



ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

Final Project

Exploratory and Inferential Analysis of the Forest Fires Dataset

Merve Karaca 2561322

Mürüvvet Eda Koyuncu 2561413

Betül Kunt 2561421

Department of Statistics, Middle East Technical University

STAT250: Applied Statistics

Prof. Dr. Berna Burçak Başbuğ Erkan

Res. Assist. Pelin Erkaya

June, 2024

CONTENTS

Abstract.....	2
Introduction.....	3
Exploratory Data Analysis.....	4
Applying the Necessary Statistical Methods and Approaches for Inference.....	7
Conclusion.....	14
References.....	15

Abstract

In the final project, selected research topic was about forest fires in Montesinho Park (in Portugal) and the variables affecting these fires. Forest fires are a natural disaster affected by different variables and seen in many parts of the world. The aim of this project is to examine the factors affecting forest fires using the forest fires sample created by observing Montesinho Park in Portugal and various statistical methods. The project started by choosing a topic; forest fires, one of today's biggest problems and the devastating effects of global warming, were chosen. Frequently recurring forest fires in Turkey had an impact on the selection of the project topic. After selecting the topic, many sites were searched for the data set to be used and the appropriate sample was found on a GitHub page(Cortez & Morasis,2007). It was checked whether any data cleaning was needed, and it was understood that it was not necessary. In this final project, research questions about forest fires are written and these research questions were examined by using various statistical (multiple linear regression, two sample test and one-way anova) methods, and answered by making the necessary interpretations, and assumptions were checked. Conclusions are drawn and final edits of the report were made.

Introduction

The data for this final project is about forest fires. In this data set there are 11 quantitative variables which are X (x-axis spatial coordinate within the Montesinho park map: 1 to 9), Y (y-axis spatial coordinate within the Montesinho park map: 2 to 9), FFMFC (index from the FWI system: 18.7 to 96.20), DMC (index from the FWI system: 1.1 to 291.3), DC (index from the FWI system: 7.9 to 860.6), ISI (index from the FWI system: 0.0 to 56.19) ,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), area (the burned area of the forest (in ha): 0.00 to 1090.84) and 2 categorical variables which are month and day in total 13 variables. The Fine Fuel Moisture Code (FFMC) is a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel. The Duff Moisture Code (DMC) is a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. This code gives an indication of fuel consumption in moderate duff layers and medium-size woody material. The Drought Code (DC) is a numeric rating of the average moisture content of deep, compact organic layers. This code is a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs. The Initial Spread Index (ISI) is a numeric rating of the expected rate of fire spread. It is based on wind speed and FFMFC. Like the rest of the FWI system components, ISI does not take fuel type into account. Actual spread rates vary between fuel types at the same ISI. ("Canadian Wildland Fire Information System", <https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi>) The research questions for this project are:

1. How do variables in Forest Fires dataset influence the Fine Fuel Moisture Code (FFMC) in Montesinho Park?
2. Is there a significant difference in the average burned area in Montesinho Park between different months of the year?
3. What is the relationship between the burned area and the temperature in Montesinho Park?

In this report group members' aims are finding the research questions' answer by applying statistical methods which are taught by instructor.

Exploratory Data Analysis

Firstly, the csv file was obtained from the github page and read by R. To check datasets' structure and the variables type(numeric,categorical,etc), `str()` command was used. It is determined that dataset consists of a data frame and there are 517 observations of 13 different variables which are 3 integers, 2 characters, 8 numeric FFMC,DMC,DC,ISI,temperature, wind,rain and area are quantitative and continuous variables. Then, to check the existence of missing values, `anyNA()` command was used. Since no missing values are detected, there is no need to deal with missing values. To calculate the summary statistics of numeric variables, `summary()` command was used.

	FFMC		DMC		ISI		Temperature		RH		Area
Minimum	18.70	Minimum	1.1	Minimum	0.000	Minimum	Şub.20	Minimum	15.00	Minimum	0.00
1st Quartile	90.20	1st Quartile	68.6	1st Quartile	6.500	1st Quartile	15.50	1st Quartile	33.00	1st Quartile	0.00
Median	91.60	Median	108.3	Median	8.400	Median	19.30	Median	42.00	Median	0.52
Mean	90.64	Mean	110.9	Mean	9.022	Mean	18.89	Mean	44.29	Mean	12.85
3rd Quartile	92.90	3rd Quartile	142.4	3rd Quartile	10.800	3rd Quartile	22.80	3rd Quartile	53.00	3rd Quartile	6.57
Maximum	96.20	Maximum	291.3	Maximum	56.100	Maximum	33.30	Maximum	100.00	Maximum	1090.84

According to output, it can be easily seen the range, the mean, the median, the first quartile and third quartile of the numeric variables, also it can be easily seen the length, the class and the mode of the categorical variables. Moreover, categorical variables' proportion and frequencies were examined.

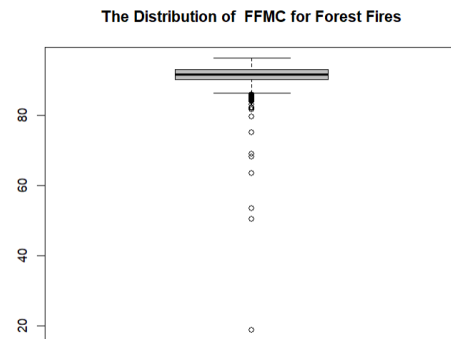
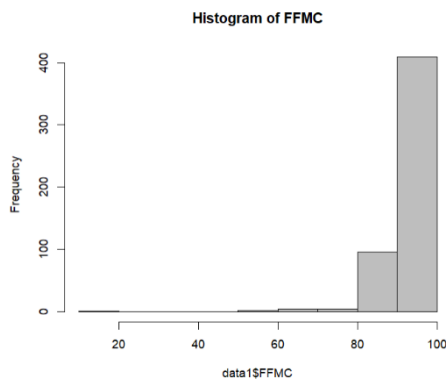
Month	Count	Percentage
January	2	0.00387
February	20	0.0387
March	54	0.104
April	9	0.0174
May	2	0.00387
June	17	0.0329
July	32	0.0619
August	184	0.356
September	172	0.333
October	15	0.0290
November	1	0.00193
December	9	0.0174

The table shows the frequency and percentage of fires by month. It also shows that August is the month with the highest frequency of forest fires and January and May have the lowest frequency.

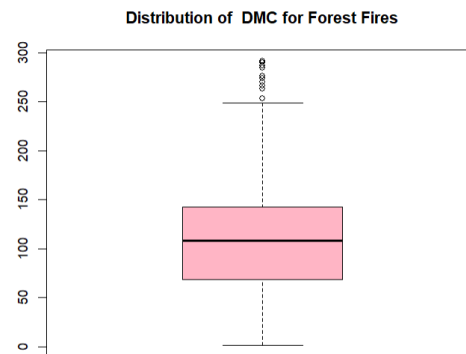
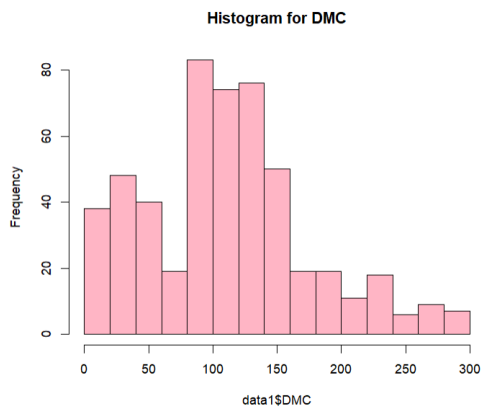
Day	Count	Percentage
Monday	74	0.143
Tuesday	64	0.124
Wednesday	54	0.104
Thursday	61	0.118
Friday	85	0.164
Saturday	84	0.162
Sunday	95	0.184

Looking at the table, we can see the proportion and frequency of fires by day, as well as the fact that Wednesday is the least common day for forest fires and Sunday is the most common day.

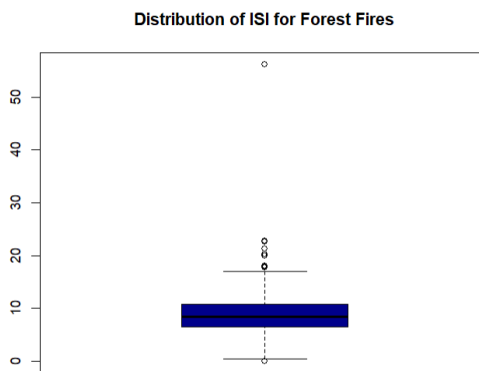
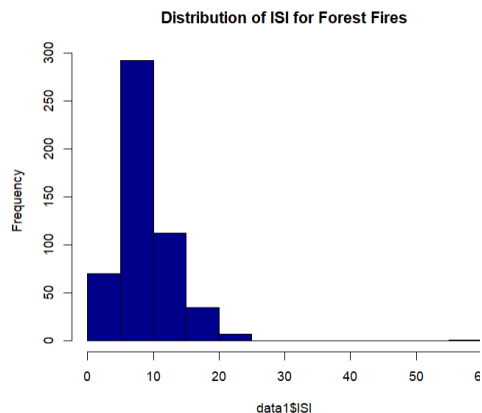
To explore the distribution of numeric variables, histogram was used. To detect outliers, box plot was used.



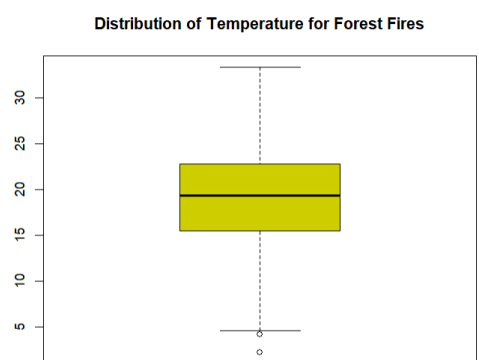
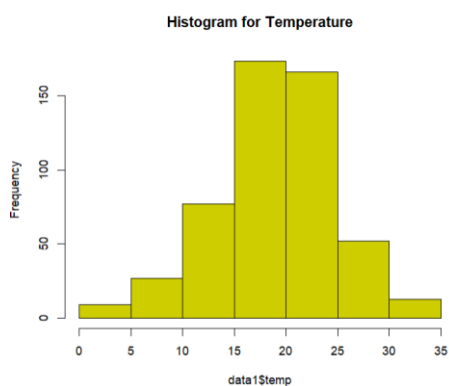
By looking this histogram, it can be said the distribution of FFMC for Forest Fires has left skewed distribution. (Mean<Median<Mode) There are too many outliers.



This histogram indicates that there is a right-skewed distribution in the DMC for forest fires. (Mean >Median >Mode) Some outliers exist.

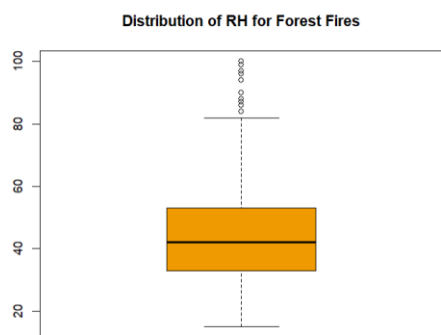
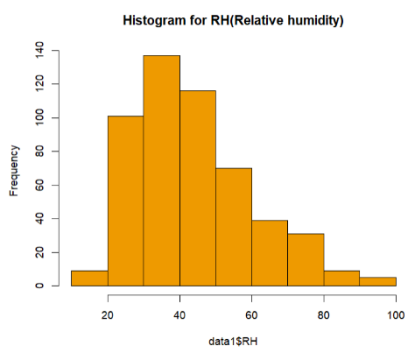


There appears to be a right-skewed distribution of ISI. (Mean>Median>Mode) The box plot was used to identify a few outliers.

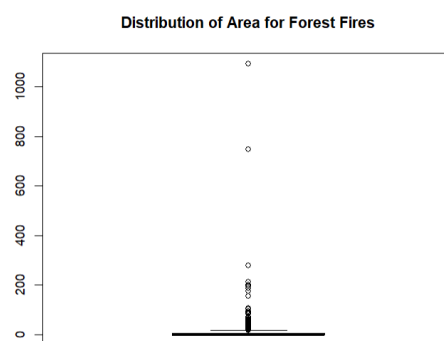
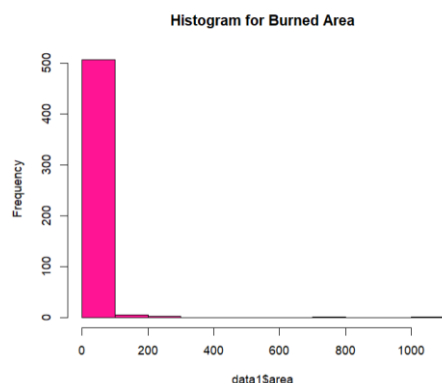


The skewness of temperature cannot be commented directly from this graph. skewness() command was used

and the result (-0.3302106) which is negative show that the distribution of temperature is left skewed. There are a few outliers.



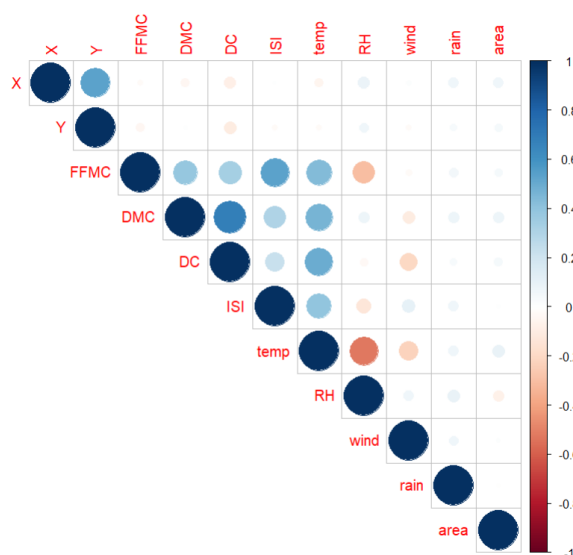
Observing this histogram, one might conclude that there is a right skewed distribution in the RH for forest fires. (Mean>Median>Mode) .Too many outliers were detected by looking the boxplot.



When looking at the histogram, a right-tailed distribution was concluded. By looking at the boxplot, it was determined that there

were a few outliers. Since the range was very wide, no comment could be made about the interquartilerange, and by looking at the summary function's outplot, it was seen that it was a very small value (6.57).

The heatmap were created to examine the relationship between numeric variables.



Analyzing values near 1 or -1, which imply high positive or negative correlations between predictors, is part of the correlation matrix assessment process.

The presence of multicollinearity may be indicated by notably high correlation coefficients between independent variables. It is imperative, therefore, to officially evaluate multicollinearity through the application of recognized statistical techniques. Using FFM as the response variable, we can conclude that ISI has the strongest

correlation between the response and covariates, indicating a moderate relationship as well. Additionally, there is a 0.68 correlation relationship between DC and DMC. Therefore, multicollinearity may lead us to suspect.

Applying the Necessary Statistical Methods and Approaches for Inference

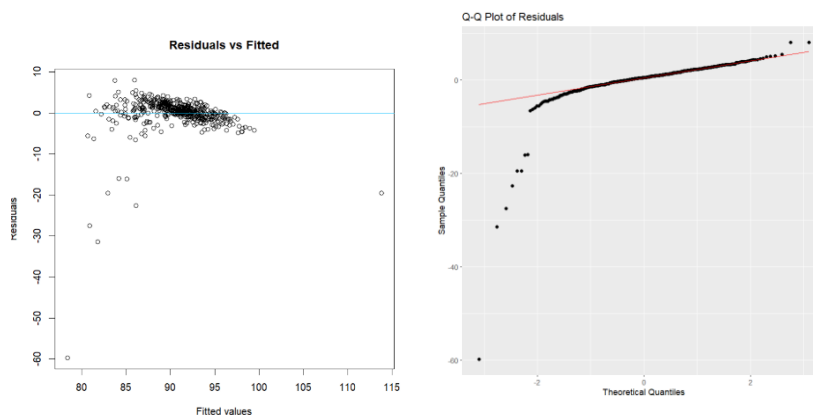
1. How do variables in Forest Fires dataset influence the Fine Fuel Moisture Code (FFMC) in Montesinho Park?

To answer this research question, a multiple linear regression model was created, with the FFMC variable being dependent and the other variables being independent.

	Estimate	Std.Error	t.value	Pr...t..	
(Intercept)	88,4895636	1,5818141	55,942	< 2e-16	The starting point for FFMC
X	0,0772852	0,0975518	0,792	0.429	prediction, known as the
Y	-0,1494434	0,1839865	-0,812	0.417	intercept, is 88.49 , assuming
DMC	0,0209661	0,0044274	4,736	2.84e-06	all other factors are at zero.
DC	0,0015668	0,0011047	1,418	0.157	DMC, ISI, and RH have
ISI	0,5072124	0,0468018	10,837	< 2e-16	noticeable effects on FFMC
temp	-0,0389173	0,0529872	-0,734	0.463	with each unit increase: 0.02 ,
RH	-0,0964844	0,0153943	-6,268	7.86e-10	0.51, and -0.10, respectively.
wind	-0,0815402	0,1124007	-0,725	0.469	
rain	0,737104	0,650585	1,133	0.258	
area	-0,0001774	0,0029909	-0,059	0.953	

Only DMC, ISI, and RH are found to significantly impact the model. This model accounts for **roughly 41% of the variation in FFMC**, which is a moderate amount. **The residual standard error, at 4.278**, indicates the typical deviation of the residuals, ideally kept low. **The multiple R-squared value of 0.4111** implies that approximately **41.11%** of the FFMC variance is explained by this model, suggesting a reasonable fit. **The adjusted R-squared, slightly lower at 0.3995**, adjusts for the number of predictors but still indicates a decent model fit. The **p-value**, smaller than 0.05, confirms the overall statistical significance of the model.

Then the assumptions were checked:



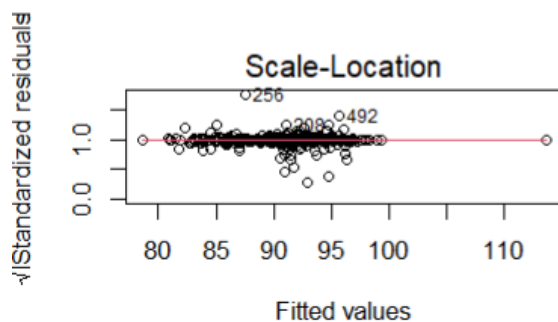
Looking at the residuals vs fitted graph, there is an evidence of a linear relationship is a horizontal line devoid of any discernible patterns. The assumption of linearity is met.

Normal Q-Q used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line. Therefore data' points follow the straight line. The assumption of normality of the residuals is met.



Scale-Location (or Spread-Location) used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.

In the plot shown on the side, the residual points do not seem to be evenly scattered, indicating that this particular assumption is not fulfilled. To address this, It was applied the inverse of the variance to meet the assumption.



Then, after the inverse of variance transformation the following plot has been created and assumption was met, residual points are almost equally spread.

Following assumptions also were checked:

X	Y	DMC	DC	ISI
1.436563	1.443851	2.267273	2.117680	1.284005
Temperature	RH	Wind	Rain	Area
2.669317	1.779244	1.143553	1.045400	1.022111

VIF results show that there is no multicollinearity problem since all values are smaller than 5.

shapiro.test()Result pvalue

FFMC <2.2e16

Since the p-value is less that significance level, FFMC is normally distributed.

lag	Autocorrelation	D.W.Statistic	p.value	Alternative.hypothesis
1	0,00863704	1,97885	0,558	rho != 0

1

The null hypothesis states that the

errors are not auto-correlated with themselves.(They are independent)

Various transformations were tried to satisfy the assumption. The summary of the model that provided the largest adjusted r square value was examined and the fitted model equation was established with the most affecting variables. Model 6 provides the greatest adjusted r squared value, as determined through the application of different transformations. The model with the greatest influencing variables was Model 11.

Variable	Estimate	Std.Error	t.value	Pr...t..
(Intercept)	81,8235478	1,4349601	57,021	< 2e-16
X	0,0562105	0,0855992	0,657	0.511691
Y	-0,12536	0,1614157	-0,777	0.437741
DMC	0,0153098	0,003912	3,914	0.000103
DC	0,0011481	0,0009688	1,185	0.236542
sqrt(ISI)	4,5625932	0,2617848	17,429	< 2e-16
temp	-0,0972593	0,0463116	-2,1	0.036214
RH	-0,0888818	0,0134914	-6,588	1.12e-10
wind	-0,1884649	0,0986023	-1,911	0.056522
rain	0,6570108	0,5708912	1,151	0.250337
area	0,0003603	0,0026242	0,137	0.890850

Variable	Estimate	Std.Error	t.value	Pr...t..
(Intercept)	79,415068	0,880294	90,214	< 2e-16
DMC	0,015378	0,002821	5,452	7.74e-08
sqrt(ISI)	4,38166	0,247062	17,735	< 2e-16
RH	-0,072971	0,010411	-7,009	7.58e-12

The fitted model is $\hat{y} = 79.415068 + \text{DMC} * 0.015378 + \text{sqrt}(\text{ISI}) * 4.381660 + (-0.072971) * \text{RH}$

2: Is there a significant difference in the average burned area in Montesinho Park between different months of the year? This research question aimed to determine whether there is a significant difference in the average burned area in Montesinho Park (Portugal) between different months of the year.

H0: There is no significant difference in the average burned area in Montesinho Park between different months of the year. ($\mu_1 = \mu_2 = \mu_3 = \dots = \mu_{12}$)

H1: There is a significant difference in the average burned area in Montesinho Park between different months of the year. (At least one pair of monthly averages is significantly different.)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
month	11	11453	1041	0.253	0.993
Residuals	505	2079412	41118		

The test showed the p-value for the month factor is 0.993, which is much higher than the significance level of 0.05 (α). (Fail to reject null hypothesis) There is no significant difference in the average burned area in Montesinho park across different months of the year. Assumptions were checked to test the reliability of this result. Firstly, Levene's Test was applied to check the homogeneity of variances.

H0: The variances of burned area across different months are not significantly different. ($\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \dots = \sigma^2_{12}$)

H1: The variances of burned area across different months are significantly different. (At least one pair of monthly variances is significantly different.)

Levene's Test for Homogeneity of Variance			
	Df	F value	Pr(>F)
group	11	0.2632	0.9918

The p-value (0.9918) is much greater than the significance level of 0.05 (α). Therefore, the null hypothesis cannot be rejected. Thus, the assumption of homogeneity is met. Secondly, Shapiro-Wilk Test was applied to check the normality of residuals.

H0: The data for the burned area in Montesinho Park for each month is normally distributed.

H1: The data for the burned area in Montesinho Park for each month is not normally distributed. (At least one month does not follow a normal distribution)

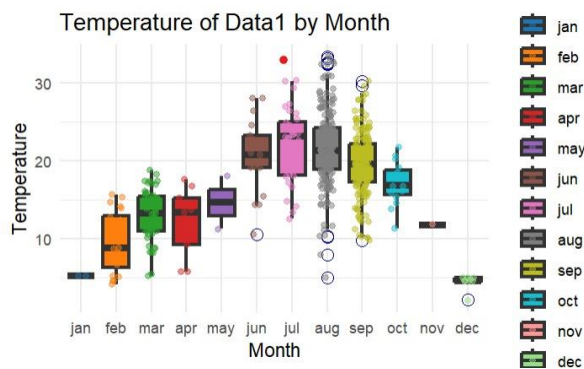
shapiro.test() Result	Pvalue
Residuals of one-way ANOVA	<2.2e16

According to this test, it can be said the assumption of the normality of residuals is not satisfied since p-value (less than $2.2e-16$) is extremely smaller than the significance level of $0.05(\alpha)$. Therefore, the null hypothesis was rejected. After determining that residuals of ANOVA are not normally distributed, it was decided that Kruskal-Wallis Test (a nonparametric alternative) which does not assume normality would be more reliable, and it was applied.

kruskal.test()	Results
Data	Burned Area by Month
Kruskal Wallis	
chisquare	23.723
df	11
p-value	0.01396

That the p-value (0.01396) is less than 0.05 indicates a significant difference in the burned area in Montesinho Park across different months.

3. What is the relationship between the burned areas and temperature in Montesinho Park?



This research question began to be answered by creating the graph on the side. The distribution of temperature by month is visualized in the graph on the side. Looking at the graph, it was determined that one of the hottest month was August and one of the coldest month was December. Then, the data

for December and August were converted into two separate samples in R using the filter method. By using area data from August and December, it was aimed to determine whether the areas burned in these two months were equal. For this, the following hypothesis was used.

μ_1 : the mean of burned area for August μ_2 : the mean of burned area for December

H_0 : $\mu_1 = \mu_2$ (There is no significant difference between the average burned area in August and December)

H_1 : $\mu_1 \neq \mu_2$ (There is a significant difference between the average burned area in August and December)

Necessary assumptions were checked before comparing two sample means. First, using `shapiro.test()`, it was determined that two samples were not normal.

Shapiro.test() Results	p.value
For August	2.2e-16
For December	0.136

According to the test result, the distribution was not normal for August data since the p-value is smaller than 0.05. However, the

sample size of data was sufficiently large enough. Therefore, the sampling distribution of sample mean was approximately normally distributed by Central Limit Theorem. Also, according to test result since the p-value is bigger than 0.05, the distribution was normal for December data. After checking normality, equality of variances was checked. In order to check equality of variances, `var.test()` is used.

σ^2_1 : the variance of burned area for August σ^2_2 : the variance of burned area for December

A t-test revealed that, contrary to expectations, there was no significant variation in the average burned area between August and December. This finding implies that even while December is one of the coldest and August is one of the hottest months, temperature may not be the only factor influencing how much land is destroyed by flames.

H_0 : $\sigma^2_1 = \sigma^2_2$ (There is no significant difference between the variance for burned area in August and December)

H_1 : $\sigma^2_1 \neq \sigma^2_2$

Var.test() Results	
F	83.379
Number of df	183
Denominator of df	8
P-value	4.5e-07
95 Percent Confidence Interval	22.48507 , 188.59161
Ratio of Variances	8.337

Since the p-value is smaller than 0.05 which is our alpha value, the null hypothesis can be rejected. Therefore, it can be said that the variances of two populations are not equal. After checking assumptions, `t.test()` was used to compare two sample means.

T.test() Results

t	-0.16934
df	119.46
P-value	0.8658
Mean of August	12.489
Mean of December	13.330

Since the p-value is bigger than 0.05, the null hypothesis was failed to reject. Therefore, it can be said that the mean of August and December is equal to each other.

Conclusion

This project used a large dataset and a variety of statistical techniques to investigate the factors driving forest fires in Montesinho Park, Portugal. The primary research issues on the variables influencing the Fine Fuel Moisture Code (FFMC), the seasonal variations in burned area, and the correlation between temperature and burned area were addressed by these analyses. Through the application of multiple linear regression, it was determined that the following factors strongly affect FFMC: relative humidity (RH), initial spread index (ISI), and drought code (DMC). This model demonstrated a decent fit, explaining roughly 53.57% of the variance in FFMC. This demonstrates how important a role these factors play in determining the park's susceptibility to fire. The analysis revealed that there was no significant difference in the average burn area in different months, as demonstrated by the ANOVA test with a p value of 0.993. This result shows that the average burned area does not change significantly from month to month. However, this test is unreliable because the normality of the residuals assumption is not met. The Kruskal-Wallis test, a non-parametric alternative, showed a significant difference in the average burned area in different months with a p value of 0.01396. A t-test revealed that, contrary to expectations, there was no significant variation in the average burned area between August and December. This finding implies that even while December is one of the coldest and August is one of the hottest months, temperature may not be the only factor influencing how much land is destroyed by flames. This research makes significant advances to our understanding of forest fires, which is becoming more and more important in light of global climate change and an increase in the frequency of fires worldwide. It does this by utilizing statistical approaches and data analysis.

References

Canada, N. R. (n.d.). *Canadian wildland fire information system: Canadian forest fire weather index (FWI) system*. Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System. <https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi>

joanby. (n.d.). Python Machine Learning Course Datasets: Forest Fires. Retrieved [June 12, 2024], from <https://github.com/joanby/python-ml-course/tree/master/datasets/forest-fires>

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,

Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

Available at: <http://www.dsi.uminho.pt/~pcortez/fires.pdf>