# Temperature & Forest Fire Analysis

2024-02-11

```
# Load and install the necessary packages for the assignment

#install.packages("Metrics")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("magrittr")
#install.packages("purrr")
#installed.packages("Hmisc")
#installed.packages("data.table")
#install.packages("ggplot2")
#install.packages("DataExplorer")
#install.packages("tidyverse")
#install.packages("quantmod")
```

```
library(Metrics)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:magrittr':
##
##     set_names
```

```
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(DataExplorer)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v readr     2.1.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x data.table::between()   masks dplyr::between()
## x magrittr::extract()     masks tidyr::extract()
## x dplyr::filter()         masks stats::filter()
## x data.table::first()     masks dplyr::first()
## x lubridate::hour()       masks data.table::hour()
## x lubridate::isoweek()    masks data.table::isoweek()
## x dplyr::lag()            masks stats::lag()
## x data.table::last()      masks dplyr::last()
## x lubridate::mday()       masks data.table::mday()
## x lubridate::minute()     masks data.table::minute()
## x lubridate::month()      masks data.table::month()
## x lubridate::quarter()    masks data.table::quarter()
## x lubridate::second()     masks data.table::second()
## x purrr::set_names()      masks magrittr::set_names()
## x Hmisc::src()            masks dplyr::src()
## x Hmisc::summarize()      masks dplyr::summarize()
## x data.table::transpose() masks purrr::transpose()
## x lubridate::wday()       masks data.table::wday()
## x lubridate::week()       masks data.table::week()
## x lubridate::yday()       masks data.table::yday()
## x lubridate::year()       masks data.table::year()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readxl)
library(readr)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:readr':
##
##     col_factor
##
## The following object is masked from 'package:purrr':
##
##     discard
```

```
library(quantmod)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
##
## ######################### Warning from 'xts' package ##########################
## #                                                                             #
## # The dplyr lag() function breaks how base R's lag() function is supposed to  #
## # work, which breaks lag(my_xts). Calls to lag(my_xts) that you type or       #
## # source() into this session won't work correctly.                            #
## #                                                                             #
## # Use stats::lag() to make sure you're not using dplyr::lag(), or you can add #
## # conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop           #
## # dplyr from breaking base R's lag() function.                                #
## #                                                                             #
## # Code in packages is not affected. It's protected by R's namespace mechanism #
## # Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning.  #
## #                                                                             #
## ###############################################################################
##
## Attaching package: 'xts'
##
## The following objects are masked from 'package:data.table':
##
##     first, last
##
## The following objects are masked from 'package:dplyr':
##
##     first, last
##
## Loading required package: TTR
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
##
## Attaching package: 'quantmod'
##
## The following object is masked from 'package:Hmisc':
##
##     Lag
```

Introduction:

I would like to reserach forest fires along with the factors that influence them. My ultimate goal is to see if they can be predicted and planned for ahead of time, althogh that may end up being a bit too advanced for the information available to me/ my current skill-set. I feel as though many people would be interested in this topic as it could help in the ongoing global warming epidemic. Because this is a predictive analysis topic I believe it can be addressed using data science.

Research Questions:

1:) Are certain countries more likely to have forest fires occur?

Do countries with a higher forestation percentage or higher average temperature face a higher risk? Are they more or less influenced if they are surrounded by water?

2:) What are the main variables in detecting forest fires?

Are there variables besides temperature that can be analyzed?

3:) Have the number of forest fires recorded yearly incresed/decreased?

Has global warming managed to increase the number of forest fires occuring or have prevenative measures been taken by most societies?

4:) Have any countries seen an increase in forestation or have they all decreased?

With the natural progression of scoieties and the increase of global population have any countries managed to increase their level of forestastion.

5:) Are they any influences on forest fires that are generally unknown or likely not thought about by the general public?

Are things like the day of the week or month of the year major factors in forest fires.

I believe that my research will most likely partially address the problem. The overall goal is to be able to predict where fires are likely to happen. I will look into the factors that are linked to this issue and chart them amongst different locations, but I ultimately think that my aim is slightly to high for my current skillset. Theres always a chance that I could prove myself wrong though.

Data Source 1:

Name: Forestfires.csv Obtained: Kaggle Link: https://www.kaggle.com/datasets/ulrikthygepedersen/forest-fires About: THis dataset will be the main dataset I use, it was created to monitor the level of forest fires, timing, coordinates, and humidity of recent times.

Data Source 2:

Name: Temperature-anomaly.csv Obtained: Kaggle Link: https://www.kaggle.com/datasets/sivashankarans/global-surface-temperature-trendc About: This dataset was created to display the low, high and median temperature of our planet since 1850.

Data Source 3:

Name: Goal15.forest_shares.csv Obtained: Kaggle Link: https://www.kaggle.com/datasets/konradb/deforestation-dataset About: This dataset was created to mark the levels of forestation around the globe and the trend of each countires forest area per capita.

I may look to add more datasets as I continue my research but I am confident in the material that these 3 currently provide.

The packages that I will look to utilize as of now are Metrics, dplyr, tidyr magrittr, purrr, Hmisc, data.table. I may look to include more packages but these are my current inclusions.

I will look to use histograms and scatterplots for the majority of my visualizations in this reserach. While these are generally the most simple visualization tools I believe that they are also the easiest when conveying data.

Questions for future steps:

The skill that I currently need to improve to best display my data/research findings is the skill we used in our most recent assignment which is creating visualization tools with variables from different datasets. I have struggled with this to date but will look to improve. Once I have progressed in this area I preduct that I will be able to handle the probable need to compare columns from different datasets and create visualization tools to accurate display the findings of my research.

References APA Format:

S, S. (2024, January 17). Global surface temperature trend(°C). Kaggle. https://www.kaggle.com/datasets/sivashankarans/global-surface-temperature-trendc

Pedersen, U. T. (2023, March 2). Forest fires. Kaggle. https://www.kaggle.com/datasets/ulrikthygepedersen/forest-fires

Banachewicz, K. (2023, July 11). Deforestation dataset. Kaggle. https://www.kaggle.com/datasets/konradb/deforestation-dataset

#_____
# Start of part 2

```
# accesses the datasets

fire_df <- read.csv("/Users/levijohnston/desktop/forestfires.csv", stringsAsFactors = FALSE)

temp_df <- read.csv("/Users/levijohnston/desktop/temperature-anomaly.csv", stringsAsFactors = FALSE)

deforest_df <- read.csv("/Users/levijohnston/desktop/goal15.forest_shares.csv", stringsAsFactors = FALSE
```

## Uses the string function to get an idea of how the structure of the datasets are currently looking

```
str(fire_df)
```

```
str(temp_df)
```

```
str(deforest_df)
```

## I will clean the data by changing the names of some of the columns to be

## easier to understand.

```
old1 <- c('iso3c', 'forests_2000', 'forests_2020', 'trend')

new1 <- c('Country', 'Forest Past', 'Forest Present', 'Trend Pattern')

old2 <- c('x', 'y', 'month', 'day', 'ffmc', 'dmc', 'dc', 'isi', 'temp', 'rh', 'wind', 'rain', 'area')

new2 <- c( "X Coordinate", "Y Coordinate", "Month", "Day", "FuelMoisture", "OrganicMoisture", "Drought"

old3 <- c('Entity', 'Code', 'Year', 'Median....', 'Upper....', 'Lower....')

new3 <- c('Entity', 'Code', 'Year', 'Average Temp', 'High Temp', 'Low Temp')

setnames(deforest_df, old=old1, new=new1, skip_absent = TRUE)
```

```r
setnames(fire_df, old=old2, new=new2, skip_absent = TRUE)
setnames(temp_df, old=old3, new=new3, skip_absent = TRUE)
```

```
## 'data.frame':    237 obs. of  4 variables:
##  $ Country       : chr  "AFG" "ALB" "DZA" "ASM" ...
##  $ Forest Past   : num  1.9 28.1 0.7 88.7 34 62.3 61.1 21.5 12.2 11.7 ...
##  $ Forest Present: num  1.9 28.8 0.8 85.7 34 53.4 61.1 18.5 10.4 11.5 ...
##  $ Trend Pattern : num  0 2.5 14.3 -3.4 0 -14.3 0 -14 -14.8 -1.7 ...
```

```
## 'data.frame':    517 obs. of  13 variables:
##  $ X Coordinate   : int  7 7 7 8 8 8 8 8 8 7 ...
##  $ Y Coordinate   : int  5 4 4 6 6 6 6 6 6 5 ...
##  $ Month          : chr  "mar" "oct" "oct" "mar" ...
##  $ Day            : chr  "fri" "tue" "sat" "fri" ...
##  $ FuelMoisture   : num  86.2 90.6 90.6 91.7 89.3 92.3 92.3 91.5 91 92.5 ...
##  $ OrganicMoisture: num  26.2 35.4 43.7 33.3 51.3 ...
##  $ Drought        : num  94.3 669.1 686.9 77.5 102.2 ...
##  $ Spread         : num  5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...
##  $ Temperature    : num  8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...
##  $ Humidity       : int  51 33 33 97 99 29 27 86 63 40 ...
##  $ Wind           : num  6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...
##  $ Rain           : num  0 0 0 0.2 0 0 0 0 0 0 ...
##  $ Area           : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
## 'data.frame':    676 obs. of  6 variables:
##  $ Entity      : chr  "Global" "Global" "Global" "Global" ...
##  $ Code        : logi  NA NA NA NA NA NA ...
##  $ Year        : int  1850 1851 1852 1853 1854 1855 1856 1857 1858 1859 ...
##  $ Average Temp: num  -0.373 -0.218 -0.228 -0.269 -0.248 -0.272 -0.358 -0.461 -0.467 -0.284 ...
##  $ High Temp   : num  -0.339 -0.184 -0.196 -0.239 -0.218 -0.241 -0.327 -0.431 -0.435 -0.249 ...
##  $ Low Temp    : num  -0.425 -0.274 -0.28 -0.321 -0.301 -0.324 -0.413 -0.512 -0.521 -0.34 ...
```

**The data is cleaned to a point that it will be easier to work with, the NA's have been**

**removed and the strings above display the sets that I will be proceeding with**

**The information that at this point is not self evident is the relationship between**

**important variables. Different ways I could look into discovering these values are with**

**combination _df's (these will not be in the submission but I have created 2 different**

**combination _df's but the histograms came out funky so I will be going from a different angle)**

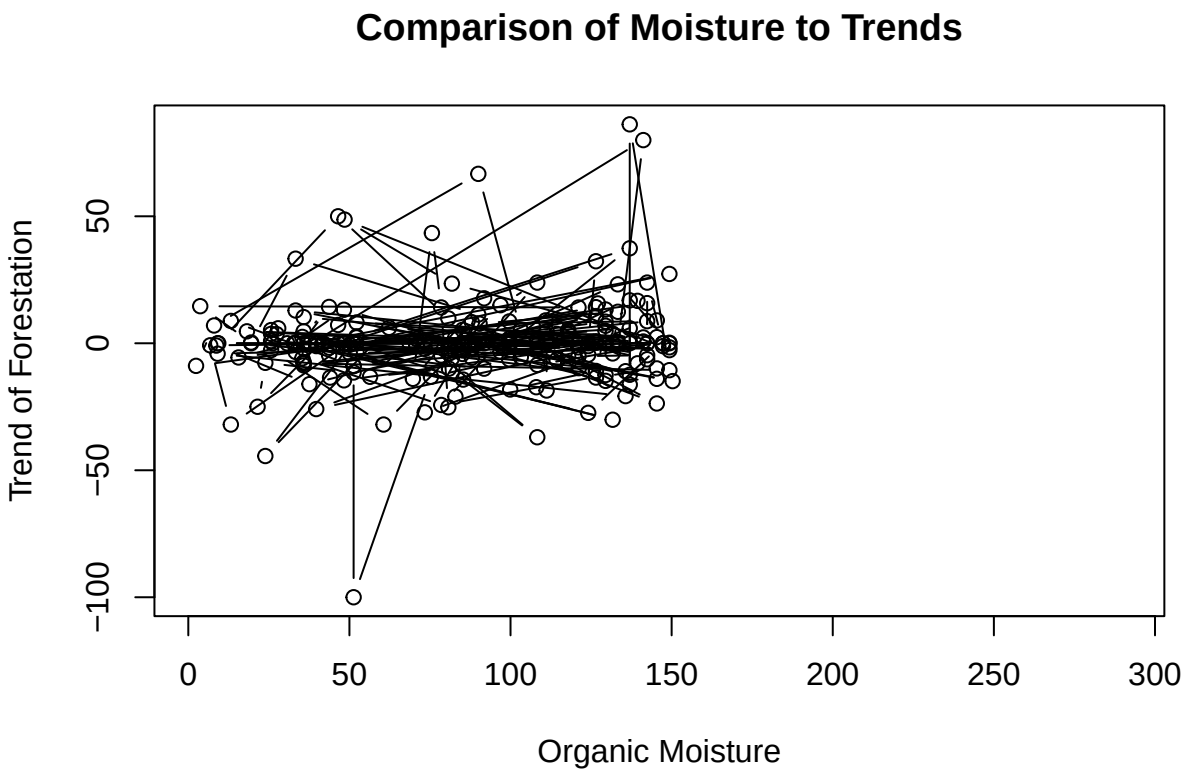
**or creating joint-partnership histograms with multiple variables.**

**My plan to 'slice-up' the data will be the second option above. Mostly plots or geom/hist**

**graphs that are easy for the viewer to interpret. A multitude of these visuals will help to**

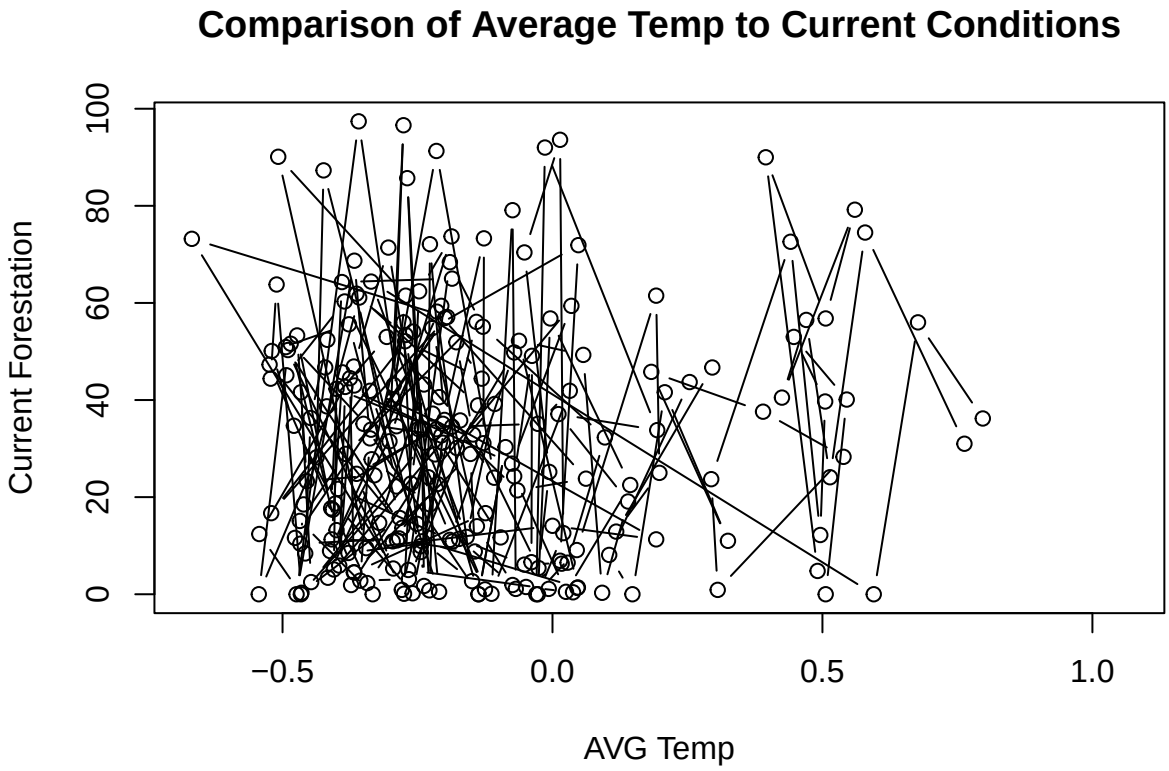
**give a better understanding of the data**

```
plot(fire_df$OrganicMoisture, deforest_df$`Trend Pattern`[1:length(fire_df$OrganicMoisture)],
     main = "Comparison of Moisture to Trends",
     xlab = "Organic Moisture",
     ylab = "Trend of Forestation",
     type = "b",
     )
```



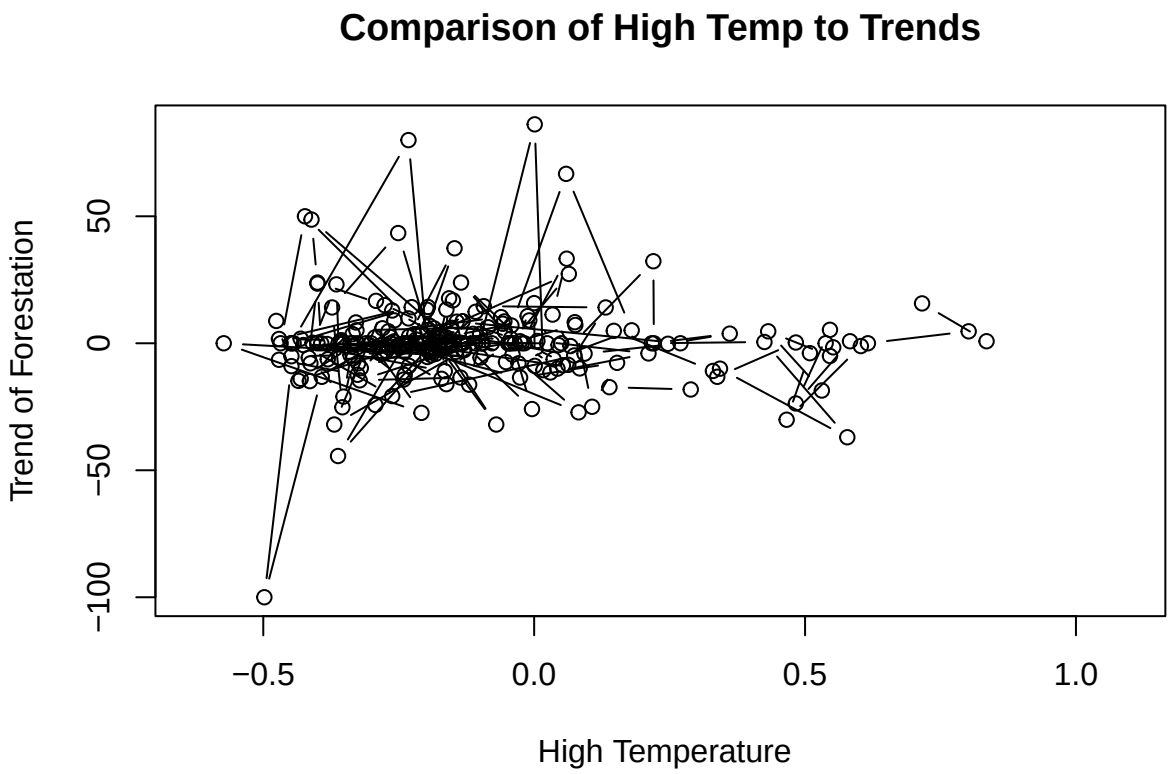

# This plot in particular shut down one of my BIGGEST hypothesis that I formed during # my initial research. I found during my initial research that Organic Moisture is natural # moisture found in areas without polution or harmful substances whilst Fuel Moisture is # essentially contaminated moisture by things such as gasoline, runoff or other unnatural # substances. Whilst the plot above does show a positive relationship between Organic Moisture # and the Trend of Forestation, the majority of the data is constant until the amount of O.M # gets to a relatively high figure, aside from a few outliers both on the positive and negative # end of this spectrum.

```
plot(temp_df$`Average Temp`, deforest_df$`Forest Present`[1:length(temp_df$`Average Temp`)],
     main = "Comparison of Average Temp to Current Conditions",
     xlab = "AVG Temp",
     ylab = "Current Forestation",
     type = "b",
     )
```



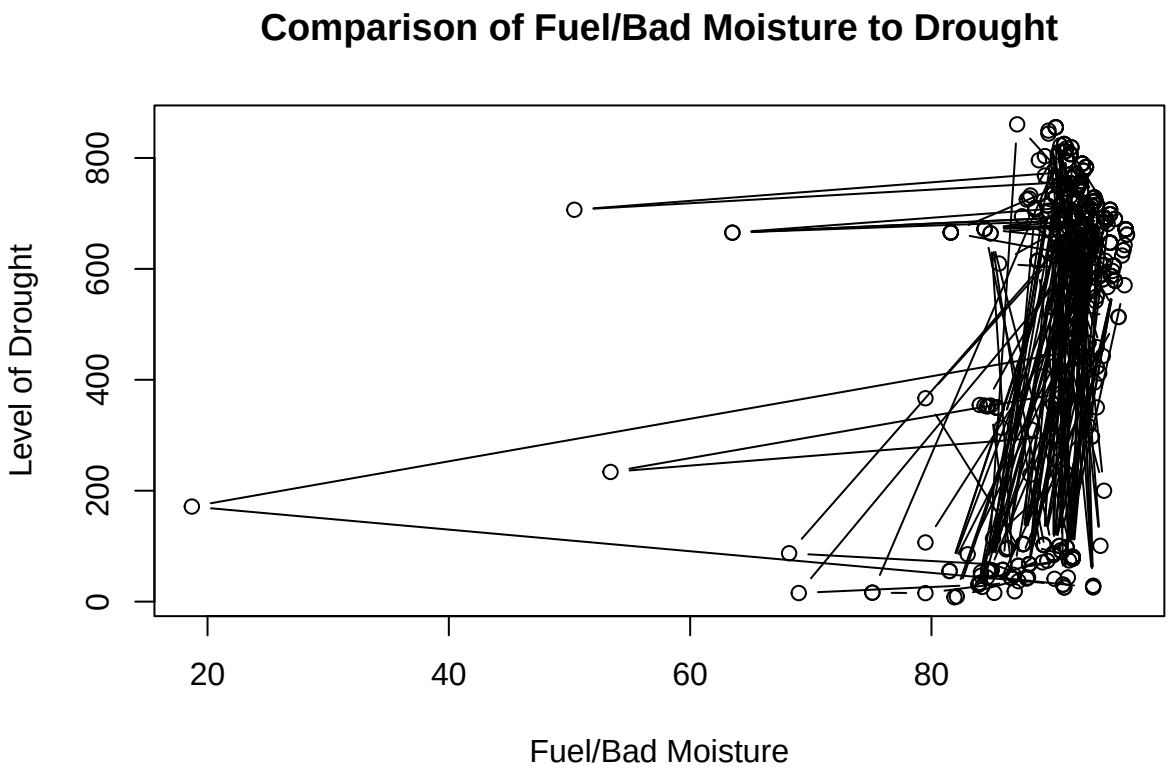

# As expected it looks as if the higher the average temperature rises the lower the levels of # forestation are. The examples in generally show an extreme dropoff right after passing 0.0 # as well

```
plot(temp_df$`High Temp`, deforest_df$`Trend Pattern`[1:length(temp_df$`High Temp`)],
     main = "Comparison of High Temp to Trends",
     xlab = " High Temperature",
     ylab = "Trend of Forestation",
     type = "b",
     )
```




# There seems to be a sweet spot of positive forestation trend in the average to slightly # below average High Temperature range. Once 0.0 is passed and the locations "high' temp is # high even amongst other 'high' temps the trend either plateaus or trends downward.

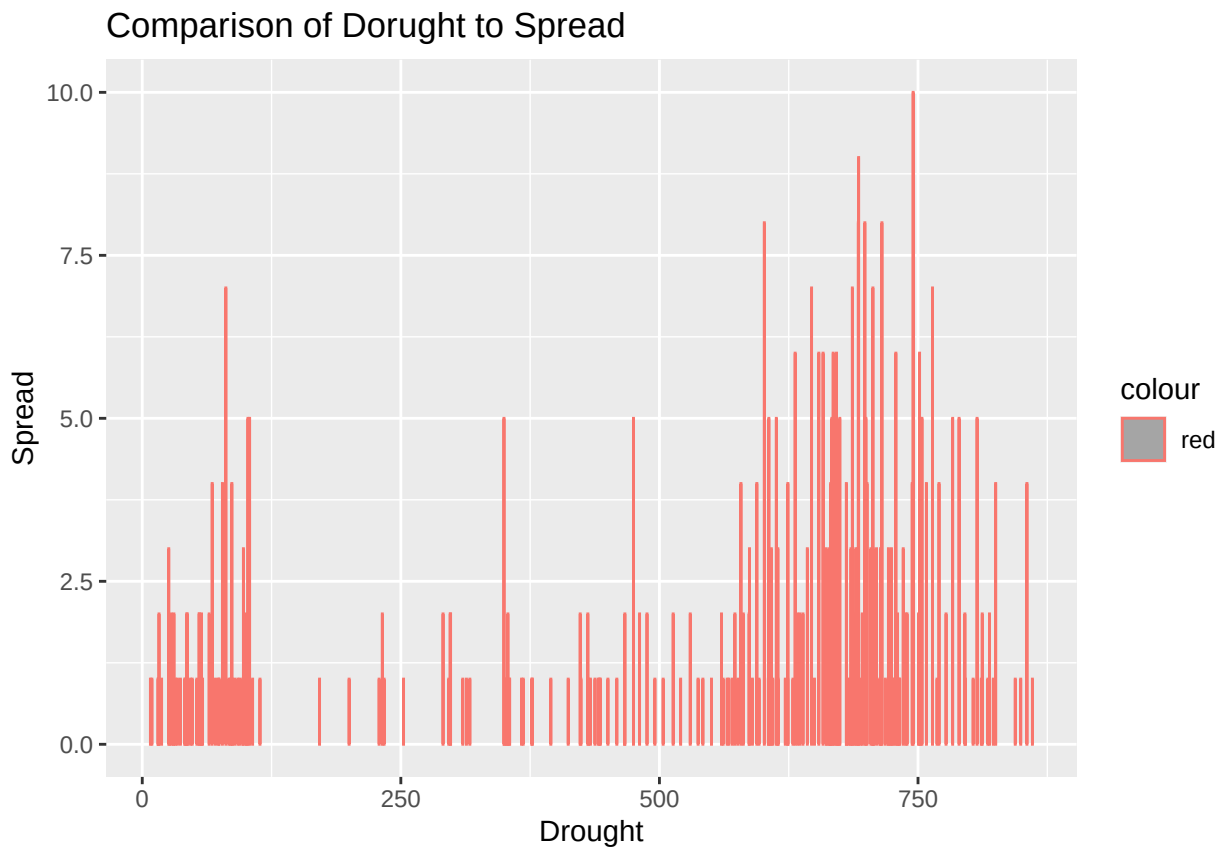
```
plot(fire_df$FuelMoisture, fire_df$Drought[1:length(fire_df$FuelMoisture)],
     main = "Comparison of Fuel/Bad Moisture to Drought",
     xlab = "Fuel/Bad Moisture",
     ylab = "Level of Drought",
     type = "b",
     )
```




# This graph is somewhat difficult to interpret, it appears as though drought only occurs when # a high lebel of Fuel Moisture gathers in an area. BUT, the level of drought that occurs once the # 80th percentile is passed is so varried that it is hard to calculate/interpret. # This will 100% be an issue that I further dive into in the next step

```
fire_df %>%
  ggplot(aes(Drought, fill = Spread, color = 'red'))+
  geom_bar(alpha = 0.5)+
  labs(title = "Comparison of Dorught to Spread",
       x = "Drought",
       y = "Spread")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```




My initial hypothesis on this analysis was that higher levels of drought would lead

to the spread (of wild fires) and it seems as though that hypothesis is right. There

does appear to be a strange spike at what appears to be the 15th percentile which is

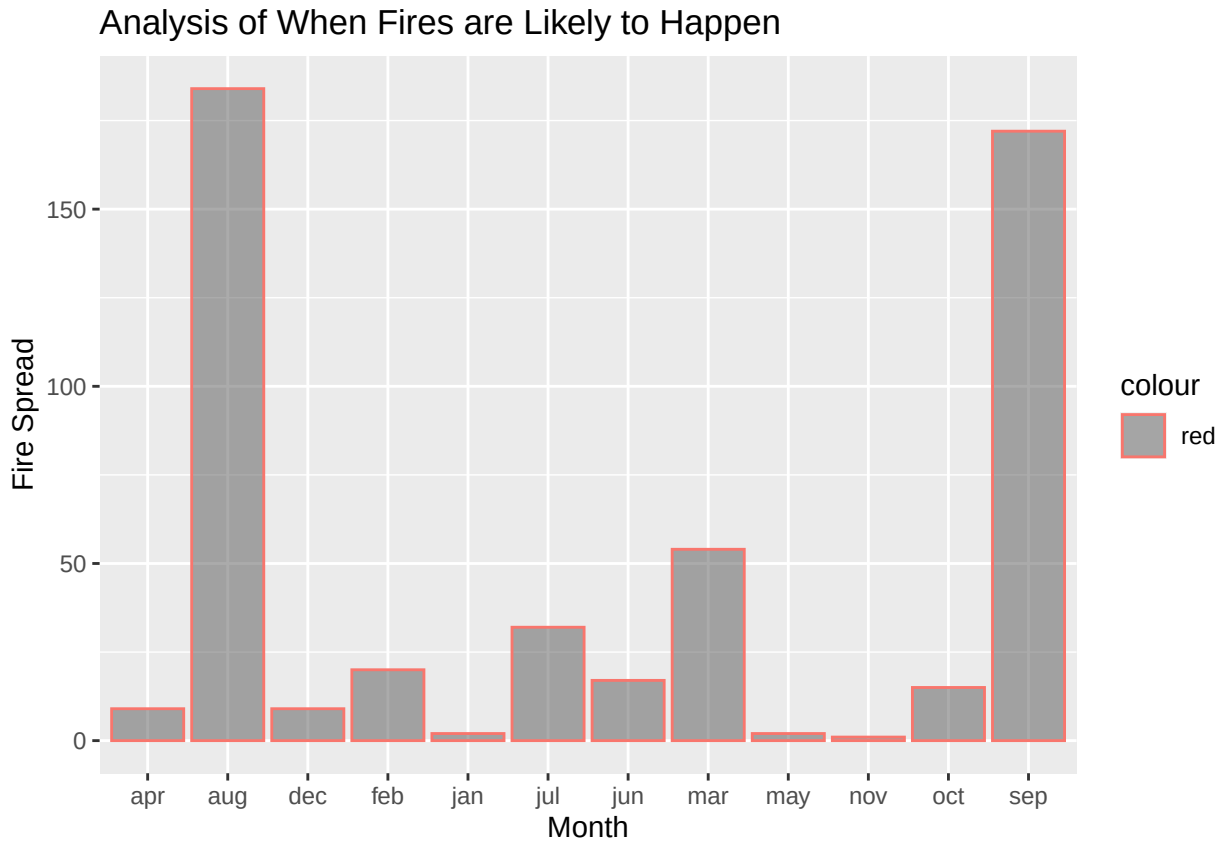
rather strange. However the majority of the spread occurs once the 60th percentile has

been passed.

```
fire_df %>%
  ggplot(aes(Month, fill = Spread, color = 'red'))+
  geom_bar(alpha = 0.5)+
  labs(title = "Analysis of When Fires are Likely to Happen",
```

```
       x = "Month",
       y = "Fire Spread")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
```




# Two massive outliers for when fires are more likely to happen with August and September much higher in numbers than any other months.Its strange that the highest months in August & September are so close to the lowest month in November. This will be useful information for when calculating later experiments.

**Interpretting the Analysis of the data and visuals above. Also addressing possible future questions involving the next steps to solving the initial hypothesis/statement.**

**After cleaning/reorganizing the datasets columns and NA values I have developed a better understanding of the datasets that I am working with which made displaying helpful visuals much more comfortable for me.**

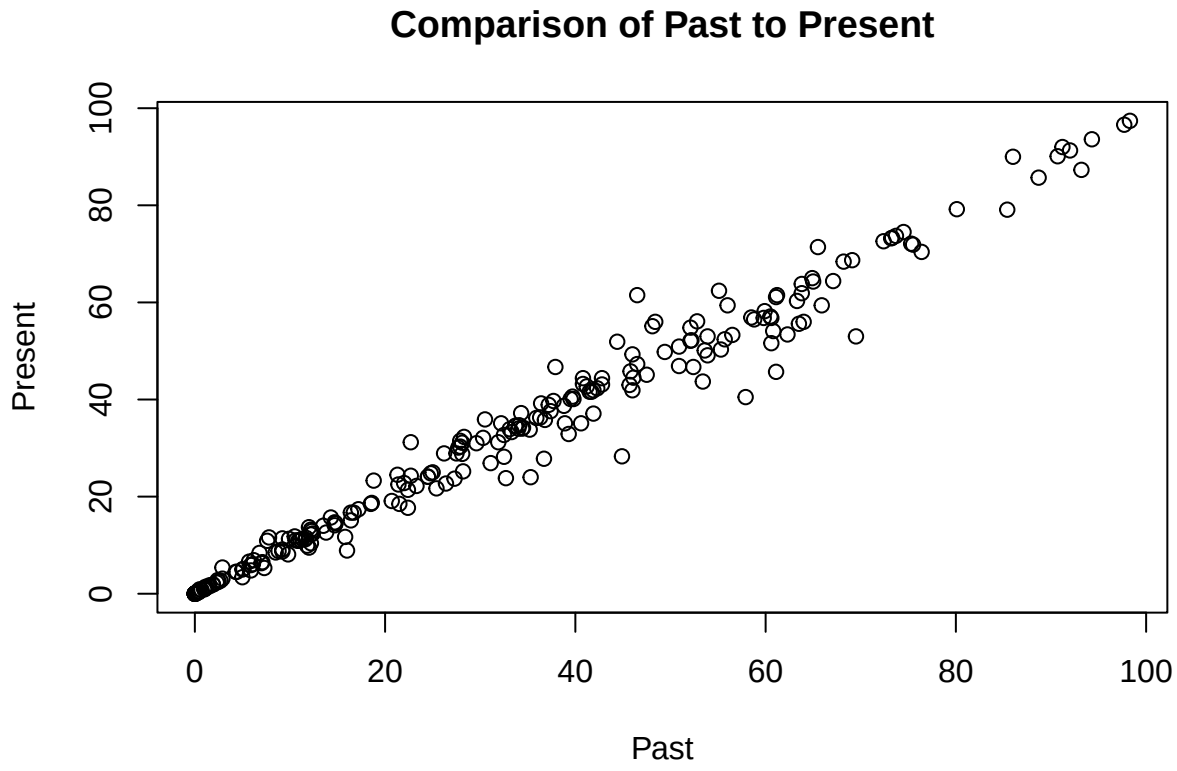
**My initial observations amongst the dataset after analyzing the information obtained through the visuals is that Fuel Moisture as a substance will be something that I need to research more thoroughly as the reading I have currently done suggested that it was a harmful subbstance that leads to the 'spread' of fires. However,in ALMOST every category analyzed above it showed to be a variable that influenced the POSITIVE result leading to increased forestation instead of increased spread.**

**The goodnews is that alot of my initial hypothesis in regards to temperature and drought in relation to spread were correct. This may seem an obvious and easy achievement but it gives me great levels of confidence that the work I am doing is accurate and won't produce false resultsat the end of the experiment.**

**Looking forward to the next questions I have in regards to achieving the initial goal/statement of being able to predict fires I have a few ideas. The main influence on these fires is the lack of rain or the increase in an areas 'drought' variable. As seen in the visuals above a multitude of factors influence an areas level of drought. The best chance at being able to predict an areas chance of spread is through drought as seen in the geom visual above. Diving into these variables I believe will ultimately give me the best chance at achieveing the initial project statement.**

**My current plan is to use machine learning with the 'Forest Past' variable and the value of the variables alongside it and analyze it against the 'Forest Present' variables and the value of the variables alongside it as well. Hopefully I can analyze the spread variable in particular of both scenarios and find the points at which the spread was high and look at the determining factors. This is the best way I could both summarize the above data for viewer and personal questions.**

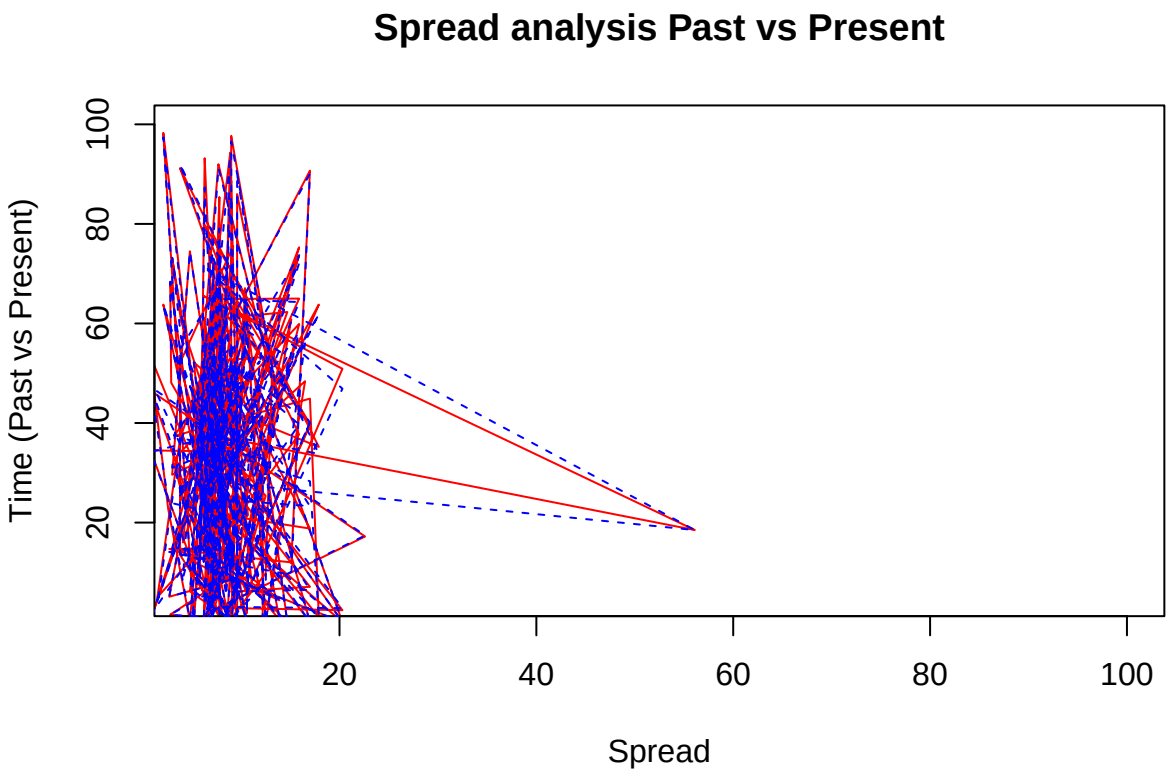
```r
plot(deforest_df$`Forest Past`, deforest_df$`Forest Present`[1:length(deforest_df$`Forest Past`)],
     main = "Comparison of Past to Present",
     xlab = "Past",
     ylab = "Present",
     type = "p"
     )
```




# Aside from a few outliers the correlation is very similar between years start/0 - 40 and 90-present. The 40+ - 85 range is where a host of positive and negative outliers have occured.

```r
plot(fire_df$Spread, deforest_df$`Forest Past`[1:length(fire_df$Spread)],
    main = "Spread analysis Past vs Present",
    xlab = "Spread",
    ylab = "Time (Past vs Present)",
    type = "l",
    col = 'red',
    ylim = c(5,100),
    xlim = c(5,100))

lines(fire_df$Spread, deforest_df$`Forest Present`[1:length(fire_df$Spread)],
      lty = "dashed",
      col = 'blue')
```


Spread analysis Past vs Present
Time (Past vs Present)
20
40
60
80
100
Spread
20
40
60
80
100

After analyzing the past and present forestry levels I tested its validity by analyzing both of these variables against their 'spread' values. The present day spread value appears to mirror the past value but upon further inspection appears to have trended in a positive trend as a heavy consolodation of 'blue' can be seen at what appears to be the 8-10 x axis/spread value range. Compared to the slightly more spread out past value consolodating in the 8-20 range.

This is the outcome I desired to obtain as I feel that this validates the data collected earlier in this study. The high correlation to the past and present variables along with the compatability seen in the pastvspresent spread analysis means that the data obtained from the Forest Past & Forest Present variables could effectively be used to create a new value labeled as "Forest Future". The initial statement/issue that I wished to address was the possibility of creating an algorithm that would allow people to identify when a wildfire could affect them and either prevent or prepare for the event.

I believe that I have addressed the original statement by creating a platform for a new variables 'Forest Future' to be created and used as a preventative measure. The method of creating this new variable is by combining Forest Past & Drought against Forest Present & Dought. I believe that this can create a variables called 'Drought Trajectory'. I discovered earlier in Part 2 of this excercise that drought is the most accurate factor in predicting spread as once the drought variable passes the 50th percentile that area's chances of experiencing a wildfire TRIPLES from a 25% chance to a 75% chance of occurance.

I beliebve that the analysis I have completed has been tested against itself in multiple variables, along with outside research on subjects such as 'Fuel moisture' that it is both accurate and helpful. Factors can also be measured to predict when a spike in drought will occur due to the correlation charts that were used for an area's natural and fuel moisture.

The implications of this data are that the initial issue is actual capable of being solved. Originally, I viewed this issue as an idea that would be interesting to understand and maybe contribute too.