Forest Fire Modeling Using Bayesian Regression Analysis

Nikki Aaron, Amanda West Beverly Dobrenz

Executive Summary

This project uses PyMC3 to calculate two Bayesian Regression Models given data on forest weather and moisture conditions -- one logistic model for whether a forest fire will occur, and one linear model for burn area

PyMCS uses Markov Chain Monte Carlo (MCMC) and Variational Inference (VI) to approximate the statistical distribution of each feature for the population from which the data has likely been drawn. These distributions can be insoected to get an idea of how much uncertainty and variation there is in the model.

The most predictive features are wind, temperature, rain, and humidity. Results show that the burn area model has less uncertainty than the occurrence model, but unfortunately both had high uncertainty. It seems that these indicators can only give a general idea of risk level.

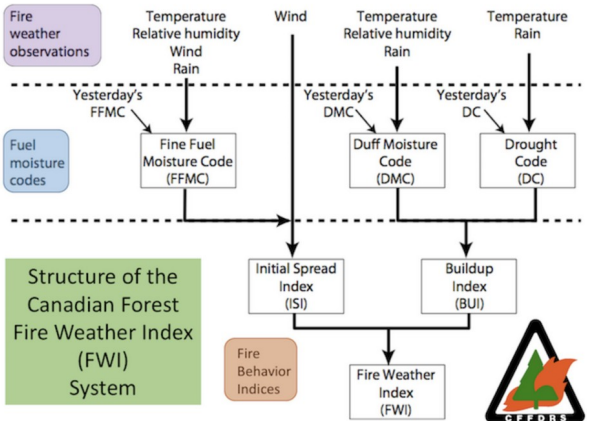
Introduction

Forest Service firefighters respond to upwards of 73,000 wildfires in the U.S each year. On average, these fires burn 7 million acres of private, state, and federal land and require a mix of ground firefighters and aerial firefighting aircraft. With wildfire costs exceeding $2.4 billion in 2017, forest fires are projected to consume two thirds of the entire budget four years sooner than originally calculated. In order to assist wildfire management decisions, this project will combine weather, climate, fuels, and fire activity data to predict the occurrence and burn area of forest fires. Specifically, this analysis can inform both the pre-positioning of firefighting assets and fire management strategies, thereby ensuring the best allocation of resources for wildfire management[1].

Data

We used three year's worth of forest fire data from the northeast region of Portugal to construct our probabilistic models. The data uses the Fire Weather Index (FWI), a Canadian system for rating fire danger. FWI moisture codes have higher numbers for drier conditions. The index is comprised of the following predictor variables with burn area as the response:

- **Area** - area of forest burned in hectares
- **FFMC** - fine fuel moisture code
- **DMC** - duff moisture code
- **DC** - drought code
- **ISI** - initial spread index
- **Temp** - temperature in Celsius degrees
- **RH** - relative humidity in %
- **Rain** - outside rain in mm/m^2
- **Wind** - wind speed in km/h
- **Weekend** - fire occurrence on Saturday or Sunday



1 https://www.fs.usda.gov/science-technology/fire/forecasting

Figure 1: Graphical Representation the Fire Weather Index[2]

Bayesian Methods Used

Bayesian methods provided a way to model the response variables from the forest fire data while also taking into account both known and suspected prior information. The first response of interest was whether a forest fire occured or not. We model this binary response with Bayesian logistic ridge regression.[3] To obtain the equation for this Bayesian logistic model, an inverse logit is used to transform Bayes Theorem so that we have a linear relationship between the log odds posterior distribution and our predictors.[4]

Additionally, to meet Ridge requirements and help with interpretability and comparison of the final model, we standardized our predictor data distributions with a mean of 0 and a variance of 1. We assumed our data came from a normally distributed population, but have no other prior information on this distribution. We also chose Gaussian prior distributions with mean 0 and standard deviation 1 for the intercept and each coefficient. Finally, we chose a Bernoulli posterior to model the binary response.

The results of the Bayesian logistic ridge regression consist of a distribution for the intercept (*beta0*) and one distribution for each of the predictors. The distribution for each of the predictors describe the slope coefficient of the predictor (*beta i*), with a dummy coding on categorical predictors. Combined, the derived distributions provide insight into the importance and uncertainty of each predictor.

The second response of interest was on the subset of data where fires occurred, and quantified the total burn area of the fire on a continuous scale. We modeled this response using Bayesian linear ridge regression and assumed a Gaussian posterior distribution. We also performed a log-transformation to ensure the zero-skewed response data supported the Gaussian posterior distribution assumption. Next, we chose a Half-Cauchy prior for the variance because these priors cannot be negative values. Finally, we mirrored our approach for the logistic model and used standard Gaussian intercept and coefficient priors. The results of this linear model consist of a set of intercept and coefficient distributions. When reviewed using both trace and forest plots, the coefficient distributions provide insight into significant factors.

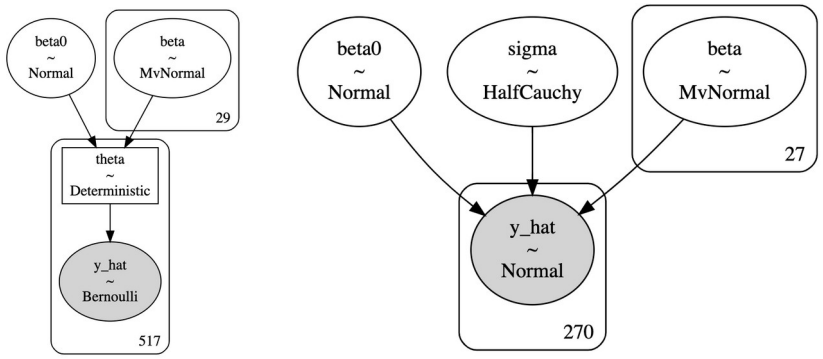
2 https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system

3 In general, lasso (L2) regression is an effective regularization method when there are a few key predictors in the data, but in our case all of the coefficients were shrunk to zero. Since our data features many moderately important predictors with multicollinearity, using ridge (L1) regularization performed better.

4 https://www.countbayesie.com/blog/2019/6/12/logistic-regression-from-bayes-theorem

Mathematical Linkage

Figure 2: Graphical Representation of Logistic (Left) & Linear (Right)



Parting from the frequentist interpretation of ordinary least squares, the Bayesian least squares regression calculates probability distributions. The loss function for Bayesian linear regression is least squared error, we assume no correlation between observations and predictors and that the error variance σ_η is constant.

Our log-transformed Bayesian linear regression is as follows:

$$\log(y) = X\theta + \varepsilon \text{, where } y \sim N(X\theta, \sigma^2_\varepsilon \mathbf{I})$$

$$\textit{where } p(\theta \vee X, y) = (p(X, y \vee \theta)p(\theta))/p(X, y)$$

In the equation above, y is the normally-distributed response vector, X is the data matrix, θ represents a random vector of the parameters, σ_ϵ is a constant value of the error variance, and ϵ is the error vector. $p(\theta \vee X, y)$ approximates to $p(\theta \vee y)$ since y is a function of the data X, and $p(\theta) \sim N(\theta_o, \Sigma_\theta$ since we use Gaussian prior and Gaussian posterior (conjugacy). We then invert σ_η to find factor λ, which we insert into the ridge equation to minimize the negative log likelihood with penalty:

–

Secondly, our the Bayesian logistic ridge regression is as follows:

$$p(y|X) = 1 / 1 + e^{-(\theta_0 + \theta_1 X)}$$

We modeled the response as Bernoulli with a logarithmic loss function and Gaussian priors. θ is a random vector of the parameters, X is the data matrix and y_i is the Bernoulli response vector that takes values 0 and 1 for $i = 1 \ldots N$. We assume there is no correlation between observations and that the variance of the Gaussian priors is constant σ_η. Similar to the linear regression, we then inverted σ_η to find factor λ, inserted this value into a ridge regression equation (negative log likelihood with penalty) and minimized.

Results

The posterior distributions in our logistic model had smaller bounds overall, with two variables not overlapping 0 at 90% HDI: *wind* and *DC (temperature, rain)*. In the linear model, three coefficients were significantly above or below 0: *DMC (temperature, humidity, rain)*, *RH (relative humidity), wind,* and *weekend*. We use trace and forest plots to understand the distribution of the data as well as the distribution

of the posteriors in both models. To estimate the distribution further we employed sampling (blue) and ADVI (red) methods (*figure 4*).

Figure 3: Forest Plots of the Coefficients for Logistic (Left) & Least Squares (Right)

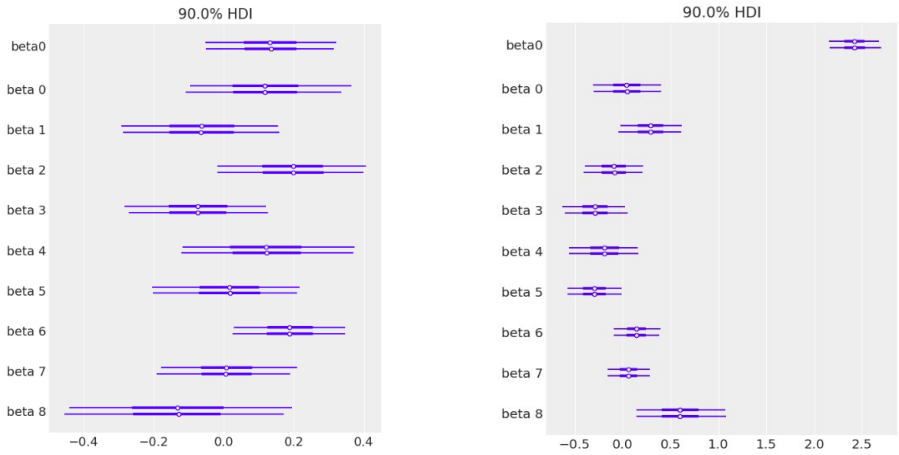
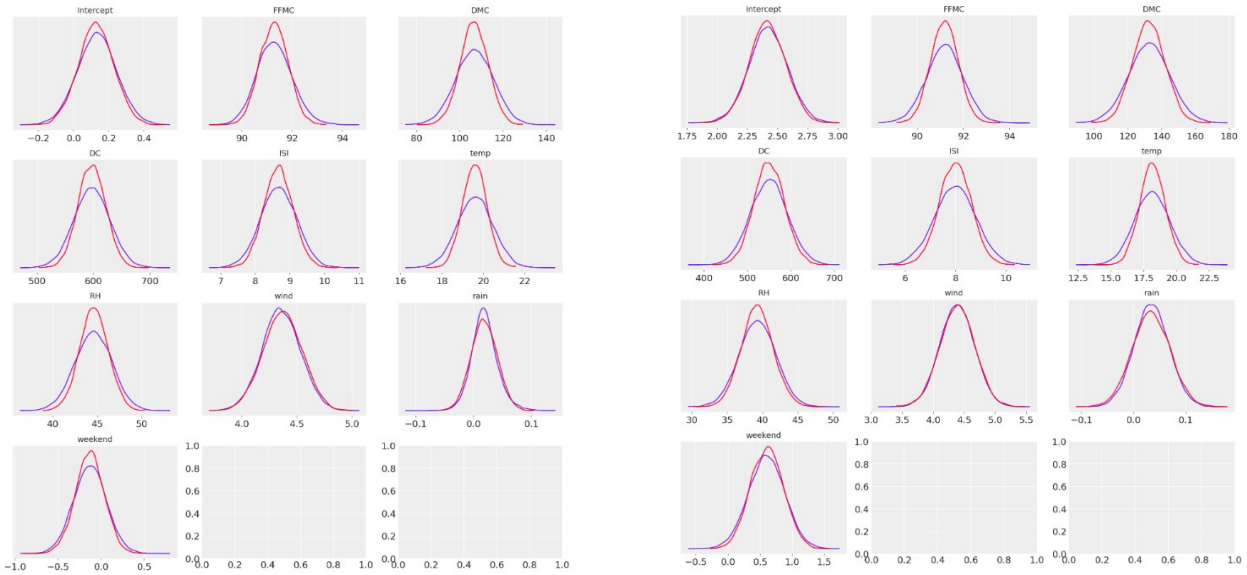


Figure 4: Trace Plots of the Coefficients for Logistic (Left) & Linear (Right)



The results of our model support common intuition and current literature about factors that influence forest fires. For example, research shows that the lower the humidity level the more vigorously a fire is able to spread.[5] Since humidity is only significant for the burn area of forest fires and not the occurrence, we can conclude that our model mirrors current forest fire literature. With this affirmation, forest fire stakeholders can use both literature and our probabilistic model to inform decision making about fire prevention and maintenance strategies.

5 https://www.auburn.edu/academic/forestry_wildlife/fire/

Conclusion

With over 73,000 wildfires each year in the U.S. alone, effectively understanding the conditions that lead to forest fires is crucial to save lives and billions in damages. Weather is notorious for its nature of changing quickly and unexpectedly, so any information that informs forest fire analysis has the potential to benefit the firefighting community. Using forest fire data from the northeast region of Portugal, we modelled wildfires using Bayesian least squares and Bayesian logistic regression. Both regression equations found wind and moisture codes related to temperature and rain to be significant factors, with humidity and weekend/weekday proving to be additional significant factors for the linear regression (fire spread).

Moving forward, we advise the collection of data about available first responders to further inform factors that influence how big a fire becomes. The linear model had much less posterior uncertainty and may be more useful. We also recommend a hierarchical model approach, where each section of the park is treated as a unique district. With this approach we can accommodate the diverse landscape of the park and pinpoint potential problem areas where forest fires are more likely to occur. It may also be useful to add timestamps for each data row to enable time-series analysis (lag) to account for days when risk has been elevated for extended periods of time.

Appendix

Table 1: Beta Value Definitions

Coefficient	Variable
Beta0	Intercept
Beta 0	FFMC
Beta 1	DMC
Beta 2	DC
Beta 3	ISI
Beta 4	Temperature
Beta 5	RH
Beta 6	Wind
Beta 7	Rain
Beta 8	Weekend

Figure 5: Log-Transformed Response

For a logistic model on whether a fire occurred, we create a binary response variable 'is_fire'.
For a linear model on a subset of data where fires did occur, we log transform the burn area response variable so that it better fits a Gaussian distribution.

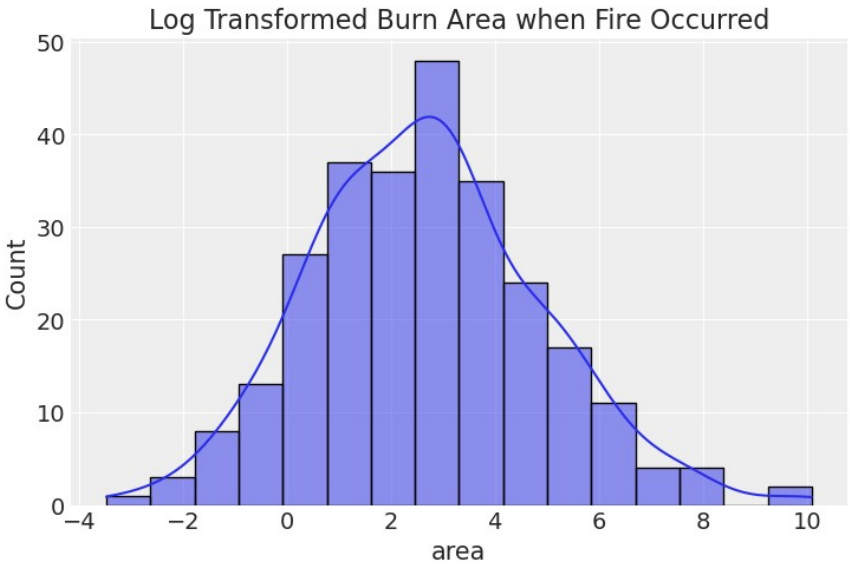


Figure 6: Kernel Density Estimation (KDE) Plots

KDE plots show the approximate distribution of data in each column from the data set. We can see that most are approximately normally distributed with a few outliers. Burn area and rainfall are skewed with many values equal to zero.

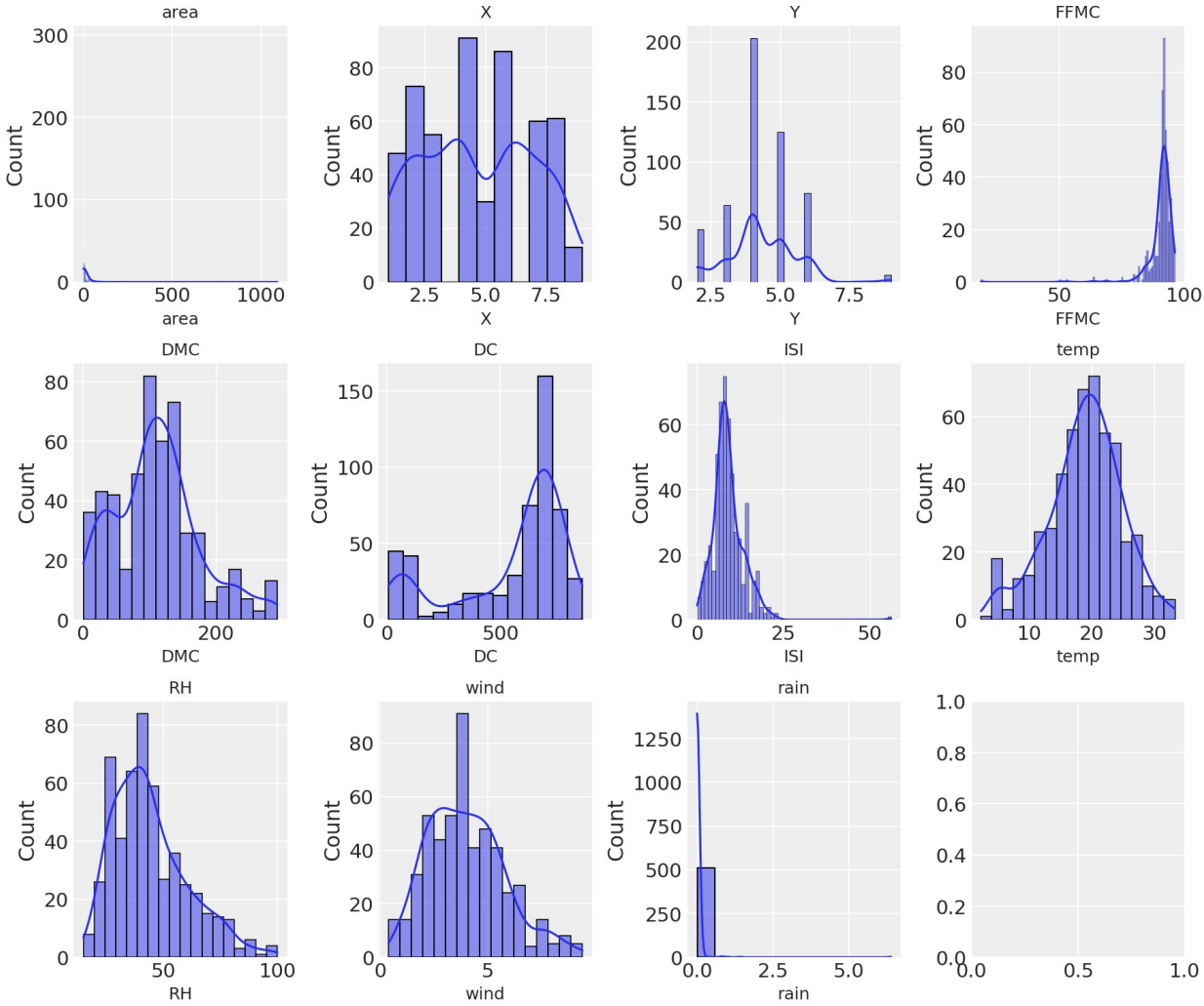


Figure 7: Violin Plots

Violin plots show the distribution of the log transformed burn area for each day of the week and month over multiple years. There does not seem to be a significant difference in burn area per day of the week, but we can see there are a lot more fires and fires with larger burn areas in the months of August and September. This makes sense because those are Portugal's hotter, drier summer months.

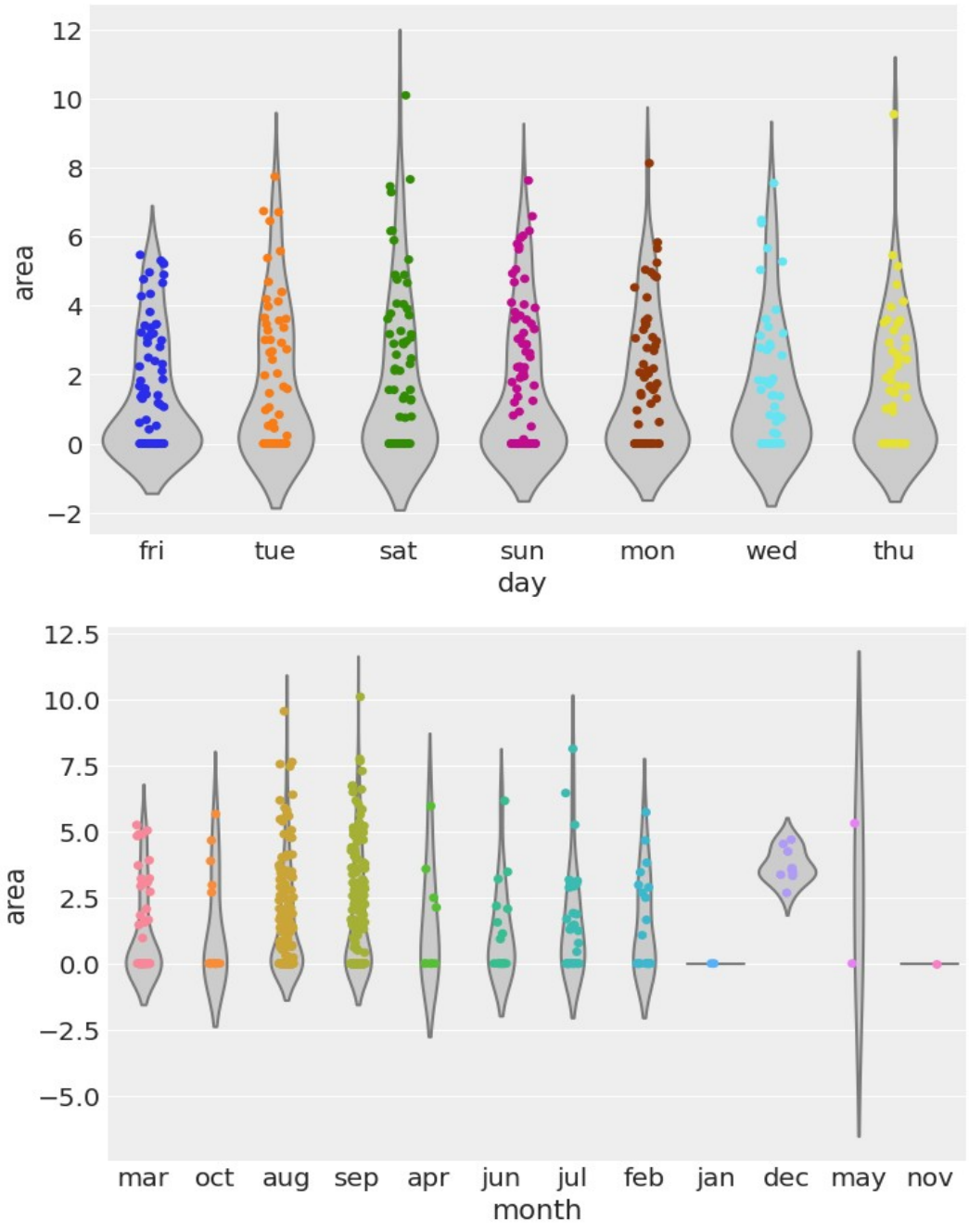
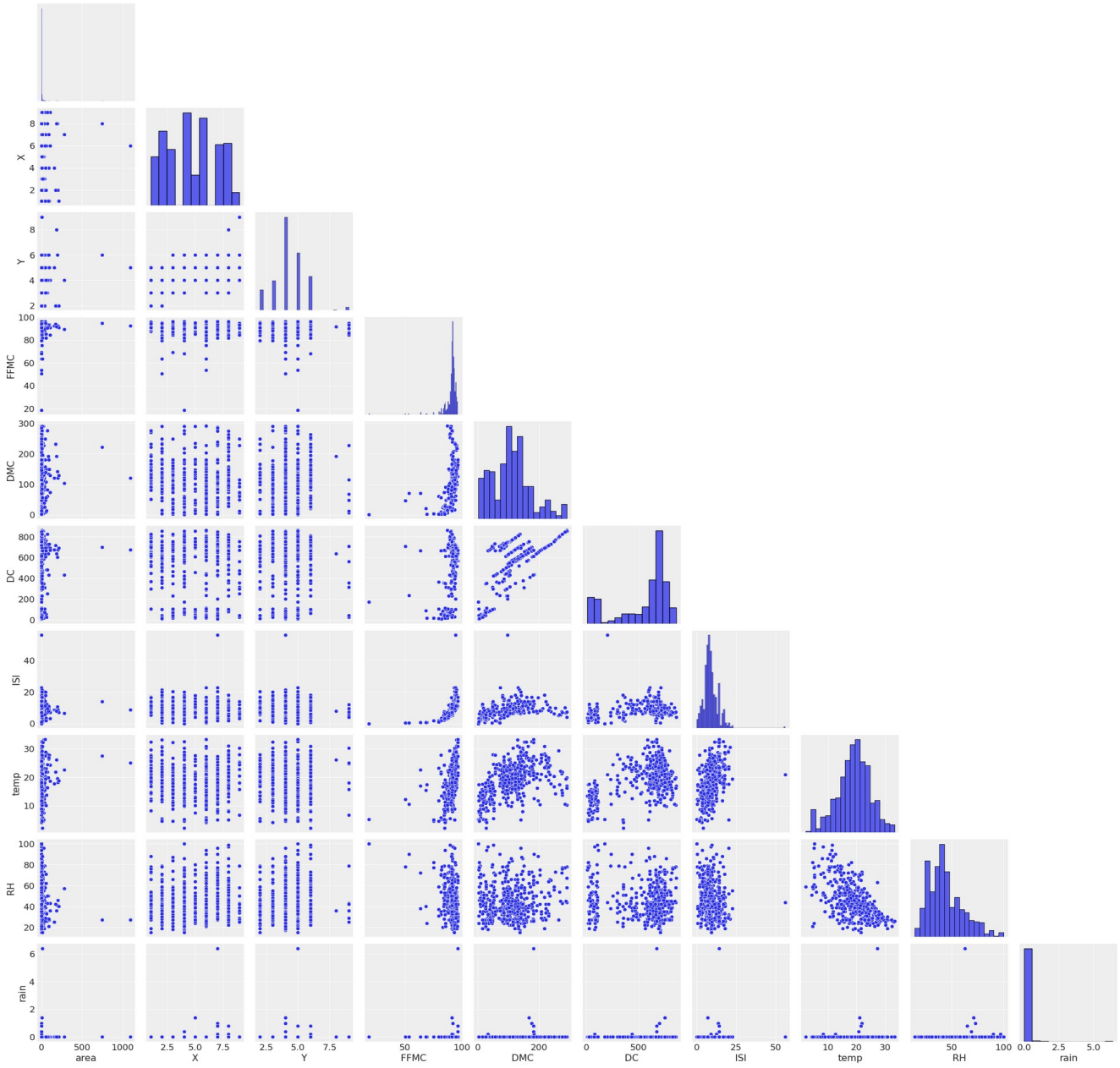


Figure 8: Correlation Plots

Pair plots were created to show correlations between pairs of variables. The only clear correlation is between temperature and relative humidity, and between some of the Fire Weather Indices. We can see that most days have no burn area (fire) at all, and two large fires stand out as outliers. We can also see that it looks like we have more fires and higher burn areas with low relative humidity, high temperature, high FFMC, and high DC.



Sources

1 - A. Rochford, 2013, “Prior Distributions for Bayesian Regression Using PyMC”, https://austinrochford.com/posts/2013-09-02-prior-distributions-for-bayesian-regression-using-pymc.html

2 - P. Cortez and A. Morais, 2008, “A Data Mining Approach to Predict Forest Fires using Meteorological Data”, http://www3.dsi.uminho.pt/pcortez/fires.pdf

3 - P. Cortez and A. Morais, 2007, UCI Machine Learning Repository, “Forest Fires Data Set”, http://archive.ics.uci.edu/ml/datasets/Forest+Fires

4 - U.S. Forest Service, “Fire Forecasting”, https://www.fs.usda.gov/science-technology/fire/forecasting

5 - W. Koehrson, 2018, “Introduction to Bayesian Linear Regression”, https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

6 - W. Kurt, 2019, “Logistic Regression from Bayes' Theorem”, https://www.countbayesie.com/blog/2019/6/12/logistic-regression-from-bayes-theorem