SDS 322E Project

Forest Fires

Team: Ela Albiston, Jennifer Amador-Gonzalez, Yihan Du, Diego Fernandez, Wenting Lu, Anna Pham, Theresa Pham

2022.12.04

Table of Contents

Introduction

Forest fires are creating economic and ecological damage while endangering human lives. Across the world, there are millions of forests destroyed each year. This issue can especially be seen in Portugal. Over 2.7 million forest hectares were destroyed between 1980 to 2005.

We want to focus on the forest fires that occurred in Montesinho Natural Park between 2003 and 2005. These forest fires were especially detrimental, with 4.6% and 3.1% of the territory affected and 21 and 18 human deaths, respectively. It is imperative to detect forest fires early in order to mitigate this issue. We want to make predictions about forest fires in order to control them. Environmental and climate conditions can be observed to see if they are conducive to forest fires. If we are able to predict forest fires in advance, there will be greater preparation and resources for management.

Data Description and Cleaning Procedures

Data

The dataset we used was obtained from the Center for Machine Learning and Intelligent Systems at the University of California, Irvine. The original dataset has 517 observations and 13 variables collected from January 2003 to December 2005 for Montesinho Natural Park.

The target variable is "Area," which is the burned area of the forest. Input variables include indexes from the Canadian danger rating system of Fire Weather Indexes (FWI) such as FFMC (Fine Fuel Moisture Code Index), DMC (Duff Moisture Code Index), DC (Drought Code Index), ISI (Initial Spread Index); other than those variables, and the dataset also include x and y-axis spatial coordinate within the park map, the month of the year, day of the week of the observed records, temperature, RH (Relative Humidity), wind speed, and outside rain.

Variable Name	Role	Level	Description	Range of Values
X	Input	Nominal	x-axis spatial coordinate within the park map	1 to 9
Y	Input	Nominal	y-axis spatial coordinate within the park map	2 to 9
month	Input	Nominal	Month of the year	Jan to Dec
day	Input	Nominal	Day of the week	Mon to Sun
FFMC	Input	Interval	Fine Fuel Moisture Code Index (FWI system)	18.7 to 96.2
DMC	Input	Interval	Duff Moisture Code Index (FWI system)	1.1 to 291.3
DC	Input	Interval	Drought Code Index (FWI system)	7.9 to 860.6
ISI	Input	Interval	Initial Spread Index (FWI system)	0.0 to 56.10
temp	Input	Interval	Temperature in Celsius degrees	2.2 to 33.30
RH	Input	Interval	Relative Humidity in %	15.0 to 100
wind	Input	Interval	Wind speed in km/h	0.40 to 9.40
rain	Input	Interval	Outside rain in mm/m^2	0.0 to 6.4
area	Target	Interval	The burned area of the forest (in ha)	0.00 to 1090.84

Table 1. Original variables - variable name, role, level, description, and range of values

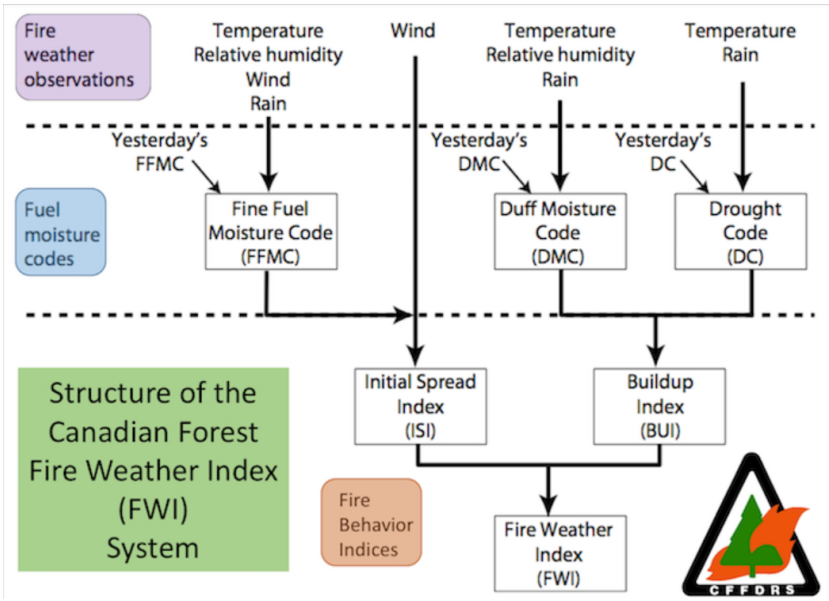

Figure 1. Structure of the Canadian Forest Fire Weather Index (FWI) System

Data Preparation

For modeling purposes, we first checked if there were missing values in the dataset because further calculations with missing values would be inaccurate and unacceptable. Then, we converted all datasets to numeric, such as the month of year and day of the week. After the conversion, it will be easier and more efficient to do the visualization.

month	day
mar	fri
oct	tue
oct	sat
mar	fri
mar	sun
aug	sun
aug	mon
aug	mon
sep	tue
sep	sat

month	day
3	5
10	2
10	6
3	5
3	7
8	7
8	1
8	1
9	2
9	6

Table 2. Convert variables to numeric

In addition, we noticed that there are 247 observations in the dataset that have a 0 with the burned area of the forest, which will highly affect the result that we conclude. Therefore, we switched the variable type of area from continuous to discrete. Then we split the area into different scales.

Variable Range	Level	Description
Area = 0	0	No fire exists
0 < Area < 50	1	Small Scale Fire
Area >= 50	2	Large Scale Fire

Table 3. Different scales of burned area

Exploratory Analysis

To have a better understanding of the data set, we need to know the relationships among variables. A good way to do this is to build a correlation heat map. The correlation heat map shows the correlation between each variable in the data set and another variable (Fig 2). The color of the heat map presents how strong the correlation is. Orange color is a negative correlation, and purple is a positive correlation. The darker the color is, the stronger the correlation is.

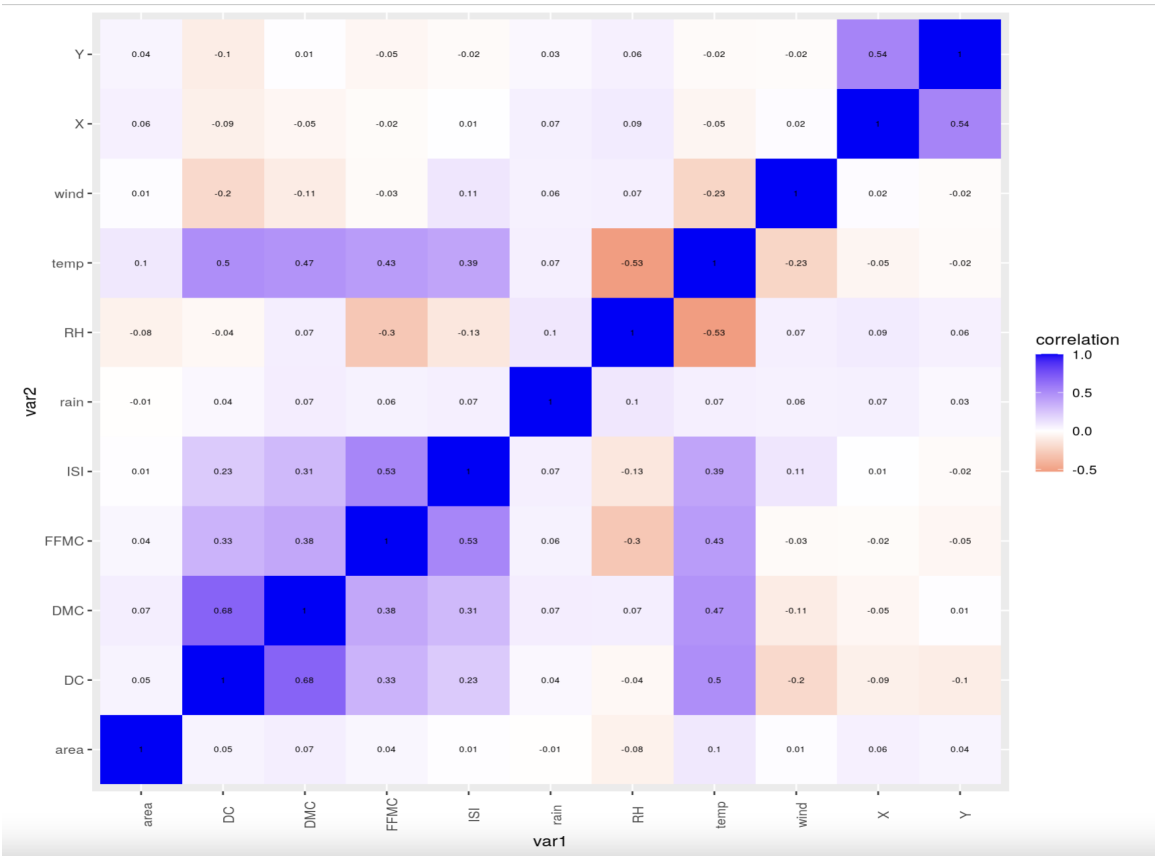

Figure 2. Correlation Heat Map

We can see a light purple with a 0.5 correlation between DC and temp, light orange with a -0.53 correlation between RH and temp, a light purple with 0.53 correlation between FFMC and ISI, and a light purple with a 0.54 correlation between X and Y (Fig. 2). Based on these four observations, we made our four hypotheses: 1) As temperature increases, the Drought Code (DC) value increases 2) Higher temperature can lead to lower humidity levels (RH) 3) The greater the initial spread index (ISI), the greater the fine fuel moisture code (FFMC) 4) Location is a factor of fire.

Hypothesis I

For hypothesis one, we want to know the relationship between DC and temperature. Due to the Canadian Forest Fire Weather Index System, DC is a drought code, which presents how dry the area is. The higher the drought code, the more extreme the drought is.

Our first step is to graph a scatter plot with a smooth line (Fig 3). We can see from the scatter plot that there is a small cluster in the lower left corner and a big cluster in the upper right corner. Both clusters' centers are not close to the positive line, so we colored the scatterplot by month. We can see that the smaller cluster in the lower left corner is green, and the larger cluster in the upper right is pink (Fig 4). The green color presents the spring months, such as February and

March. The pink color presents the fall months, such as August and September. We can see from Fig 4 that there are more pink observations than the other colors. Therefore, we made a frequency table for the month's observation (Table 4). It is clear that August and September have more observations than the rest of the months. 2+15+1+9+20+54+9+2+17+32= 161 < 172 (September) < 184 (August).

Then we remake the scatter plot for temperature and drought code by month instead. From Fig 5, we can see that each month does not have a strong relationship between temperature and drought code. An obvious observation is that August has a wide range of temperatures, from 5 to 35 degrees, and the slope of the line is flat. In summary, there is no strong evidence that as temperature increases, the drought code increases. It is possible that drought codes change to seasonal changes.

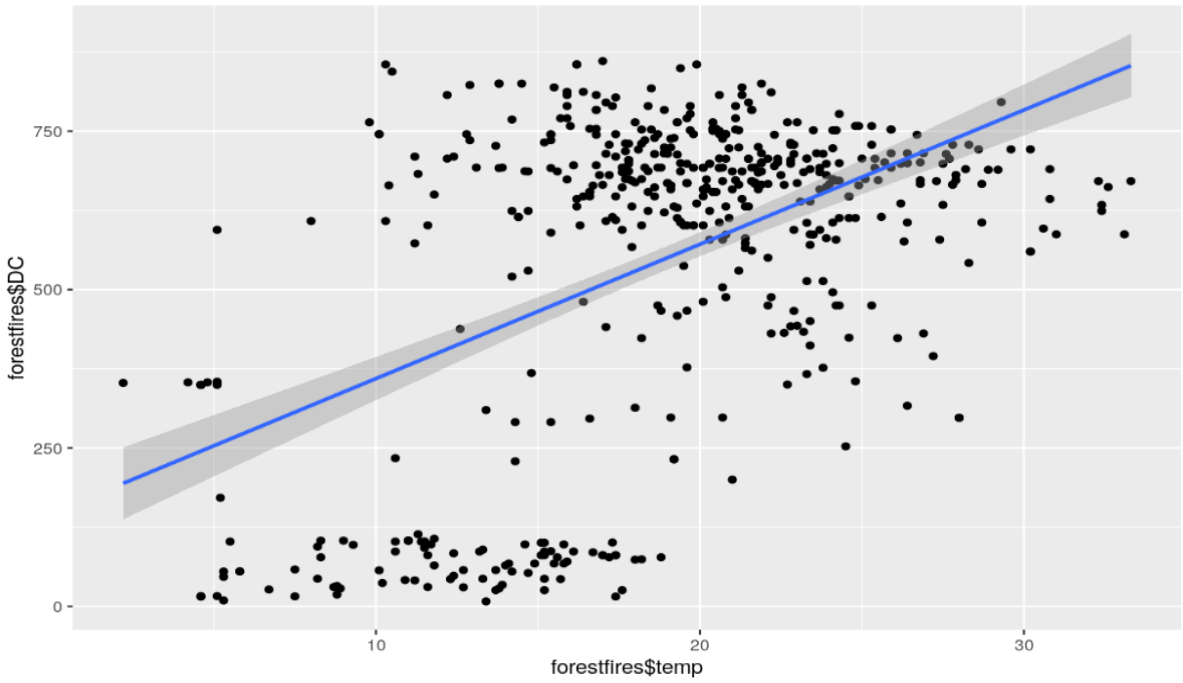

Figure 3. Scatter plot for Temperature and Drought Code with a smooth line.

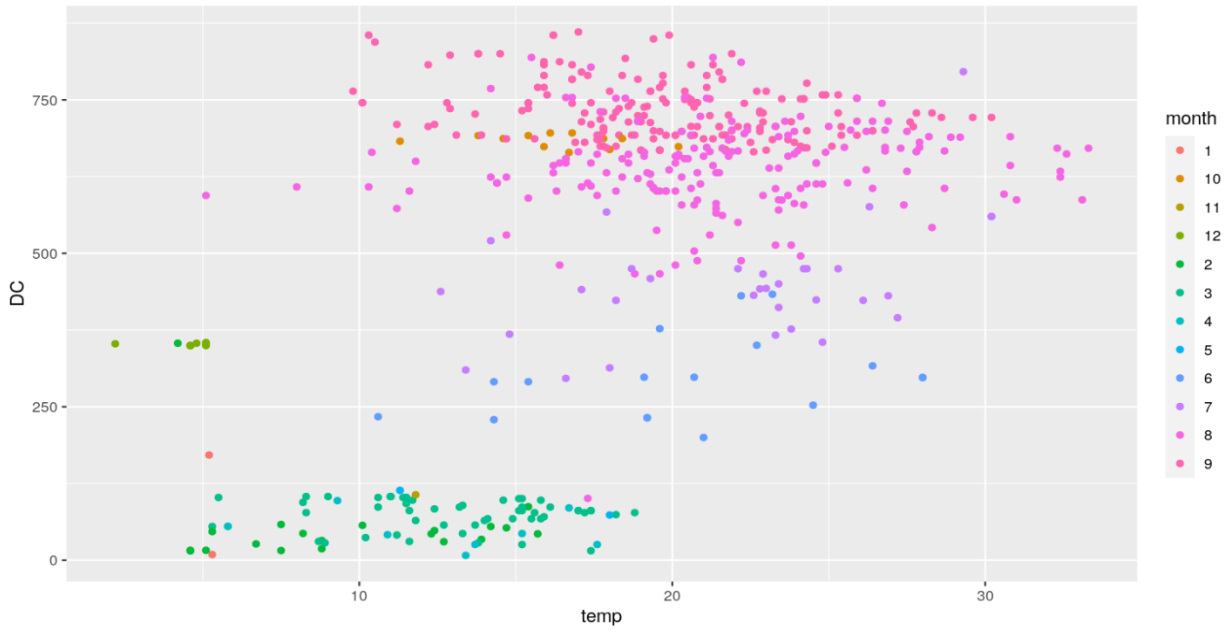

Figure 4. Scatter plot for Temperature and Drought Code with Month Colored.

month <chr>	n() <int>
1	2
10	15
11	1
12	9
2	20
3	54
4	9
5	2
6	17
7	32
8	184
9	172

Table 4. Observations table by month

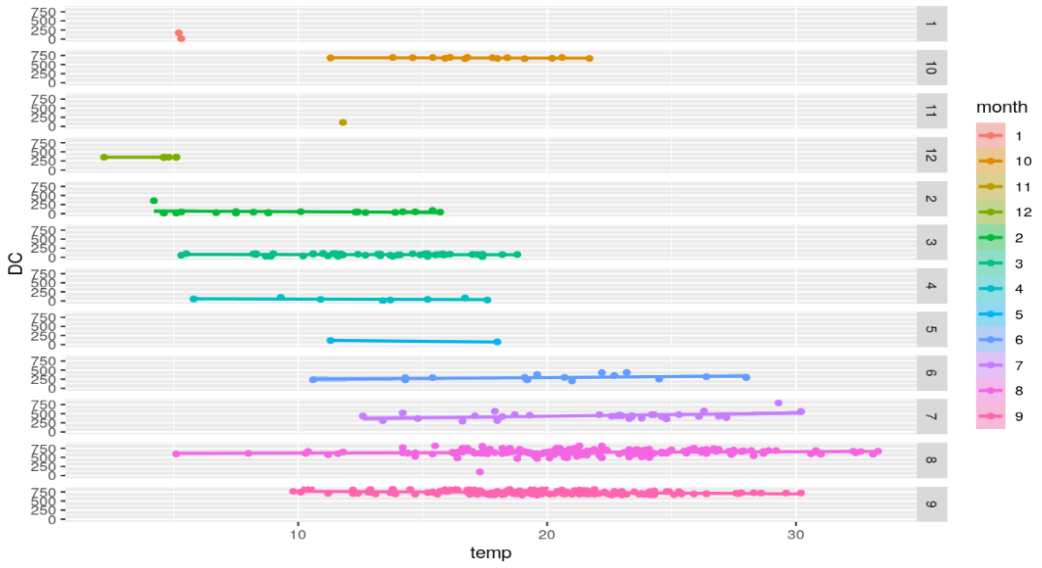

Figure 5. Scatter plot for each Month's Temperature and Drought Code with Month Colored

Hypothesis II

One of the hypotheses that we would be exploring with the variables of our data is whether higher temperatures can lead to lower humidity levels when looking at the two variables: Relative Humidity and Temperature.

After conducting this, we found that Relative Humidity in our x-axis and Temperature in the y-axis of our graph leads to a negative linear relationship between the variables (Fig 6). The results showed that as relative humidity increases, temperature decreases. This would be vise-versa that as temperature increases, relative humidity decreases, as we see by the inverse relationship between these two variables, which proves our initial hypothesis to be true.

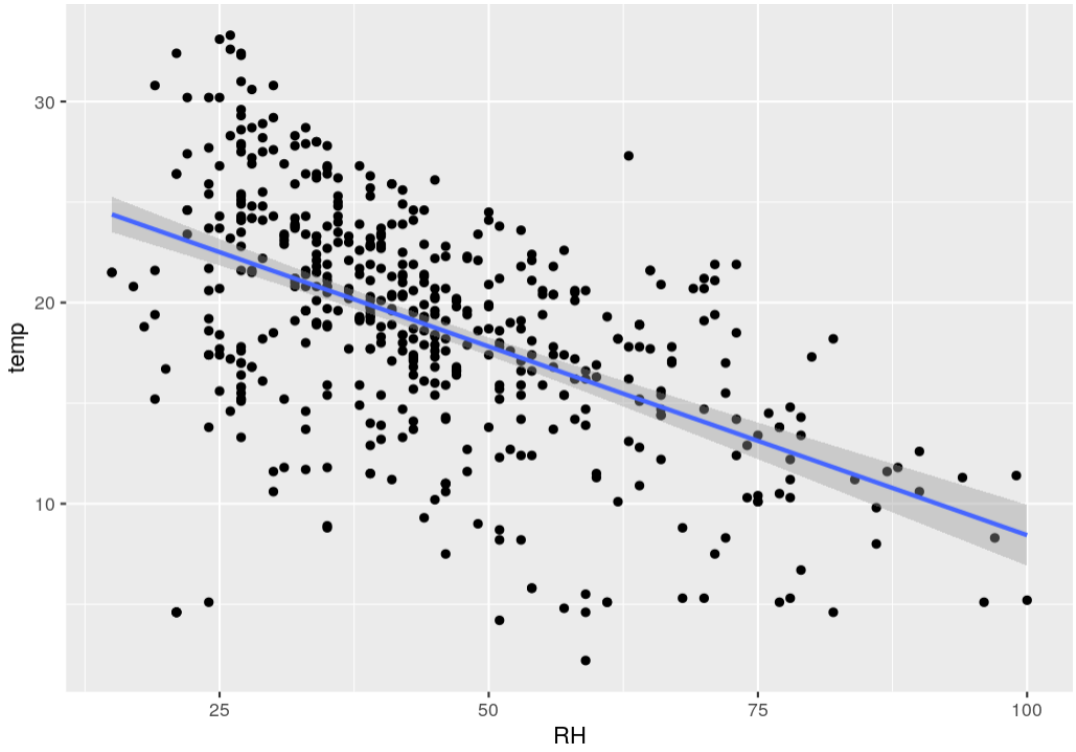

Figure 6. Linear regression model for Relative Humidity and Temperature

Hypothesis III

In our third hypothesis, we try to analyze whether a greater Initial Spread Index (ISI) leads to a greater Fine Fuel Moisture Code (FFMC).

Using these two numeric variables and conducting the same test for seeing the relationship between relative humidity and temperature, we were able to see a fitted positive linear correlation when using ISI as our x-axis and FFMC as our y-axis (Fig 7). This fitted correlation is part of what seems to be an exponential increase relationship between the two variables, with a few significant outliers. Nevertheless, our hypothesis continues to be proved by the linear regression model of these two variables.

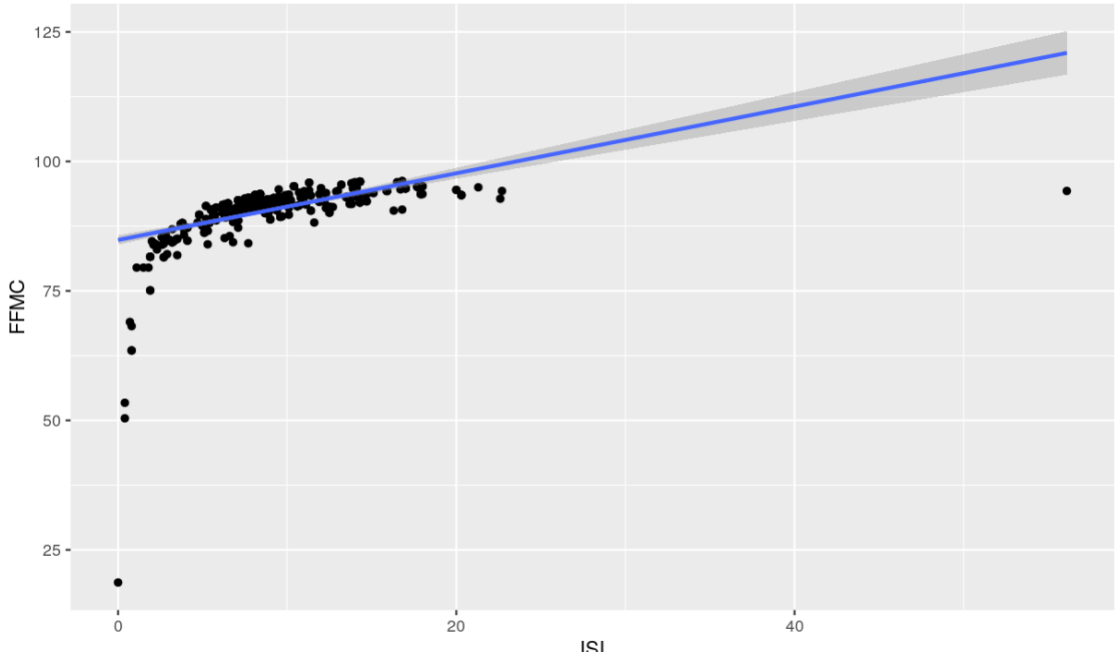

Figure 7. Linear regression model for Initial Spread Index (ISI) and Fine Fuel Moisture Code (FFMC)

Hypothesis IV

To figure out whether the location is a factor of forest fire burning area, we first map each observation into a map with its X and Y coordinates (Fig. 8). We found the data collection map online (Fig. 9). The map we form by the X and Y axis matches the park map. We want to know if each location has a different burning area. Therefore, we colored each observation by the group of burning areas with green for 0, blue for 1, and red for 2 as described in data cleaning. It is clear to see from Fig. 8 that the southern border has a larger burned area than the rest.

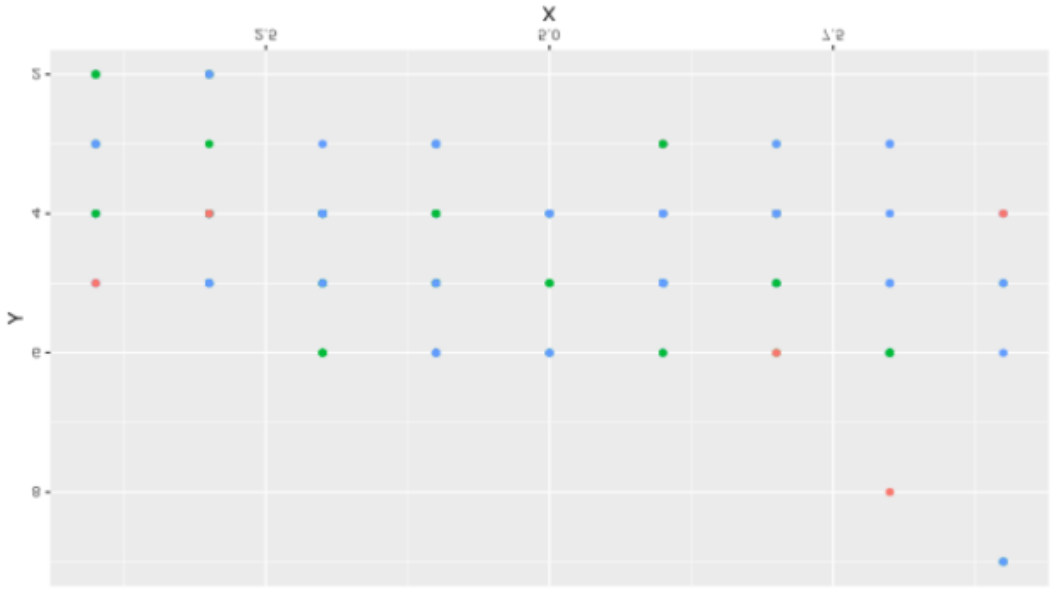

Figure 8. Location Map Presenting the Severity of each observation

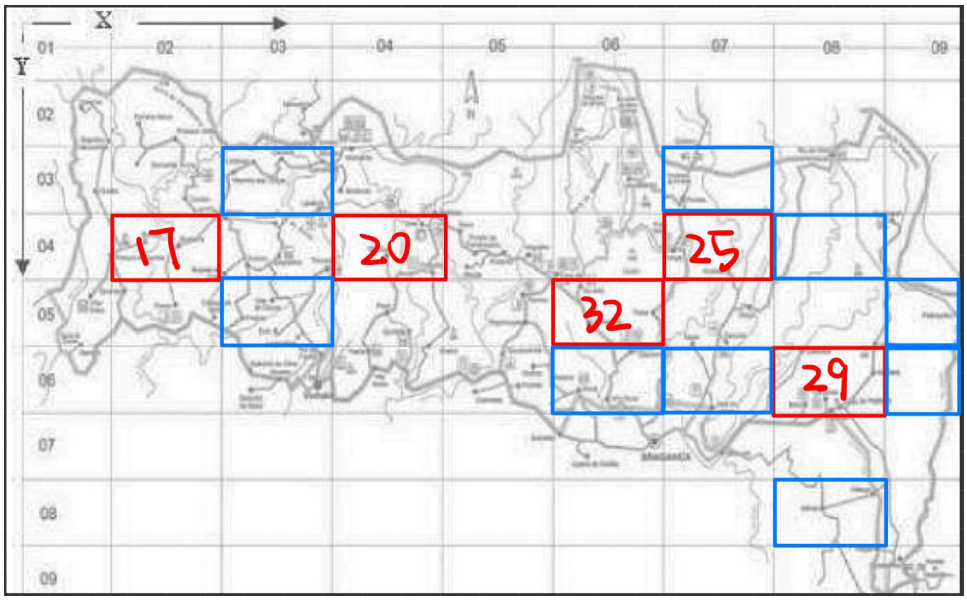

Figure 9: Location Map

By using the dataset with only 207 observations which exclude the area = 0, and summarizing the x and y-axis frequency, we drew the graph above. The number within the rectangle indicates the frequency of forest fires collected from observed data. The blue sections are areas with fewer fire accidents. In contrast, the red sections are areas with more fire accidents, with a top on top indicating the frequency. It is clear to us that the blue sections surround the red sections. Several

causes may result in this behavior, including but not limited to timely first-time response, weather, geographic factors, and different vegetation coverage.

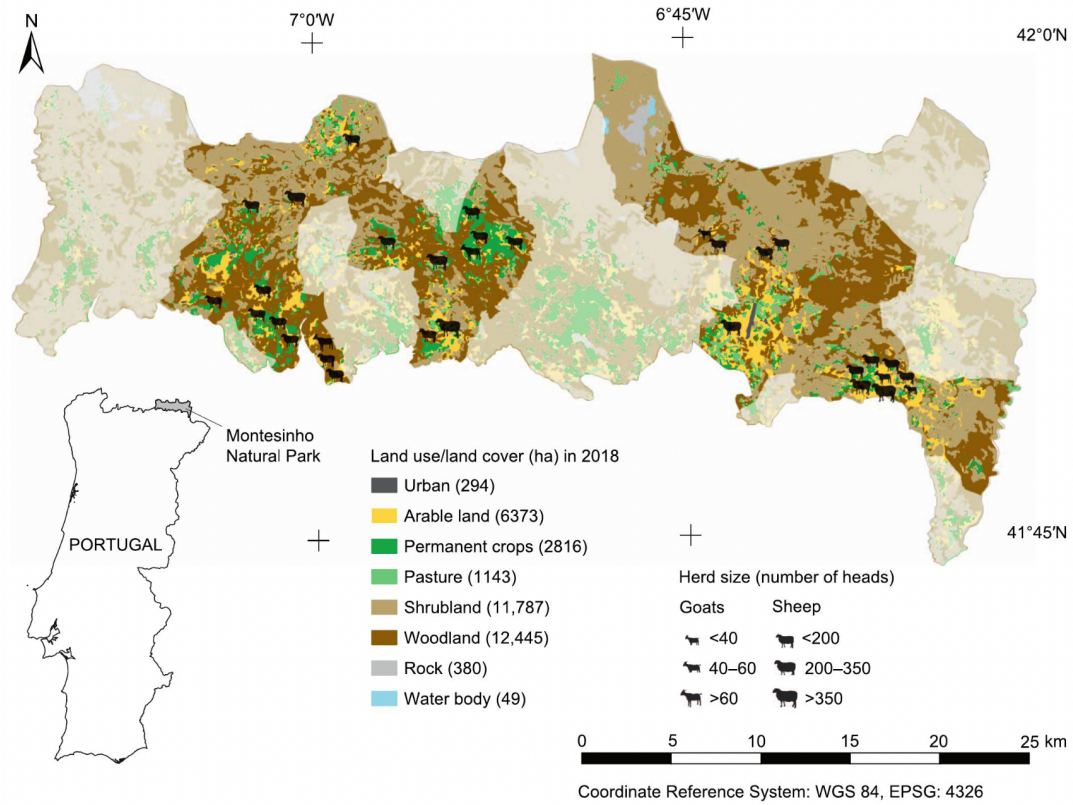

Figure 10: Montesinho Natural Park Vegetation and Herd Coverage Map

Clustering

We further analyzed our data via clustering. We decided to use the PAM clustering algorithm to cluster our data since it is more robust than K-means, and there is no concern about computation time and memory storage due to our data being relatively small. In order to determine the number of clusters in our data, we tested the silhouette width for values of k from 2 through 10.

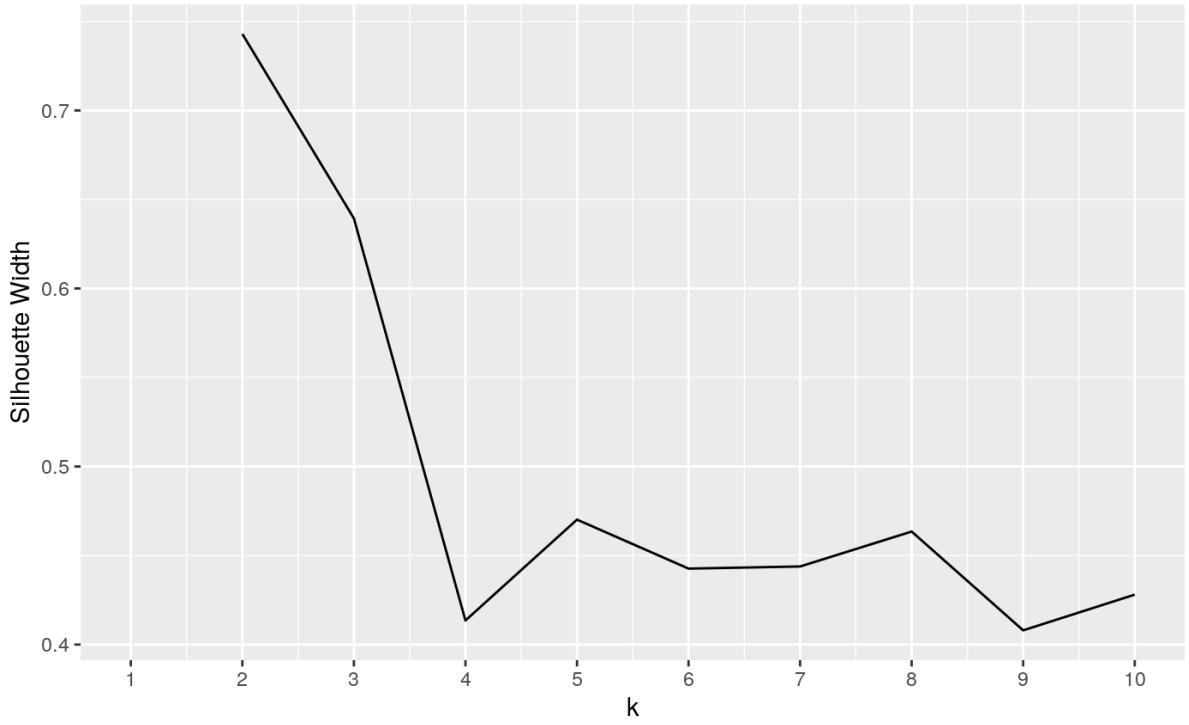

Figure 11. Silhouette width for clusters of 2 through 10

Figure 11 shows each tested k value for the PAM algorithm and its corresponding silhouette width. The silhouette width for a k of 2 surpassed 0.7, which is much larger than the silhouette width for any other value of k, indicating that it is most appropriate to cluster our data into two clusters. After clustering our data, we created a generalized pairs plot to visualize the similarities and differences of each variable across the two clusters.

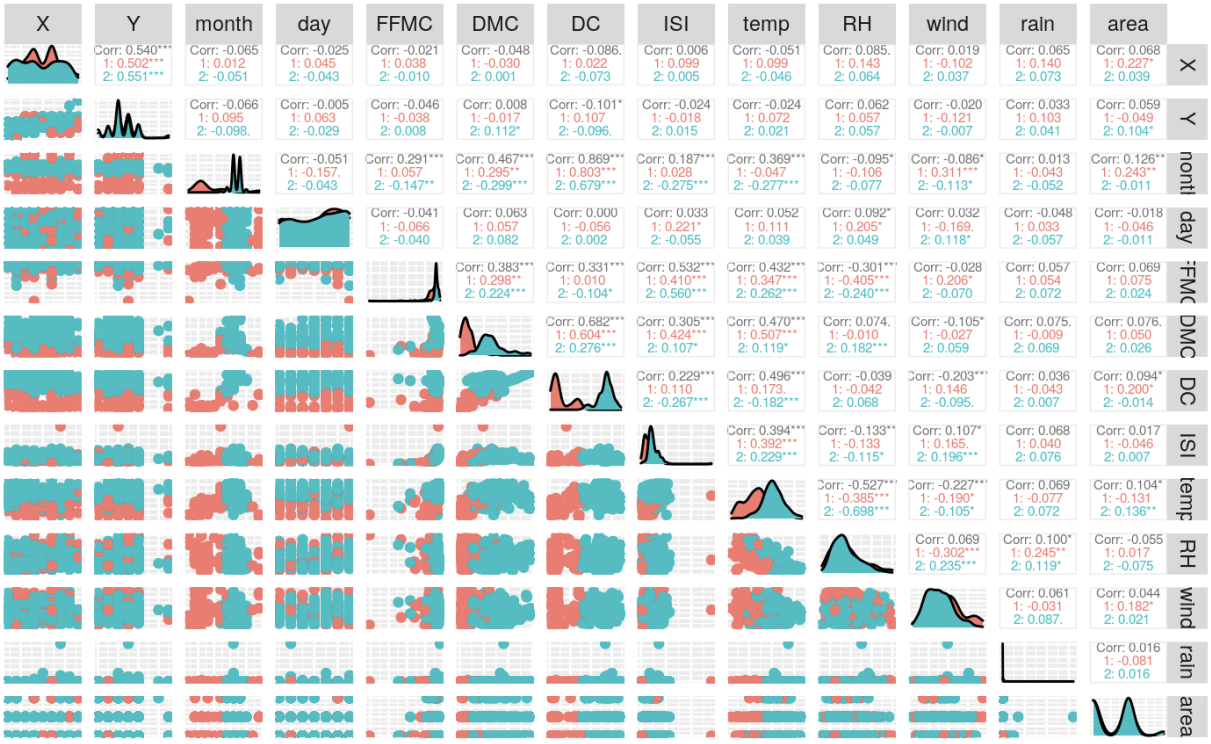

Figure 12. Generalized pairs plot for all 13 variables in dataset

Based on Figure 12, we determined that the two clusters are most different on DC and DMC and similar on relative humidity and wind speed. Interestingly, the clusters seemed to suggest a seasonality to the occurrence of forest fires.

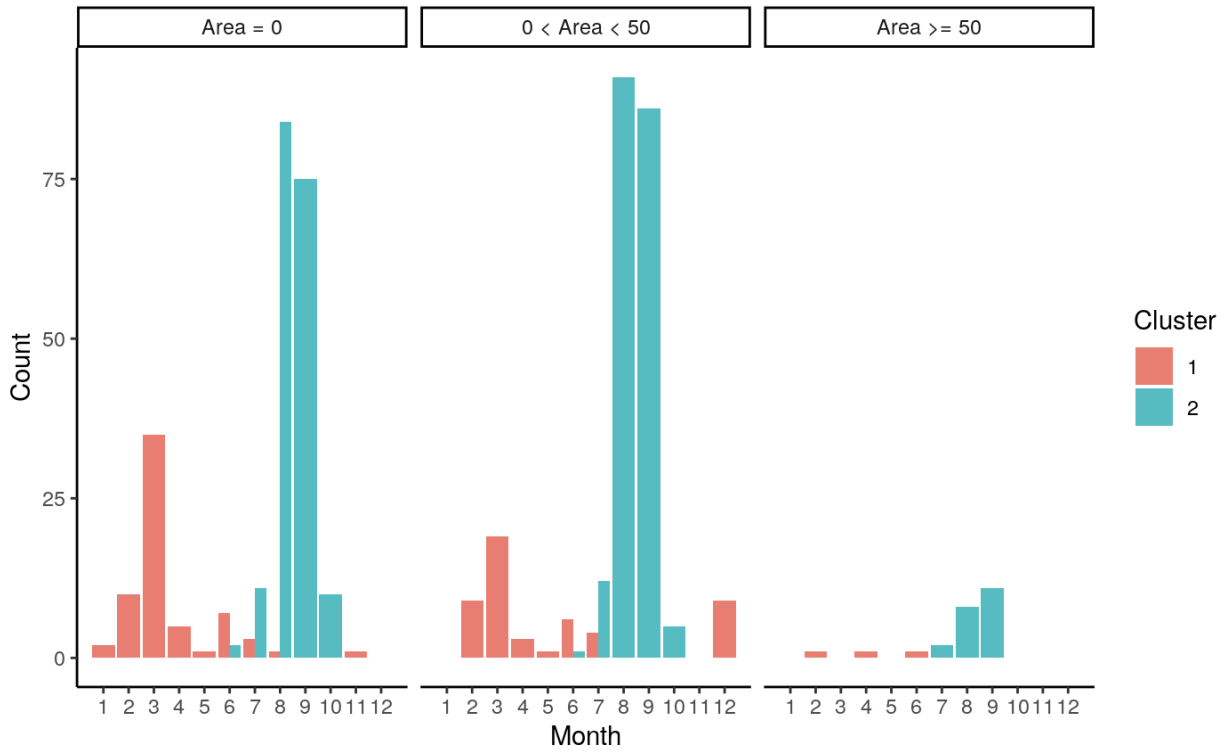

Figure 13. How clusters are spread across each month

To explore the seasonal aspect of forest fires further, we created a plot to see how each area level is distributed across all 12 months and which cluster these observations belong to. In the figure above, observations clustered into cluster 2 are heavily concentrated in August and September, suggesting that these months are part of the season when forest fires typically occur.

Modeling

Our goal is to generate a regression model that predicts the burned area of a forest fire from the meteorological predictors in our data. We started with simple linear regression and chose temperature as our only predictor since in the correlation heatmap, the temperature has the highest correlation with area among the other variables (0.1). Upon examining our initial plot of temperature versus area, we discovered two additional outliers of significantly large area values along with the zero area observations across all temperatures.

We ran a linear regression on both the full data and inlier data (Fig 14), in which we excluded all the zero areas and large areas, or outliers. As expected, temperature and area have a positive linear relationship; however the coefficient, or slope, was still smaller than we expected.

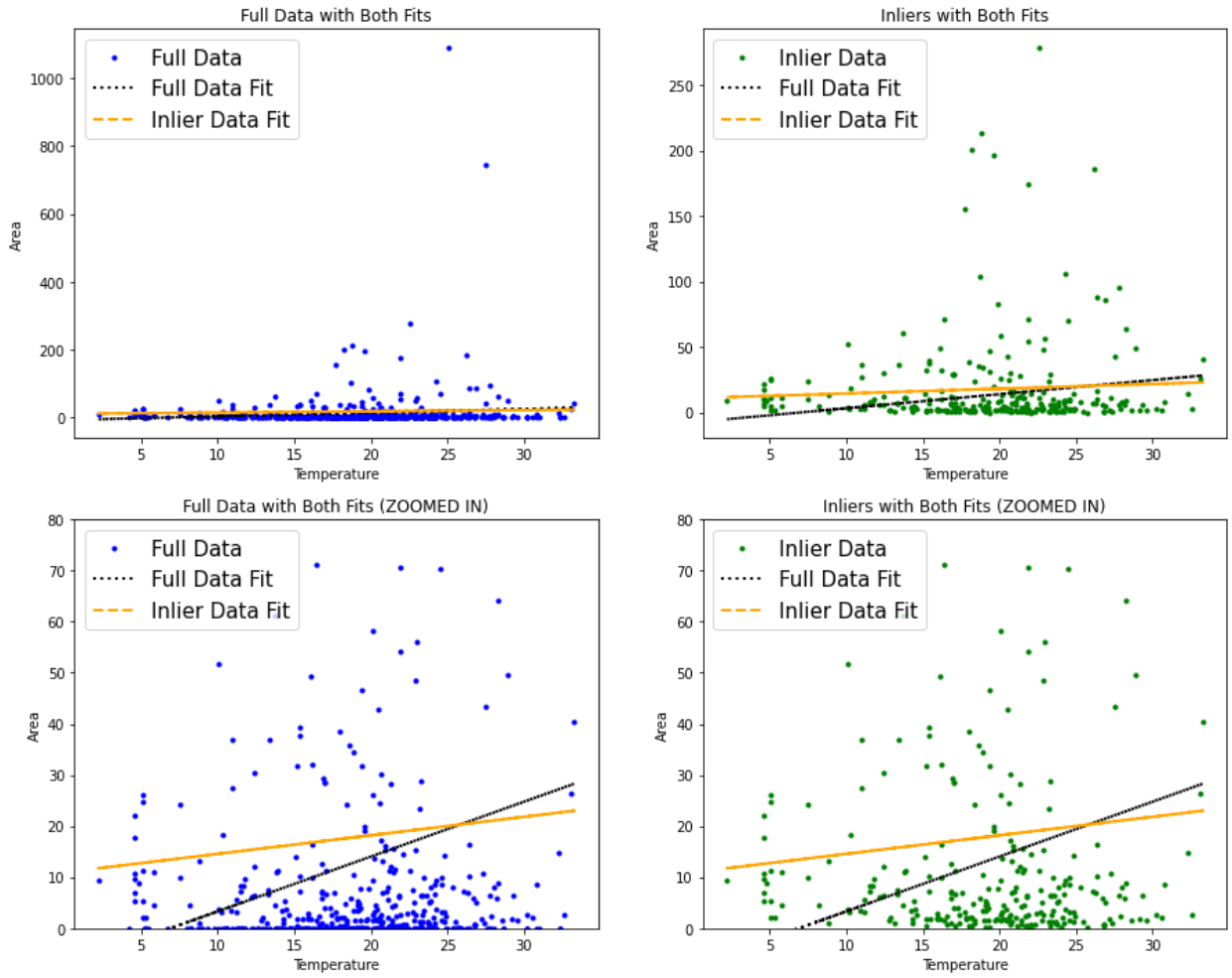

Figure 14. Simple linear regression with temperature to predict area on full and inlier data

We found the mean squared error to compare the accuracy of both models. The linear regression on the full data has an MSE of 4005.508, while the linear regression on the inlier data has an MSE of 1275.5608. It is clear that by removing the outliers from the data, we were able to create a model with less error.

Although we created a linear model to predict the area of a forest fire from temperature alone, this simple linear model does not seem to be optimal for predicting the area of a forest fire. In our regression, we observed a high p-value for the temperature of 0.3, thus temperature alone does not have a significant relationship with area in the inlier data. More than likely, multiple independent variables contribute to the size of a natural fire. We tested this by running three more complex regression models using training and test sets with all 12 predictors as features to identify the optimal regression model to predict area. We only used the inlier data we identified from the simple linear regression to run these new models because we have shown that excluding outlier data runs more accurate regressions.

We began with multiple linear regression on the training set, including all 12 variables, and graphed diagnosis plots to compare predicted and true area (Fig 15).

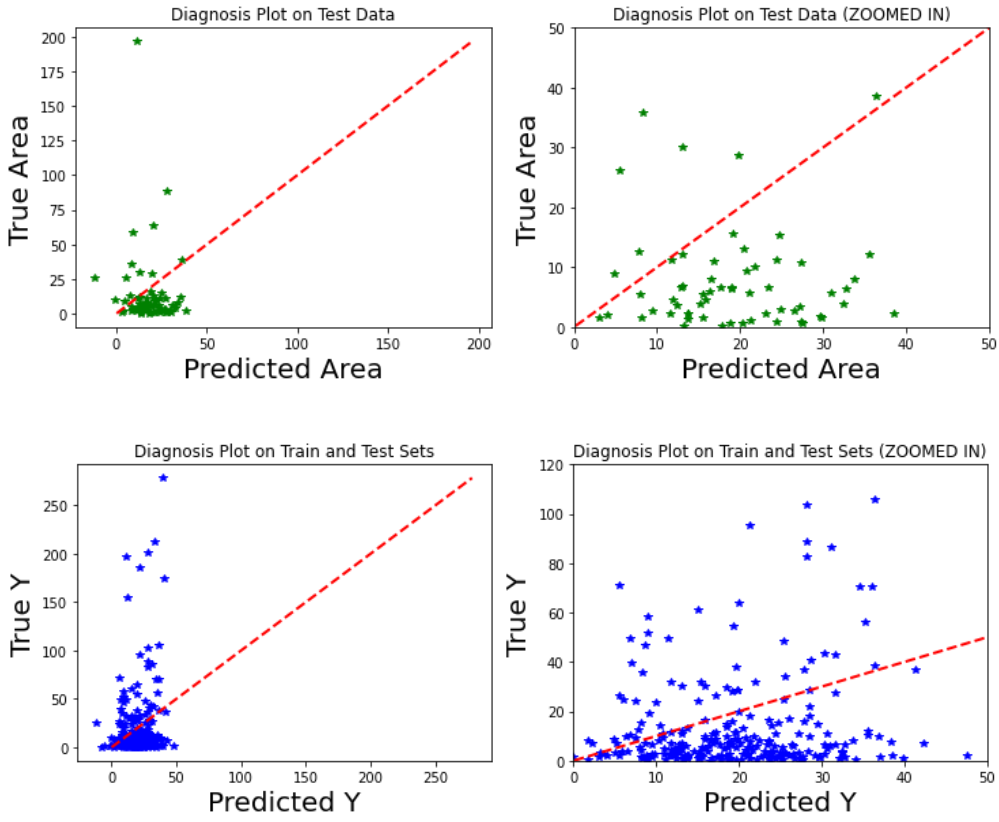

Figure 15. Multiple linear regression diagnosis plots

Then, we considered if the data was not linear, so we used a decision tree regressor to predict the area from all 12 predictors on the inlier data (Fig 16). Decision tree regression is a supervised machine learning algorithm that breaks the data into layers of subsets using true or false classifications. To avoid overfitting, we set the maximum depth of the decision tree to 15.

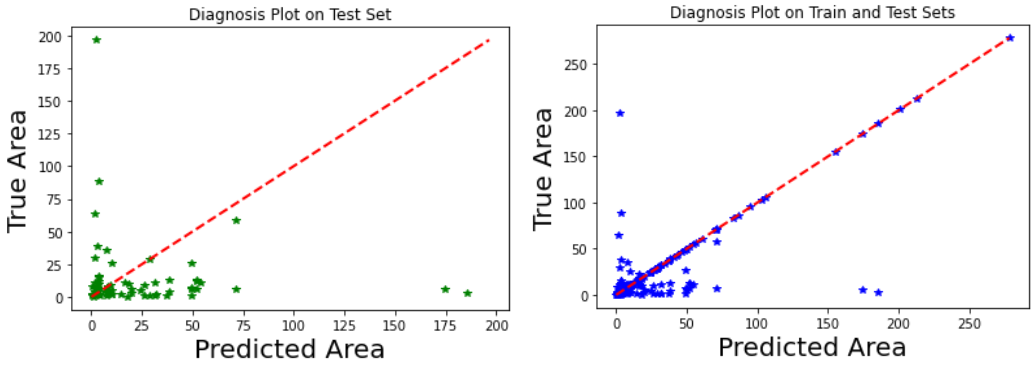

Figure 16. Decision tree regressor diagnosis plots

And lastly, we attempted to formulate one last model using a random forest regressor (Fig 17), with the expectation of predictions with less error than those of the previous decision tree regression. This is because the random forest algorithm consists of multiple decision trees and adds randomness among the trees, so the predictions should be more accurate. Similar to the decision tree model, we avoided overfitting by setting a maximum depth of 10 for the random forest algorithm.

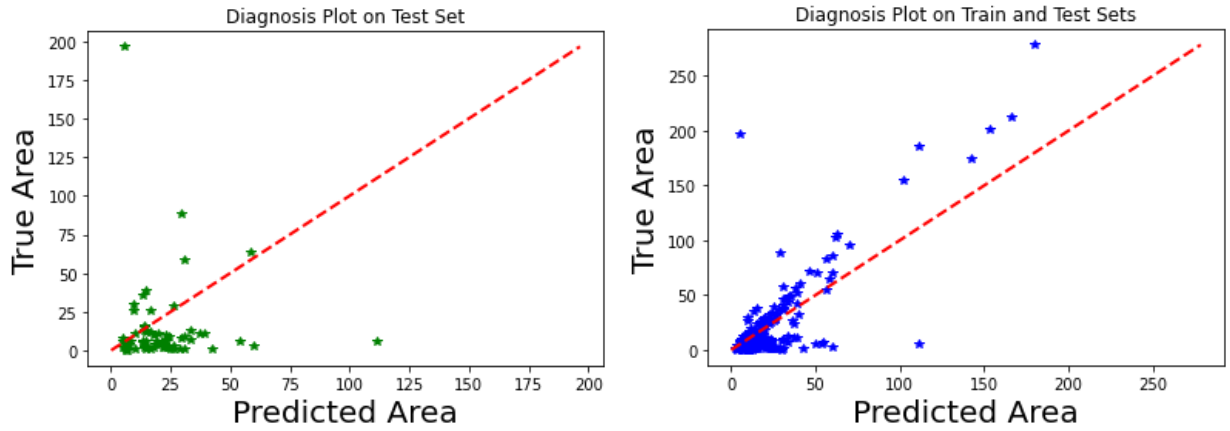

Figure 17. Random forest regressor diagnosis plots

Discussion

For Hypothesis I, we did a scatterplot in ggplot() to visualize the relationship between temperature and drought code. We observed that the drought code changes by seasonal instead of by temperature. Our findings suggest that the national park manager should pay greater attention to the fire risk in the Fall. However, it would be a better analysis if we had more evenly distributed observations by month since August and September data are way more than the rest. For future applications, the monitor should collect the data at the same frequency, and then we can have a better conclusion on whether Fall is at a higher risk of fire than Spring.

For hypothesis II, after conducting the linear regression model analysis and fitting the data to the regression line, we found that the model could predict and confirm our initial hypothesis, concluding that higher temperatures can lead to lower relative humidity levels. Some potential limitations to this analysis could include the appropriate ways in which these variables were collected and whether we can consider their accuracy.

When looking at hypothesis III, it is seen that running a linear regression model given the numbers in our analyzed variables, Initial Spread Index (ISI), and Fine Fuel Moisture Code (FFMC), some significant outliers lead to the linear model not fitting all of the points of the data. Despite this, a relationship between the two variables can be clearly seen and still confirm our initial hypothesis. A potential limitation to this could have been drastic changes in the data collection within the two variables that led to the outliers in the data.

For Hypothesis IV, a location map is built by the X and Y axis from each element in the data set. It is clear to see that the red dots are in the south (Fig. 8), which are the large-scale burned areas. With consideration of wind direction, the wind is coming from the south, and that explains why the south has a larger burned area. The wind helps the fire spread. For application, the national park manager should pay more attention to the south border, and the firefighters should practice arriving at the south border in a short time in spring. Therefore, when the fire happens in Fall, they can react to the fire quickly so that the burned area can be reduced. In addition to the relationship between the location and the scale of the fire, we also use the summarize function to conduct the relationship between location and the frequency of the forest fire. Based on the observation and compared to the Montesinho Natural Park map, we can conclude that the reason that causes the high frequency of forest fires is the different vegetation coverage. The locations with a higher frequency of forest fires are most likely surrounded by shrubland or woodland, which will easily result in tremendous fire.

Based on our clustering of the data, Montesinho park is most at risk for the occurrence of a forest fire, especially more devastating forest fires, during August and September. This means that resource management and those who actively work to fight forest fires should prepare accordingly for these months.

To assess the performance of our three regression models to predict area from all 12 features, we calculated the mean squared error of each. We found that the decision tree regression had the most error, and the multiple linear regression had the least error. Both the random forest regression and multiple linear regression had lower mean squared errors than the simple linear regression model with temperature alone, proving that we were able to create more accurate models to predict area when including multiple variables as features.

We then visualized the SHAP values for all three regression models to determine how much each variable contributed to the predictions of each model (Fig 18). Upon observing the SHAP summary plot of features and their impact on the model output for all three regressions, we identified the Y coordinate, DC (drought code), ISI (initial spread index), and RH (relative humidity) as features that had a consistently high impact on prediction across all of our regression models. This further explains why predicting area from temperature alone did not create an accurate prediction model since temperature does not carry a relatively high SHAP value.

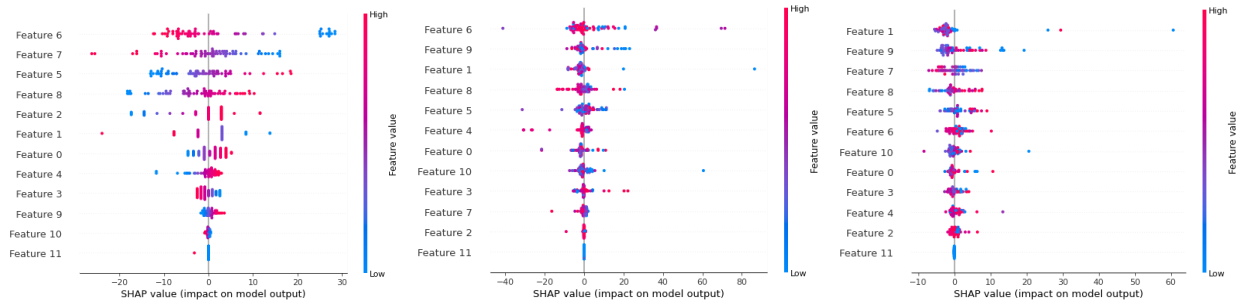

Figure 18. From left to right: SHAP value summary for multiple linear regression, decision tree regression, and random forest regression

Our most accurate model for predicting area was the multiple linear regression with an MSE of about 908.407. However, the model could still be improved since several factors are left unaddressed, and other predictions can be made other than the area measurement itself. We could create subsets of the predictors instead of using all 12 only to include features that have a notable effect on the area. We could also address possible interactions between certain variables. Since there were so many zero area values that we excluded, another next step could be to run logistic regressions to determine if a fire does (area > 0) or does not occur (area = 0), and then if a fire is a small scale (0 < area < 50) or large scale (area >= 50).

Limitations

While the data collected did serve useful in analyzing the factors contributing to a wildfire, a strong prediction of where the next one may occur requires more information. In the future, it would be beneficial to collect data based on the type of vegetation in the area of past wildfires and firefighting intervention, which would provide more context behind the time elapsed to respond to a fire and the strategy used to fight it. Additionally, outlier detection techniques were not used when collecting the data in this data and need to be addressed because forest fires are rare occurrences. Another limitation of our results is that the data we are using is all collected from Montesinho park in Portugal, which has a Mediterranean climate. We cannot safely generalize them to the rest of the world because different parts have different climates.

Conclusion

With this project, we were able to find answers to our hypotheses to help determine the locality of the next wildfire. First, it was discovered that a positive relationship between DC and Temp cannot be determined with this specific data set. Next, the group found a negative linear relationship between humidity levels and temperature. We also noticed a positive exponential relationship between the initial spread index and the fine fuel moisture code. Lastly, when checking if location alone was a factor of wildfire occurrence, it appeared that the line of the park has the least frequency of wildfire but the largest burned area. Through the regression analysis, we wanted to see if we could predict the area of a forest fire from our given predictors.

After modifying the data and running several different regression models, we found that a multiple linear regression that included all 12 predictors achieved the most accurate prediction results with the smallest MSE. We also identified the Y coordinate, drought code, initial spread index, and relative humidity to be consistently strong contributors to predicting area with our regression models.

In all, this data can help protect the forest by making people aware of the contributors of a forest fire and when to be most alert. Park managers should look closely at the highest fire rate area to have sufficient fire control resources at those times. The proposed model is useful for improving firefighting resource management.

Acknowledgments

Name	Percentage Contribution
Ela Albiston	100 %
Jennifer Amador-Gonzalez	100 %
Yihan Du	100 %
Diego Fernandez	100 %
Wenting Lu	100 %
Anna Pham	100 %
Theresa Pham	100 %

Bibliography

Fire weather index (FWI) system. NWCG. (n.d.). Retrieved December 6, 2022, from https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9

Castro, J., Castro, M., & Gómez-Sal, A. (n.d.). *Changes on the climatic edge: Adaptation of and challenges to pastoralism in Montesinho (northern portugal)*. BioOne Complete. Retrieved December 6, 2022, from https://bioone.org/journals/mountain-research-and-development/volume-41/issue-4/MRD-JOURNAL-D-21-00010.1/Changes-on-the-Climatic-Edge--Adaptation-of-and-Challenges/10.1659/MRD-JOURNAL-D-21-00010.1.full