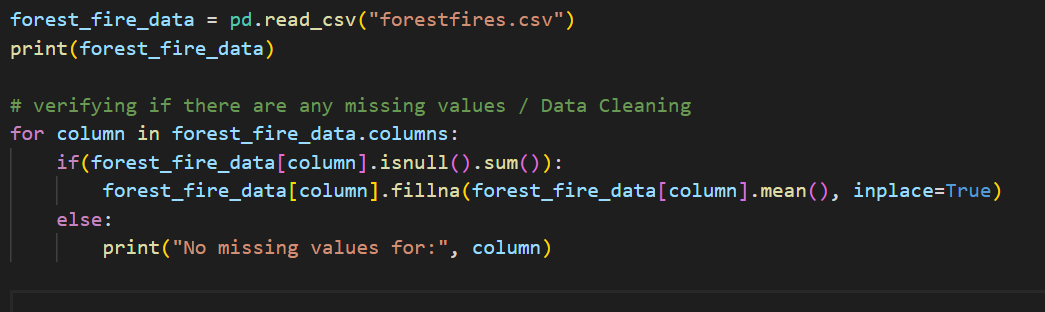
Relative humidity (RH) is the ratio of the amount of moisture in the air to the amount of moisture necessary to saturate the air at the same temperature and pressure. Relative humidity is expressed in percent

The wind speed, the amount of rain from that day and the area which was engulfed by flames.

The aim of the project is to analyze the data collected, to present it in a way that the average people can understand it, to highlight some interesting observations that could be done on the dataset and to explore the correlation between the variables.

Problem 1

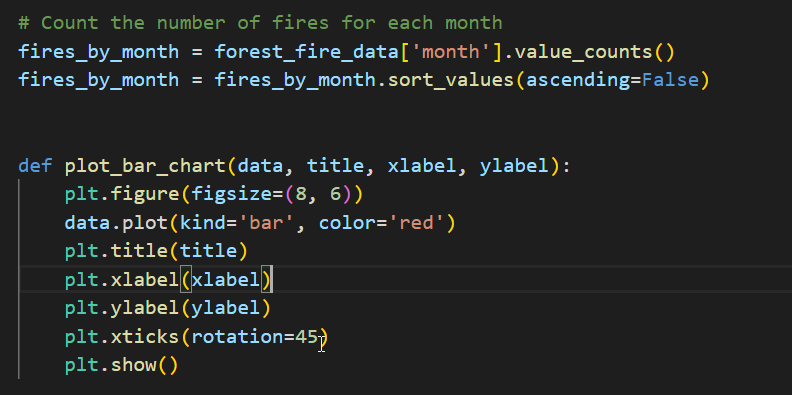
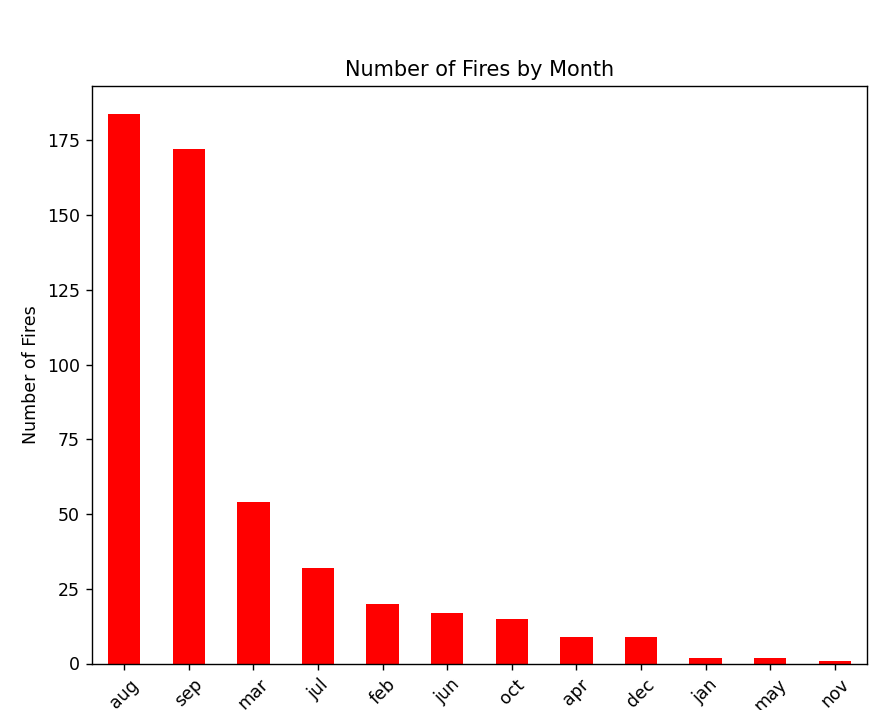
The first action was to extract the dataset into a pandas Data Frame so it would be easy to manipulate the data. After that, to ensure that the processing of the data will be done properly, there was a check for missing values and they were replaced.

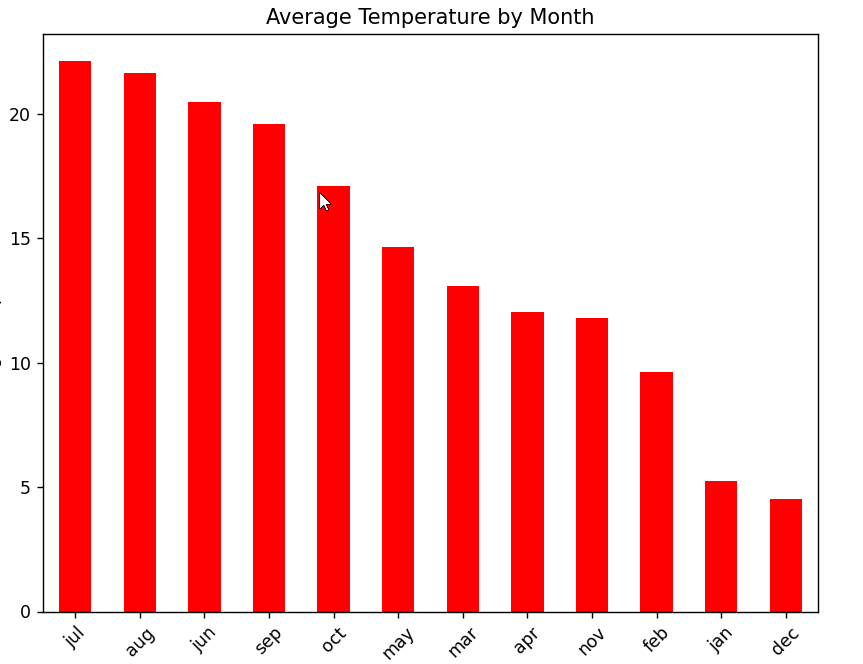


Problem 2

The second problem consists of visualizing the distribution of forest fires by month to understand the frequency of fire occurrence throughout the year. It was used a bar chart that was created with the help of the matplotlib package to visualize the data, also the data was sorted and a function was created to aid in the future creation of other bar charts in the project.

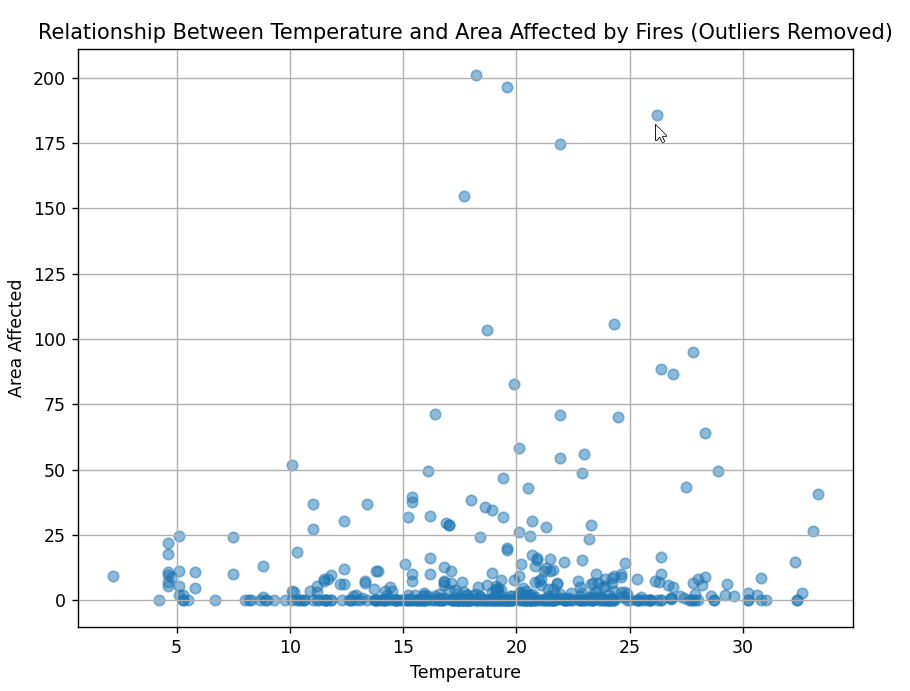
It is clearly visible that august and September are the months with the most fires.

The average temperature for each month was also plotted so we can observe more clearly the environmental conditions of each month. And as we can see, in the month prior (June, July) to the ones that have the most forest fires, the temperatures were significantly hotter, leading to a drier environment.



Problem 3

The next problem was to investigate the relationship between temperature and the area affected by fires, while identifying and handling outliers to ensure accurate correlation analysis. The standardization was made using StandardScaler, from the packet sklearn to prepare the data for outlier detection and correlation analysis. The outliers were identified by using z-score and were removed from the dataset and then the calculation for the correlation coefficient between temperature and area affected by fires was made, all these were prepared using the pandas package. Finally, the plot was made with matplotlib.

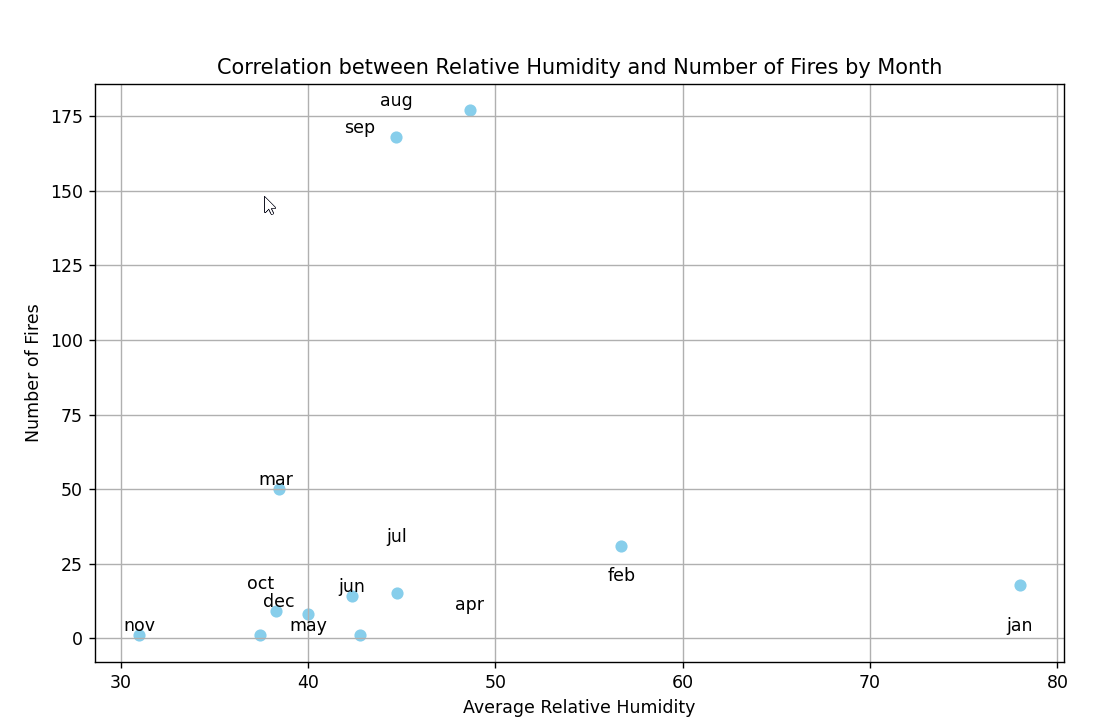


The correlation coefficient is 0.0697 and suggests a weak positive correlation between temperature and the area affected by fires after removing outliers.

While the correlation is positive, indicating that higher temperatures tend to be associated with slightly larger fire areas, the strength of the relationship is relatively low.

This finding suggests that temperature alone may not be a strong predictor of fire severity, and other factors such as humidity, wind speed, and fuel moisture content may play significant roles.

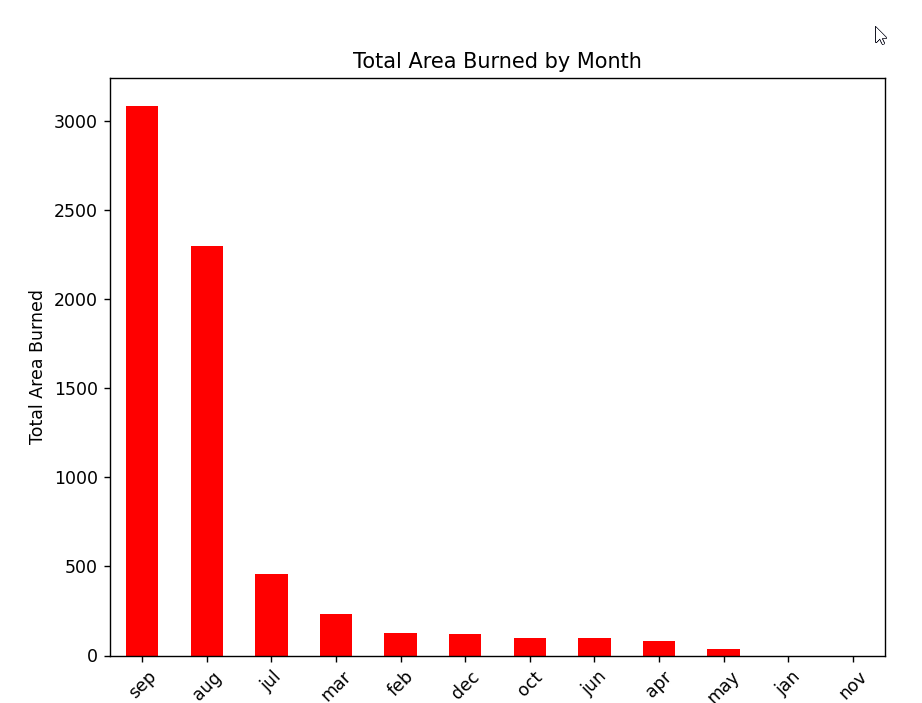
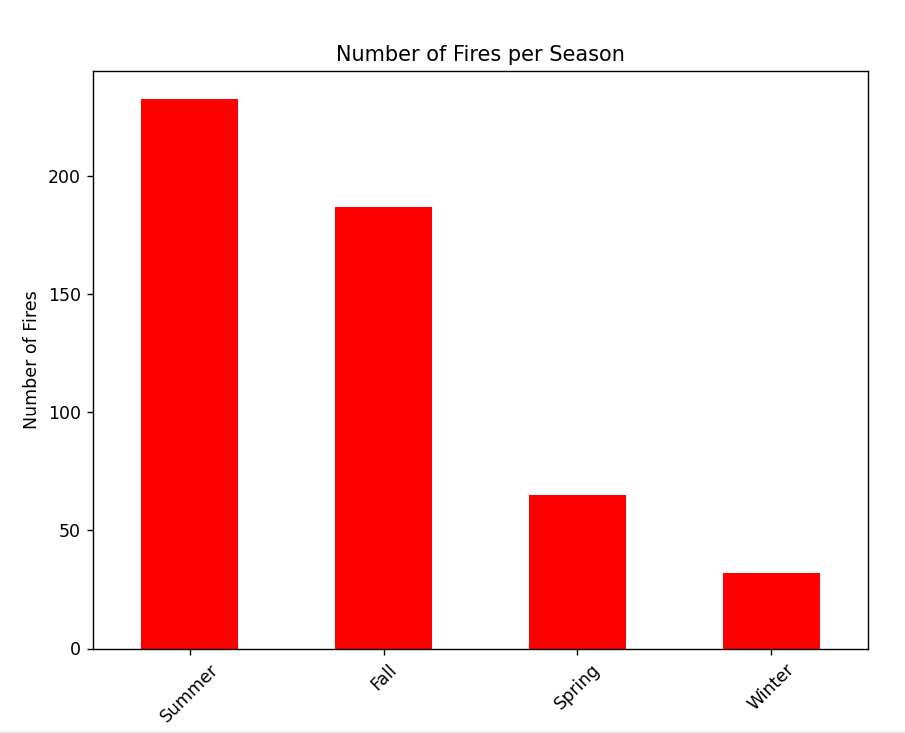
The correlation between the humidity and the number of fires was also made and plotted

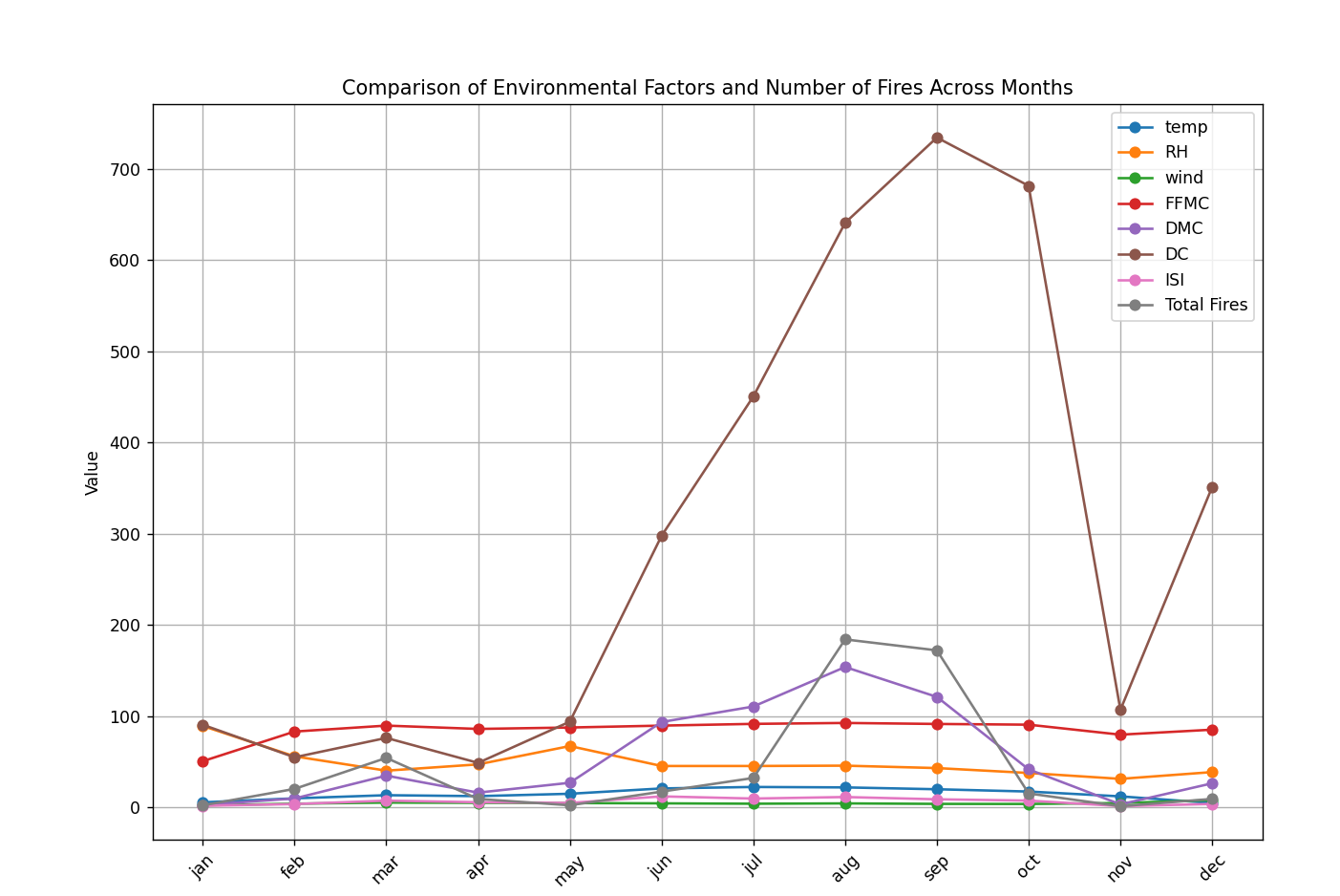


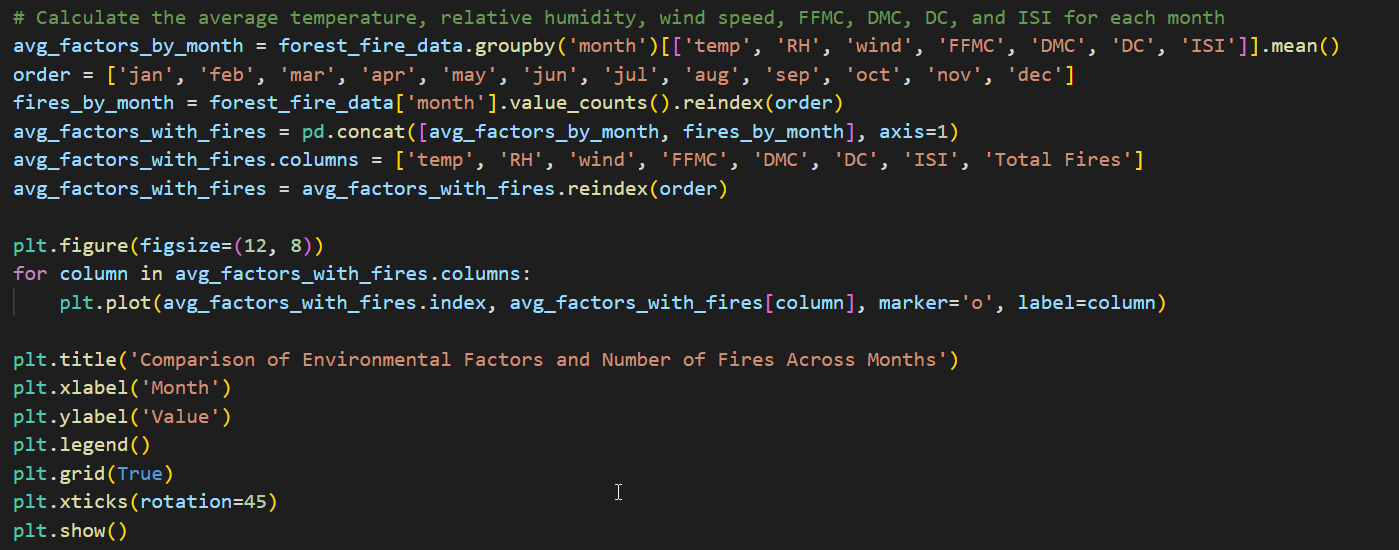
The correlation coefficient is -0.1015, so it indicates a weak negative correlation between relative humidity and the number of fires by month. While higher relative humidity levels may be associated with slightly fewer fires, the correlation is weak, suggesting that other factors also influence fire occurrence.

Problem 4

More manipulation of the data was made so we can display and understand it better, the majority was made through the pandas and the matplotlib packages. There were plots made for the area burned each month, the number of fires in each season and a representation of the average of every variable through the months.







There can be seen an increase of the duff fuel moisture and the drought code while the fire increased in number, so we explored the correlation between the 2.

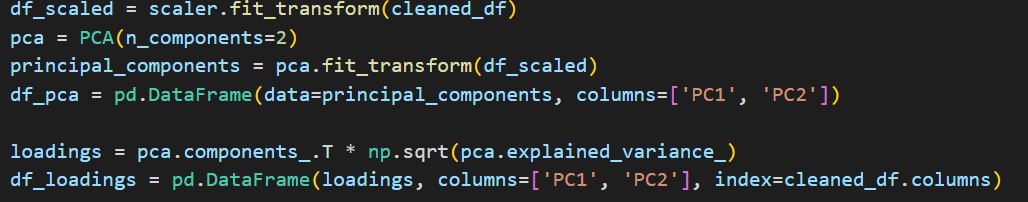
The positive value (0.073) indicates a weak positive correlation between DMC and the number of fires. This suggests that there is a slight tendency for the number of fires to increase as the DMC increases, but the relationship is not very strong.

Problem 5

Now that we have some understanding of the variables we will compute the Principal Component Analysis to reduce its dimensionality and visualize the relationships between variables.

PCA is applied to the dataset after standardizing the features. The PCA algorithm calculates the principal components, which are linear combinations of the original variables. Loadings are computed to understand the contribution of each original variable to the principal components.

The analysis is done using sklearn for PCA, pandas for manipulating the dataset and matplotlib for representation.



PC1 appears to represent a combination of weather and environmental factors that influence fire behavior.

Variables with high negative loadings on PC1, such as FFMC, DMC, DC, ISI, temp, and rain, suggest that lower values of these variables contribute more to PC1. This implies that conditions like lower fuel moisture, temperature, and rainfall, along with higher fire danger indices, are associated with higher PC1 scores. Conversely, variables with positive loadings on PC1, namely RH and wind, indicate that higher relative humidity and wind speed contribute positively to PC1 scores.

PC2 captures variability related to fire occurrence and environmental conditions. High negative loadings for temp and area on PC2 imply that lower temperatures and smaller fire areas contribute more to higher PC2 scores. This suggests that conditions associated with smaller fire sizes and cooler temperatures drive higher PC2 scores. In contrast conditions characterized by higher fuel moisture, fire danger indices, relative humidity, wind speed, and rainfall are associated with higher PC2 scores, potentially indicating higher fire activity or different environmental conditions.

Problem 6

At last, it was applies a multiple linear regression on the dataset using sklearn.

