The Intensity of Forest Fires Throughout the Year

TD2101 Data Science Fundamentals

Anis Hazirah binti Mohamad Sabry

Student ID: 1211300373

Table of Contents

1. Introduction

For any forest dwellers and visitors, there always exists the chance of being unfortunate enough to be caught in a tragedy known as a forest fire. As the flames wreck havoc and destruction upon human life and the natural environment, one might wonder how things might turn out that way. By which it means, what factors come into play to cause such an event. This project aims to study the intensity of forest fires throughout the year and the elements that play a role in it.

The reasons such a subject should be studied is because of how something that is potentially preventable can easily disrupt our ecosystem, wildlife habitats, and destroy human lives. Besides that, the forest is filled with a plethora of natural resources. Damage caused by forest fires may lead to a significant economic impact on communities dependent on forest resources as a means of living. By studying the underlying nature and patterns of the factors that contribute to forest fires, we can create better strategies for prevention and early detection of these hazardous events.

A study on the relationship between forest fires, the Fire Weather Index (FWI) indices, and weather conditions include exploratory data analysis, feature understanding, feature relationship, feature selection, and model construction and comparison. The initial steps will uncover the underlying patterns found within these variables. From our findings, we will then be able to make prediction models based on key features. Several models are created and a comparison of performance will be performed in order to find the most accurate method of forecasting forest fires.

1.1. Project objectives

The project aims to analyse the relationship between a forest fire and the elements that contribute to it which are the Fire Weather Index (FWI), and the weather conditions. From this analysis, we uncover the patterns and nature of forest fires, allowing us to understand the correlation between the FWI, weather conditions, months of the year, and the fires themselves. The intensity of the fires can then be ranked based on the area burned and the

FWI index, allowing us to discover the intensity of a forest fire that leads to a large amount of area burned.

1.2. Project Output

We aim to provide insights and information on weather conditions that have the most influential impact on a forest fire, and understand its relationship with the FWI as a means of better comprehending forest fire behaviour. Additionally, we aim to forecast which months of the year would a forest fire most likely occur. Identification of this period allows us to take a more precautious approach and provide additional resources to be prepared in forest fire prevention.

2. Exploratory Data Analysis

The dataset was collected from UCI's Machine Learning Repository. The dataset, 'Forest Fires Data Set', contains information of forest fires that have occurred within Montesinho park located in the northeast region of Portugal. It provides spatial coordinates of the fires that occurred within the Montesinho park, the month and day of the event, FWI indices information, weather conditions such as the temperature, relative humidity (RH), wind speed, and rain, and the area of the forest burned in hectare (10,000 square metres).

Additional information for the FWI and weather factors:

FWI: Forest Fire Weather Index system information:

- FFMC - 1-2cm deep, 16 hour time lag, moisture content of litter (twigs, leaves, etc)
- DMC - 5-10cm deep, 12 day time lag, moisture content of decomposed organic material underneath litter
- DC - 10-20cm deep, 52 days time lag, dryness of soil
- ISI - how fast a fire spreads, FFMC + wind

Weather information:

- temp - temperature in Celsius degrees: 2.2 to 33.30
- RH - relative humidity in %: 15.0 to 100
- wind - wind speed in km/h: 0.40 to 9.40
- rain - outside rain in mm/m2 : 0.0 to 6.4

The dataset is a multivariate dataset containing 517 instances and 13 attributes. It was donated on 29th February 2008, and contains a '.csv' file and '.names' file. For the purpose of this project, only the 'forestfires.csv' file was used.

After importing the dataset, a cleanup procedure involving dropping duplicate entries and removing rows containing null values is done as preparation for analysis. Thus, leaving us with 513 instances.

2.1. Feature Understanding

To understand the dataset, we will first dive into a rudimentary analysis of the dataset as is, collecting the statistical values of the dataset.

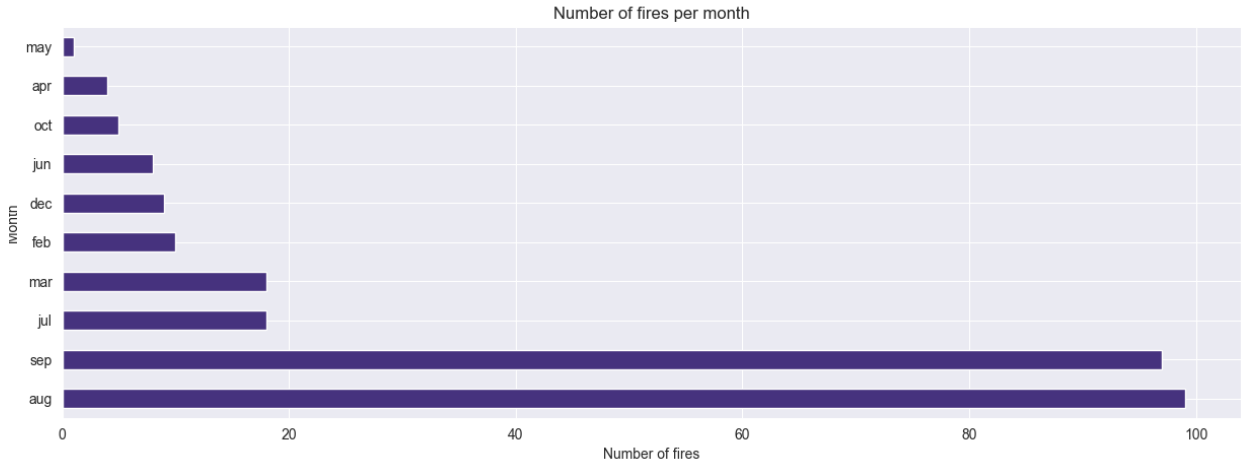

Figure 2.1: Number of fire occurrences a month

Figure 2.1 showcases the count of fires that occur within a month. The data is found by counting rows where the 'area' variable is more than 0, collecting only fires that have caused significant damage. To give more accurate numbers to this, the following is the occurrences of fires in a month found in the dataset:

Month	**Fire count**
August	99
September	97
July	18
March	18
February	10
December	9
June	8

October	5
April	4
May	1

Figure 2.2: Fire count per month

Based on the given information, the months of August and September have a tendency to garner the highest count of forest fires. While the months of October, April, and May are safest from these catastrophes.

In seasonal terms, we can say that the seasons of spring (March, April, May) and winter (December, January, February) tend to not have as many instances of forest fires. It may be likely that their weather conditions may not be as vulnerable to being a target of forest fires. Seasons of summer (June, July, August) and autumn (September, October, November) come right after one another, potentially having higher chances of a forest fire happening due to the temperature rise during the summer. Human activities during more active months may also lead to higher chances of a forest fire taking place. Further exploration of why some months are more susceptible to forest fires than others will be performed via finding the pattern of the FWI and weather conditions for those months.

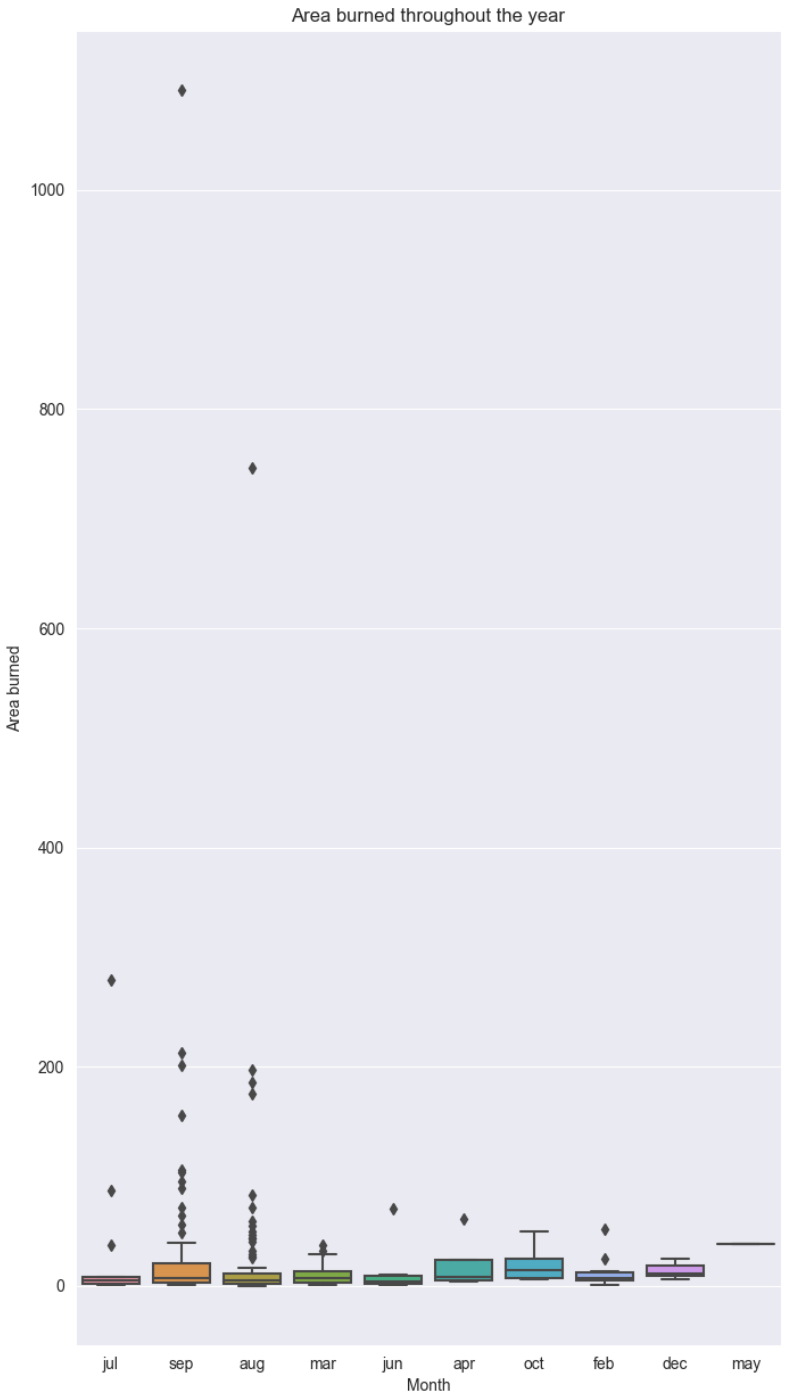

Figure 2.3: Boxplot of area burned by the month

We see that in September there is an outlier of over 1100 hectares of burned area. While in a month with lesser fire occurrences, July, there's an outlier of nearly 300 hectares of burned area. These abnormalities will not be removed from the dataset, but will play a role in investigating the causes of such intense peaks of fire.

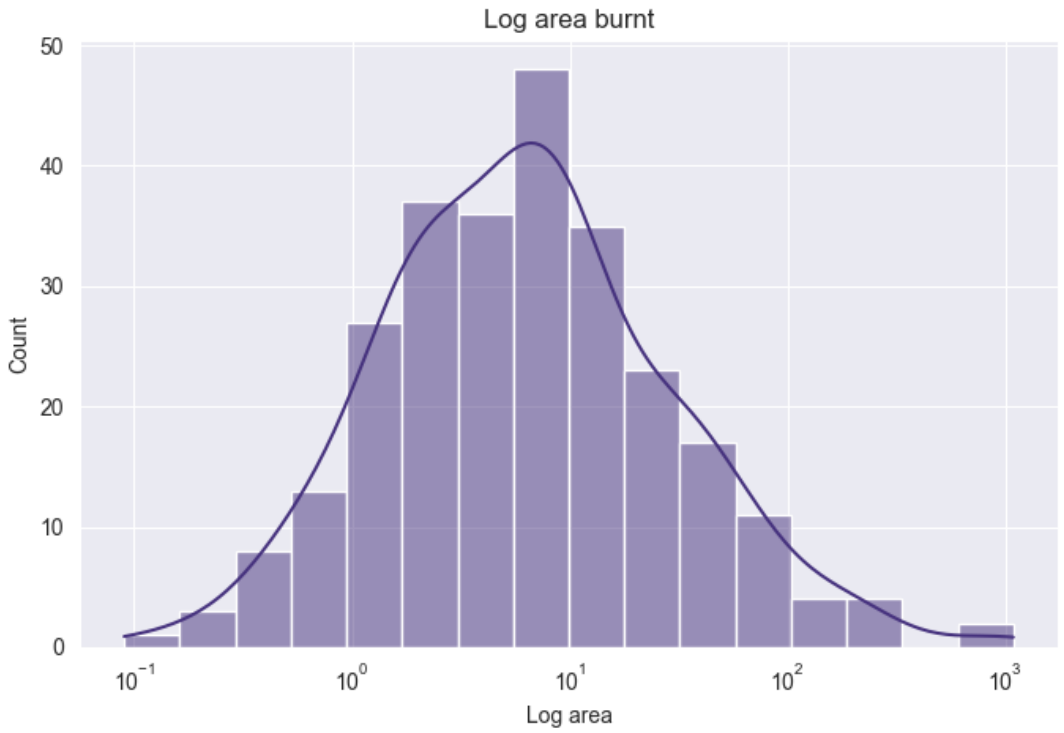

Figure 2.4: Histogram of log area

To see if our area is normally distributed, a histogram is created. Using log(area), we can see that the data is normally distributed and no additional transformation is required.

2.1.1. Analysis of weather conditions and the FWI by the month

To gain more insights on the factors that may contribute to the rise of fires in certain months, the FWI and weather conditions are examined.

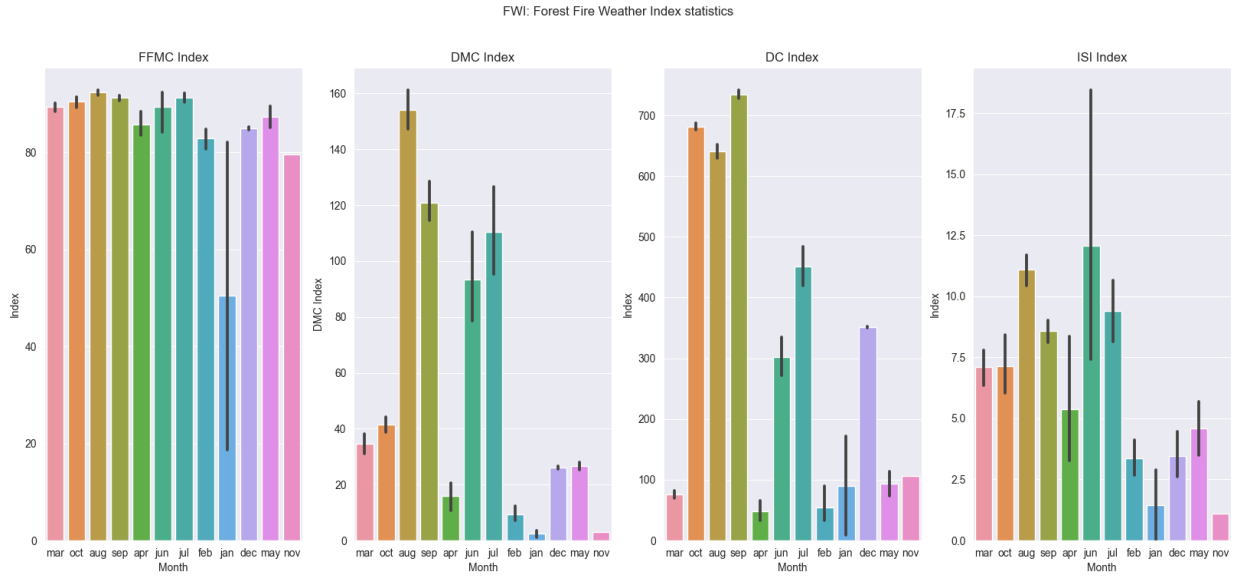

Figure 2.5: FWI indices scoring by the month

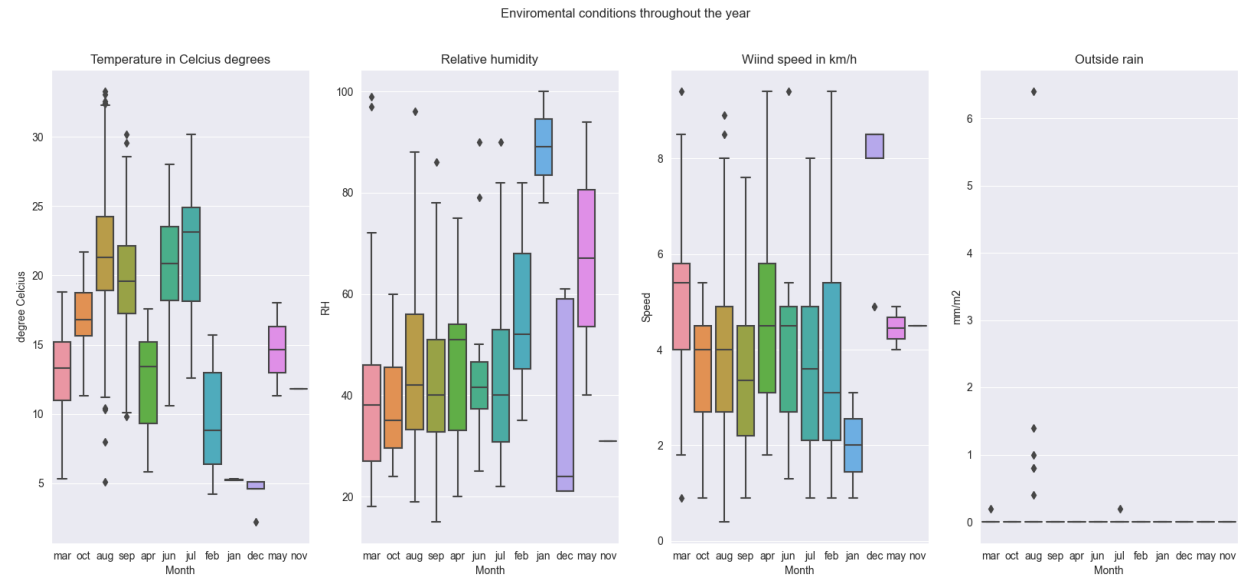

Figure 2.6: Weather conditions by the month

By the FWI indices, we can see that records of scores for the months of August, September, June, and July are consistently high numbers across all 4 indices. Other things of note is that despite how rain is present during the month of August, it still manages to catch more forest fires in comparison to the other months. We can also note how the months of January and December, months with 0 area burned, have similar weather conditions, with only relative humidity being an exception.

From here, it is clear that seasonal changes affect the environmental conditions and the FWI indices, potentially leading to more outbreaks of forest fires.

2.2. Feature Relationships

To investigate our findings even further, tests are performed to see the correlation between each variable and the area burned.

2.2.1. The correlation between the FWI, weather conditions, and area burned

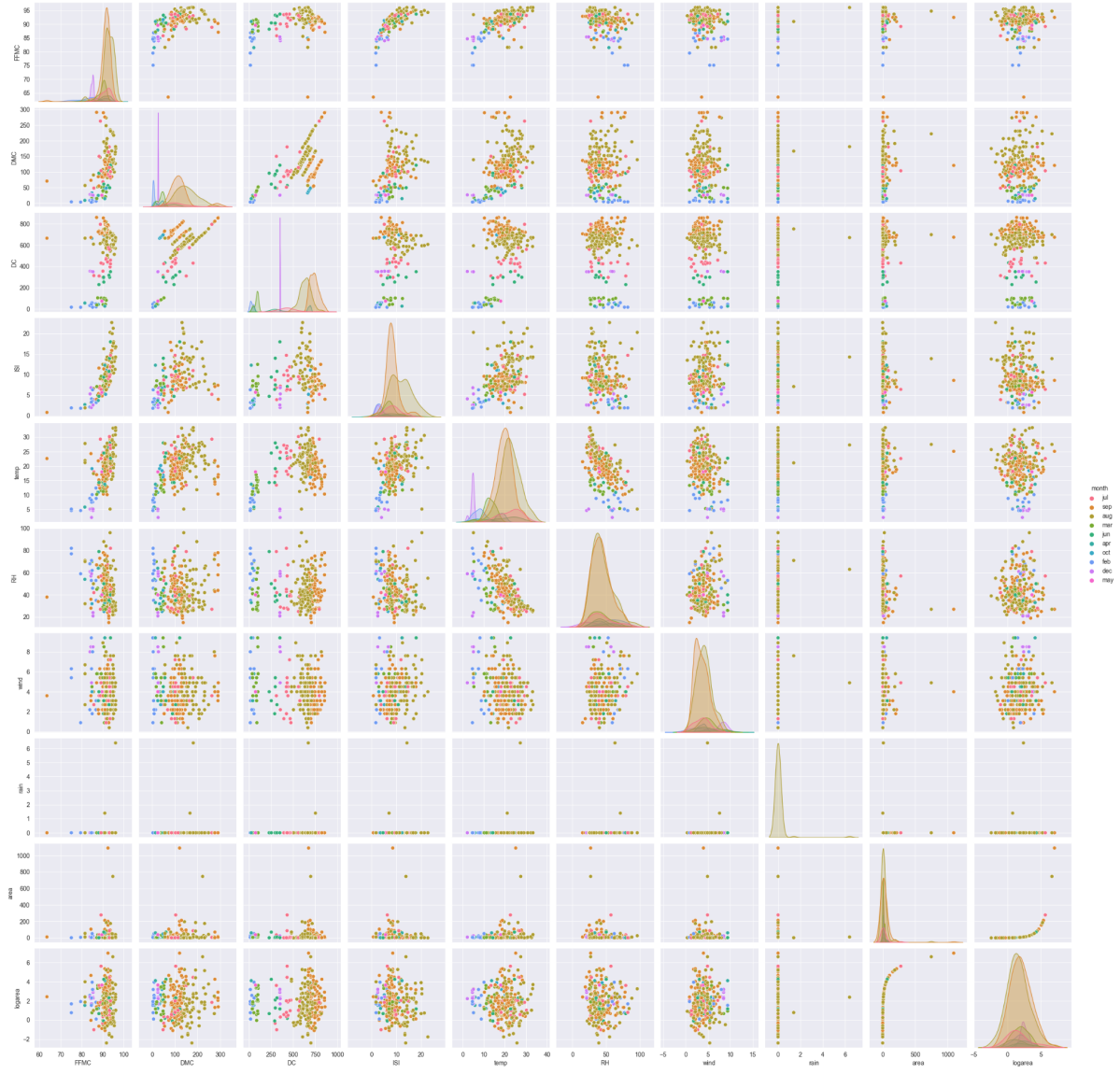

Figure 2.7: Pairplot of FWI indices, weather conditions, area, and log area

From the above diagram, we are able to make out linear relationships between some of the variables, with only rain being an exception to the correlation between them. To further investigate, a heatmap is made of the correlation.

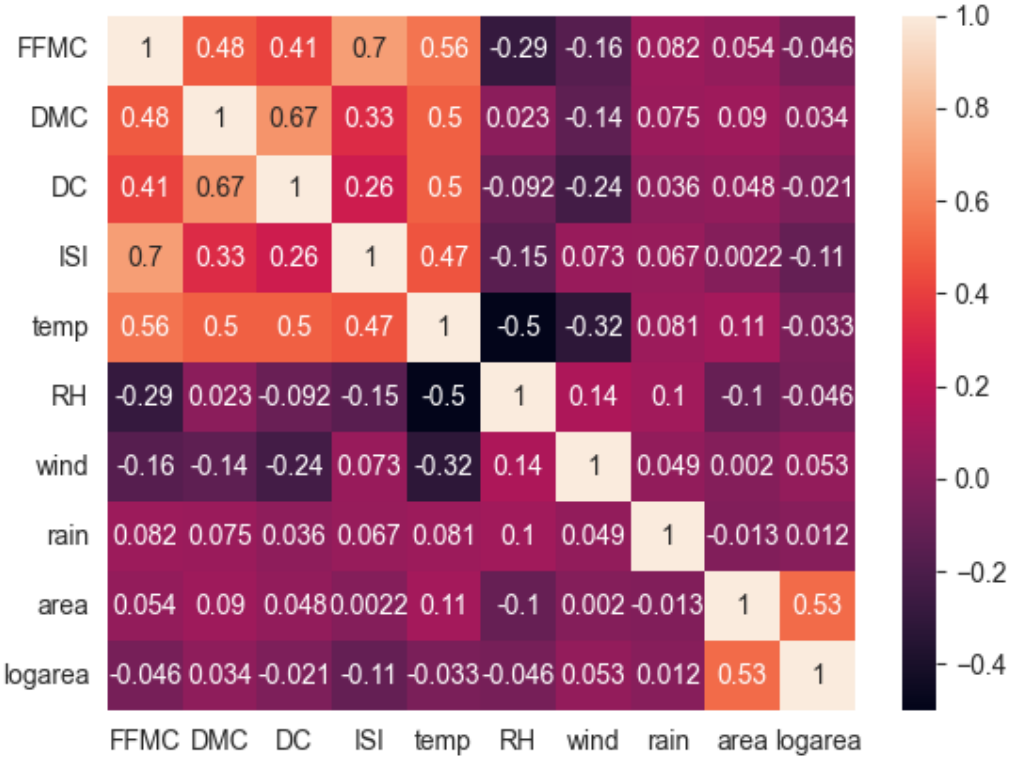

Figure 2.8: Heatmap of FWI indices, weather conditions, area, and log area

From the heatmap we can conclude that the weather condition that has the most correlation with the FWI indices is temperature, while the rest do not have as high of a correlation with it. This strong positive correlation indicates the temperate may contribute to drier conditions, increasing the probability and severity of a forest fire.

We can observe that most of the elements that come into play for a forest fire do play a role in the area burned, with the exception of relative humidity and rain pertaining to lesser scores than the other variables. Although they may not directly manipulate the outcome of a forest fire, they may interact with other variables that would influence a forest fire. As we can see in the heatmap, the highest correlation RH has is with the wind, while the highest correlation rain has is with RH.

2.2.2. How the amount of area burned is influenced by the FWI indices and weather conditions

Figure 2.9: Average weather conditions and FWI indices by the month

In the figure above we can clearly note that a spike during May in the average area burned occurred at the same time there was a peak in the relative humidity of the area, in the same instance where wind speeds were declining from its peak. By the FWI indices, the DMC and DC had a spike during the same month as well, while the ISI declined during that month.

A similar trend occurs during the month of September, where a spike in the average area burned showed a decrease in the wind speeds and the ISI index. We can conclude that the decrease in wind speeds during certain months coincides with an increase in the average area burned, implying that forest fires have more capacity to spread faster during times of calmer winds. It is worth noting that during this period, the RH value is lower than the month prior.

A spike in burned area can also be seen in the month of December, coinciding with an increase in RH and wind speed, but a decrease in temperature. As December is during winter, colder months may explain the drop in temperature. In terms of the FWI indices, each has an increase during the month of December as well.

2.2.3. Grading forest fire intensity by the FWI

In the actual FWI system, the fire intensity is ranked on the difficulty of controlling the fire based on the fire intensity and fire-fighting capabilities. The intensities are then classed as ‘low’, ‘moderate’, ‘high’, ‘very high’, and ‘extreme’. In our project, we will be using our own sum-based approach in order to get a FWI score. The scores will then be categorised into their own rank based on the amount of area burned via the .rank() function.

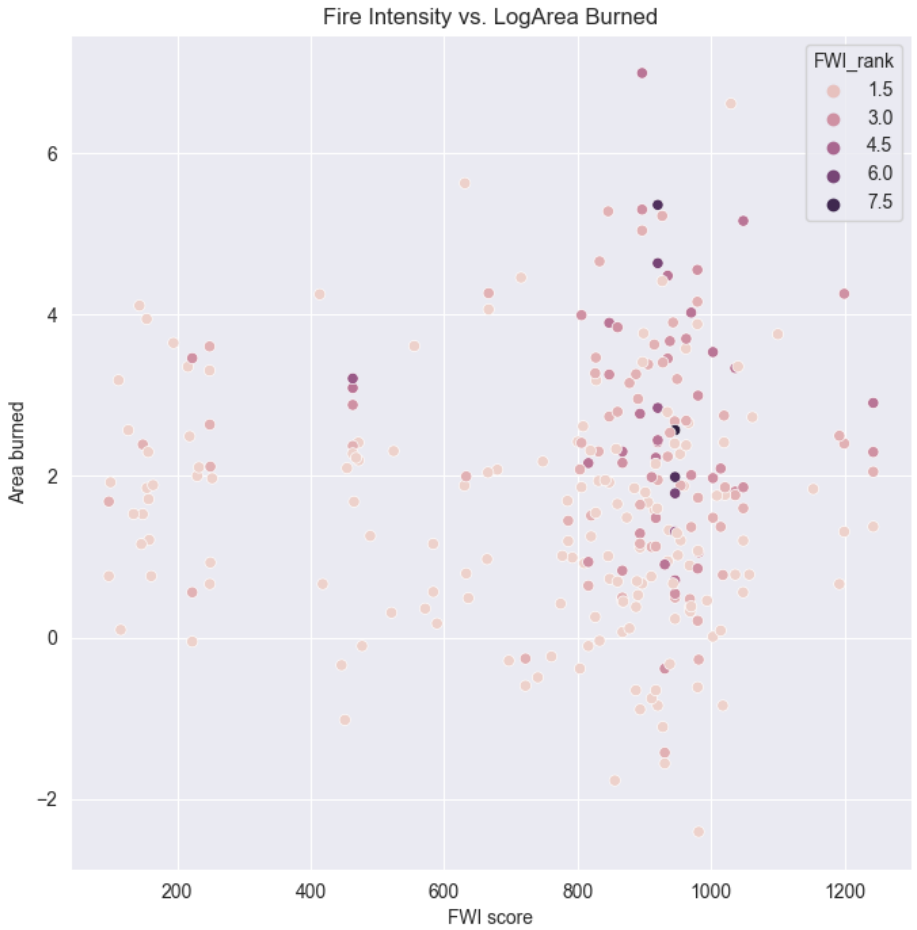

Figure 2.10: Fire intensity ranking based on the sum of FWI indices and the log area burned

Rank	Fire counts	Min score	Max score	Mean score	Min area	Max area	Mean area
1	154	97.6	1243	724.095	0.09	746.28	15.699
2	62	97.6	1243	842.889	0.24	196.48	22.854
3	29	222.6	1243	911.572	0.68	200.94	26.492
4	16	463.1	1243	920.831	2.03	1090.84	1011.390
5	3	463.1	946	776.400	3.71	24.77	15.227
6	2	920.1	946	933.050	5.97	103.39	54.680
7	2	920.1	946	933.050	7.31	212.88	110.095
8	1	946	946	946.000	13.06	13.06	13.060

Figure 2.11: Breakdown of the rankings

For this study, lower intensity fires will be classed from ranks 1 to 2, moderate fires will be between 3 and 5, while high intensity fires are 6 and above, with 8 being the highest intensity of a forest fire.

We deduce that most intense fires are not often correlated to a large burned area. Most likely attributed to fast responses from people ensuring the fire is kept under control and managed before it damages more forest area. High FWI scores do not immediately translate to high intensity fires either, most likely due to environmental elements that play into the fire's intensity and spreadability. Fires classed as high intensities have an average scoring between 900 and 950.

Month	Ranking
April	2
August	4

December	5
February	2
July	2
June	1
March	3
May	1
October	4
September	8

Figure 2.12: Maximum intensity rank by the month

There is no surprise that September has a maximum ranking of 8, where its most extreme area burned was over 1100. The maximum ranking of fire intensity by the month gives us another intriguing insight- in months where fires rarely occur, there is a chance of it becoming one that proves itself to be alarming. In the month of December, fires rarely occur yet somehow it has a maximum rank of 5. Based on the data we have seen, it's not surprising for December to have a high rank, as the weather conditions and FWI indices have shown to be much higher compared to its neighbouring months.

3. Feature Selection

In the original dataset, the output ‘area’ had undergone a ln(x+1) transformation function. The exact transformation was applied to areas of values more than 0, and the results were placed in another variable in the dataframe called ‘logarea’.

For the purpose of the project, the ‘x’, ‘y’, and ‘day’ variables will not be in use. Features that will be used for our model are month, FFMC, DMC, DC, ISI, an additional variable added called ‘FWI_score’ from the sum of all the FWI indices, the ‘FWI_rank’ that we had created to determine fire intensities, temperature, relative humidity, wind, rain, area, and logarea.

Most of the data used in the modelling are entries where the area is more than 0, examining only data of areas with a significant impact in regards to our forest fire study. To ensure our data is fit for modelling, the month feature, which is a categorical variable, will go through a label encoder. The entire data will then go through a standard scalar function to standardise the dataset and enhance model prediction.

4. Model Construction and Comparison

4.1. Prediction of area burned based on weather conditions

We gather the environmental factors (temperature, RH, wind, and rain) into our X variable and have our target feature, logarea, as our y variable. The dataset is split by 70% training and 30% testing. A linear regression is used as our method of solving this problem. The statistics of the linear regression is as follows:

Mean absolute error: 38.1197876825502

Coefficients:

Temperature: 1.79176125

RH: -0.02445396

Wind: 2.93163738

Rain: -5.25848272

Intercept: -22.928626165888627

Determination coefficient: 0.0024710520747246445

From here we can see that the temperature and wind are positive influences on the area burned, while RH and rain will decrease the area burned.

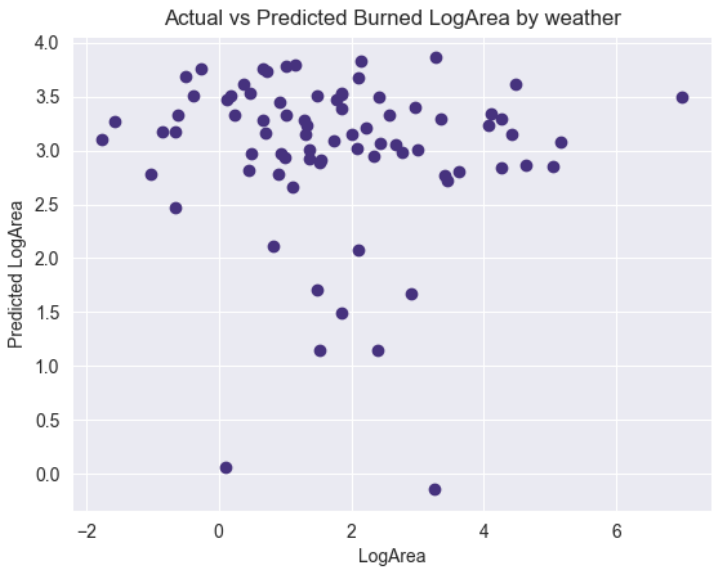

Figure 4.1: Predicted vs actual values of burned log area by weather

The scatter plot does not show a strong linear relationship, indicating that weather may not hold as much of an influence in predicting the chances of a large forest fire happening.

4.1.1. K-Means approach

The weather variables and the area burned will instead be clustered together. The ideal K-value is K = 5 based on the elbow test as shown below.

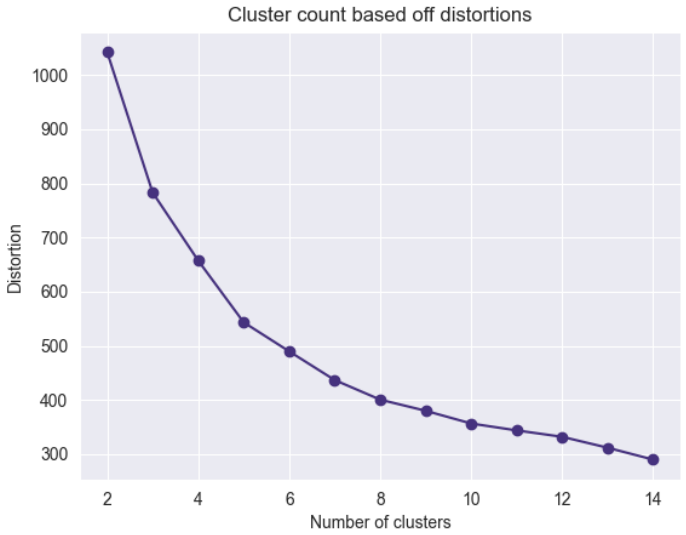

Figure 4.2: KMeans elbow test for weather vs logarea

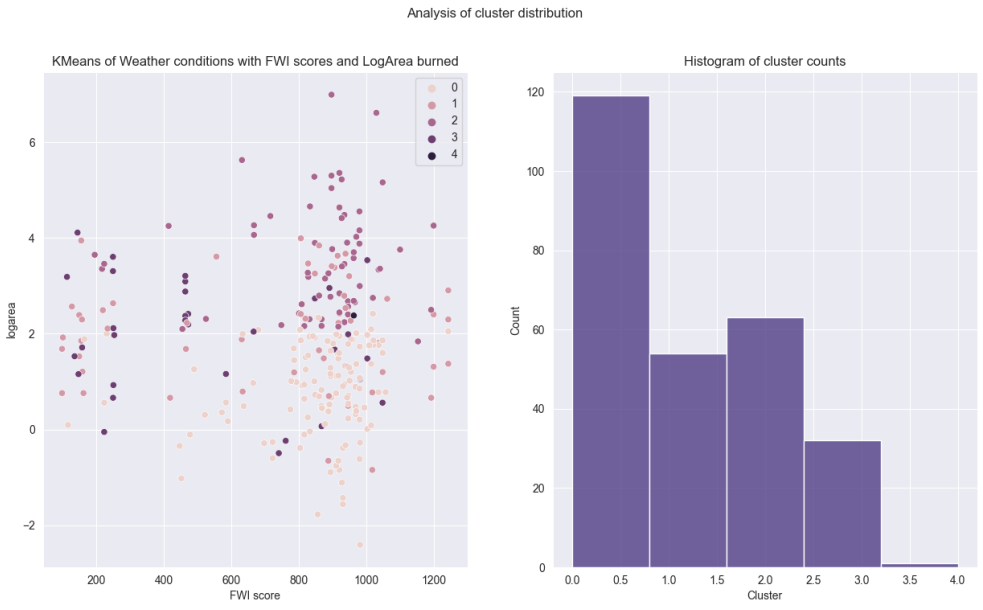

Figure 4.3: Analysis of clusters for K=5 of weather vs logarea

As we can see from the scatter plot above, the rankings of more intense fire conditions are not identical to weather conditions of more extreme nature. The spread of intense weather conditions goes towards even the lower ends of the graph, where lesser area was burned and a lower FWI score was recorded.

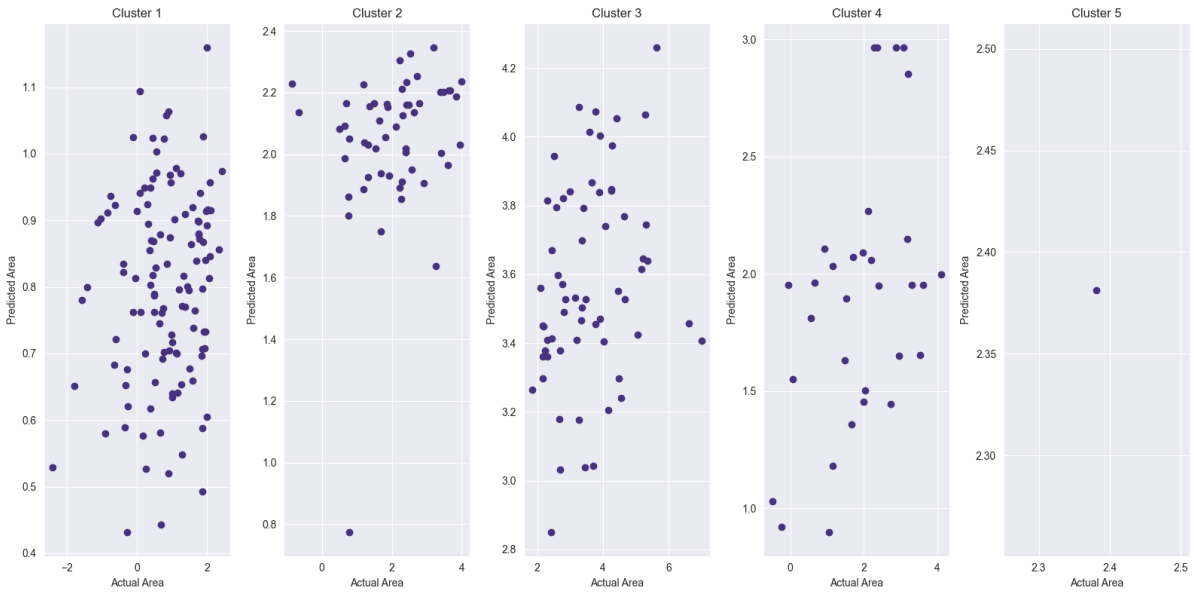

Figure 4.4: Breakdown of each cluster prediction

Cluster	Temperature	RH	Wind	Rain	Logarea
1	0.3518346	-0.28355069	-0.43397965	-0.072784	-0.68139179
2	-0.96091479	1.52260479	0.28366172	-0.00770695	0.13803178
3	0.60754547	-0.614141	-0.29768726	-0.072784	1.13839521
4	-0.92331395	-0.34574	1.70815102	-0.072784	0.0486634
5	1.29176361	1.27643592	0.41930981	15.99195081	0.35578029

Figure 4.5: Cluster centres for weather vs logarea

From the K-Means approach, we can assume that:

- Clusters 1-2 are of normal weather conditions

- Clusters 3-4 are of modest weather conditions
- Cluster 5 is of intense weather conditions

Clusters of a more moderate weather condition tend to have higher FWI scores, and a higher area burned. On the other hand, more intense weather conditions do not indicate a higher means of area burned. This may be the cause of weather conditions dampening the spreadability or intensity of a fire in the case of the rain factor and higher wind speeds.

From a combination of these two models, we can conclude that forest fires that result in large areas being burned tend to happen within weather conditions of cluster 3, where the average logarea burned is the highest among the rest of the clusters. Indicating that the most ideal weather conditions for a fire to occur is during the rise of temperature, and the decline of other variables.

4.2. Prediction of area burned based on FWI indices

The same approach is done with the FWI indices. Below is the result of our linear regression model.

Mean absolute error: 1.3019885542583811
Coefficients:
FFMC: -0.03239497
DMC: 0.00317338
DC: -0.00050459
ISI: -0.01258019
Intercept: 4.874096034843026
Determination coefficient: 0.00025684738712572486

By using FWI indices instead of weather conditions, we are able to achieve a much smaller MAE score.

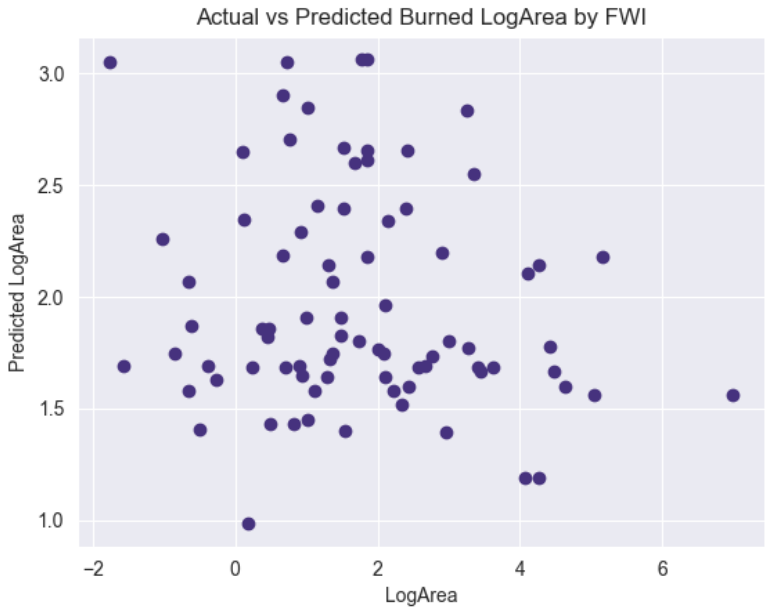

Figure 4.6: Predicted vs actual values of logarea burned by FWI indices

Compared to our linear regression with weather conditions, the FWI scatterplot shows a more promising linear relationship between the FWI indices and the logarea burned.

4.2.1. K-Means approach

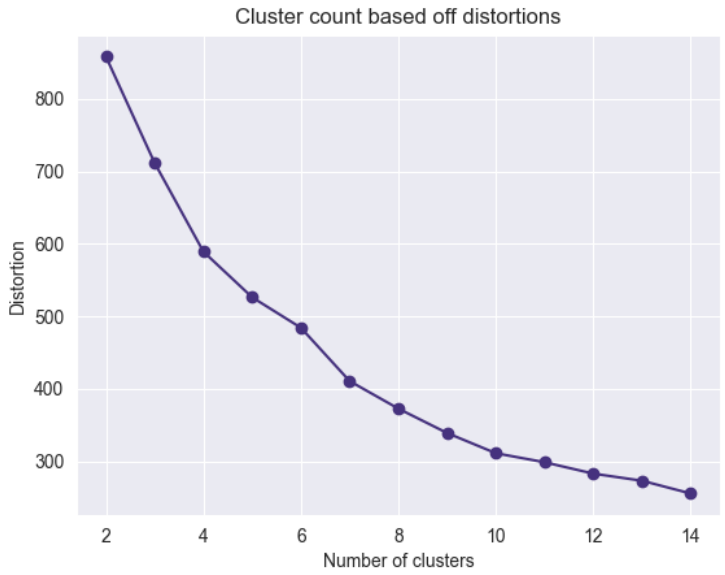

Figure 4.7: KMeans elbow test for FWI indices vs logarea burned

From our elbow test, we will proceed with using KMeans with K=6.

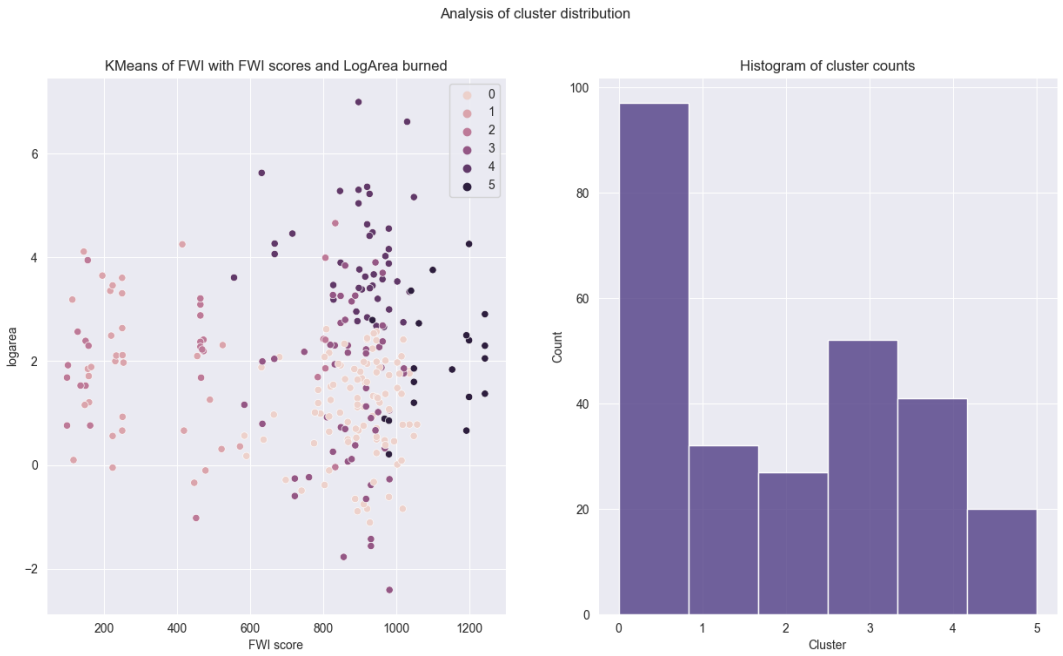

Figure 4.8: Analysis of cluster for K=6 of FWI vs logarea

We observe that the cluster distribution is much more similar to the ranking distribution.

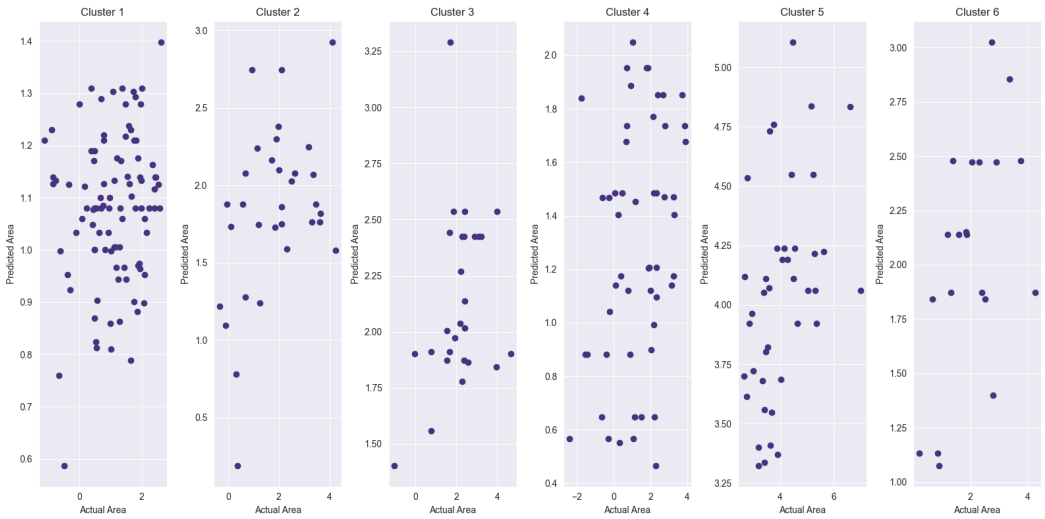

Figure 4.9: Breakdown of each cluster prediction

Cluster	FFMC	DMC	DC	ISI	Logarea
1	0.18022058	0.00826638	0.46949986	-0.17918279	-0.4994913

2	-3.30389782 e-01	-1.13560277 e+00	-1.90604979 e+00	-4.04927627 e-01	-4.48960261 e-04
3	-2.29237936	-1.39610209	-1.12299404	-1.5177311	0.196322
4	0.81281225	0.39310011	0.21593312	1.59368949	-0.36188344
5	0.24138038	0.21217521	0.42927079	-0.10157503	1.44643997
6	0.14112434	2.20459086	0.84721607	-0.36950616	0.13391145

Figure 4.10: Cluster centres for FWI vs logarea

Clusters with more extreme FWI indices can be found on the higher end of the log area burned. Cluster 2 shows peculiar centre values compared to the rest of the clusters, having values that are extremely low. Clusters with a negative ISI value but positive values in FFMC, DMC, and DC tend to indicate a potentially larger area being burned as found in clusters 5 and 6.

4.3. Investigating weather conditions and FWI scoring

To see if weather conditions would have an effect on the FWI scoring, a linear regression model and a random forest regressor model is used.

The linear regression results are as follows.

Coefficients:

Temperature: 26.6871067

RH: 3.9896828

Wind: -8.90108992

Rain: -11.24916436

Intercept: 126.44319495449611

Determination coefficient: 0.4108976587636197

Mean absolute error: 185.24635917452773

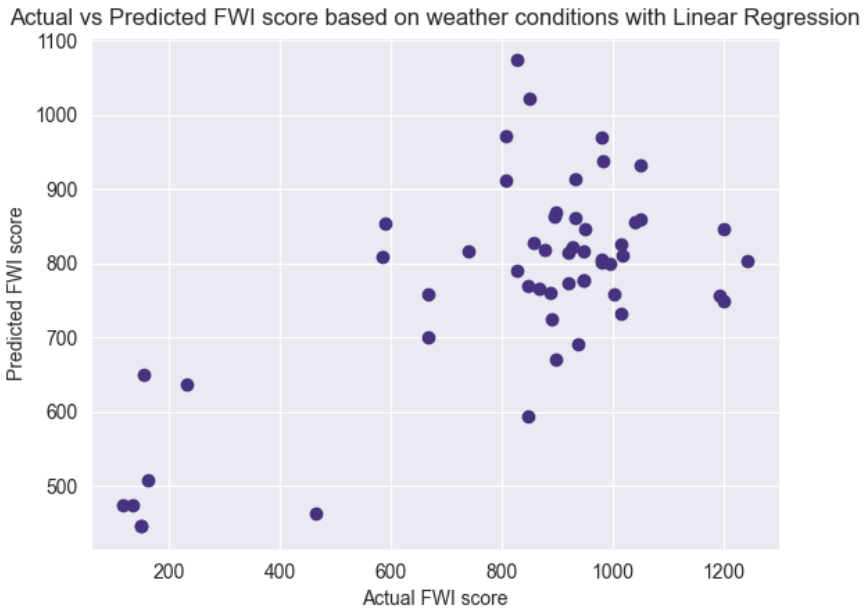

Figure 4.11: Actual vs Predicted FWI scores based on weather conditions by Linear Regression

As the MAE score is of high value, the same test was done using a random forest regressor. With random forest, we get a MAE score as low as 142.77111111111114. Although not ideal, it is still lower than the MAE achieved from linear regression.

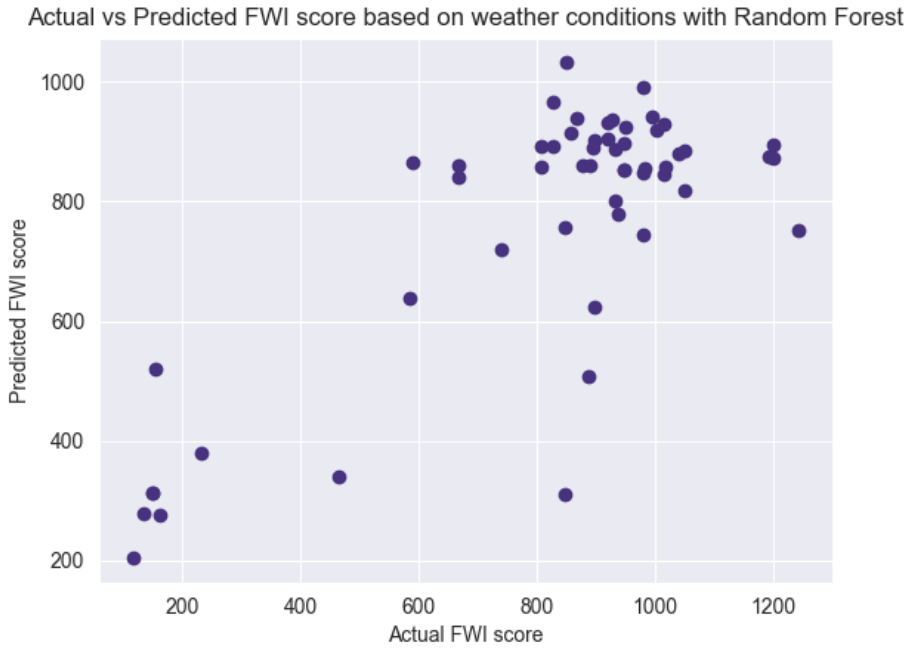

Figure 4.12: Actual vs Predicted FWI scores based on weather conditions by Random Forest

Based on the predictions, the weather indices do not appear to be an efficient way to predict possible FWI scores. This is most likely due to the fact that the FWI indices are of weather conditions that accumulate overtime, and not an instance of an event such as the rain or the current wind speed of the day. An idea for a more accurate prediction would be to gather the average weather data of a month, and proceed to predict the possible FWI scoring from that data. As opposed to trying to predict it based on one time instances.

4.4. Forest fire predictions by the month

To predict the forest fire counts in a month, logistic regression and a gradient boosting classifier was used. These two models were used as they are most ideal in prediction with categorical variables. The features FWI indices, FWI score, FWI rank, weather conditions, and the area burned will be placed into our X variable.

Linear regression accuracy: 0.7901234567901234
Gradient boosting accuracy: 0.8765432098765432

Linear Regression		**Gradient Boosting Classifier**	
September	34	**August**	33
August	32	**September**	32
July	4	**July**	5
March	4	**February**	4
February	4	**March**	3
October	2	**December**	3
December	1	**June**	1

Figure 4.13: Table of predictions of forest fire occurrences in a month

Based on the results, we observe that both August and September are likely candidates for months with a high occurrence of forest fires. This is not surprising considering the extreme weather conditions and the FWI scoring during those months.

In both of the models, each of them predicted a month that was not present in the other prediction. The linear regression predicted the month of October, which does not appear in the gradient boosting classification. While the gradient boosting classification predicted the month of June. Neither of these predictions are entirely false as both of these months have had an instance of a forest fire occurring.

5. Deployment

A publication of this project is hosted on streamlit with the following URL: https://forestfire-1211300373.streamlit.app/

The original code in the Jupyter notebook was added into a Python script with Streamlit functions, such as using ‘st.write()’ instead of ‘print’ and using ‘st.pyplot()’ in order to showcase graphs inside the Streamlit. The altair library was used in order to make the plots more interactive in Streamlit. Besides that, additional tabs were added as an easier way to navigate the entirety of the research project as we can see in the figure below.

Figure 5.1: Front page of the website

In order to host the code into Streamlit, an account was required in order to create a public URL that can be viewed publicly in the internet domain. After creating an appropriate Streamlit Python script, all that was left to do was simply host it by adding the Github repository of the project into the application. An additional text file called ‘requirements.txt’ is needed in the case of installing libraries that were used inside of the file. Once the required libraries is fully installed and the application loads the Python script, the project will then be live on Streamlit.

Streamlit is interesting in that it allows a tooltip customisation. This allows users to be able to dive deeper into data points and find the exact values for each of them, allowing a more intricate understanding of the project.

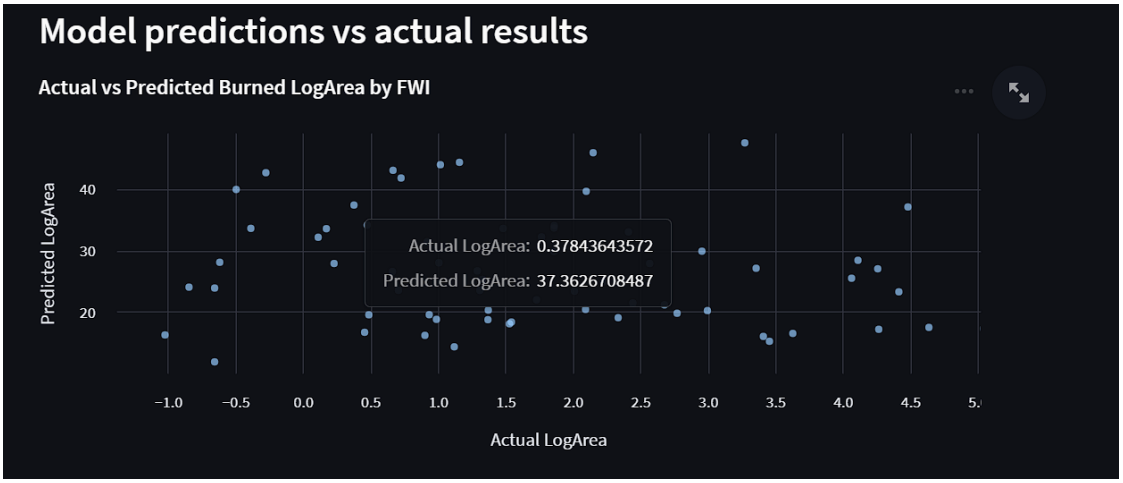

Figure 5.2: Example of information on datapoints

Included in the website is a prediction tool where users may input values for variables for environmental factors, or FWI indices and predict whether or not there is a chance of a fire occurring due to those factors.

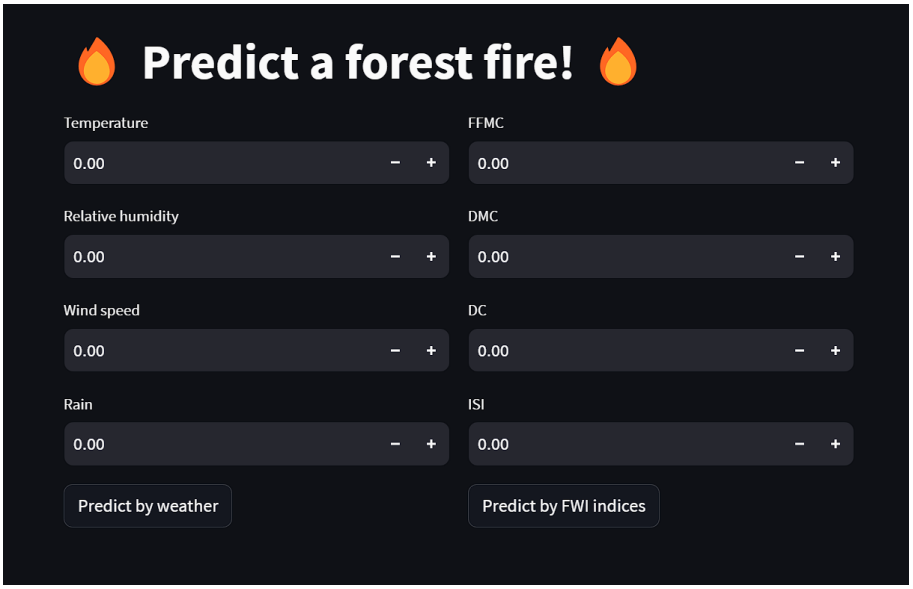

Figure 5.3: Prediction pages

6. Conclusion

Through analysing the forest fires dataset, we find that the months of August and September are candidates with high forest fire occurrences. The months of October, April, and May on the other hand, have the lowest fire occurrence count. Higher scores of FWI were shown in the months of August, September, June, and July.

Correlation between each variable and burned area indicates that most of the factors, apart from relative humidity and rain, play a role in influencing the chances of a forest fire occurring. In trying to predict the intensity of a forest fire, all the variables must be taken into account as relying on FWI scores alone shows us that not all high FWI scores translates into fires of high intensity.

Various models including linear regression, random forest, and K-Means clustering were used as an approach to finding the relationship between the variables and area burned, along with predicting the likelihood of fires happening and the extent of the damage that would occur. Our findings suggest that weather conditions are not a reliable predictor in trying to predict a forest fire. Through K-Means, we discover that most fires tend to take place in moderate weather conditions, as intense weather conditions of rain and RH dampen the chances of a forest fire becoming more intense.

A similar procedure was done with FWI indices in place. From this, we uncovered a more linear relationship between the FWI indices and the log area burned. The results of the clustering from K-Means is also similar to the rankings of fire intensity, indicating that they have an impact on the level of intensity a fire is.

As FWI indices are based on weather conditions, we can draw a conclusion that periods of more rain and RH may indicate lesser FWI scoring. Thus, reducing the intensity of possible forest fires that might occur. It is advised that additional support and observation for forest fire prevention is required during periods of hotter temperatures and the months following them.

Although seasons were not a factor with great emphasis in this project, a conclusion can still be drawn based on the months most often associated with those seasons. From this

study, we find that the seasons of summer and autumn have a higher potential than other seasons to attract forest fires.

To conclude, our research finds that both weather conditions and FWI indices play a significant role in the count of fire occurrences and the intensity of a forest fire. However, neither can be used as a standalone measure, as both are required in order to make a more accurate judgement and prediction.