UC San Diego

Analyzing Wildfire Trends in the US from 1992-2015

MATH 189 Project

Ana Dominguez

Charlie Gillet

Joseph Guzman

June 9, 2024

Professor. Vishwanath

1 Introduction

1.1 Purpose

As of 2024, wildfires in the United States continue to increase in frequency due to the ongoing impacts of climate change. This crisis is growing in its effect on our ecosystems, infrastructure, and lives. By investigating these patterns, we aim to provide insights that can inform more effective wildfire prevention and mitigation strategies, ultimately protecting our ecosystems and infrastructure.

The question arises: What trends and correlations can we identify that will allow for forecasting the frequency of wildfires and provide data-driven solutions? We aim to contribute to this understanding by analyzing data from over 1.8 million wildfires recorded across the US. By examining the frequencies and regional variations of wildfires, we seek to uncover patterns and trends that can inform disaster preparedness and mitigation strategies.

1.2 Data

Our dataset, obtained from Kaggle, contains a spatial database of over 18 million wildfires that occurred in the United States over a 24-year period. The core features of the dataset include the discovery date, fire size, and a point location at least as precise as the Public Land Survey System (PLSS) section (1-square-mile grid). We imported the data as a .sqlite file and converted it to a pandas DataFrame for analysis.

We filtered out unnecessary variables by removing columns such as commanding agency names. Since the data has been validated by multiple agencies, we assumed there were no outliers. Null values were present in only one column, the date of containment, but this did not affect our analyses. Additionally, we focused on fires over 1,000 acres to exclude smaller fires that are not of major concern to governing agencies.

2 Exploratory Data Analysis

2.1 Data Insights

Our exploratory data analysis will focus on identifying relationships between different wildfire attributes, such as duration versus cause, using simple linear regression and correlation computations. We also compared frequencies of fires within mainland regions and seasons. This preliminary analysis will help us uncover significant trends that will inform our main analysis.

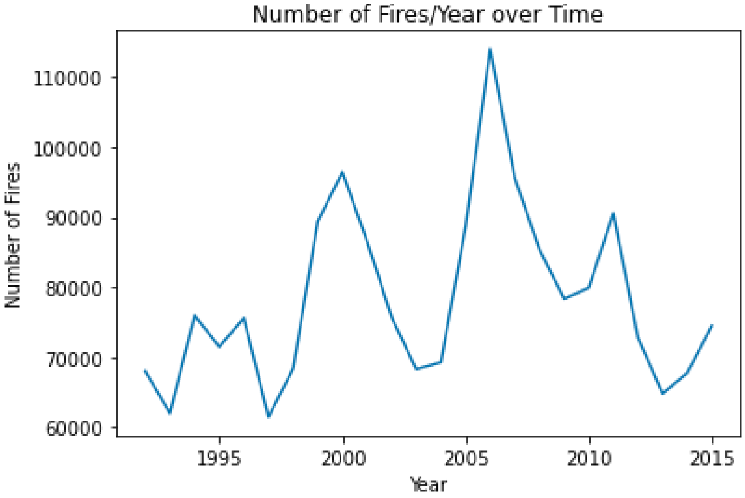


Figure 1: Number of Fires based on year over time.

For the entire US, the number of fires per year does not outwardly appear to have any trend, with high fluctuations dominating the trend (Based on Figure 1). The percent change over time is 99 percent, which agrees with the graph showing no clear change over time. While the entire US may not be experiencing more major fires, there are further seasonal and regional trends that may appear.

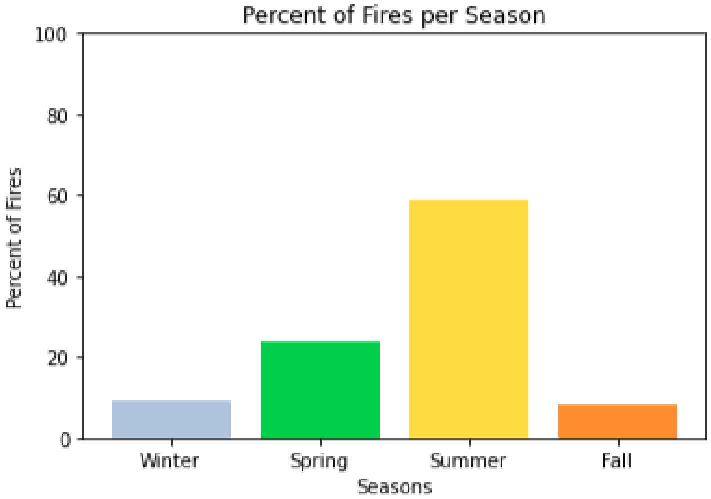


Figure 2: Fire percentages per Season.

So, we aimed to visualize the percentage of fires per season to identify any climate season trends. Based on the visualization, we see that the majority of fires, about 60%, happen during the summer months, with spring coming in second with about 25% of fires. (Based on Figure 2)

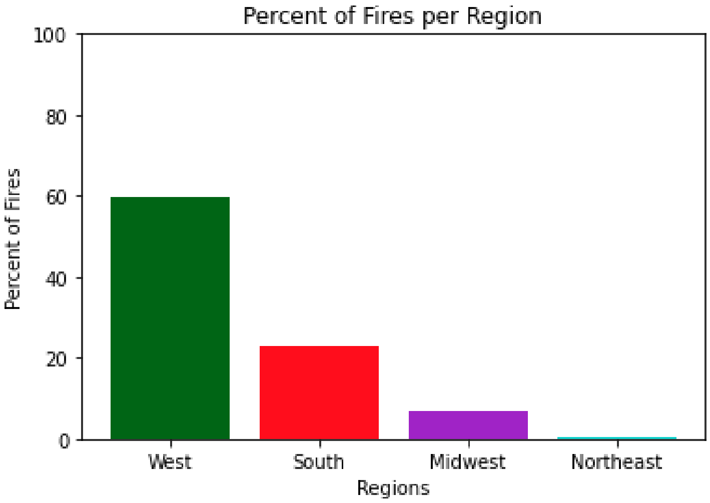


Figure 3: Fire percentages per season.

We then wanted to identify any trends based on location, using the US Census Bureau definition of mainland US regions. Based on the visualization, we can infer that the majority of the fires, about 60%, happen in the western region of the US. (Based on Figure 3).

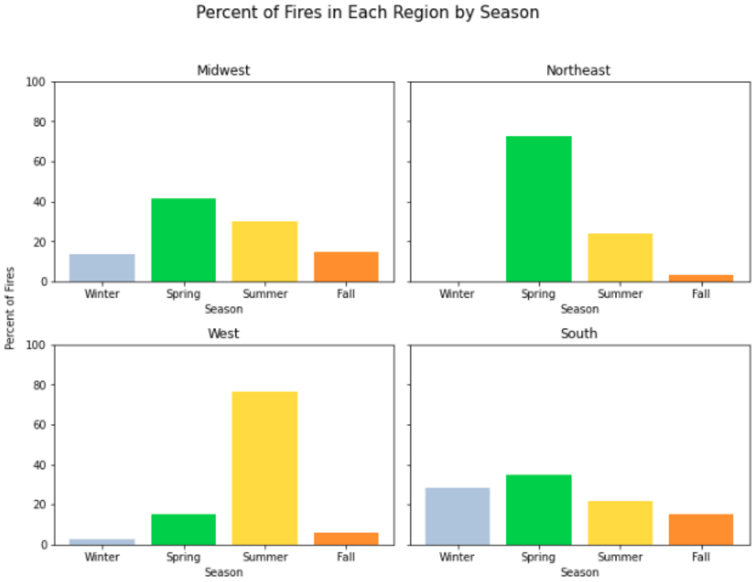


Figure 4: Fire percentage for each region based on season.

Currently, we understand that the majority of fires occur during the summer in the western region of the US. But how do wildfires compare across different regions by season? Based on Figure 4, the visualization describes each region's seasonal fire percentages. This is important because it helps us identify underlying trends that previous visualizations may have overlooked. For the entire US, most fires occur in the summer, but this visualization reveals that this is primarily due to a summer peak in the Western US. In other mainland US regions, there is a springtime peak in fires instead.

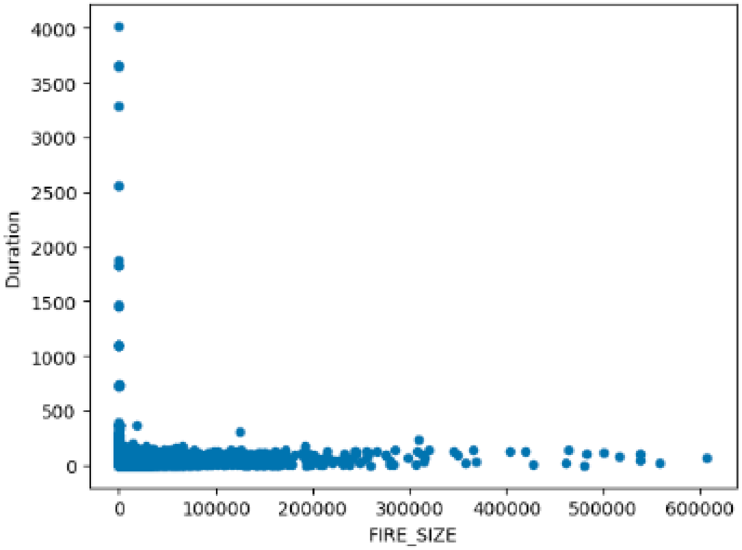


Figure 5: Fire duration over fire size.

Moreover, we wanted to investigate if there were any correlations between duration and fire size. According to the visualization, there doesn't appear to be a strong correlation between the two variables, as most of the data points lie either horizontally or vertically. (Figure 5) However, most of the data is clustered within the 0-500 range for both duration and fire size. Further exploratory data analysis is needed to confirm this observation, but for now, the majority of the data seems scattered.

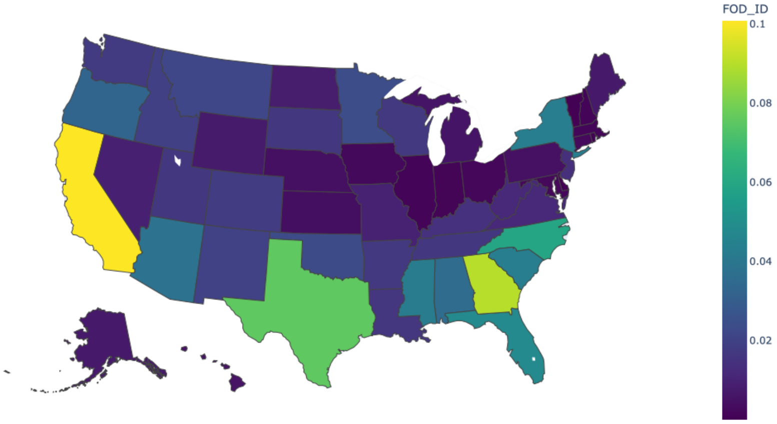


Figure 6: A geospatial heat-map of the United States map, detailing the amount of fires per state.

Lastly, we wanted a geospatial visualization (Figure 6) on how the data can be represented towards the amount of fires per state. Within the visualization, the heat-map is based on the levels of fire where 0.1 is considered, the most number of fires given the proportion of all fires.

2.2 Data Cleaning

Since our dataset has 1.8 million rows, and most of these rows represent very minor wildfires, we will filter the DataFrame to only contain wildfires in class F and class G. The classes are on a scale from A-G increasing in size, where class F represents 1000+ acres burned, and class G represents 5000+ acres burned. Over 80% of the total acreage burned in the dataset is by class F and G wildfires, and this reduces the size of our dataset to about 11,000 rows, which is much more reasonable to work with. The distribution of total acreage burned by wildfires from each class is visualized below:

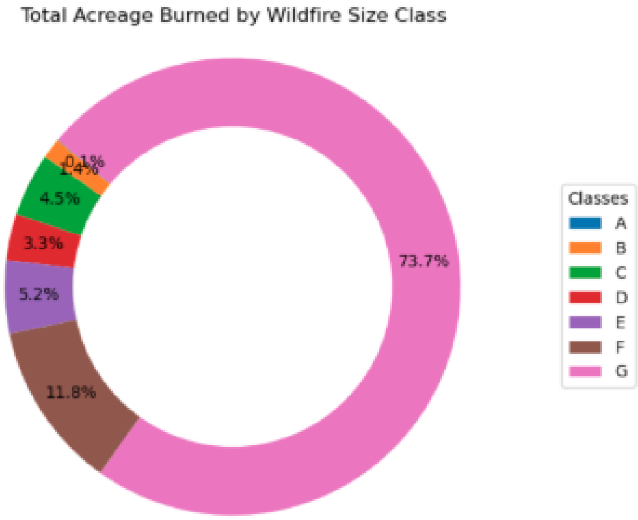
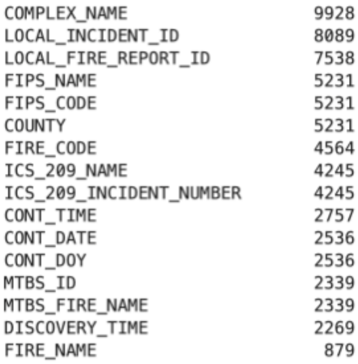


Figure 7: Total acreage burned by wildfire size class.

Now, we still have many columns with a major proportion of missing values, and most of these are unlikely to be helpful for prediction due to their high missingness, so these columns will be cleaned out of the dataset. These columns are also not ones that we would believe to be helpful predictors, conveniently. After removing those columns, the dataset is now reduced from 39 to 23 columns. Here are the following columns that contained missing values, with the corresponding amount of missing values:



```
COMPLEX_NAME              9928
LOCAL_INCIDENT_ID         8089
LOCAL_FIRE_REPORT_ID      7538
FIPS_NAME                 5231
FIPS_CODE                 5231
COUNTY                    5231
FIRE_CODE                 4564
ICS_209_NAME              4245
ICS_209_INCIDENT_NUMBER   4245
CONT_TIME                 2757
CONT_DATE                 2536
CONT_DOY                  2536
MTBS_ID                   2339
MTBS_FIRE_NAME            2339
DISCOVERY_TIME            2269
FIRE_NAME                  879
```

Figure 8: Missing Values

For predicting fire size, there are some columns that are not viable to use as independent variables, or that are bound to cause multicollinearity, for the following reasons:

- FIRE_SIZE_CLASS : This is determined by the fire size, so it should not be used for prediction.
- OBJECTID, FOD_ID, FPA_ID: These are unique IDs assigned to each reported wildfire, so they do not have predictive value.
- SOURCE_SYSTEM_TYPE and other variables regarding the reports: The variables having to do with how the wildfires are reported are not useful because they only tell us about the aftermath, they do not tell us anything about how or where the wildfire was caused.
- DISCOVERY_DATE: This column is redundant to FIRE_YEAR and DISCOVERY_DOY, so it will be excluded.
- STAT_CAUSE_CODE: This column is redundant, we will use STAT_CAUSE_DESCR instead.
- OWNER_CODE: This column is redundant, we will use OWNER_DESCR instead.
- STATE: Since we are including LATITUDE and LONGITUDE, including the state is somewhat redundant, as we already have a measure of location.

The columns that may be potential useful predictors are the following:

- FIRE_YEAR (discrete quantitative): The year of the fire can tell us about the size of the fire if there is a certain yearly pattern.
- DISCOVERY_DOY (discrete quantitative): Since fires appear more frequently during different times of year in certain regions, this could tell us about the size of wildfires.
- STAT_CAUSE_DESCR (categorical): This column tells us the cause of the fire, which is likely to relate to the size of the fire.
- LATITUDE and LONGITUDE (discrete continuous): The location of the fire is likely to have something to do with the size of the fire. In our EDA, we saw most of the top 200 largest fires occur in Alaska or the western U.S., for instance.
- OWNER_DESCR (categorical): This column contains the name of the primary owner or entity who managed the land at the fire's point of origin when the fire started. This could be useful if certain owners have areas of land that are more prone to fires.

At this point, our dataset only contains 6 predictors. Having such a small amount of predictors has some pros and cons. The model will be simple which is a good counter to overfitting; however, having this few predictors could potentially cause the model to not be highly accurate.

In regards to outliers, there are only 4 numerical columns: FIRE_YEAR, DISCOVERY_DOY, LATITUDE, and LONGITUDE. None of these columns could have any outliers. This is because all of these columns' values are within a fixed range: years 1992-2015, days 1-365, and the latitude/longitude boundaries of the United States.

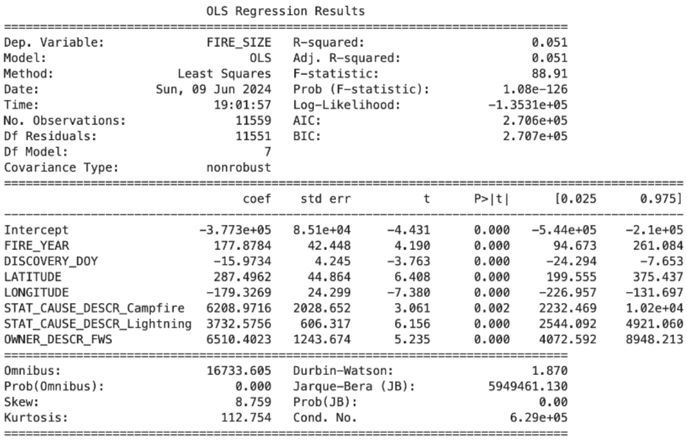
3 Statistical Analysis

3.1 Predicting Fire Size through Regre ssion

Our primary goal is to provide meaningful insights through statistical analysis of the wildfire dataset. One valuable insight would be the ability to forecast potential wildfires. We figured that creating a linear regression model to predict the sizes of fires would be effective, as predicting the size of a wildfire based on factors such as location or the date of the fire could help answer our question. In our EDA, we noticed that there were some notable trends relating to the locations and dates of fires. For instance, fires in the U.S. are more common during the spring and summer seasons, and the west and south regions of the U.S. have a much higher quantity of wildfires than the midwest and northeast regions. Seeing these correlations led us to believe that a linear regression model could be used to forecast wildfires.

3.2 Creating Regression Models

For our first attempt at creating a model, we will use backward selection to find the combination of variables that produces the smallest, or nearly the smallest BIC.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              FIRE_SIZE   R-squared:                       0.051
Model:                            OLS   Adj. R-squared:                  0.051
Method:                 Least Squares   F-statistic:                     88.91
Date:                Sun, 09 Jun 2024   Prob (F-statistic):          1.08e-126
Time:                        19:01:57   Log-Likelihood:            -1.3531e+05
No. Observations:               11559   AIC:                         2.706e+05
Df Residuals:                   11551   BIC:                         2.707e+05
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                  -3.773e+05   8.51e+04     -4.431      0.000   -5.44e+05    -2.1e+05
FIRE_YEAR                    177.8784     42.448      4.190      0.000      94.673     261.084
DISCOVERY_DOY                -15.9734      4.245     -3.763      0.000     -24.294      -7.653
LATITUDE                     287.4962     44.864      6.408      0.000     199.555     375.437
LONGITUDE                   -179.3269     24.299     -7.380      0.000    -226.957    -131.697
STAT_CAUSE_DESCR_Campfire   6208.9716   2028.652      3.061      0.002    2232.469    1.02e+04
STAT_CAUSE_DESCR_Lightning  3732.5756    606.317      6.156      0.000    2544.092    4921.060
OWNER_DESCR_FWS             6510.4023   1243.674      5.235      0.000    4072.592    8948.213
==============================================================================
Omnibus:                    16733.605   Durbin-Watson:                   1.870
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5949461.130
Skew:                           8.759   Prob(JB):                         0.00
Kurtosis:                     112.754   Cond. No.                     6.29e+05
==============================================================================
```

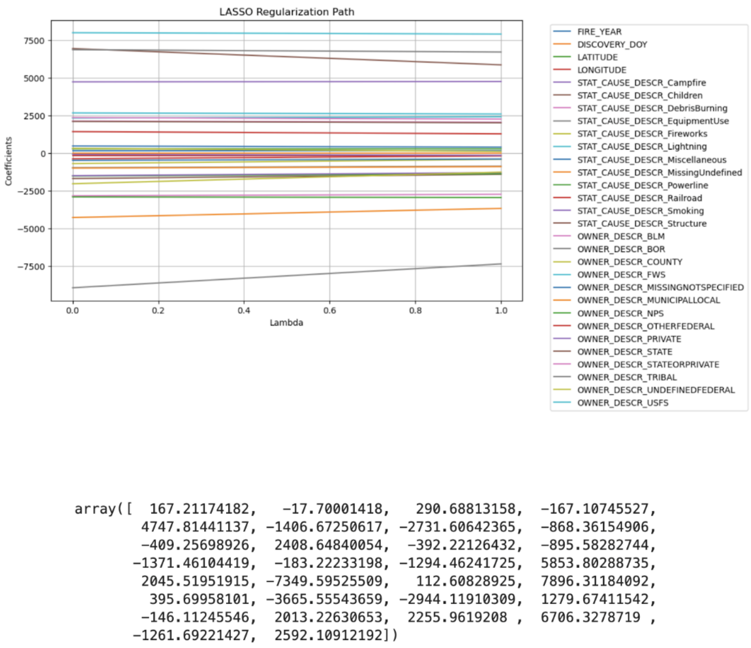
Figure 9: OLS Summary

This model, with formula

```
FIRE_SIZE ~ FIRE_YEAR + DISCOVERY_DOY + LATITUDE + LONGITUDE + STAT_CAUSE_DESCR_Campfire
+ STAT_CAUSE_DESCR_Lightning + OWNER_DESCR_FWS
```

has the lowest BIC of all possible models, with a value of 2.707e+05. However, we can immediately tell that this model is not very strong, especially because the adjusted R-squared is only 0.051. This means that almost 95% of the variance in the data is not explained by this model. In addition, the high condition number of 6.29e+05 means that there is still a high amount of multicollinearity in the covariates used.

Since the previous approach was unsuccessful, we can try to select which variables are viable predictors by using shrinkage methods. Here, we test values of λ ranging from 0 to 1 in increments of 0.01 to see which variables reach 0 and which do not. The purpose of this is to find a good value of λ to use for Lasso regression. The following graph is difficult to read, so an array of the smallest coefficients for each variable are included below the graph.



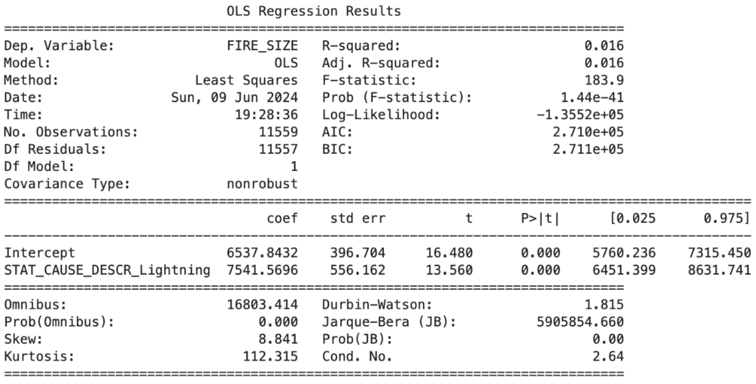
```
array([  167.21174182,   -17.70001418,   290.68813158,  -167.10745527,
        4747.81441137, -1406.67250617, -2731.60642365,  -868.36154906,
        -409.25698926,  2408.64840054,  -392.22126432,  -895.58282744,
       -1371.46104419,  -183.22233198, -1294.46241725,  5853.80288735,
        2045.51951915, -7349.59525509,   112.60828925,  7896.31184092,
         395.69958101, -3665.55543659, -2944.11910309,  1279.67411542,
        -146.11245546,  2013.22630653,  2255.9619208 ,  6706.3278719 ,
       -1261.69221427,  2592.10912192])
```

Figure 10: LASSO Path and cofficients

As we can see, none of the variables reach a coefficient of 0 for any λ value. This means that none of these variables can be ruled out immediately as insignificant through this method. Because of this, we will use $\lambda = 1.0$, as the regularization parameter.

Now, we will try to use Lasso regression instead of ordinary least squares to see if we can find a better model. Additionally, we will try to use Ridge regression to see if it works better than Lasso regression or ordinary least squares. As it turns out, both Lasso and Ridge regression lead to the model with the following formula:

```
FIRE_SIZE ~ STAT_CAUSE_DESCR_Lightning
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              FIRE_SIZE   R-squared:                       0.016
Model:                            OLS   Adj. R-squared:                  0.016
Method:                 Least Squares   F-statistic:                     183.9
Date:                Sun, 09 Jun 2024   Prob (F-statistic):           1.44e-41
Time:                        19:28:36   Log-Likelihood:            -1.3552e+05
No. Observations:               11559   AIC:                         2.710e+05
Df Residuals:                   11557   BIC:                         2.711e+05
Df Model:                           1
Covariance Type:            nonrobust
===========================================================================================
                                 coef    std err          t      P>|t|     [0.025     0.975]
-------------------------------------------------------------------------------------------
Intercept                   6537.8432    396.704     16.480      0.000   5760.236   7315.450
STAT_CAUSE_DESCR_Lightning  7541.5696    556.162     13.560      0.000   6451.399   8631.741
==============================================================================
Omnibus:                    16803.414   Durbin-Watson:                   1.815
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5905854.660
Skew:                           8.841   Prob(JB):                         0.00
Kurtosis:                     112.315   Cond. No.                         2.64
==============================================================================
```

Figure 11: OLS results

Both Lasso and Ridge regression give us the same model which only contains one covariate, which represents the "Lightning" category of the STAT_CAUSE_DESCR column. This model gives an accuracy that is even worse than the OLS model, with an R-squared of 0.016. The BIC is about the same as the BIG for the OLS model, and the only upside of this model is that there is a very low condition number of 2.64. However, it is very obvious that this model is unviable, because it only uses one of the one-hot encoded columns of STAT_CAUSE_DESCR.

3.3 Underlying Assumptions

After performing our regression models and seeing their low accuracy, we may be able to find an explanation for this low accuracy by checking the 5 primary assumptions of linear regression: Linearity, independence, homoscedasticity, normality, and multicollinearity.

1. Linearity: The relationship between the independent and dependent variables is nearly linear.

   We tested this by plotting the relationships between predictors and the response variable (fire size). Most of the relationships do not all seem linearly correlated, but we had hoped that the variables that are not highly correlated with the response variable would have been weighted low in our model. For example, below is a plot of FIRE_YEAR versus FIRE_SIZE, which does not seem to have a strong correlation. However, there is a slight upward trend in fire size in the later years. This is not highly strong evidence to assume linearity, which is a major sign of why our model achieved low accuracy.

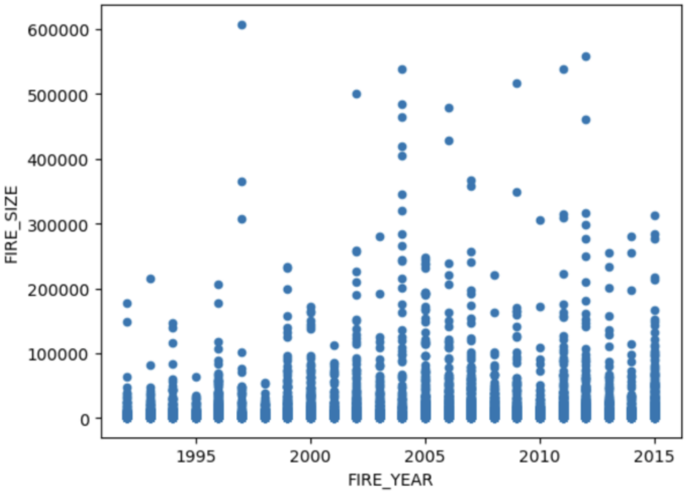


Figure 12: Fire Size over Fire Year

2. Independence: Observations are independent of each other.

   We believe it was reasonable to assume independence in this dataset because it consists of individual wildfire incidents recorded over different times and locations. Running the .duplicated() method on our DataFrame showed zero duplicated rows. The independence assumption was probably not one of the primary reasons for our model's low accuracy.

3. Homoscedasticity: The residuals have constant variance.

   We checked for homoscedasticity by plotting the residuals versus the fitted values. The plot (below) shows a pattern, suggesting that the variance of the residuals is not constant across different levels of the fitted values. We can reasonably say that this assumption is violated, which could have been a cause of our model's low accuracy.

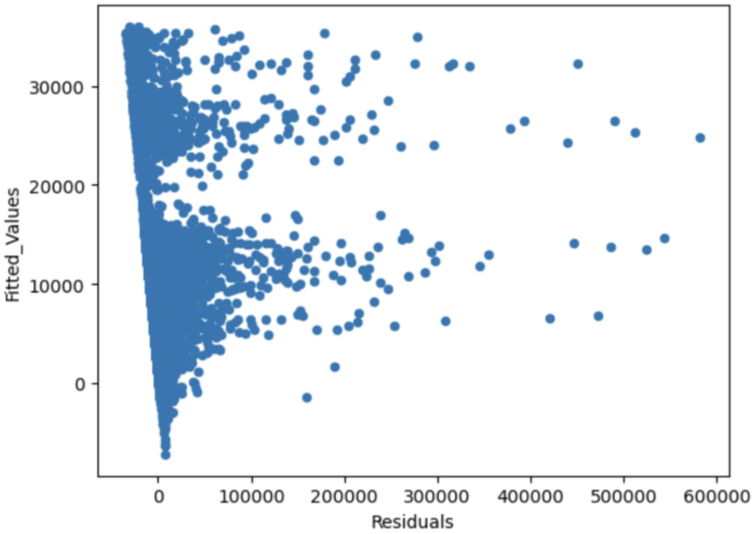


Figure 13: Fitted Values over Residuals

4. Normality: The residuals are normally distributed.

   In making our OLS model, we assumed that the residuals would be normally distributed. However, this is clearly not the case, as shown in the Q-Q plot below. The residuals are clustered towards the lower end of the fire size range. This is another likely cause of our model's low accuracy.

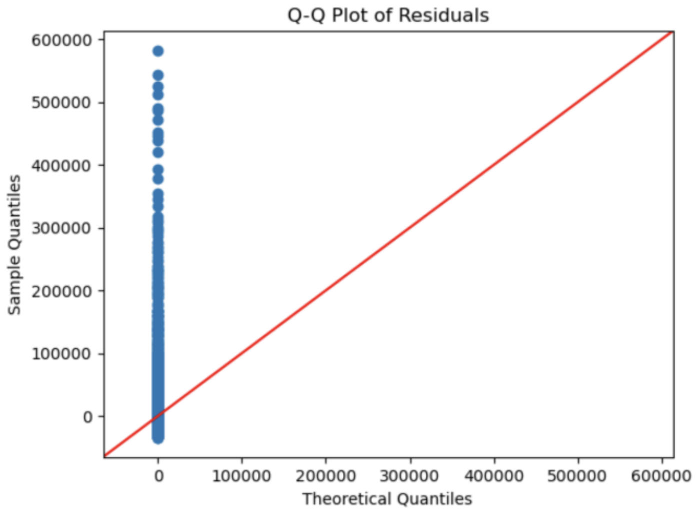
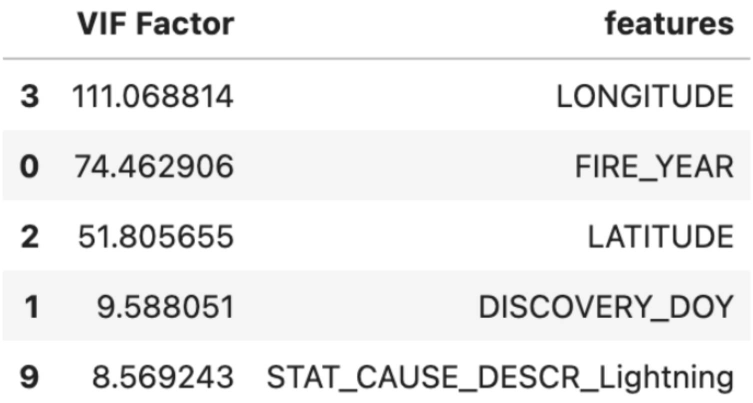


Figure 14: Theoretical vs Sample Quantities

5. Multicollinearity: Independent variables should not be too highly correlated with each other.

   To deal with the multicollinearity we found in our first OLS model, our idea was to see which covariates have a high variance inflation factor and remove those from the model. Here are the five covariates with the highest VIFs:



	VIF Factor	features
3	111.068814	LONGITUDE
0	74.462906	FIRE_YEAR
2	51.805655	LATITUDE
1	9.588051	DISCOVERY_DOY
9	8.569243	STAT_CAUSE_DESCR_Lightning

Figure 15: VIF factors and features

As we can see, nearly all of the important features here have very high VIFs, especially FIRE_YEAR, LATITUDE, and LONGITUDE, with VIFs of about 74, 51, and 111 respectively. A VIF of over 5 to 10 is considered problematic for causing multicollinearity, so this assumption is violated. This is evidence that our ordinary least squares model was unreliable, but we did not have any other variables to work with, so we proceeded to find the results.

3.4 Interpretation of Results

After trying to find a model that minimizes BIC using ordinary least squares, Lasso regression, and Ridge regression, we have fairly reasonable evidence to say that FIRE_SIZE is probably not a variable that can be predicted with high accuracy. This result is not surprising given that four of the five primary assumptions for linear regression were violated prior to the creation of our model. Despite the model's low accuracy, however, there are still some things to take away from this.

Our primary takeaway from our model's low accuracy is that it does not confirm that wildfires are completely unpredictable, but rather that this particular dataset we used was not viable for linear regression. To forecast wildfires with high accuracy, we would need a new dataset, or we would have to find advanced ways to manipulate the dataset that we have not learned yet. In the future, looking for new datasets would likely be the most reasonable way to fit a more accurate model, since manipulating this dataset enough to fit an accurate model might be too tedious and time consuming.

4 Hypothesis Testing

In order to determine if changes in frequency of fires over time was significant (and not just random natural variation), we performed a hypothesis test using bootstrapping in order to get the percent significance of each value. The null hypothesis was that there was no change over time, so 0% change, and the alternative hypothesis was that change did not equal 0%. Quantiles were calculated from the bootstrap method, which allowed us to include natural variation in the generation of our null hypothesis values. They are also referred to as percentiles/significance.

Entire US seasonally:

- Winter: percent increase by 109%, significant to the 92 percentile
- Spring: percent increase by 18%, significant to the 78 percentile
- Summer: percent decrease by 2%, significant to the 56 percentile
- Fall: percent decrease by 62%, significant to the 98 percentile

This means we can say with high confidence (>90%) that fires are becoming more common in the winter and less common in the fall.

Within regions of the US:

- West: percent decrease by 23%, significant to the 99 percentile
- South: percent increase by 120%, significant to the 97 percentile

- Midwest: percent increase by 55%, significant to the 93 percentile
- Northeast: no change (0%), significant to the 80 percentile

This means we can say with high confidence (>90%) that fires are increasing in the south and the midwest, and decreasing in the west.

Within each region, seasonally (only values with high confidence > 90%):

- West: fires in the summer increase by 18% and fires in the fall decrease by 60%
- South: cannot make any conclusions with > 90% confidence, however, fires in the fall decrease by 64% with 89% confidence
- Midwest: fires in the fall decrease by 37% and increase by 60% in the spring
- Northeast: unable to say due to low number of fires being too few to divide up by season (only 21% of values in this dataset were not 0)

5 Conclusion

We mostly kept this project related to what we originally wrote in our Project Proposal. We ended up using the same dataset, explored several plots of the data, and performed hypothesis tests relating to regional and temporal changes. One aspect of this project that we did not initially write about was the inclusion of regression modeling, including ordinary least squares, Lasso regression, and Ridge regression. Our proposal did not mention this because we were not at that stage of our project at the time of writing the proposal. Once we finished our exploratory data analysis, we figured that attempting to forecast wildfires would further enhance our project.

Our research on wildfires over the past 24 years has yielded alarming results regarding their frequency and trends. Fires are becoming much more frequent in the winter and less frequent in the fall, indicating that fires are shifting from fall to winter. Regionally, fires are becoming more common in the South and in the Midwest, and less common in the West. Regions are individually experiencing the US-wide decrease in fall frequency, and the West and Midwest are experiencing increases in the summer and spring, respectively. What all of this means is that there are changes in regional and seasonal frequencies, many quite large, that government agencies need to be highly attentive to. Fires put a huge strain on governments, taking up a lot of manpower, time, and money, and with changes in frequencies only in 24 years being up to 120%, the unpredictability of wildfire can be disastrous for small communities.

We believe an effective and highly accurate model can assist policymakers and government agencies in making data-driven decisions to improve wildfire prevention and ecological health, as well as to take timely precautionary measures and have the resources to fight fires when and where they occur. More data is needed in order to create this kind of model. Unfortunately, our models had high multicollinearity and not enough data, but they are an example of the kinds of models that could be used by policymakers in the future. What can be communicated immediately to policymakers is changes in frequency and locality over time, especially those with very high confidence levels. Over such a small time period there were large changes in these factors, and it is crucial for policymakers to be as up to date as possible when preparing and battling the fires of the current year.

Future steps for our analyses include gathering more data, improving model accuracy, and more precisely analyzing where and when fires are occurring in the US. Ultimately, our analyses contribute to a highly important subject in today's age- how wildfires are changing and what kinds of steps we can take in order to adapt.