California Wildfire Analysis

Shih-Chi Chen, Kevin Xu, Selam Berekat, Wyatt Workman

5 June 2022

Abstract

In recent years, California has experienced an increase in number of wildfires. Scientists believe that this is a result of human-caused climate change. The first part of our analysis aims to determine which counties and regions of California have been impacted by wildfires more so than others. Also, we will examine different causes of wildfires and which counties or regions of California were more susceptible to fires of a certain cause. In the second part of the analysis, we will examine weather data from different parts of California from different years to examine if there are any noticeable difference in weather patterns in high fire years versus low fire years. Finally, the T test of standardized regression will be used to do the important factor analysis by coefficients.

1 Project Goals and Questions of Interest

The primary goal of our analysis is to answer five questions of interest:

1. Which California counties and regions have experienced the most number of wildfires between 1992 and 2015?
2. What are the primary causes of these wildfires?
3. Are certain regions more susceptible to wildfires of a certain cause?
4. Are there any discernible weather patterns in high and low fire years for different regions of California?
5. Which variables are statistically significant in regard to fire size?

To answer these questions, we aim to gather our data from non-tabular sources. This will allow us to use our data-scraping skills, which are necessary in many industry settings. Our data on wildfires will be obtained from a SQL database file, and we will use an API available through the National Oceanic and Atmospheric Administration (NOAA) to obtain pertinent weather data. Additional information on the data for each part of our analysis can be found in sections 2.1 and 3.1. Note that our data can be found in the .csv files in our Canvas submission.

2 Analyzing Wildfire Patterns

In this section, we will describe our methodology regarding our analysis of the wildfire data. We begin by describing the source and the nature of the data used. Next, we will show some exploratory visualizations that will be used to gain insight on the different parameters of interest regarding the wildfires. Then, we will examine some geospatial plots to visualize the locations of different types of wildfires. Finally, we will interpret our results explain their significance.

2.1 Data Description

Our wildfire data for this project was obtained from a SQLITE database file that contains data from 1.8 million wildfires that occurred in the United States between 1992 and 2015. The data were obtained from various federal, state and local fire reporting systems.

For the scope of our project, we opted to focus our analysis on fires that occurred in California. This decision was made in part because we realized that analyzing all of the fire data would slow down computation times significantly, and that this would lead to very broad research questions. We extracted our data using a SQL query to select variables of interest, and filtered by observations with "CA" listed as the location.

2.2 Exploratory Plots

Figure 1 (below) shows that there were more fires in 2006 and 2007 than any other year in our data set. The main causes of fires are lightning in 2006 and miscellaneous in 2007. In addition, lightning appears to have caused the most number of wildfires in each year. According to Wikipedia, the major factors to the extreme fire were drought in Southern California, hot weather, and the unusually strong Santa Ana winds, with gusts reaching 112 mph (180 km/h) in 2007 [11].

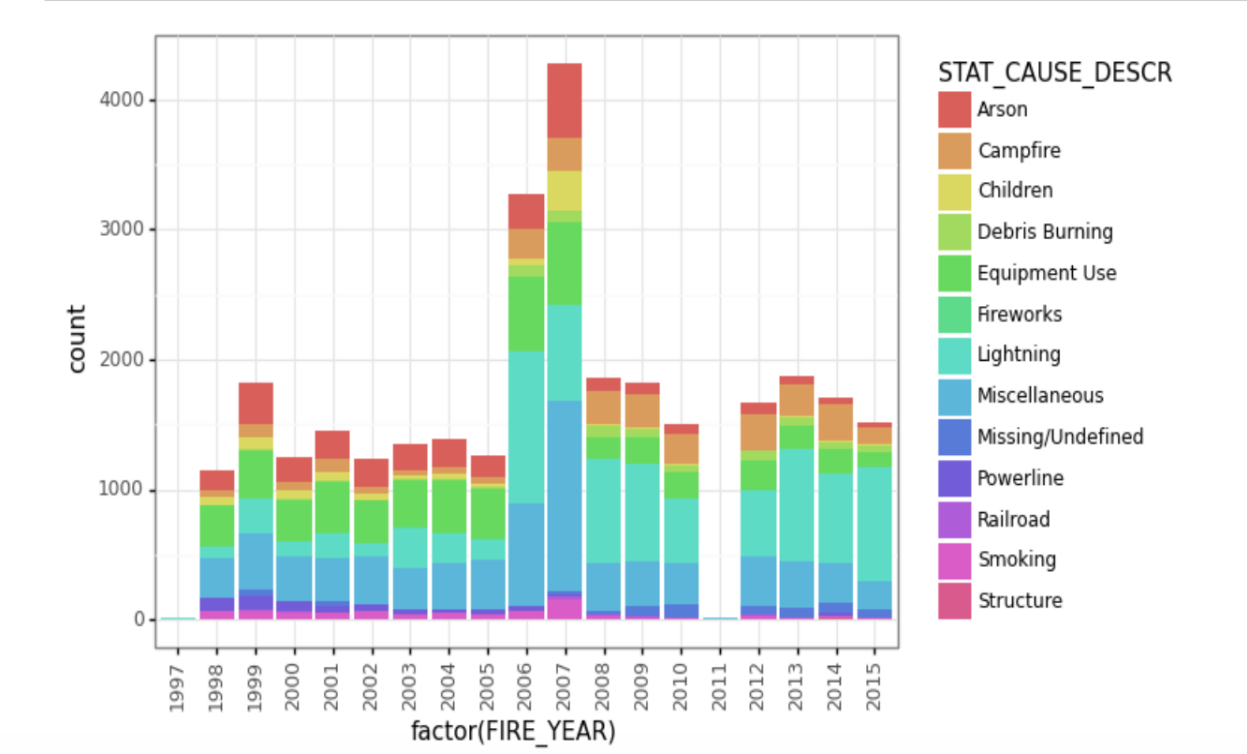


Figure 1: Fire Years vs Description of the cause of the fire. Additional details on how this plot was produced can be found in the STA160-Final Project Plots Cindy.ipynb

Figure 2 (below) shows that Riverside had the most fires of all counties. In Particular, most of the fire size class is of class A; the final fire size is less than or equal to 0.25 acres. Furthermore, those counties with more fires tend to be in southern California, such as Riverside, Los Angeles and San Bernardino.

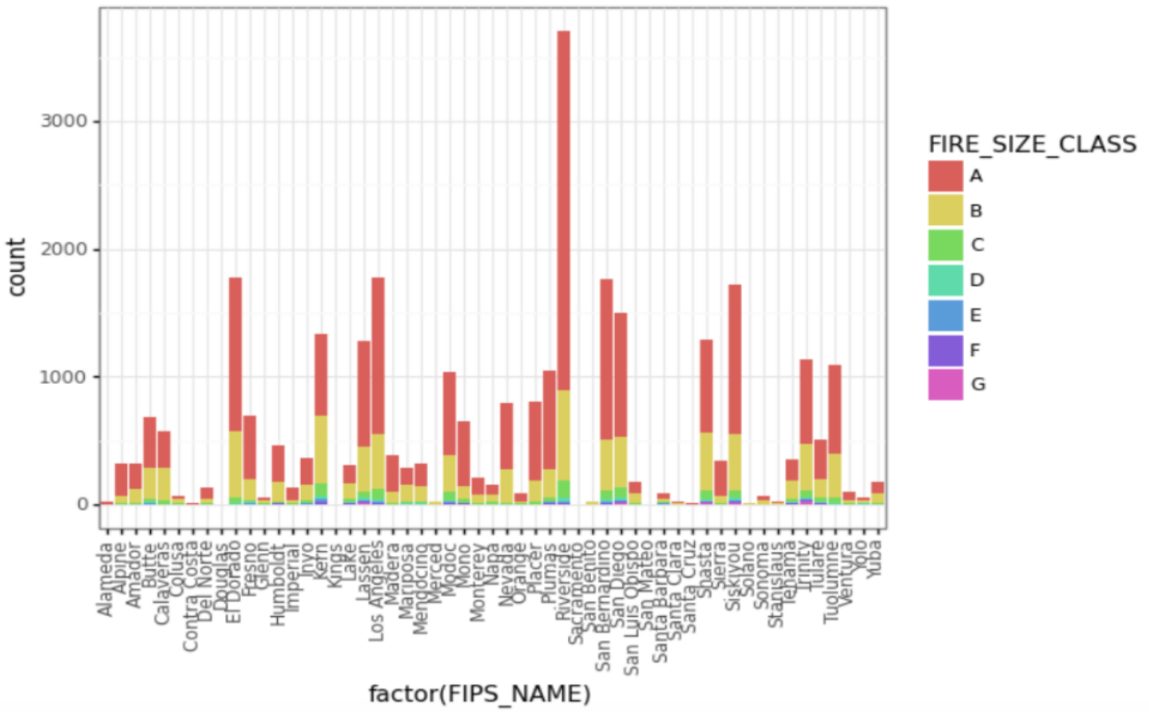


Figure 2: FIPS Name vs Fire size class. Additional details on how this plot was produced can be found in the STA160-Final Project Plots Cindy.ipynb

Figure 3 (below) shows that the largest fire size was 315,578.8 acres on August 12, 2012 in Lassen. The cause of this fire was due to lightning. Since the fire size was large, the burned area covered California and Nevada. In addition, the burned area in California made this fire the second-largest wildfire in California since 1932. Figure 3 also shows that the fire size tends to become larger after 2005 [10].

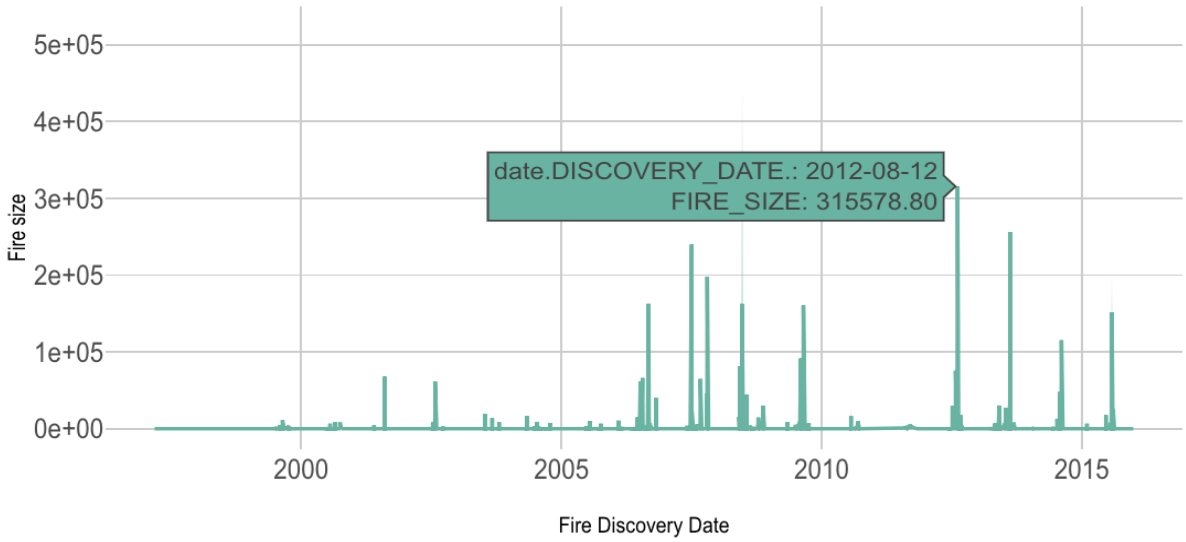


Figure 3: FIPS Date vs Fire size. Additional details on how this plot was produced can be found in the sta160-Final Cindy.Rmd

Finally, Figure 4 (below) shows that most small size fire continuous days are less than 50. And for those large size fires tend to have more continuous days to put out fires. However, there are some outliers that show up in the plot. For example, there are two fires not large but with largest continuous days more than 300 days. Besides, a little linear relationship between these two variables is found if the outliers are excluded.

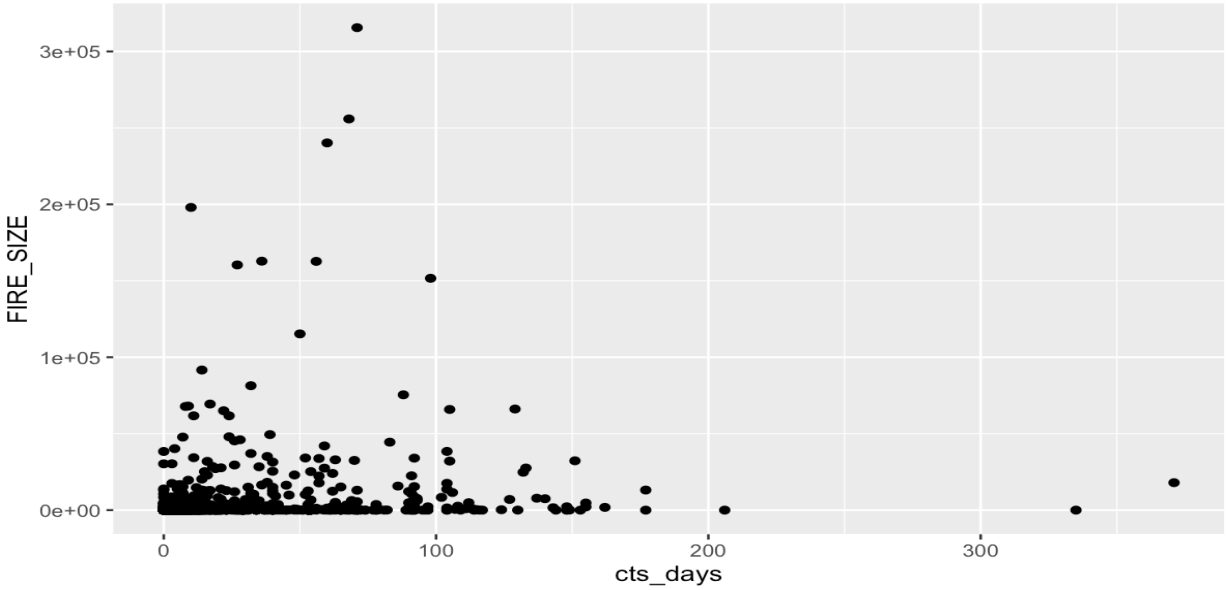


Figure 4: Fire continuous days vs Fire size. Additional details on how this plot was produced can be found in the sta160-Final Cindy.Rmd

2.3 Geo-spacial Visualizations

Next, we turn our attention to geospatial visualizations. The objective of producing these plots is to find patterns and make observations as to where certain wildfires occur. Specifically, we wanted to examine the causes of wildfires more closely, and determine if certain regions experienced more wildfires due to a specific cause.

The first step in producing these visualizations was to produce a plot of California, with all of the county boundaries, in order to visualize fires by county. This was accomplished by using a data file known as a shapefile, a data file that can be imported as a dataframe and contains a "geometry" field that describes each county as a geometric object in terms of latitude and longitude. Various shapefiles can be found on government websites. The shapefile used in this project can be found on the Open Data Portal on CA.gov. Figure 5 (below) shows the resulting visualization from this shapefile data.

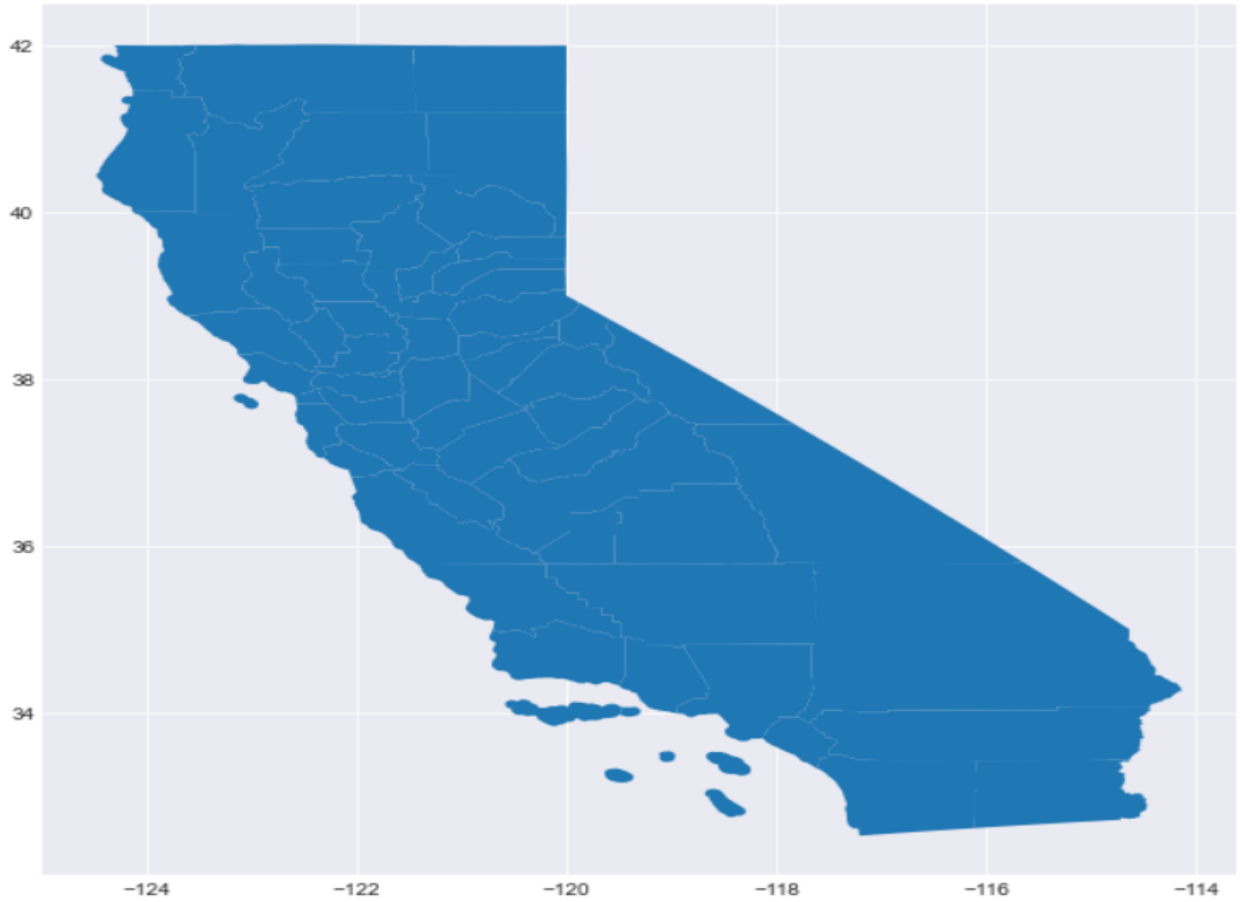


Figure 5: Visualization of California produced from a shapefile. Note that the x and y axes are scaled according to latitude and longitude. See the fire visualizations.ipynb for more information.

Having established this plot of California, it is now possible to plot each fire in its respective county. This required the use of the Shapely and Geopandas packages. The Shapely package was used to create a geometric object from the latitude and longitude of each fire from our wildfire data that was obtained from the SQL database. The Geopandas library was used to store these geometries in a pandas dataframe.

Next, we began the process of visualizing all of these fires on a map of California. First, we were interested to see which counties experienced the most number of wildfires. To answer this question, we produced a chloropleth map, which shows count data by differing color gradients (figure 6, below)

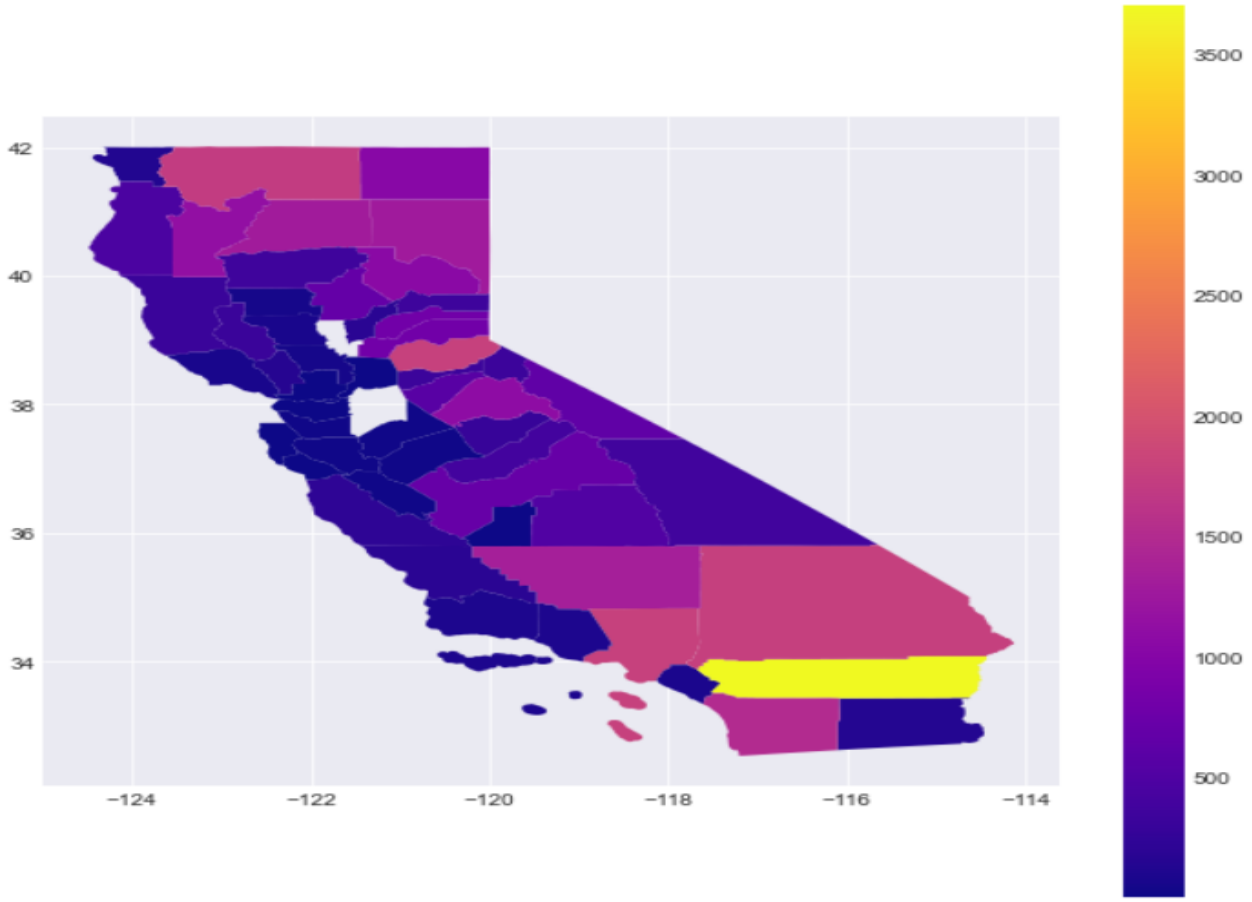


Figure 6: Chloropleth map showing number of wildfires by county. Additional details on how this plot was produced can be found in the fire visualizations.ipynb notebook

From this plot, we can obtain a general idea of which counties experienced the most and least numbers of fires. Of equal importance, we can see which regions of the state were more greatly impacted by wildfires during the time period our data was recorded (1992-2015). First, it appears that a majority of the fires have occurred in the Northeast, Foothill, and southern regions of the state, while relatively few wildfires have occurred in the coastal and delta regions of the state. Another result that we found interesting from this plot is that Riverside county was the county that experienced the most number of fires (the county colored yellow in the southern part of the state). It should be noted that there were two counties that our dataset had no wildfire data. These are the counties that are uncolored in figure 6 above.

Finally, we visualized the location of each fire on the map of California. Specifically, we created additional dataframes for each fire cause of interest and plotted their location. In doing this, our goal was to identify if certain regions or counties experienced more fires due to a particular cause. Due to the fact that there are 13 different fire causes in our dataset, we will only present a few of the visualizations our group found to be the most interesting. For reference, these fire location visualizations can be found in the fire visualizations.ipynb notebook.

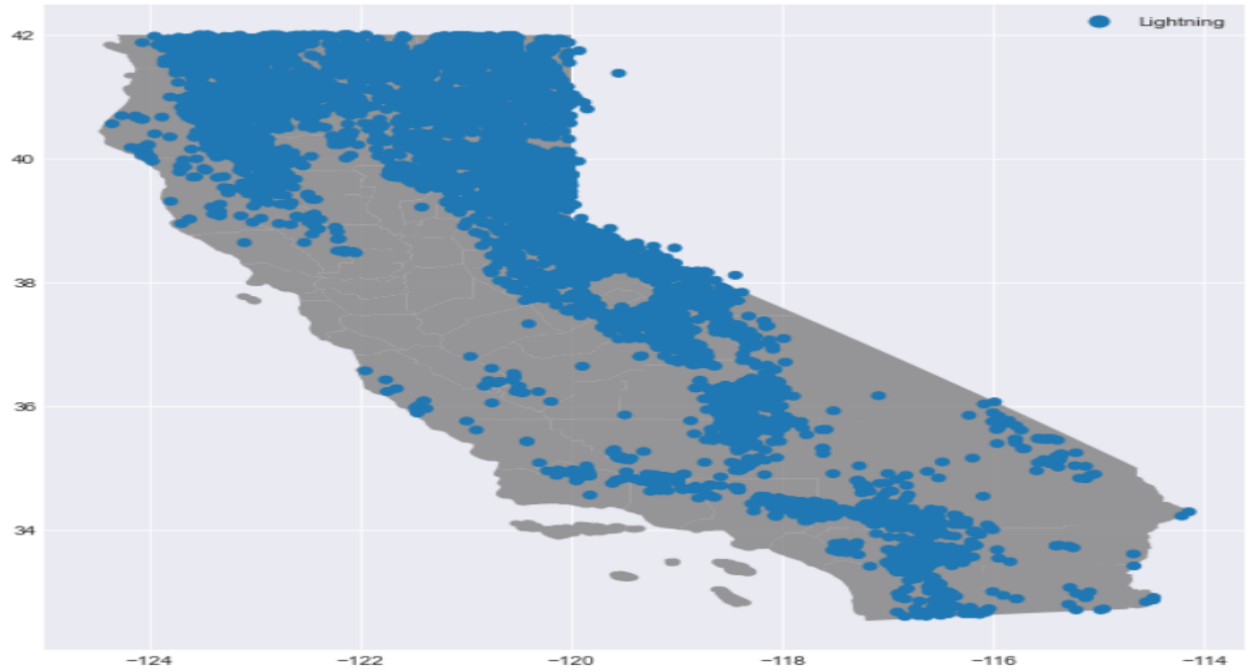


Figure 7: Lightning-caused wildfires. We can see that this primarily affected the Sierra Nevada range and the northern part of the state. Lightning also tends to cause fires in more forested parts of the state. Additional details on how this plot was produced can be found in the fire visualizations.ipynb notebook

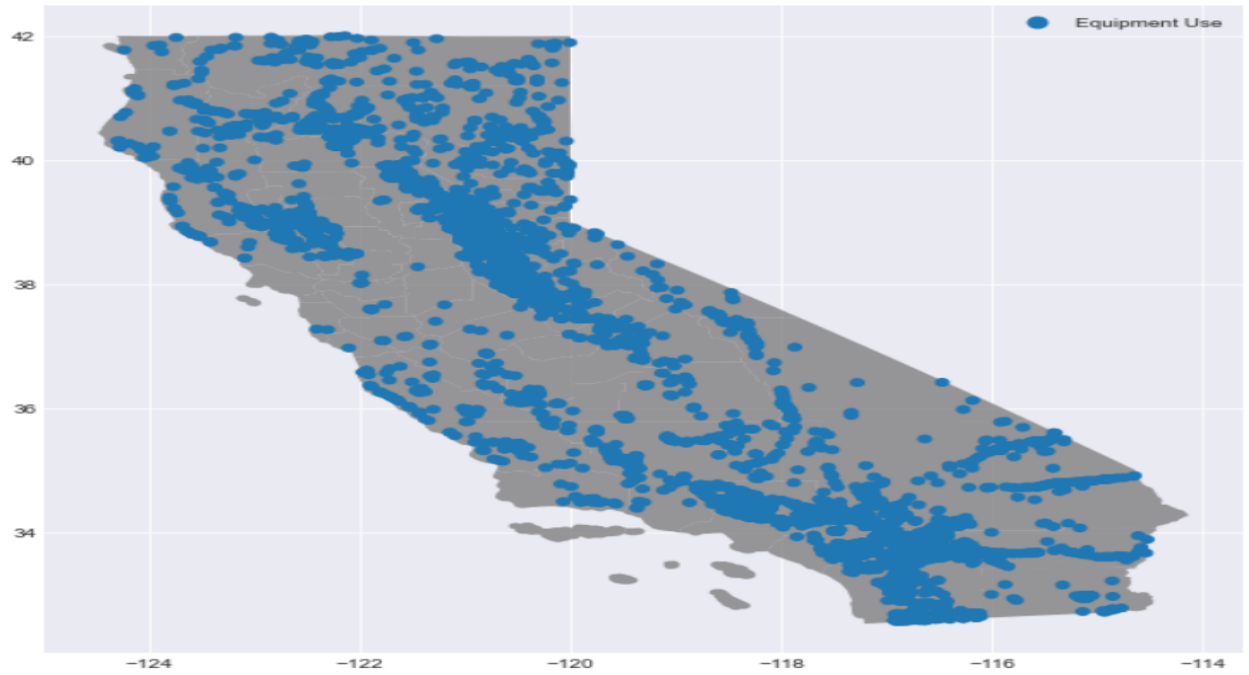


Figure 8: Wildfires caused by equipment use. There is a larger number of wildfires in the southern part of the state as a result of equipment use, as well as in the central valley. This is likely due to agricultural work. Additional details on how this plot was produced can be found in the fire visualizations.ipynb notebook

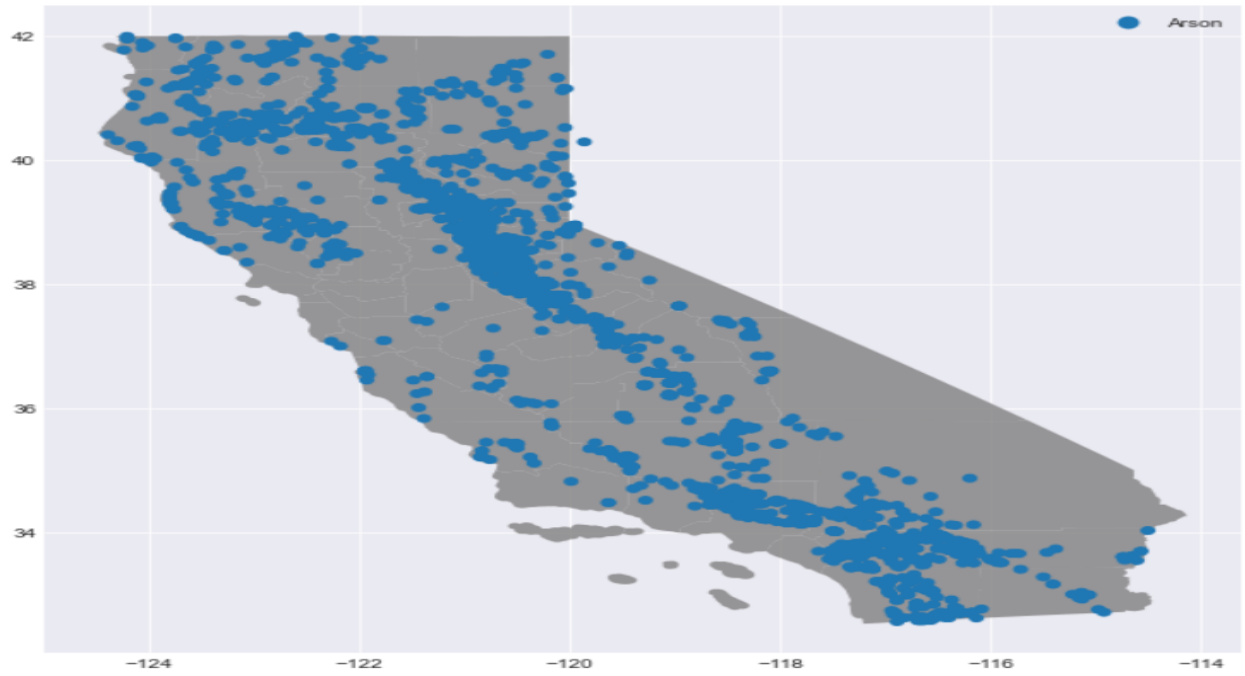


Figure 9: Wildfires caused by arson. We can see that most fires caused by arson occurred in the northern Sierra Nevada ranges, and in the southern part of the state. Additional details on how this plot was produced can be found in the fire visualizations.ipynb notebook

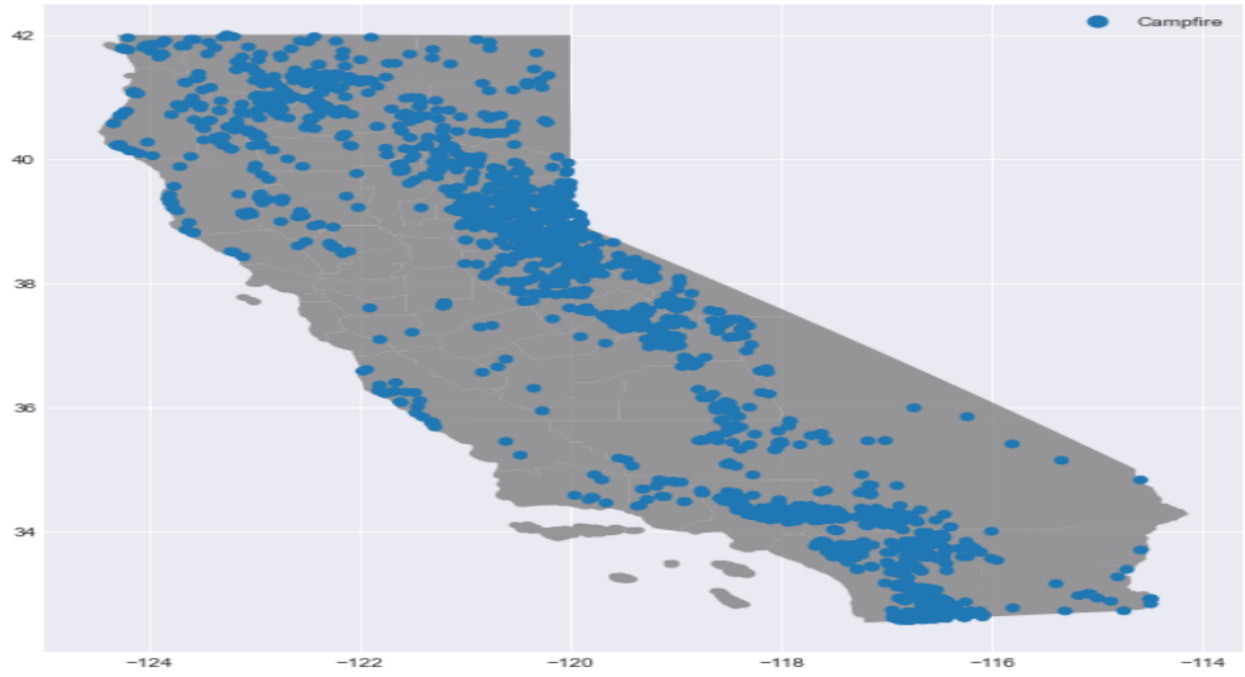


Figure 10: Wildfires caused by campfires. Again, these fires primarily occurred in the southern part of the state, and the Sierra Nevadas. Additional details on how this plot was produced can be found in the fire visualizations.ipynb notebook

2.4 Interpretations and Next Steps

From the exploratory plots and geospatial Visualizations, it can be seen that most of the fire occurs in Northern and Southern California, especially in Riverside county, which experienced the most number of fires. In addition, we found that most of the wildfires in our dataset were caused by lightning strikes. Next, in order to perform a more in-depth analysis on other weather-related factors that lead to wildfires, weather data from NOAA will be combined with our wildfire dataset.

3 Weather Patterns in Wildfire-Prone Areas

Having determined the regions and counties that experienced more wildfires, we turn our analysis to weather data. In this section, our goal is to examine weather data from different counties and different years in which there were more or less wildfires, and if there are any discernible differences between counties. Specifically, we will focus our analysis on data from the years 2003, 2006, 2007, and 2010, for the following 6 counties:

- El Dorado County
- Riverside County
- Mendocino County
- Merced County
- Imperial County
- Siskiyou County

We were specifically interested in the years 2006 and 2007, because they had the most number of wildfires in our dataset. In addition, we picked counties from different regions of the state to gain an idea on how weather patterns differ by regions, and in counties in these regions that had a large and small amount of wildfires.

3.1 NOAA API Description

National Oceanic and Atmospheric Administration ("NOAA") is the reliable source used to scrape weather data. According to the visual plots of the fire dataset, there has been a significant amount of fire in the state of California from the year 2006 to 2007. In this project, the analysis was narrowed to six counties with significant fire for the years 2006, 2007, 2003, and 2010. The importance to access the weather data from NOAA was to compare the weather observation data with the fire dataset for the selected counties and years.

The procedure used to scrape the weather data was to first request a unique token from the website. Upon the permission to access their data, the selection preference was based on a daily summary of the weather for the specific location FIPS code of the counties, and the specific years are 2006, 2007, 2003, and 2010. This includes the observation type of data that were maximum temperature (tmax), minimum temperature (tmin), rainfall precipitation (PRECIP), wind speed, and average temperature (average). Following, the scraped data were merged and stored in the python panda's data frame for ease of analysis. Below is the descriptive summary of the combined data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12183 entries, 0 to 12182
Data columns (total 23 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Unnamed: 0          12183 non-null  int64
 1   Date                12183 non-null  object
 2   wind_speed          12183 non-null  float64
 3   FIPS_NAME_x         12183 non-null  object
 4   FPA_ID              315 non-null    object
 5   FIRE_CODE           315 non-null    object
 6   FIRE_NAME           315 non-null    object
 7   FIRE_YEAR           315 non-null    float64
 8   DISCOVERY_TIME      315 non-null    float64
 9   STAT_CAUSE_CODE     315 non-null    float64
 10  STAT_CAUSE_DESCR    315 non-null    object
 11  date(CONT_DATE)     315 non-null    object
 12  FIRE_SIZE           315 non-null    float64
 13  FIRE_SIZE_CLASS     315 non-null    object
 14  LATITUDE            315 non-null    float64
 15  LONGITUDE           315 non-null    float64
 16  FIPS_CODE           315 non-null    float64
 17  FIPS_NAME_y         315 non-null    object
 18  STATE               315 non-null    object
 19  tmax                315 non-null    float64
 20  tmin                315 non-null    float64
 21  Precip              315 non-null    float64
 22  average             315 non-null    float64
dtypes: float64(12), int64(1), object(10)
memory usage: 2.1+ MB
```

Figure 11: Brief summary about the Data Frame that includes number of columns, column names, data type of each column, non-null values and memory usage. See 01.web.scrape.ipynb for more information.

3.2 Exploratory Analysis

We begin our exploratory analysis by examining the average temperatures in our dataset scraped from NOAA. Figure 12 (below) shows the distribution of average temperature for each county.

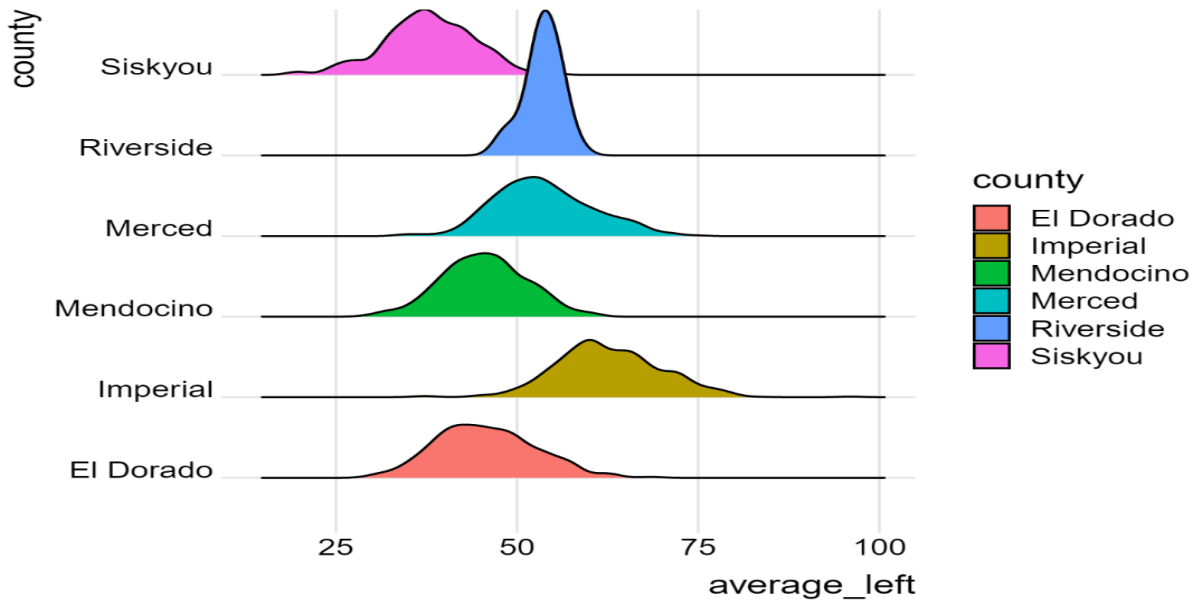


Figure 12: Distribution of average temperature between 2003 and 2010. See STA160 allcom plots.rmd for more information.

Although there appear to be slight differences in mean temperature for each county, it does not appear to be statistically significant, as all of the distributions are reasonably close to each other, and they all overlap to some degree. In addition, when we examine the distributions with respect to the geographic location of each county, we see that counties in northern California (El Dorado, Mendocino, Merced, and Siskiyou), tend to have cooler average temperatures, and are reasonably close to each other. The southern California counties (Riverside, Imperial) tend to have warmer average temperatures, and are also reasonably close to each other.

Next, we examine the distribution plots of average precipitation for each county (figure 13, below):

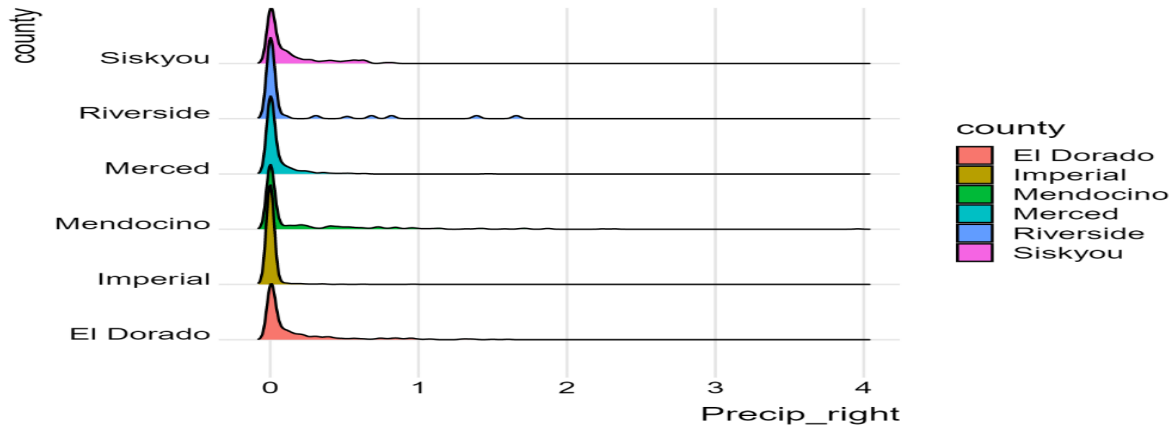


Figure 13: Distribution of average precipitation between 2003 and 2010. See STA160 allcom plots.rmd for more information.

From these plots, we see that this time period was quite dry, although there do not appear to be any significant differences between counties.

Finally, we observe the distribution plots of average wind speed in each county (figure 14, below):

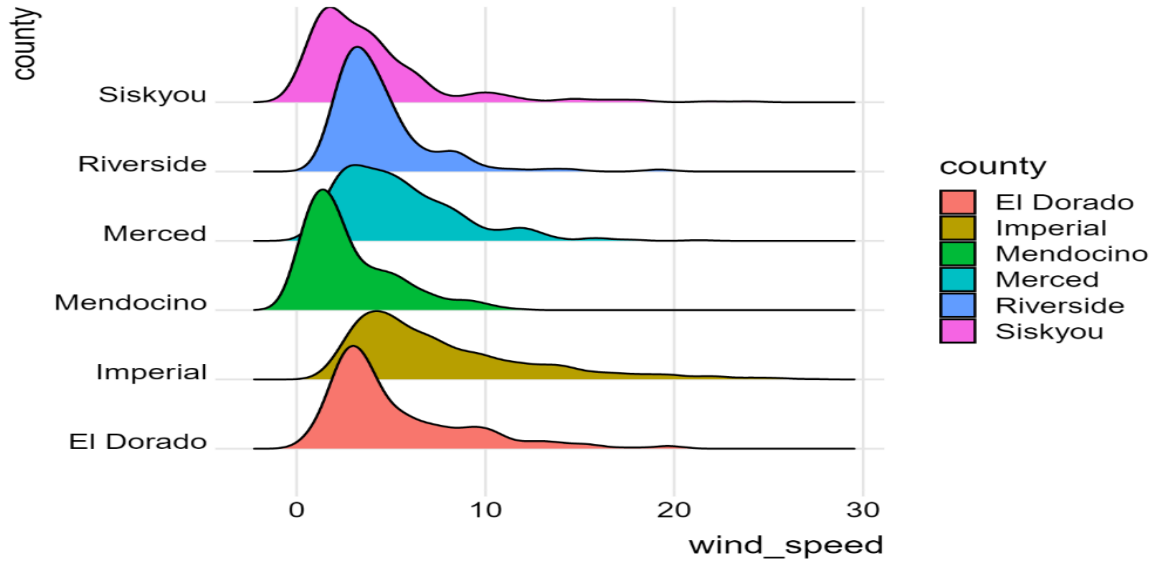


Figure 14: Distribution of average wind between 2003 and 2010. See STA160 allcom plots.rmd for more information.

Again, it does not appear that there are any significant differences in average wind speed, due to the overlap of each distribution. However, by looking at region rather than at the county level, we see that southern counties had a higher average wind speed than the northern counties.

Next, we visualized wind speed for each fire size class in our dataset, as well as for each fire cause. These can be seen below in figures 15 and 16, respectively.

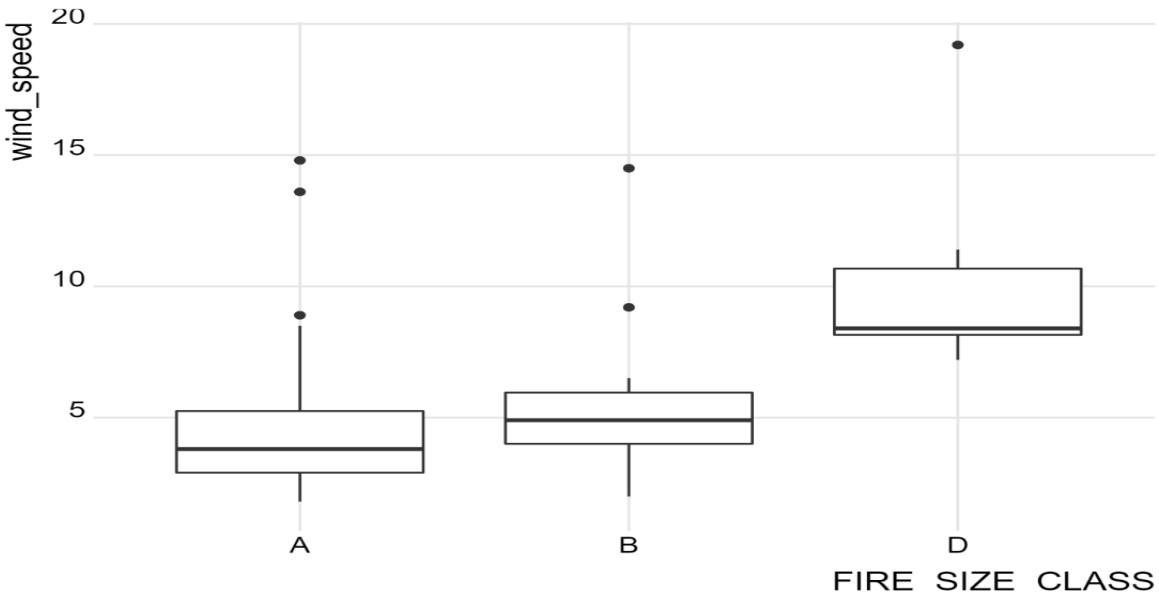


Figure 15: Boxplot of average wind speed for each fire size class, between 2003 and 2010. See STA160 allcom plots.rmd for more information.

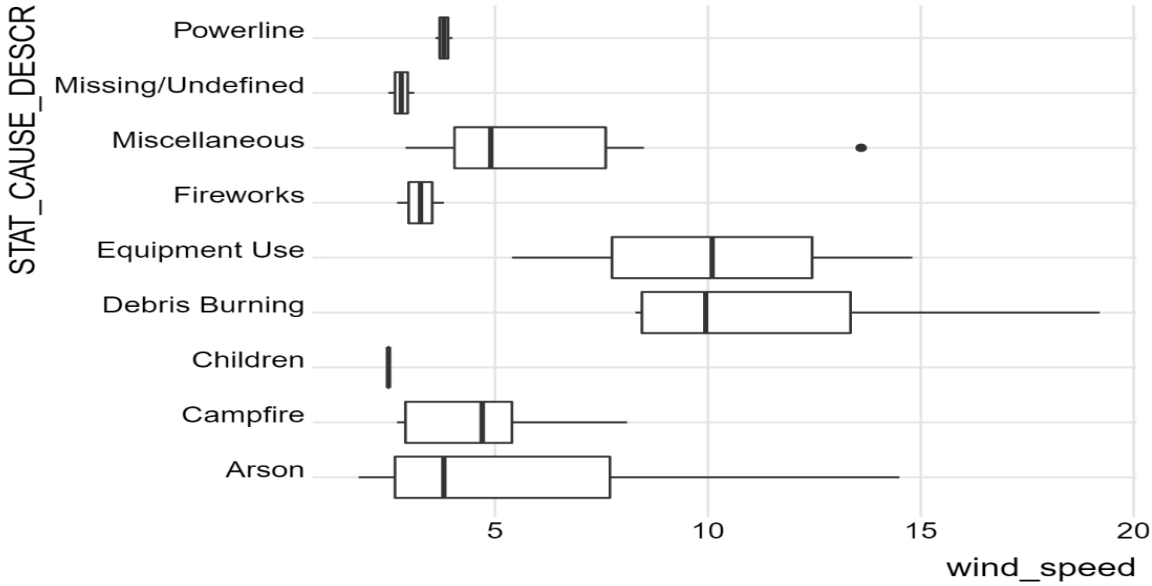


Figure 16: Boxplots of wind speed by cause of wildfire, between 2003 and 2010. See STA160 allcom plots.rmd for more information.

From figure 15, we see that the wind speed on fire discovery dates of class D wildfires tend to be much higher than smaller fires class A and B, of which there appears to be little difference. The fact that the box for class D fires has no overlap with boxes for fire classes A and B leads us to believe that there is a significant difference in wind speed on discovery dates of wildfires that turn out to be class D wildfires versus class A and B wildfires.

Figure 16 yields similarly interesting findings. We observe that wind speeds for fires caused by equipment use and debris burning tend to happen on windier days than any other fire. Further, given that none of

the other box plots overlap with equipment use or debris burning, we can conclude that, on the basis of this data, there is a significant difference in wind speed for fires caused from equipment use or debris burning than for any other cause in this dataset.

Finally, given that we found earlier that there appeared to be differences in average temperature between regions (figure 11), we decided to produce a box plot of average temperature by fire size class, to see if a significant difference between classes exists. Figure 17 below shows the resulting plot.

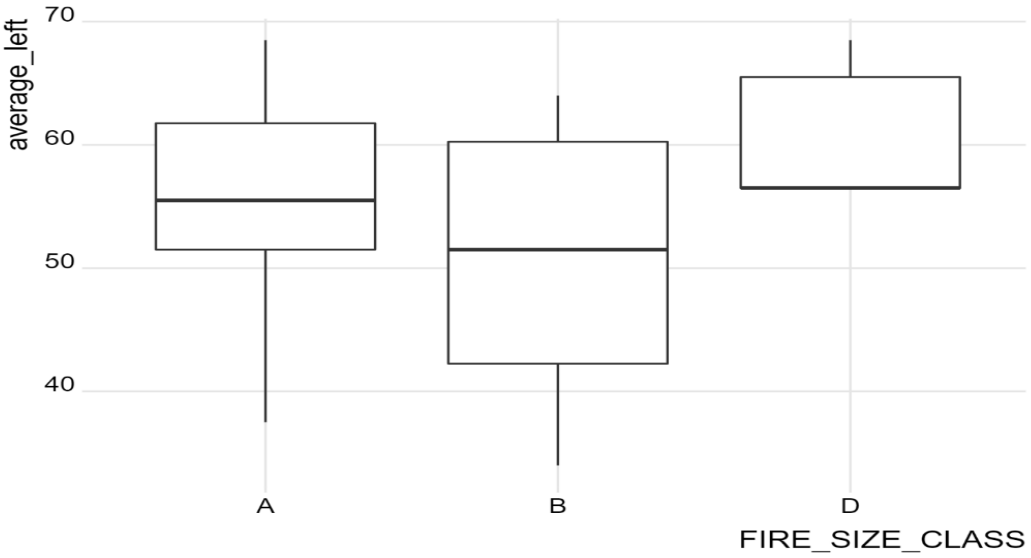


Figure 17: Boxplots of wind speed by cause of wildfire, between 2003 and 2010. See STA160 allcom plots.rmd for more information.

Since there is overlap between each box for each fire class, we conclude that, on the basis of our data, that average temperature on the wildfire discovery date does not differ between different fire size classes.

3.3 Hierarchical Clustering

Hierarchical clustering is a agglomerative (bottom-up) clustering method in which the general structure of data points can be observed from a distribution-free perspective. It observes the closeness between data points based on the L2 norm of k continuous variables in $\mathbb{R}^k$ space.

Procedure:

: Only continuous variables will be used to compute distance between variables. We consider weather information: 1) average weather, 2) precipitation (rainfall/humidity), 3) wind-speed.In the initial

case the number of clusters is equal to the number of data points. (cluster centroid (mean) is the location of data point) Next, clusters with the smallest dissimilarity metric, among all pairs of clusters, are conjoined together using Ward's method. Ward's method uses a sum of squares dissimilarity metric rather than the conventional L2 norm. It locates the centroid between two clusters and computes the sum of squares of the combined cluster minus the sum of the sum of squares of each cluster individually. This process of cluster conjoinment continues until all clusters conjoin into one big cluster. Ward's distance metric is shown below:

$$\mathrm{SS}(C) = \sum_{j \in C} \|X_j - \overline{X}_C\|^2 = \frac{1}{2|C|} \sum_{j \in C} \sum_{k \in C} \|X_j - X_k\|^2,$$

with $\overline{X}_C$ denoting the centroid of the X_i in C

Observations: We investigate the hierarchical clusters on three instances of data:

- All weather data from January 2003 to May 2010.
- 10 percent random subsets of all weather data from January 2003 to May 2010.
- Daily county weather patterns for all day within one week of fire discovery.

1) Investigating all weather data from January 2003 to May 2010:

- This is shown in Figure 18 below.
- We can see that there exists two major clusters within our data. As the two clusters form their respective claves at highly differing L2 norms, we can see that one of the clusters is much 'tighter' than the other. Also, the centroids of the two clusters are quite separated. This may give indication that weather data (temperature, precipitation, wind speed) may only be effective at identifying certain classes of fire size.

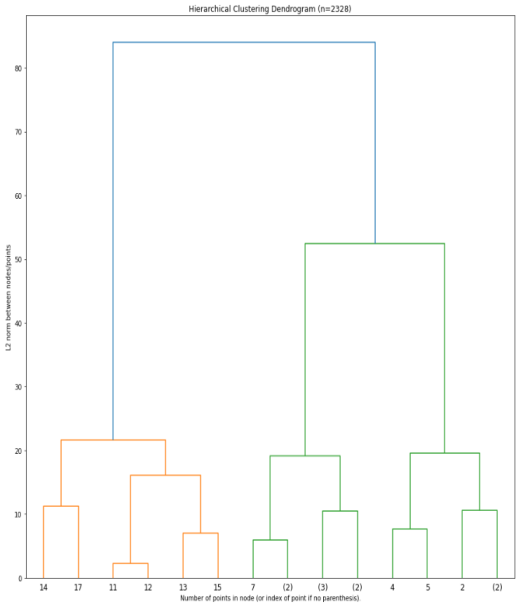


Figure 18: Dendrogram of all weather data from January 2003 to May 2010. See weather setup.ipynb for more information.

2) Investigating 10 percent random subsets of all weather data from January 2003 to May 2010.

- This is shown in Figure 19 and 20 below.
- The 10 percent random subsets of weather data show that there exists two or three major clusters of weather data. This may be indicative of three major classes of weather condition that may give rise to fires.

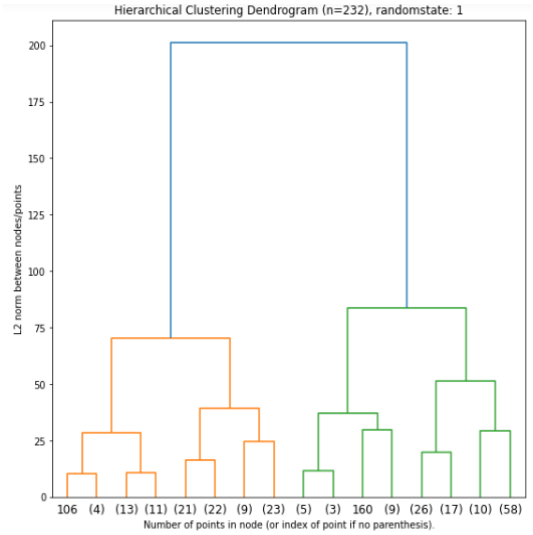


Figure 19: Dendrogram of 10 percent random subsets of weather data from January 2003 to May 2010. See weather setup.ipynb for more information.s

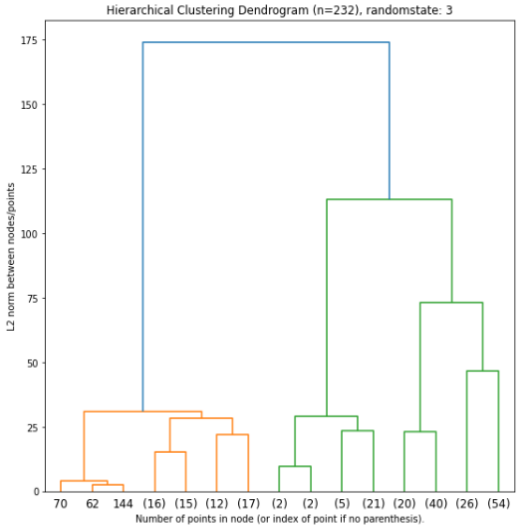


Figure 20: Dendrogram of 10 percent random subsets of weather data from January 2003 to May 2010. See weather setup.ipynb for more information.

3) Investigating daily county weather patterns for all days within 1 week of fire discovery.

- This is shown in Figure 21, 22, and 23 below.
- We expect that there will be two major hierarchical clusters indicating the difference between class A and class B weather structures. (with significantly more observation belonging to the cluster containing most class A observations). (19 A fires, 10 B fires, 2 D fire.)
- We take a semi-supervised learning approach where we cluster weather patterns of a location in a 15 day date range with the 8th day being the fire discovery date. With this procedure, we will have knowledge of the fire-class associated with certain daily weather information. For instance, we can identify the wind, temperature, and precipitation information of a location that will have a class A fire in 3 days, and we can assign class A to that data point.
- Interpretation: As the fire discovery date approaches, weather patterns in euclidean space do not become much more differentiable between class A and class B fires. This implies that the predictability of fires severity from natural causes are not significant from the variables temperature, wind speed and precipitation. Similarly, the weather patterns following fire discovery do not given strong indication of the class of a given fire either. Note that due to the limited overlap of available weather data during the periods that fires occurred, we were only able to use 31 fire cases to conduct this hierarchical clustering analysis. Due to the limited sample size, our findings may not be representative of weather structure of fire classes at scale.
- The dendrograms below are associated with the aggregated county weather patterns 3 days before fire discovery, on the day that a fire was discovered, and 3 days after fire discovery.

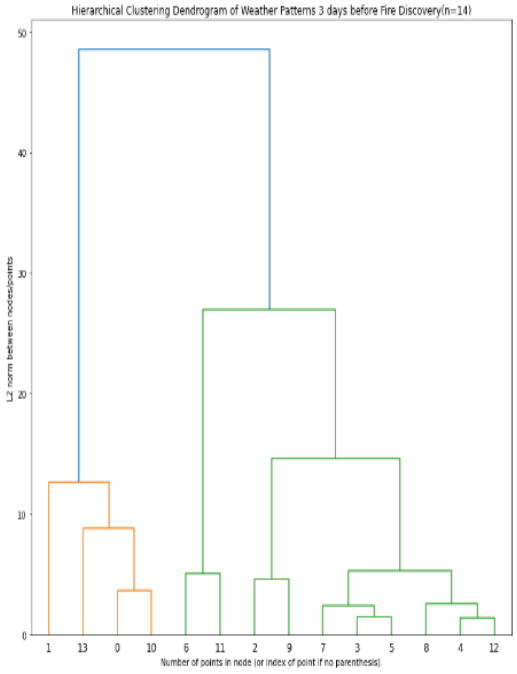


Figure 21: Dendrogram of county weather data three days before fire discovery. (n=31) See weather setup.ipynb for more information.s

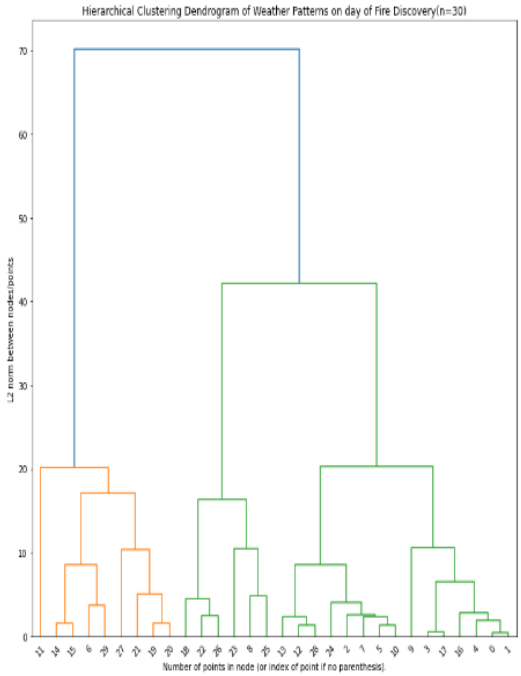


Figure 22: Dendrogram of county weather data on the day of fire discovery. (n=31). See weather setup.ipynb for more information.

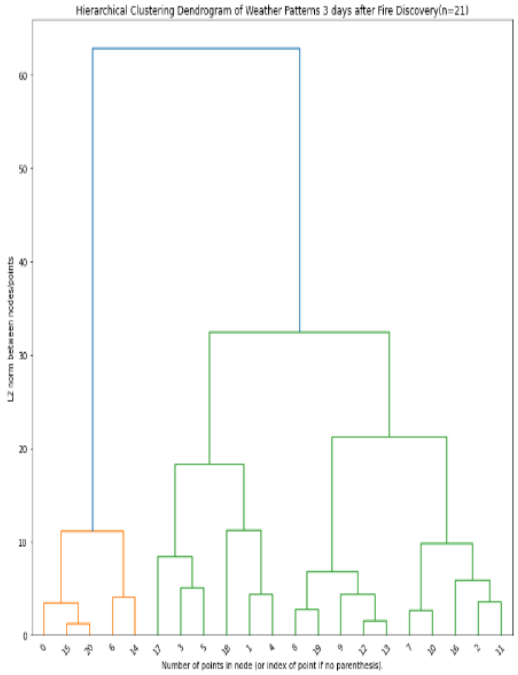


Figure 23: Dendrogram of county weather data three days after fire discovery (n=31). See weather setup.ipynb for more information.

3.4 Regression Analysis

In order to find which factors have a significant effect on the fire size, multiple and standardized regression will be used to analyze p-values on variables. In this study, we only focus on analyzing the following continuous variables associated with fire size: Fire YEAR, Precip right (Rainfall), average left(average temperature),wind speed, cts day(fire continuous days), LATITUDE, LONGITUDE.

3.4.1 F Test of Multiple Regression Analysis

State the hypothesis:

$$H_0 : \beta_0 = \beta_1 = \beta_2 = ...\beta_7 = 0$$
$$H_a : At\ least\ one\ of\ \beta_k \neq 0$$

Since the F-stat's p-value is small (8.837e-08), we would reject the null hypothesis and conclude that at least one of these predictor variables and fire size have significant relationship.

3.4.2 T Test of Standardized Regression (Important factor analysis by coefficients)

In order to find which factor has the most significant influence on fire size, standardized regression analysis is needed because each predictor variable has different units.

After the standardized regression is conducted, the t-test p-values on Figure 24 shows that most of variables have small p-values. Especially, Continuous days, Wind speed and Latitude have the smallest p-values since their absolute test statistics are the largest. For fire continuous days, the fire size increases 0.6626 as fire continuous day increases one day. For wind speed, the fire size increases 0.3752 as wind speed increases one unit. On the other hand, for Latitude, the fire size decreases 1.25 as latitude increases one unit. This result is the same as we previously showed on the geospatial Visualizations. Northern California (large latitude) has fewer fires than South California (small latitude).

To sum up, according to these standardized regression coefficients, Continuous days and Wind speed are the most significant influence on fire size with a positive relationship to the fire size while Latitude are the third highest significant influence on the fire size with a negative relationship to the fire size, respectively.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.807e-15  9.027e-02   0.000 1.000000
scale(FIRE_YEAR)    4.340e-02  1.111e-01   0.391 0.698446
scale(wind_speed)   3.752e-01  9.701e-02   3.868 0.000443 ***
scale(Precip_right) -1.479e-01  1.025e-01  -1.443 0.157558
scale(average_left) -4.164e-02  1.233e-01  -0.338 0.737593
scale(cts_days)     6.626e-01  1.086e-01   6.099 5.13e-07 ***
scale(LATITUDE)    -1.250e+00  3.635e-01  -3.439 0.001493 **
scale(LONGITUDE)   -1.168e+00  3.707e-01  -3.152 0.003266 **
```

Figure 24: Standardized regression summary. Additional details on how this summary was produced can be found in the sta160-Final Cindy.Rmd

3.5 Time Series Analysis

Time series analysis is a type of data analysis where the set of data points is taken at a specified time or usually at equal intervals. We use the analysis to predict future unobserved values based on the observed data values. It uses just one variable "time" to extract meaningful statistics and other important characteristics. Also, it is applicable to perform business forecasting, understand/analyze past behavior, plan for future events, and evaluate current accomplishments.

The goal of this project is to compare the climate change pattern in the six counties from the years 2006 to 2007 and 2003. However, due to a lot of missing observations for the years 2006 and 2007, the focus of the analysis will only be on the year 2003.

The research question is to check whether assumptions are met to fit the right model that predicts the monthly wind speed rate in the counties. The procedure is to first check whether the time-series data are stationary. This means the data should have a constant mean, constant variance, and an auto-covariance that should not depend on time. In addition, the transformation of data can be applied to remove noise and meet the stationarity assumption. This follows by computing the sample autocorrelation and partial-correlation (pacf). Finally, we fit the model and select the optimal model based on the selection criterion of Ljung-Box statistic that contains the smallest AIC value.

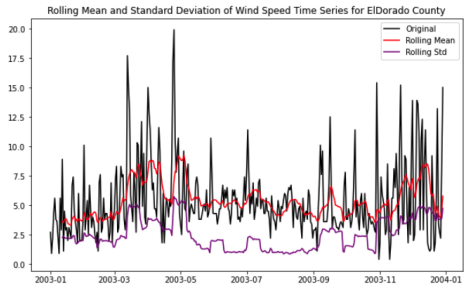
We first explore the data using exploratory graphics of a time series plot. This helps to determine whether the assumption of time series holds or not. The first important assumption is stationarity. It determines if the mean, variance, and autocorrelation of the time variable are constant over time. This means the overall behavior of the data should remain constant and not the same. This helps to apply the right models to forecast future behaviors.

Following to it, Augmented Dickey-fuller test is a statistical significance test applied for time series data to compute test statistics, critical value, and p-value to infer about the given set of data. This method provides a negative number to help determine the result of the hypothesis test. Auto-lag is a function to calculate the difference using the correlation matrix between values that are at one time period apart. In this case, we have a lag of a 12-month time period apart. The function helps in finding the good quality statistical models for the given data using an AIC estimator for predicting error. This applies by selecting an AIC with lower values as an optimal/better fit.

Hypothesis Test:

$$H_0 = \text{Given data is not stationary}$$
$$H_a = \text{Given data is stationary}$$

(1) El-Dorado County



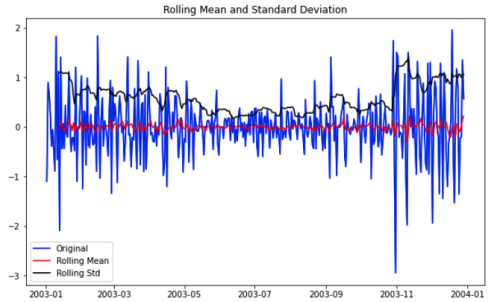
(a) Time Series Plot with Rolling Mean and Standard Deviation for El Dorado County

```
Result of Dickey-Fuller test
Test Statistic                -6.227550e+00
p-value                        5.045358e-08
#Lags Used                     4.000000e+00
Number of Observation Used     3.580000e+02
Critical Value (1%)           -3.448749e+00
Critical Value (5%)           -2.869647e+00
Critical Value (10%)          -2.571089e+00
dtype: float64
```

(b) Statistical Test Result

Figure 25: Initial Exploratory Graphics and Statistics. See 03.timeseries.2003.windspeed.ipynb for more information.

In the above plot, we observe the behavior did not remain constant and thus stationarity does not hold. The statistical result indicates the data is not stationary because the test statistic (-6.22e+00) is less than the critical value (-2.86e+00) at a 5% significance level.



(a) Transformed Time Series Plot with Rolling Mean and Standard Deviation for El Dorado County

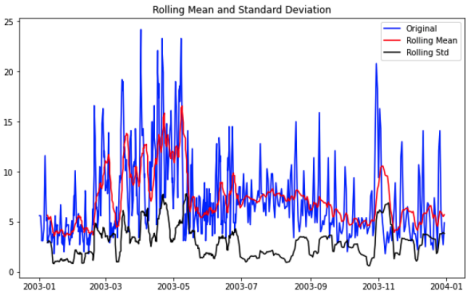
```
Result of Dickey-Fuller Test:
Test Statistic                -7.711616e+00
p-value                        1.259459e-11
#Lags Used                     1.700000e+01
Number of Observation Used     3.440000e+02
Critical Value (1%)           -3.449503e+00
Critical Value (5%)           -2.869979e+00
Critical Value (10%)          -2.571266e+00
dtype: float64
```

(b) Statistical Test Result

Figure 26: Transformed Exploratory Graphics and Statistics. See 03.timeseries.2003.windspeed.ipynb for more information.

We applied log transformation to the given data, and the output shows no improvement. We then shifted the time interval by one and finally, we had stationary data. Also, we can refer to the test statistics are greater than the critical value. This indicates we reject the null at the 5% significance level in favor of the alternative that is data is stationary.

(2) Imperial County



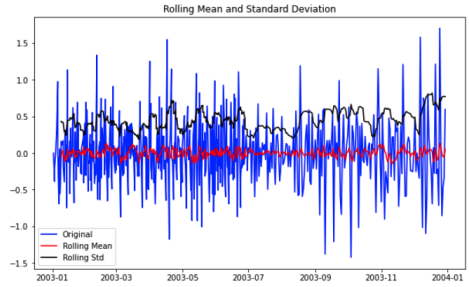
(a) Time Series Plot with Rolling Mean and Standard Deviation for Imperial County

```
Result of Dickey-Fuller test
Test Statistic                  -4.907685
p-value                          0.000034
#Lags Used                       6.000000
Number of Observation Used     507.000000
Critical Value (1%)             -3.443314
Critical Value (5%)             -2.867258
Critical Value (10%)            -2.569815
dtype: float64
```

(b) Statistical Test Result

Figure 27: Initial Exploratory Graphics and Statistics., See 03.timeseries.2003.windspeed.ipynb for more information.

The plot indicates data is not stationary because the behavior does not remain constant and thus stationarity does not hold. According to the statistical test result, we fail to reject the null because the test statistic (-4.29) is less than the critical value (-2.86) at a 5% significance level. This indicates data is not stationary.



(a) Transformed Time Series Plot with Rolling Mean and Standard Deviation for Imperial County

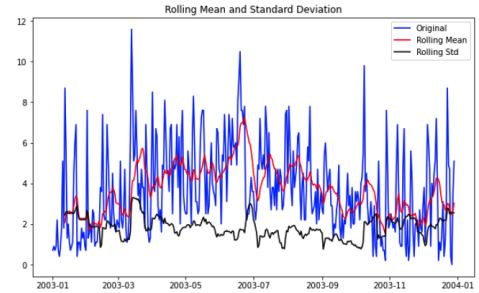
```
Result of Dickey-Fuller Test:
Test Statistic                -1.037978e+01
p-value                        2.154155e-18
#Lags Used                     1.300000e+01
Number of Observation Used     4.990000e+02
Critical Value (1%)           -3.443523e+00
Critical Value (5%)           -2.867350e+00
Critical Value (10%)          -2.569864e+00
dtype: float64
```

(b) Statistical Test Result

Figure 28: Transformed Exploratory Graphics and Statistics., See 03.timeseries.2003.windspeed.ipynb for more information.

So that we correct the assumption, log transformation is applied and also shifting the time interval by one. This corrects the assumption and the statistical test indicates we reject the null at 5% significance level. This is because the test statistics (-1.03e+01) is greater than critical value (-2.86e+00). Also, we can observe both mean and variance in the plot demonstrate are constant with time.

(3) Mendocino County



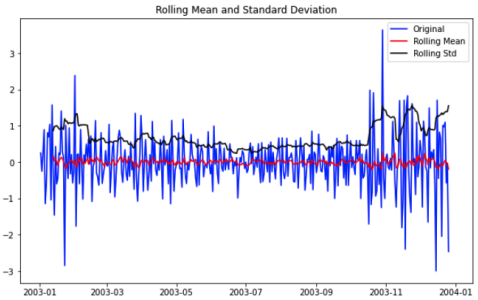
(a) Time Series Plot with Rolling Mean and Standard Deviation for Mendocino County

```
Result of Dickey-Fuller test
Test Statistic                -5.980645e+00
p-value                        1.842046e-07
#Lags Used                     4.000000e+00
Number of Observation Used     3.580000e+02
Critical Value (1%)           -3.448749e+00
Critical Value (5%)           -2.869647e+00
Critical Value (10%)          -2.571089e+00
dtype: float64
```

(b) Statistical Test Result

Figure 29: Initial Exploratory Graphics and Statistics, See 03.timeseries.2003.windspeed.ipynb for more information.

The plot indicates data is not stationary because the behavior does not remain constant and thus stationarity does not hold. According to the statistical test result, we fail to reject the null because the test statistic (-5.98e+00) is less than the critical value (-2.869) at a 5% significance level. This indicates data is not stationary.

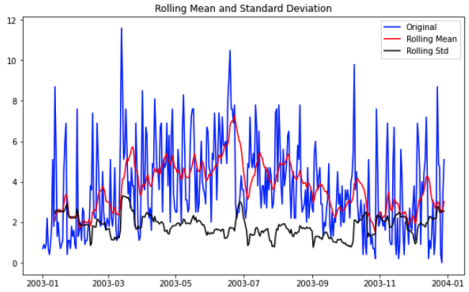


(a) Transformed Time Series Plot with Rolling Mean and Standard Deviation for Mendocino County

Figure 30: Transformed Exploratory Graphics and Statistics., See 03.timeseries.2003.windspeed.ipynb for more information.

So that we correct the assumption, log transformation is applied and also shifting the time interval by one. This corrects the assumption and the statistical test indicates we reject the null at 5% significance level. Also, we can observe both mean and variance in the plot demonstrate are constant with time.

(4) Riverside County



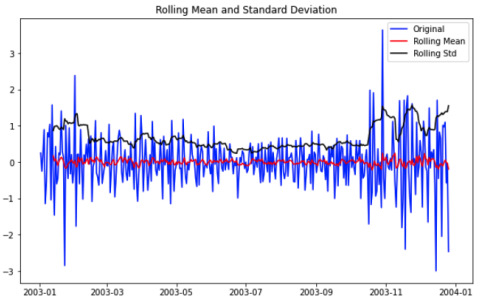
(a) Time Series Plot with Rolling Mean and Standard Deviation for Riverside County

```
Result of Dickey-Fuller test
Test Statistic                -5.980645e+00
p-value                        1.842046e-07
#Lags Used                     4.000000e+00
Number of Observation Used     3.580000e+02
Critical Value (1%)           -3.448749e+00
Critical Value (5%)           -2.869647e+00
Critical Value (10%)          -2.571089e+00
dtype: float64
```

(b) Statistical Test Result

Figure 31: Initial Exploratory Graphics and Statistics, See 03.timeseries.2003.windspeed.ipynb for more information.

The plot indicates data is not stationary because the behavior does not remain constant and thus stationarity does not hold. According to the statistical test result, we fail to reject the null because the test statistic (-4.60) is less than the critical value (-2.864) at a 5% significance level. This indicates data is not stationary.

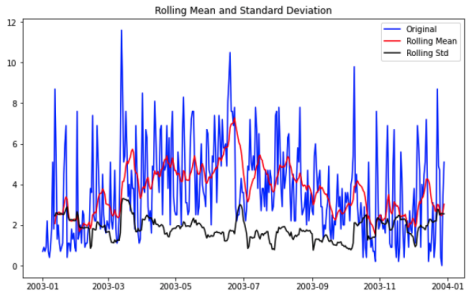


(a) Transformed Time Series Plot with Rolling Mean and Standard Deviation for Riverside County

Figure 32: Transformed Exploratory Graphics and Statistics., See 03.timeseries.2003.windspeed.ipynb for more information.

So that we correct the assumption, log transformation is applied and also shifting the time interval by one. This corrects the assumption and the statistical test indicates we reject the null at 5% significance level. This is because the test statistics (1.05e+01) is greater than critical value (-2.864). Also, we can observe both mean and variance in the plot demonstrate are constant with time.

(5) Siskiyou County



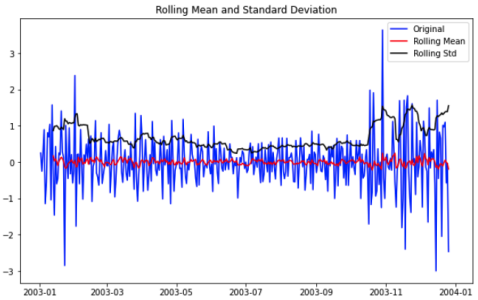
(a) Time Series Plot with Rolling Mean and Standard Deviation for Riverside County

```
Result of Dickey-Fuller test
Test Statistic                -5.980645e+00
p-value                        1.842046e-07
#Lags Used                     4.000000e+00
Number of Observation Used     3.580000e+02
Critical Value (1%)           -3.448749e+00
Critical Value (5%)           -2.869647e+00
Critical Value (10%)          -2.571089e+00
dtype: float64
```

(b) Statistical Test Result

Figure 33: Initial Exploratory Graphics and Statistics, See 03.timeseries.2003.windspeed.ipynb for more information.

The plot indicates data is not stationary because the behavior does not remain constant and thus stationarity does not hold. According to the statistical test result, we fail to reject the null because the test statistic (-5.20) is less than the critical value (-2.86) at a 5% significance level. This indicates data is not stationary.

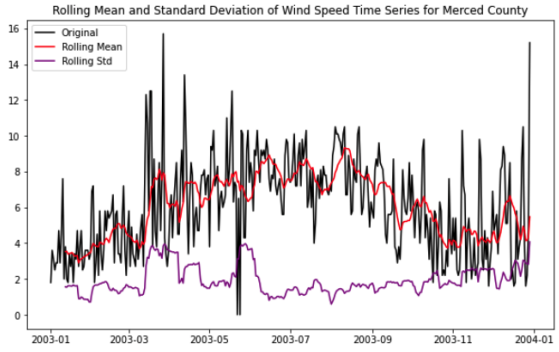


(a) Transformed Time Series Plot with Rolling Mean and Standard Deviation for Riverside County

Figure 34: Transformed Exploratory Graphics and Statistics., See 03.timeseries.2003.windspeed.ipynb for more information.

So that we correct the assumption, log transformation is applied and also shifting the time interval by one. This corrects the assumption and the statistical test indicates we reject the null at 5% significance level. This is because the test statistics (-1.06e+01) is greater than critical value (-2.865e+00). Also, we can observe both mean and variance in the plot demonstrate are constant with time.

(6) Merced County



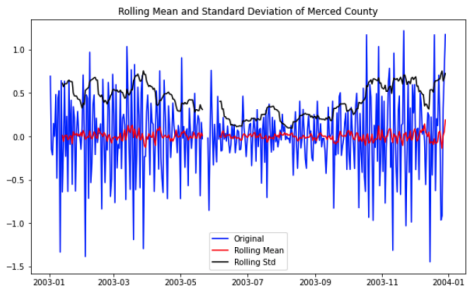
(a) Time Series Plot with Rolling Mean and Standard Deviation for Merced County

```
Result of Dickey-Fuller test
Test Statistic                  -3.359299
p-value                          0.012427
#Lags Used                       8.000000
Number of Observation Used     354.000000
Critical Value (1%)             -3.448958
Critical Value (5%)             -2.869739
Critical Value (10%)            -2.571138
dtype: float64
```

(b) Statistical Test Result

Figure 35: Initial Exploratory Graphics and Statistics, See 03.timeseries.2003.windspeed.ipynb for more information.

The plot indicates data is not stationary because the behavior does not remain constant and thus stationarity does not hold. According to the statistical test result, we fail to reject the null because the test statistic (-3.35) is less than the critical value (-2.86) at a 5% significance level. This indicates data is not stationary.



(a) Transformed Time Series Plot with Rolling Mean and Standard Deviation for Merced County

Figure 36: Transformed Exploratory Graphics and Statistics., See 03.timeseries.2003.windspeed.ipynb for more information.

So that we correct the assumption, log transformation is applied and also shifting the time interval by one. This corrects the rolling mean but not the variance. Therefore, we fail to reject the null at 5% significance level. Also, we can observe the variance in the plot is not constant with time.

4 Conclusion

This report compares different practical data science approaches such as exploratory visualization, geospatial visualizations, extracting datasets from SQL databases and web scraping, exploratory analysis, hierarchical clustering, regression analysis, and time series analysis. There were two datasets compared and the first dataset is about wildfires in California between the years 1992 and 2015. And the second dataset is about weather climate in California for the years 2003, 2006, 2007, and 2010. This research aims to find out which counties and regions experienced the greatest number of fires in the year between 1992 and 2015. What the primary causes of those wildfires was? If any of the primary causes triggered more wildfires in certain regions/counties? If there are any discernible weather patterns for different regions that experienced high and low fire seasons? And finally, to distinguish if variables are statistically significant regarding the fire size?

From the experimental result of both exploratory and geospatial visualization, the findings were that wildfires were more prone in the year between 2006 and 2007. The county that experienced the greatest number of fires during that time was Riverside. According to the geospatial visualization, most counties in Southern California are delicate to having wildfires. Even though there are other factors such as equipment use, arson, etc. that can cause a wildfire, however, the most important cause of wildfires in California is lightning strikes. Also, in the year of 2012 California experienced the second-largest wildfire in Lassen city which was 315,578.8 acres that impacted both California and Nevada.

Following the above interesting findings, the research expanded by including weather climate scraped data for the years with the highest number of fires and for the counties that were most prone to the wildfire. The focus was narrowed to examine the maximum temperature, minimum temperature, average temperature, precipitation, and wind speed weather for the counties El-Dorado, Mendocino, Merced, and Siskiyou in the years 2003, 2006, 2007, and 2010. The results suggest that counties in northern California (El-Dorado, Mendocino, Merced, and Siskiyou), tend to have cooler average temperatures. On the other hand, counties in southern California (Riverside, Imperial) tend to have warmer average temperatures and higher average wind speeds than the northern counties. Comparing visually the different types of fire sizes, the fire size of class D tends to have a higher wind speed above 8mph compared to the other types of fire sizes (fire sizes class A and B).

In addition, to examine the closeness between data points of similar types, hierarchical clustering was applied. The experiment indicates that weather data of type temperature, precipitation, and wind speed may only be effective at identifying certain classes of fire size. On the other hand, investigating only for selected 10 random subsets suggests three major classes of weather conditions that may give rise to fires.

Also, investigating daily county weather patterns for all days within one week, the Euclidean space for the weather patterns does not become much more differentiable between class A and class B fires. This implies that the predictability of fire severity from natural causes is not significant from the variable's temperature, wind speed, and precipitation. Similarly, the weather patterns following fire discovery do not give a strong indication of the class of a given fire either. Due to the limited sample size, our findings may not be representative of the weather structure of fire classes at scale.

To determine which variables from the wildfire and weather dataset are statistically significant to cause a wildfire, multiple regression analysis and standardized regression analysis were applied. For selected seven predictor variables namely fire year, precip right, average left, wind speed, cts day, latitude, and longitude. A multiple regression analysis F statistically significant results in rejection of the null hypothesis and the smaller p-value (8.837e-08) indicate a significant association between at least one the seven predictor variables and the dependent variable fire size. Following this, to determine the most significant factor affecting fire size, a standardized regression analysis was applied by fixing the scale of the predictor variable. This experiment results in smaller p-values for each variable indicating northern California counties have fewer fire sizes compared to Southern California with higher fire sizes. Also, both continuous days of fire burning and wind speed variables have the highest significant influence on the fire size. This confirms the result from the visualizations is valid.

Finally, a time series analysis was performed to build a model that can forecast weather climate change for future events. However, due to the limited data observations for selected counties and the selected years. The analysis checks the stationarity assumption for the wind speed climate data only. Both explanatory plots and Dickey-Fuller statistically significant test were performed for each county in the year 2003. Most of the data required transformation to meet the assumption. However, since the results can be misleading, we have not performed the implementation of the model.

References

[1] CA geographic boundaries. California Open Data. (2019, October 23). Retrieved from https://data.ca.gov/dataset/ca-geographic-boundaries

[2] GGPLOT2 barplots : QUICK START GUIDE - R software and Data Visualization. STHDA. (n.d.). Retrieved from http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

[3] Kassambara, Visitor, amp; Amao. (2017, November 17). Plot one variable: Frequency graph, density distribution and more. STHDA. Retrieved from http://sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more

[4] Nathan, P. (2020, June 2). Visualizing Geospatial Data in python. Medium. Retrieved from https://towardsdatascience.com/visualiz geospatial-data-in-python-e070374fe621

[5] National Centers for Environmental Information (NCEI). (n.d.). Climate Data Online: Dataset Discovery. Datasets | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Retrieved from https://www.ncdc.noaa.gov/cdo-web/datasets

[6] National Centers for Environmental Information (NCEI). (n.d.). Climate Data Online: Web Services Documentation. National Climatic Data Center. Retrieved from https://www.ncdc.noaa.gov/cdo-web/webservices/v2dataTypes

[7] Stewart, R. (2018, November 1). GeoPandas 101: Plot any data with a latitude and longitude on a map. Medium. Retrieved from https://towardsdatascience.com/geopandas-101-plot-any-data-with-a-latitude-and-longitude-on-a-map-98e01944b972

[8] Tatman, R. (2020). 1.88 million US wildfires. 1.88 Million US Wildfires. Retrieved from https://www.kaggle.com/datasets/rtatman/ million-us-wildfires

[9] Time Series. the R Graph Gallery. (n.d.). Retrieved from https://r-graph-gallery.com/time-series.html

[10] Wikimedia Foundation. (2021, July 21). Rush fire. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Rush$_Fire$

[11] Wikimedia Foundation. (2022, May 23). 2007 California wildfires. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/2007$_C$

[12] Zeglis, C. (2019, December 2). How to visualize data on top of a map in python using the geoviews library. Medium. Retrieved from https://towardsdatascience.com/how-to-visualize-data-on-top-of-a-map-in-python-using-the-geoviews-library-c4f444ca2929