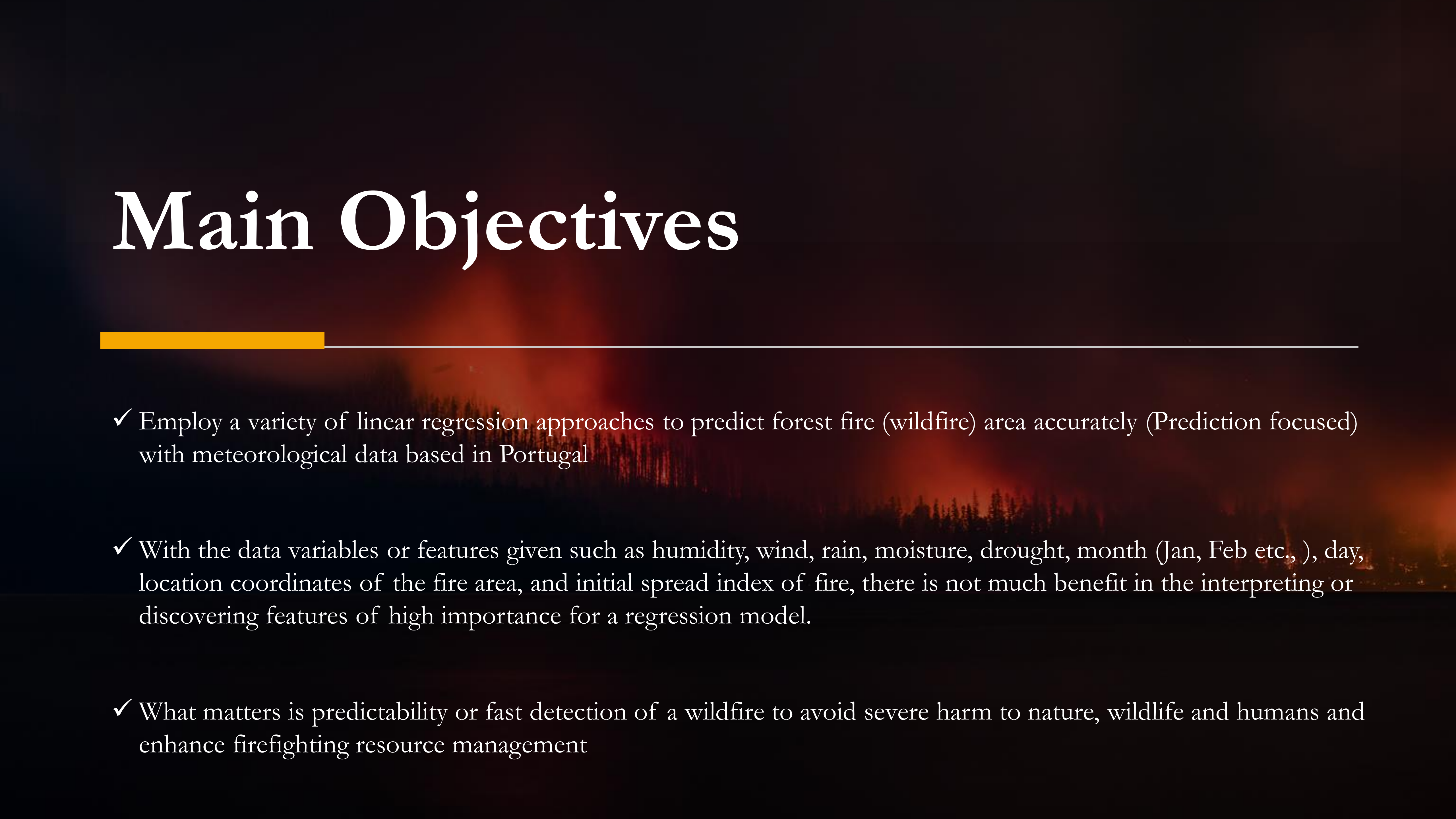
Forest Fires

Areeb Shafqat

22nd November 2021



Main Objectives

- ✓ Employ a variety of linear regression approaches to predict forest fire (wildfire) area accurately (Prediction focused) with meteorological data based in Portugal

- ✓ With the data variables or features given such as humidity, wind, rain, moisture, drought, month (Jan, Feb etc.,), day, location coordinates of the fire area, and initial spread index of fire, there is not much benefit in the interpreting or discovering features of high importance for a regression model.

- ✓ What matters is predictability or fast detection of a wildfire to avoid severe harm to nature, wildlife and humans and enhance firefighting resource management

Data Summary

All feature variables:

- X–axis spatial coordinate
- Y-axis spatial coordinate
- Month of the Year
- Day of the week
- Fine Fuel Moisture Code (Moisture Conditions)
- Duff Moisture Code (DMC)
- Drought Code (DC)
- Initial Spread Index (ISI)
- Temperature (Celsius Degrees)
- Relative Humidity
- Wind Speed (Km/h)
- Outside Rain (mm/m2)

- Data is available at https://archive.ics.uci.edu/ml/datasets/Forest+Fires
- There are 517 observations and 13 columns (2 categorical which are Month & Day and the rest numerical)
- No null and duplicated observations
- Area burnt (Target variable) is measured in hectares from 0.00 to 1090.84
- The target is very much skewed towards 0.0 and hence a logarithm transform will be performed later.

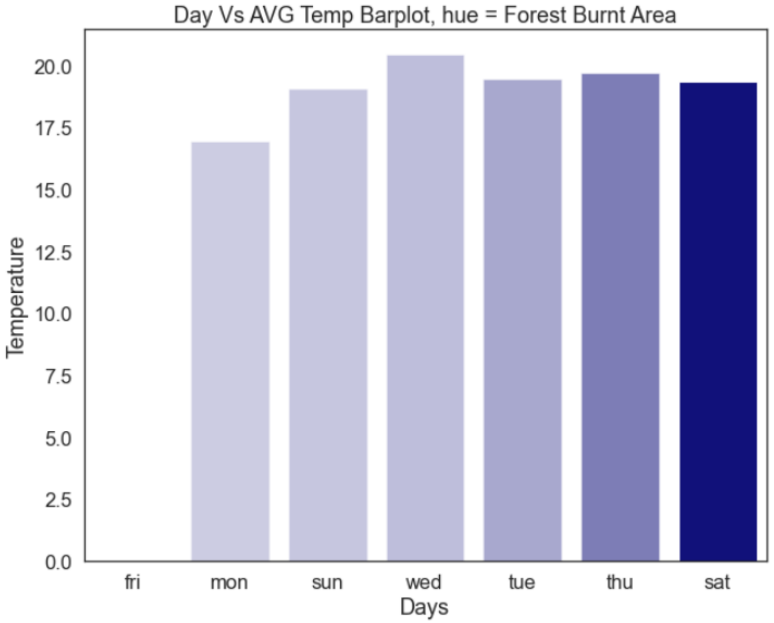
Data Exploration and Engineering

Generated Insights

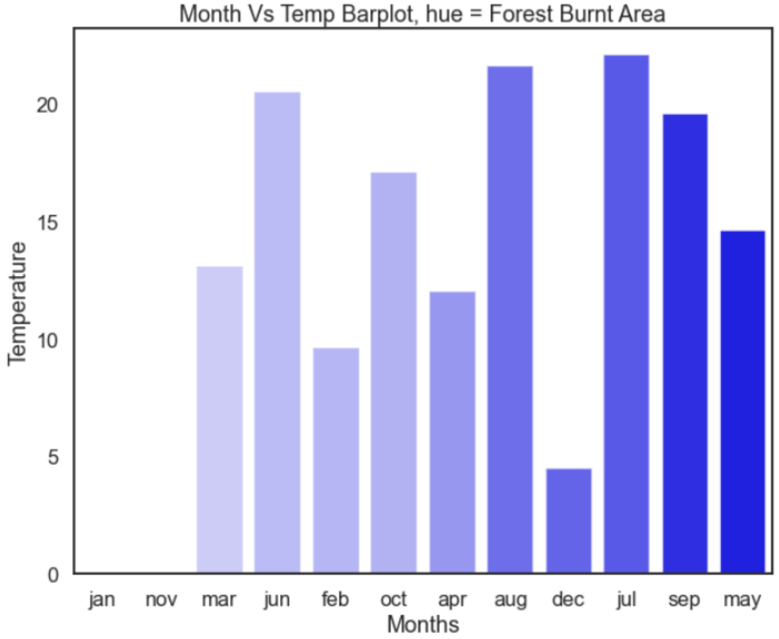
- From the first plot, we can infer that Saturday and Thursday are days where there is most likely to be a larger forest wildfire relative to other days of the week.

- Based on both the plots of temperature against months and days, it is evident that a higher temperature does not necessarily lead to larger area forest fires where for example, contrary to what one might expect, December has one of the highest forest burnt area on average (The color concentration represents the mean area of forest fire for a month or day).

- The bottom plot indicates that September, July and May are the months to look out for when preparing for wildfires.



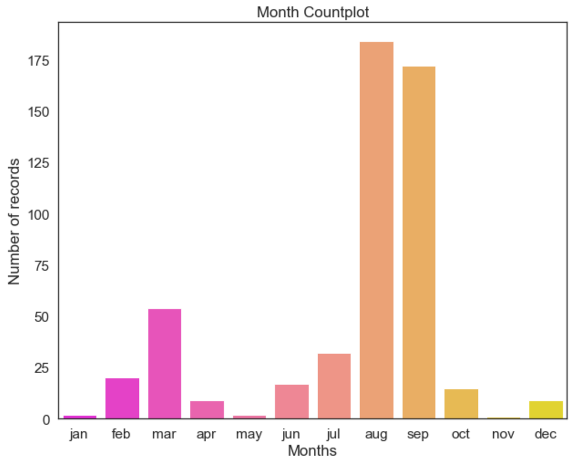
Day Vs AVG Temp Barplot, hue = Forest Burnt Area

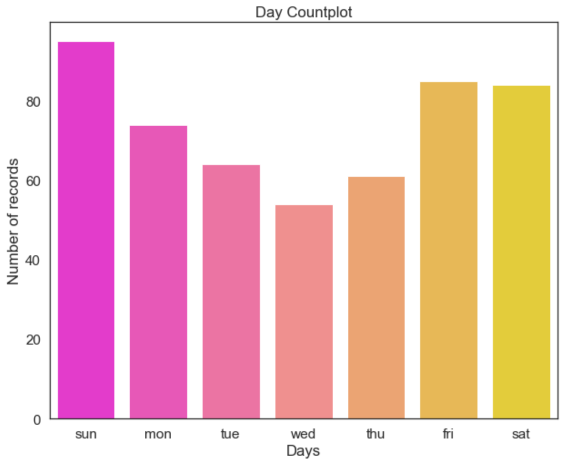


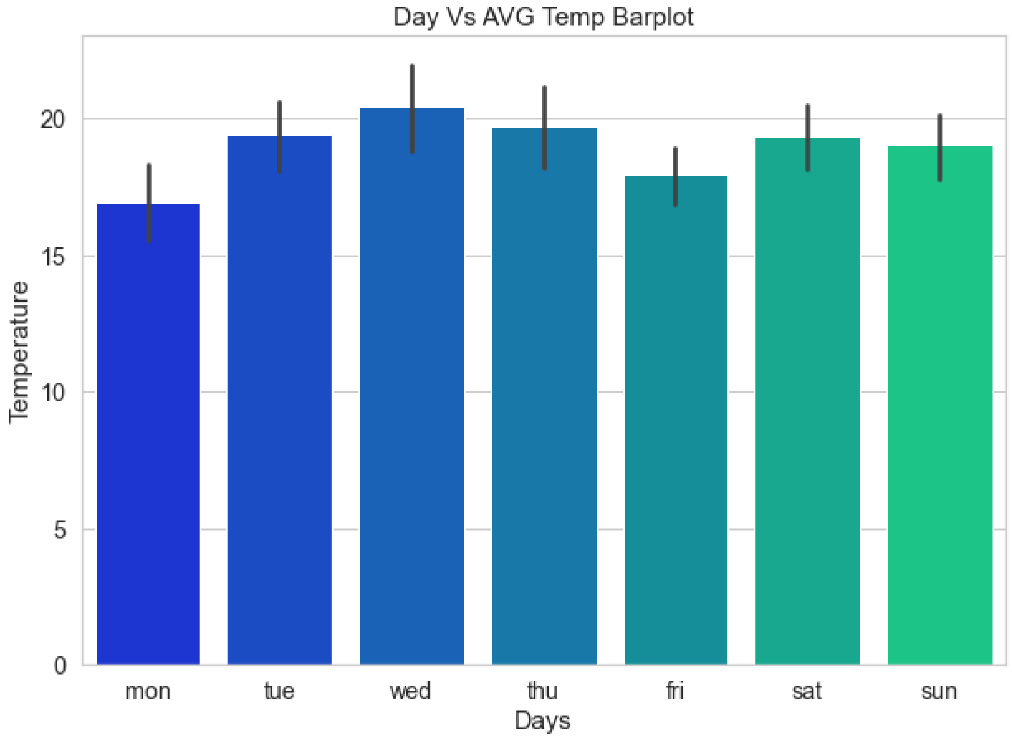
Month Vs Temp Barplot, hue = Forest Burnt Area

Generated Insights

- The plots shown here indicate the number of times a forest fire event was recorded and the temperature variation between days
- The greatest number of wildfires recorded were on Sundays, Saturdays and Fridays
- The most frequent months in which forest fires occurred were August and September although this does not mean that these fires were the most severe which was discussed on the previous slide.

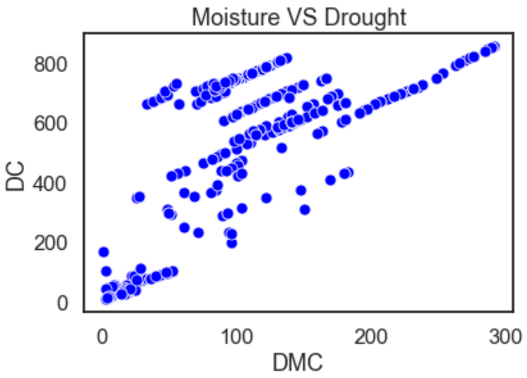


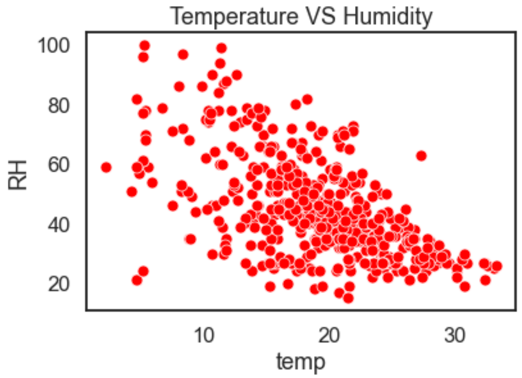


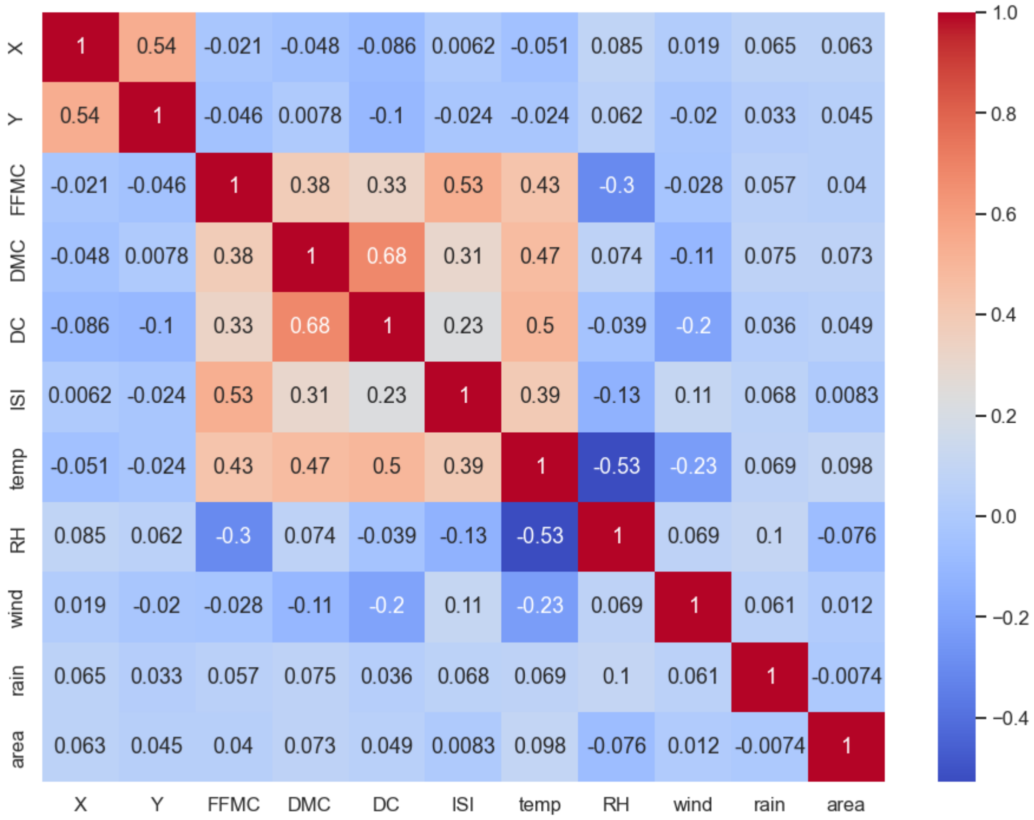


Generated Insights

- Looking at relationships between variables, we find that temperature and humidity have a negative correlation which is intuitive since higher temperatures result in lower humidity's.
- As for Duff Moisture (DMC) and Drought, these have a strong positive correlation where DMC indicates fire extinguishment
- Other correlations include between the X and Y coordinate, temperature and DC, and ISI and FFMC







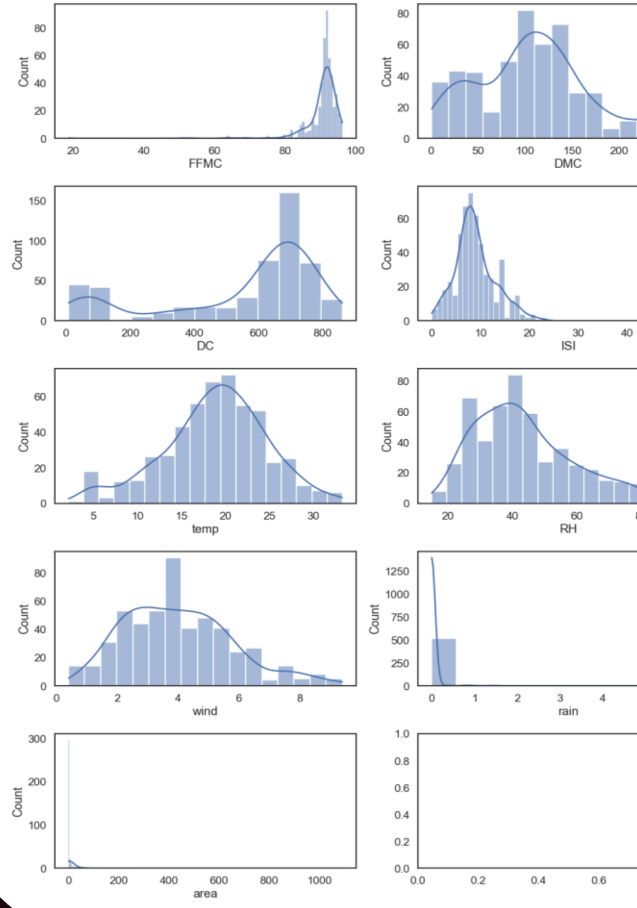
	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1	0.54	-0.021	-0.048	-0.086	0.0062	-0.051	0.085	0.019	0.065	0.063
Y	0.54	1	-0.046	0.0078	-0.1	-0.024	-0.024	0.062	-0.02	0.033	0.045
FFMC	-0.021	-0.046	1	0.38	0.33	0.53	0.43	-0.3	-0.028	0.057	0.04
DMC	-0.048	0.0078	0.38	1	0.68	0.31	0.47	0.074	-0.11	0.075	0.073
DC	-0.086	-0.1	0.33	0.68	1	0.23	0.5	-0.039	-0.2	0.036	0.049
ISI	0.0062	-0.024	0.53	0.31	0.23	1	0.39	-0.13	0.11	0.068	0.0083
temp	-0.051	-0.024	0.43	0.47	0.5	0.39	1	-0.53	-0.23	0.069	0.098
RH	0.085	0.062	-0.3	0.074	-0.039	-0.13	-0.53	1	0.069	0.1	-0.076
wind	0.019	-0.02	-0.028	-0.11	-0.2	0.11	-0.23	0.069	1	0.061	0.012
rain	0.065	0.033	0.057	0.075	0.036	0.068	0.069	0.1	0.061	1	-0.0074
area	0.063	0.045	0.04	0.073	0.049	0.0083	0.098	-0.076	0.012	-0.0074	1

Data Cleaning Steps

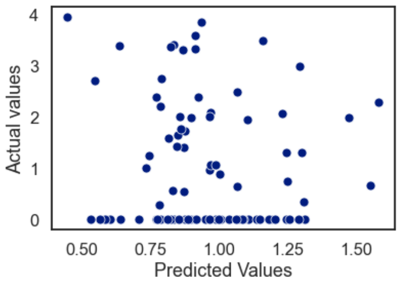
Removal of all outliers with a z-score or standard deviation above 4

Convert categorical columns like Months and Days to numerical e.g., (Jan – 0, Feb – 1,…) , (Sun – 0, Mon - 1…)

To remove much of the skewness (>0.75) of all variables including area, a natural logarithm transform was performed on all the data (np.log1(data))



Linear Regressions



Polynomials Features = 2

Standard Scalar

Same Train Test Splits

Vanilla

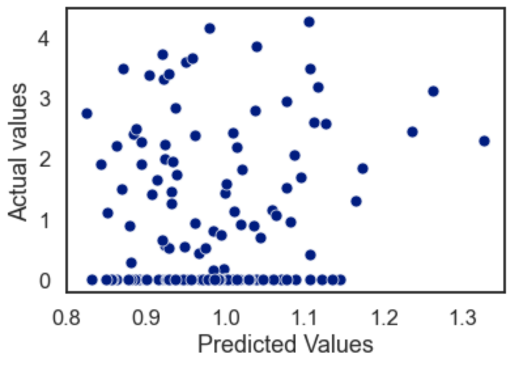
Metric	Value
RMSE	1.247
R2 - score	-0.061

Feat	Importance
FFMC	-3.694
Month	0.499

Ridge ✓

RidgeCV

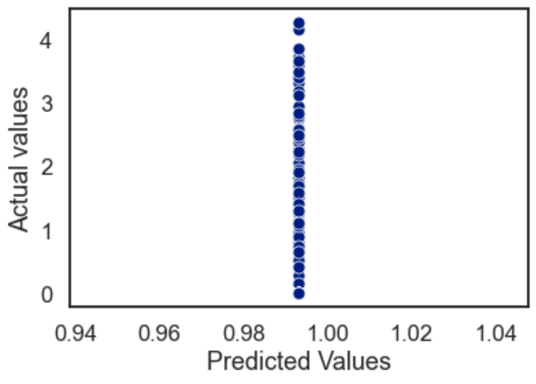
Metric	Value
RMSE	1.2195
R2 - score	0.00913



Lasso

LassoCV

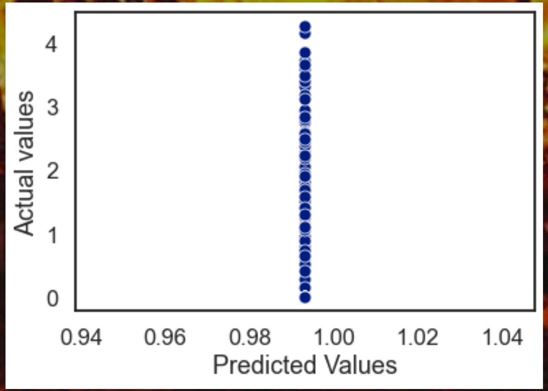
Metric	Value
RMSE	1.226
R2 - score	-0.000862



ElasticNet

ElasticNetCV

Metric	Value
RMSE	1.226
R2 - score	-0.000862



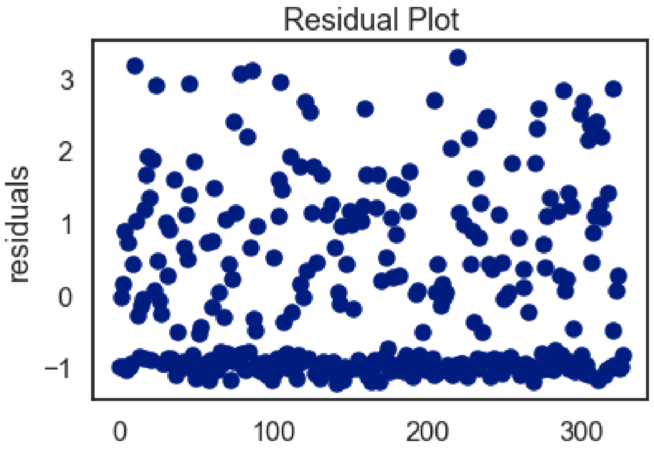
Flaws and Recommendations

- All R squared terms are extremely low which indicates that the independent variables do not explain much of the variation in the area of wildfire variable (dependent)

- This is mainly because the residuals or the errors between predictions and actual values do not follow a normal distribution

- The above result is found through using D’Agostino K^2 Test where an output p-value lower than 0.05 would indicate the distribution of residuals is not normally distributed and hence a linear regression cannot be used for this dataset. The plot of residuals from Ridge Regression, which performs the best on metrics such as RMSE and R_2 squared is shown on the right:

- Using models such as Random Forests and Support Vector Machines would be the ideal choice to go with in this case along with potentially adding data from other countries which could enhance the model tenfold



P-value = 5.64376948e-09

Thank You!