**Wildfire Analysis and Management**

Sai Swaroop Tunuguntla (811300334)

Manuteja Medarametla (811286643)

Homeswar Rao Nelakurthi (811285989)

Group Number: 2

Contents

[Introduction 1](#_Toc1302013453)

[Answers To Research Questions 1](#_Toc930665871)

[Data Preparation 1](#_Toc2053483960)

[Exploratory Data Analysis (EDA): 1](#_Toc2033971963)

[Feature Engineering 1](#_Toc262339428)

[Distribution of Acres Burned 1](#_Toc776445491)

[Correlation Matrix 1](#_Toc2094087588)

[Visualization 1](#_Toc1879471208)

[Distribution of Acres burned. 1](#_Toc1909644687)

[Number of wildfires over the years 1](#_Toc859994463)

[Geographical distribution of wildfires 1](#_Toc471786701)

[Correlation Analysis 1](#_Toc1251627271)

[Trends 1](#_Toc1202200026)

[Geographical Distribution 1](#_Toc488933529)

[Geographical Distribution of Wildfires 1](#_Toc503437248)

[Top 10 Counties by Number of Wildfires 1](#_Toc1542782393)

[Machine Learning Models 1](#_Toc386423630)

[General Uses of ML in Wildfire Management 1](#_Toc1452344692)

[Model Training Validation 1](#_Toc1718020602)

[Results and Evaluation 1](#_Toc225716437)

[Interpretability and Insights 1](#_Toc281368832)

[Team Collaboration 1](#_Toc1924038139)

[Future Directions 1](#_Toc287628730)

[Q and A 1](#_Toc982344991)

[References 1](#_Toc1183594711)

# Introduction

We have selected a dataset from Kaggle, "California Fire Incidents," as our primary source. The dataset will serve as a reliable source of information regarding the geography, date, taxonomy, area burned, the number of acres burned, and the cause of wildfires in California. The fact that this dataset, in detail, characterized the fire incident rate in CA helps greatly in considering the analytics of such causes and patterns. With massive wildfires occurring more often and more severely than in the past, fueled by climate change and human activities, this dataset is designed especially for studying the widespread consequences of wildfire on the environment, human life, and safety.

One of our research goals is an intensive exploration of the link between wildfire and air quality. This is a critical environmental concern due to the increasing frequency of big fires that cause vast destruction. Specifically, our focus is the air quality dynamics of wildfires, identifying the quick and long-term impacts of particulate articles and other pollutants upon their dispersion before and during such events. Data will allow us to capture and analyze any modifications in air quality metrics that occurred during, let us say, wildfire events. Such studies may yield answers to how fast air quality goes back to its former levels after a wildfire is over. We may also explore whether repeated wildfire events contribute to air quality degradation. These findings serve as the basis for learning about various approaches to mitigate the adverse consequences of air pollution by wildfires.

By exploiting the "California Fire Incidents" dataset, this course would provide practical advice to guide policymakers, firefighters, and area residents. Recognizing the mechanisms through which wildfires affect breathing air quality makes it clear how important this is for government and civil service people, who must ensure the environment is safe and healthy (Buechi, 2021). Additionally, our conclusions will be helpful in the ongoing conversation on climate change, environmental protection, and hurricane readiness. The state of the voice can provide evidence-based advice on reducing the health risks induced by air pollution produced during wildfire events. Thus, our research analysis will help communities, policymakers, and stakeholders acquire the skills required for managing and responding to the growing problems of wildfires in California.



Figure 1: Top 5 Rows

# Answers To Research Questions

**What Changes Can Be Observed In Air Quality During And After A Wildfire?**

Wildfires can do many bad things to air quality during and after the wildfire itself. One of the most severe consequences of wildfires is the almighty rise of toxic air pollutants such as PM (particulate matter), CO (carbon monoxide), NOx (nitrogen oxides), and VOCs (volatile organic compounds). They may drastically affect visibility in addition to evoking health issues, especially among individuals who are often troubled by breathing and cardiovascular problems (Zhao, 2020). However, the given period is accompanied by increased attention and air quality monitoring by people, communities, and government authorities; people and communities start taking preventive measures such as staying indoors, employing air purifiers in houses, and wearing masks. Such joint endeavors can help prevent the health-related hazards arising from the air pollutants mentioned during wildfires.

After the detonation of the wildfire, the pace of the pollution control process begins, and it can vary according to the size of the fire, weather conditions, and the area's geographical composition. Precipitation, for example, can remove pollutants from the air, balance the air quality, and reduce the pre-wildfire level. Some instances also cause significant steps in forest management and many policies to make the current fire situation less critical and lesser in severity. Eventually, the outcome will be an improved air quality (Tymstra, 2020). Also, in the post-disaster period, society becomes more capable of implementing responses at hand from the community's side. This can be practiced by adopting air quality improvement measures such as strengthening air quality monitoring systems and improving the overall public health preparedness to minimize further impacts of bushfires on air quality and public health wellness.

**How Long Does It Take For The Air To Return To Its Pre-Wildfire State?**

The period in which the air quality is expressed of its ambrosopheric level can be very heterogeneous depending on a great variety of factors such as the number of acres and intensity of the WildfireWildfire, type of materials burned, weather conditions, and surrounding environment capability to dissipate pollutants. Except in some cases where the environmental conditions prevent ready smoke dissipation, air quality improves when the fire is put out. Meanwhile, this is a quick procedure, having visible signs of improvement determined within 2-4 weeks after the fire's end. However, if the weather conditions are helping and, for instance, there is rain, that will wash away the particulate matter from the air (Tymstra, 2020). On the other hand, if it is a significant fire or an area with minimal air circulation, especially when it is windy, more time is required for air quality to return to normal.

Wildfires' aftermath commonly enhances air quality monitoring and improvement projects, often based on developing air quality management strategies (Zhao, 2020). This inclusive approach, by helping to speed up air quality restoration, also tightens the bonds between residents of the whole community for future air quality issues. Furthermore, the rise in studies of the air quality of wildfires can enhance comprehension regarding the effects of such fires on air quality. As a result, more effective policies can be implemented to protect from wildfires and manage wildfires, eventually giving rise to healthier air quality standards. Joined efforts that coordinate between governmental bodies, environmental groups, and the community could be an avenue for not just the journey towards restoring wildfires - plaguing air quality. However, it can also aid in efforts towards environmental well-being and sustainability.

**Is There A Cumulative Effect on Air Quality When Wildfires Occur More Frequently?**

It is critically important for us to know about the possible soil damage from wildfires, which happen more regularly, as well as to understand the long-term impact of these natural events on the natural settings and health of the community. Wildfires burning at high frequencies may maintain the level of pollutants in the air for a longer time, with particulate matter becoming the most common part of the mixture. These pollutants may be present in the air with significant concentrations in places with specific climate conditions that experience regular fires. However, such a challenge also has a positive element in it, which, by being effectively solved and with the improvement of air quality controlling strategies, can, in turn, hinder the spread of pollution. By focusing on sustainable land resource management practices, improving early-stage detection systems of wildfires, and investing money in air purification technologies, we can enormously reduce the adverse outcomes of frequent wildfires on air quality (Tymstra, 2020). These actions can help maintain air standards to accommodate busy wildlife activity even amid the escalating wildfire moves due to climate change and other factors.

More brutal wildfires represent an engine that operates the research and innovation on air quality measurement and management. Fire scientists and ecologists use satellites and machine learning to predict the likelihood of wildfire causes and the extent of the air impact. The accumulation of this information yields critical data that can help design better plans for wildfire management and air quality conservation. There is also an increase in the participation of community members in proactive measures in dealing with the wildfire problem, which could also help mitigate the adverse effects of wildfires on air quality. The fate of frequent fires could be solved with the joint work of government bodies, research facilities, and the community, which we can use to ensure a safe environment for our children and benefit from the results in the long term.

# Data Preparation

In order to perform a comprehensive field study with the data of the California Fire Incidents dataset, we first have to structure the data and find a summary of a few main statistics. The dataset from Kaggle that we will be employing covers diverse aspects of wildfires in California, including the size of the area burned, locations of management units, dates of incidents, etc., among others. We have looked into columns with many missing values like AirTankers`, `CrewsInvolved,` and `StructuresDestroyed.` It is vital to consider this since it discloses the dataset's limited dimension concerning certain essential aspects when dealing with the gaps in the data. Alternative solutions comprise imputation where necessary and companies' ability to avoid variables in the research without affecting the credibility of the information about the target audience.

The results of the statistical sum-up exhibited the presence of sharp periodicity in tables like `AcresBurned` and `StructuresDestroyed,` which had significant standard deviation rates. The existence of diversely distributed objects became evident with the indication of outliers, and the nexus came forth upon further examination. Moreover, the `AcersBurned` column ranged from `0` to 410,000+ acres and had a highly right-skewed distribution, showing that most incidents were more minor. The nature of the data, which consists of the skewness to the extent that some outliers are present, makes using the median to represent the central tendency and inter-quartile range realizations along the dispersion line more appropriate.

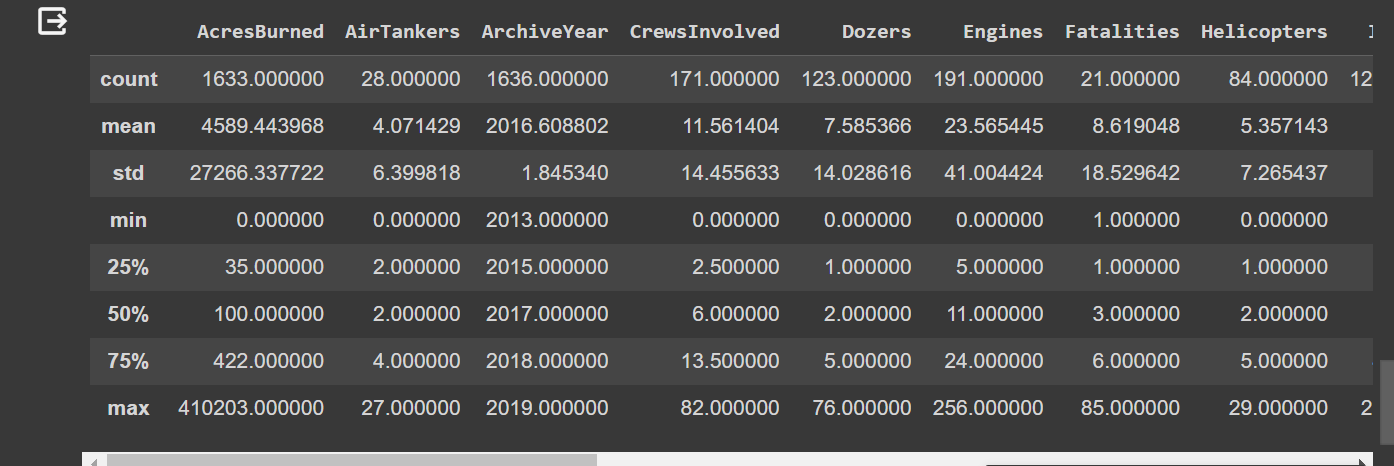


Figure 3: Statistical Overview

In confronting the nonlinear shapes and the outliers, the visualization methods were used to evaluate patterns within the data set. Histograms and the KDE plots for all the quantitative columns showed a short skew, confirming that a larger share of incidents were concentrated within smaller square distances of the area burned and the properties destroyed. At the same time, a few extreme events were very catastrophic. These visualizations validated our initial thoughts of having bouts of variability and outliers. They grew bigger by providing us with an artistic and intuitive portrayal of wildfire incidents, making comprehending the magnitude and impact easier.

# Exploratory Data Analysis (EDA):

Since it is evident that the problem areas are data anomalies, missing values, and outliers, a multifaceted solution needs to be applied. If a missing value is present and not much, then imputation of data is likely to be considered for good distribution characteristic of the dataset. Contextually speaking, columns with missing values lose the chance to participate in specific study samples. Outliers would be classified as extreme values or data entry errors using statistical summaries or visualization; using such a tool, we will filter out these outliers and only keep the genuine ones. What is a genuine outlier will be retained for analyses where it would be relevant in the context of affecting the results, and transformations or robust statistical methods could be used instead to mitigate the spread in other analyses. This careful and detailed approach to data exploration and preprocessing produces a robust, reliable interpretation and analysis framework. It ensures that the data-driven observations are both robust and relevant.

A blue lines on a white background

Description automatically generated

Figure 4: Annual data from 1992 to 2015

This chart shows annual data from 1992 to 2015 with the vertical axis possibly representing a count or measurement related to wildfires. The heights of the bars fluctuate significantly over the years, with a notable peak around 2006, suggesting an increase in the measured variable during that year.

* In 2006 maximum incidents of wildfires took place
* Around 10,000 - 15,000 incidents of wildfire take place every year.

A blue line graph on a white background

Description automatically generated

Figure 5: Line graph 1995 to 2015

A time series line graph from 1995 to 2015 shows data points that peak sharply in 2005 and then decline. This could indicate the measurement of a specific factor related to wildfires, such as area burned or number of incidents, showing a significant event or change in 2005.

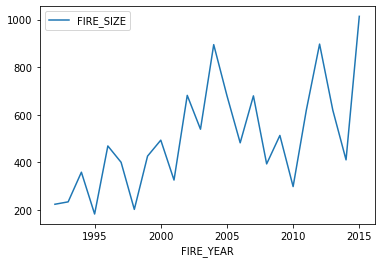


Figure 6: Line graph 1995 to 2015 indicates fire size

Above is a line chart that plots the number of wildfire incidents reported in a given timeframe every year. Each bar presents the sum of all the wildfires that burned during a year and reveals their frequency from one year to another on this timeline. On the other hand, the area burned shows no obvious pattern of increasing or decreasing wildfire occurrences, so it is quite uncertain whether the number of fires varies from year to year with or without a common tendency.

Another line chart, also spanning from 1995 to 2015, shows the variable labeled as "FIRE\_SIZE." The graph has a somewhat erratic pattern with peaks and troughs, indicating fluctuations in the size of wildfires over the years.

# Feature Engineering

Feature Engineering is an essential step in machine learning that consists of selecting, humanizing, or artificially building new features from the raw data to enhance the performance of machine learning models. With this feature, the model can mine the data and fetch the information that one may not be able to see in the raw data. So, it is an additional intelligent part of the program that enables the model to learn. Basic actions of feature engineering can play a critical role in more complex models that achieve higher accuracy on unseen data, saving process time in the long run. In the specific case of assessing the impact of wildfires on air quality, feature engineering identifies and separates variables that could increase pollution and thus help produce better analyses that are more detailed and valuable.

In this article, we attempt to evaluate fire-related air pollution and use some of the features from the "California Fire Incidents" dataset. Critical values like `AcresBurned,` `ArchiveYear,` `Latitude,` `Longitude,` and `Started` (date and time when the fire commenced) are also contained in this dataset. Further on, we will integrate the following characteristics: `Duration` (determined from the start to the finish date), `Season` (obtained from the starting date), and `SizeCategory` (categorizing the fire into four groups: Less than 10,000 acres, 10,000 to 100,000 acres, more than 100,0 These elements are selected as per their impact on air quality score as is likely these big and long fires are more likely to cause the impact on air quality. The spread of the contamination into the air could be determined by the climatic conditions and the time at which the air pollutants travel.

The reasoning behind such selections is based mainly on direct and indirect aerial that wildfires cause. The size and duration of the fire (`AcresBurned` `Duration`) are thought to relate to the amount of pollution released, so air quality is reduced. The spatial location (`Lat` and `Long`) will be used to examine the changes in air quality related to `fire incidents` with the nearest air monitoring stations. The `Season` feature can help investigators investigate seasonal aspects of air quality variation. By selecting and designing specific instruments, we seek to solve the issues of interest, giving knowledge about the essence of the air quality shift in the times of fires, the air quality recovery period, and the overall effect of recurrent fires on the air quality. This analysis would enable authorities to formulate and refine certain decisions that would ensure the safety of the citizens from acute illnesses because of air quality deterioration.

### Distribution of Acres Burned

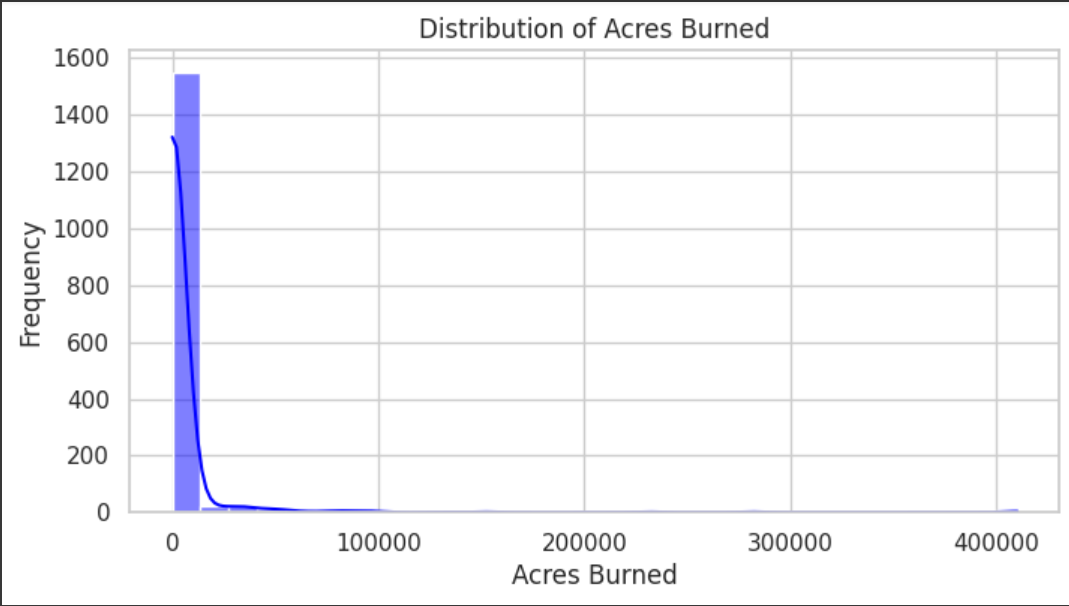


Figure 4: Acres Burned

The histogram demonstrates the trend of acres burned in fire-burning routines. Most wildfires barely cover a small area, which is worth noticing if the high number of the smallest values is considered. However, these random number events, also called outliers, violate the linearity assumption of the regression.

This simple count plot illustrates the yearly occurrence of reported fire incidents. The number of recalcitrant fires varies yearly, and anecdotal evidence suggests that this variability could be connected to weather conditions fluctuations, fire management stakeholders' policies, and land use transformation.

### Correlation Matrix

A screenshot of a computer

Description automatically generated

Figure 6: Correlation Matrix

Heatmap is a simple visualization technique that shows the correlation between variables in a dataset. This visualization helps researchers understand the relationship between two variables, especially the possibility that the number of acres burned socially correlates with air tankers, crews, etc.

# Visualization

One of the research questions that we will concentrate on is the impact of wildfires on air quality. This necessitates the visualization of spatial relationships between fire attributes (such as size, location, and duration) and the impact that they might have on air pollution. Since air quality data is not directly in the "California Fire Incidents" dataset, we will show the wildfire data information through visual forms. However, we could infer possible adverse effects on air quality. This involves maps of wildfire size distribution, time distribution of wildfires across years, and expedition of wildfire distribution over land. With the aid of these visualizations, one can gain an understanding of the size of wildfires, their recurrence frequency, and the scope that they have for affecting the air quality in different regions, as well as in different conditions (Zhao, 2020).

## Distribution of Acres burned.

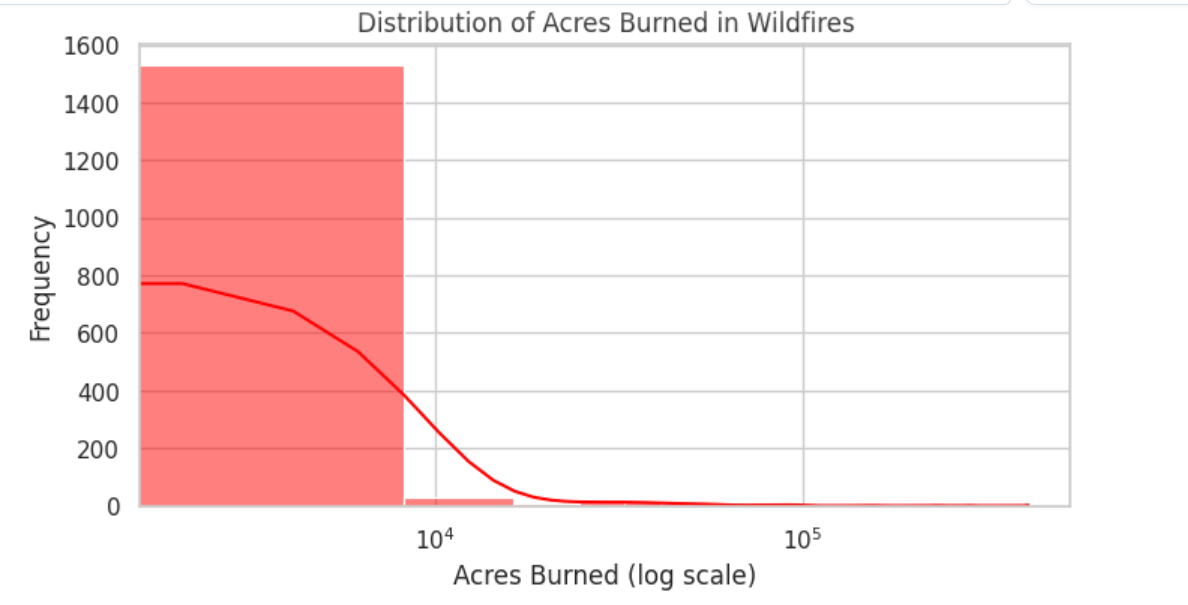


Figure 7: Acres Burned

The graph above deals with the logging of burnt acres, and the x-axis is used in logarithmic form to show that wildfires cover a vast extent. This indicator emphasizes the fact that most wildfires torched only an area of a few acres. However, there are also cases of significant wildfires that can cause extreme destruction, burning hundreds of thousands of acres or more. Therefore, the bigger size contributes to worse air quality, which is one of the most significant and substantial volumes of smoke and pollutants in the air (Zhao, 2020).

## Number of wildfires over the years

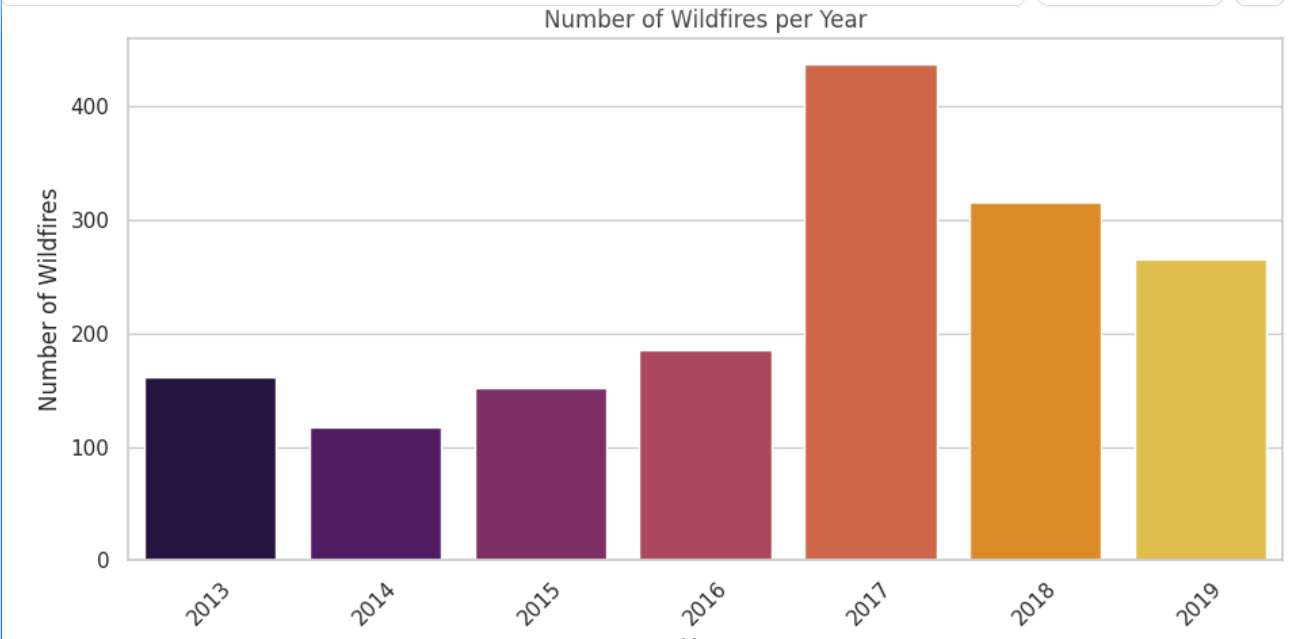


Figure 8: WildFires Per Year

The bar graph depicts the number of wildfire incidents per year from 2013 to 2019. This visualization aids us in learning the temporal distribution/patterns of wildfires, and also, later on, we can identify any trends that may have become prominent over the years. There is little room for geniality on the other side of the table. There is always some variance in the figures. However, it is notable that reporting standards changes, wildfire detection capabilities, and natural variability in weather conditions should be accounted for.

## Geographical distribution of wildfires

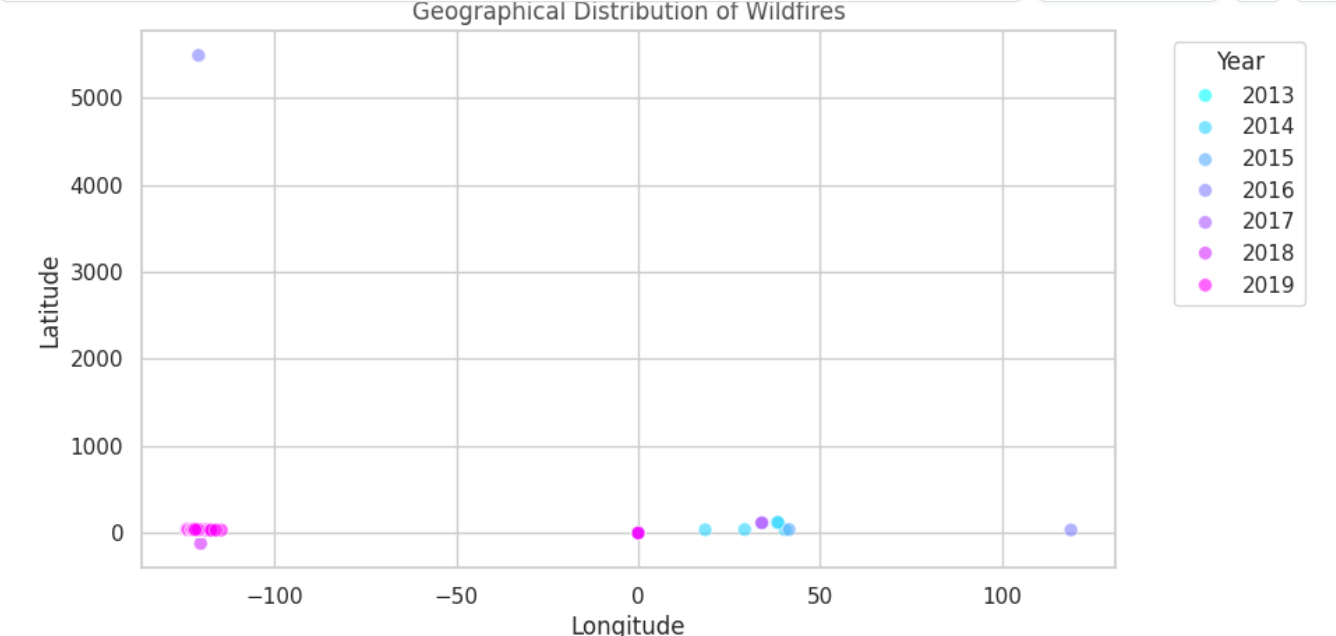


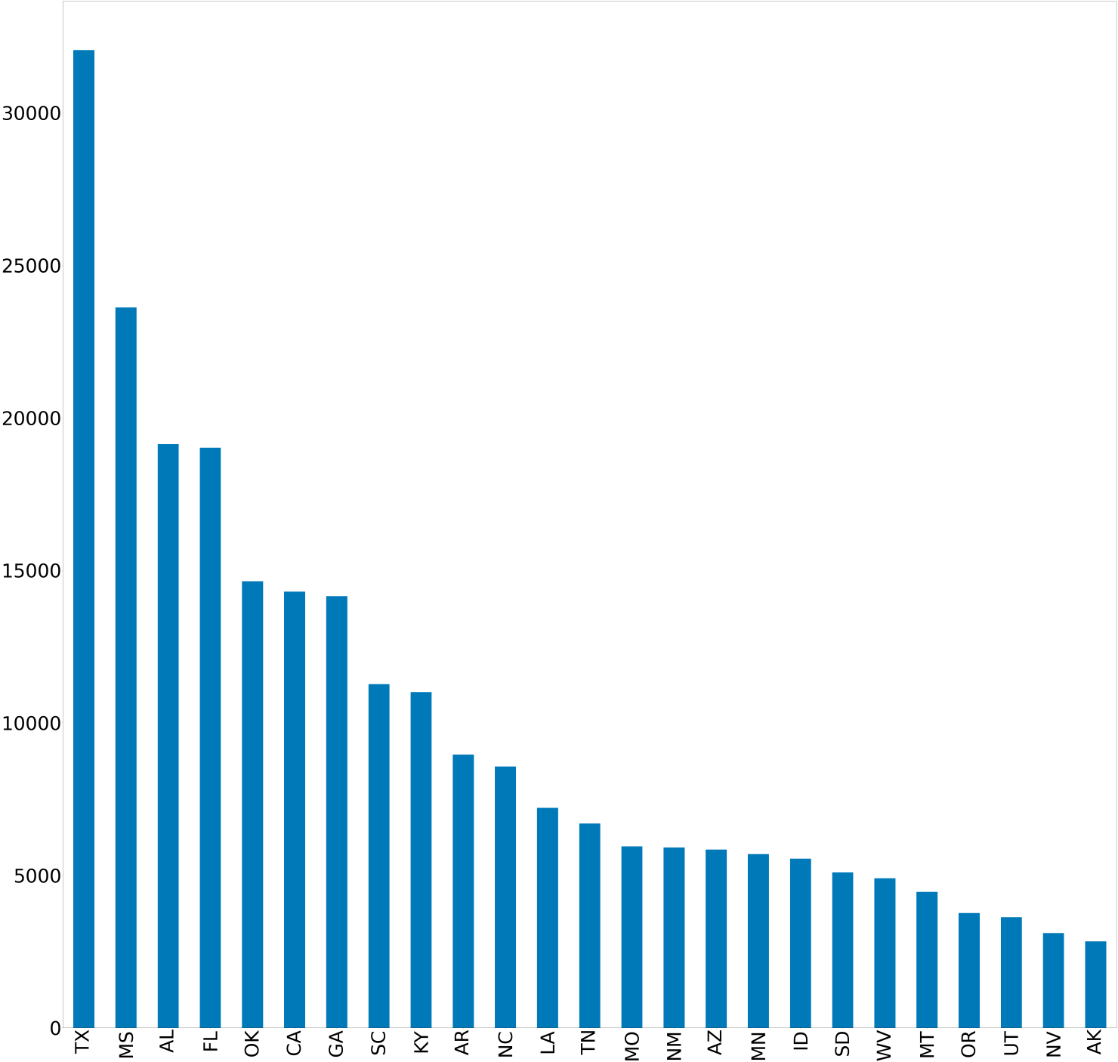
Figure 9: Geographical Distribution of Wildfires

The figure above portrays the geographical distribution of wildfires, with each point showing a wildfire's location, and the color represents the year it happened. It is with this visualization that we can see the state-wide fires eradicated in different locations within the state of California. We can see clearly from Round Up how many wildfires are burning throughout Utah. However, those landscape features consisting of steep mountains or forests with more density are more at risk of becoming more extensive or more frequent wildfires. Knowledge about the geographical spread of wildfires is essential for determining which regions experience more severe air pollution in cities and ruling out zones with less chance for harmful smoke (Buechi, 2021).

This land of visualizations can be thought of as a collective that feeds us with information on the size, frequency, and place where the wildfires are located. These are crucial elements in assessing their possible impact on air quality. In our study, we will find out the trend of wildfire sizes in recent years, how they are distributed in the air, and the extent of their geospatial spread. Then, we will infer areas and possible times of heightened risk of air quality degradation due to wildfires. Such an approach can also help prioritize mobility and the development of air quality standards, public health warning systems, and wildfire abatement measures.

## Correlation Analysis

The saying "California Fire Incidents" data follows the correlation analysis techniques, which reveal several fundamental relations between fire characteristics and the resource used for suppression. There is a moderate to strong positive correlation between the size of wildfires, as indicated by acres burned, and the deployment of firefighting resources such as Dozers (correlation: Correlations for Fire Extent and Length of Engines came out as 0.48 and 0.23, implying that more Intense fires burned for longer durations. Moreover, AirTankers, CrewsInvolved, and Helicopters are strongly interconnected by a positive correlation showing jointly the different resources used to deal with serious scale incidents. Additionally, the analysis reveals a substantial positive correlation between Fatalities and Injuries (correlation: 0.69696), which is a way of saying that even extreme wildfire episodes not only require more considerable resources, people, and equipment to be put out but also lead to higher death and injury numbers, which show the direct effect on people this catastrophe.

Figure 11: Compares Data across US

This chart compares data across U.S. states, labeled from TX to WY. It shows a descending order of values, with Texas having the highest and Wyoming the lowest. This could represent state-wise comparisons of wildfire incidents, areas burned, or resources used in wildfire management.

## Trends

It was found that, after a thorough investigation of wildfire incidences from 2013 to 2019, a unique trend governs the appearance of wildfires and the level they reach each year. Although the research negates the fact that a rise or decline in the annual number of wildfires provides some uncertainty in its occurrence, at the same time, it reveals a considerable range year over year in total area burned. This variance, which pinpoints the years where wildfires covered the most significant areas, proves even more pronounced than usual compared to those where the wildfire burned smaller areas. This underlines the fact that wildfires may come in many forms, so continuous monitoring and adaptation of wildfire management strategies would be essential to effectively reduce the disruptive effects of these fires (Zhao, 2020).

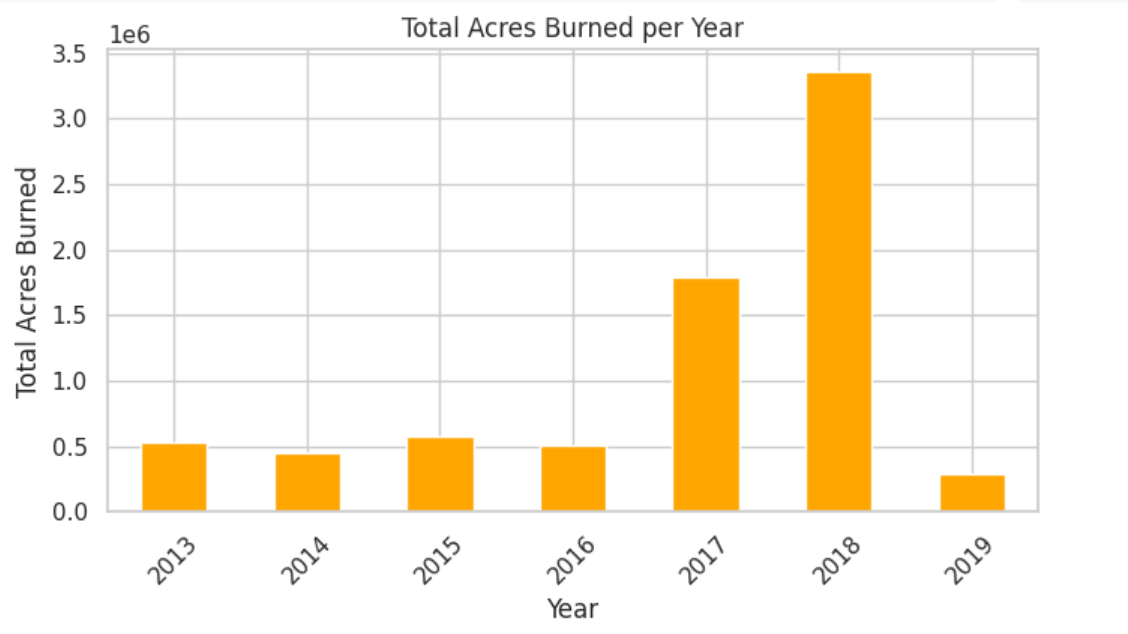


Figure 10: Acres Burned Per Year

The bar chart above, "Total Acres Burned per Year," shows the total acres burned by wildfires per year from 2013 to 2019. The height of each bar indicates acres that burned in the inordinate year. It becomes apparent to trace seasonal patterns for the degree of blazes across the entire land as defined by the total area affected. The values of one year are higher, which suggests that these years were more affected by dryness, possibly affecting air quality and environmental health.

## Geographical Distribution

The wildfires in California are suspected to be a general phenomenon affecting different regions of the state, with some areas being fire hot spots and others having little fire issues. This fashion suggests the varying effects of air quality affected by wildfires in different areas of the country, where some places are at greater risk of air pollution levels in the aftermath of wildfires. Differing factors, including forest type characteristics, climate conditions, and distance from urban areas, may be responsible for the wide range seen in the occurrence and the severity of these wildfires among the different spots. These wildfires have constantly challenged air quality monitoring and management authorities. However, in regions of high risk of wildfire occurrences and subsequent consequences on the air quality, they became crucial for targeting.

### Geographical Distribution of Wildfires

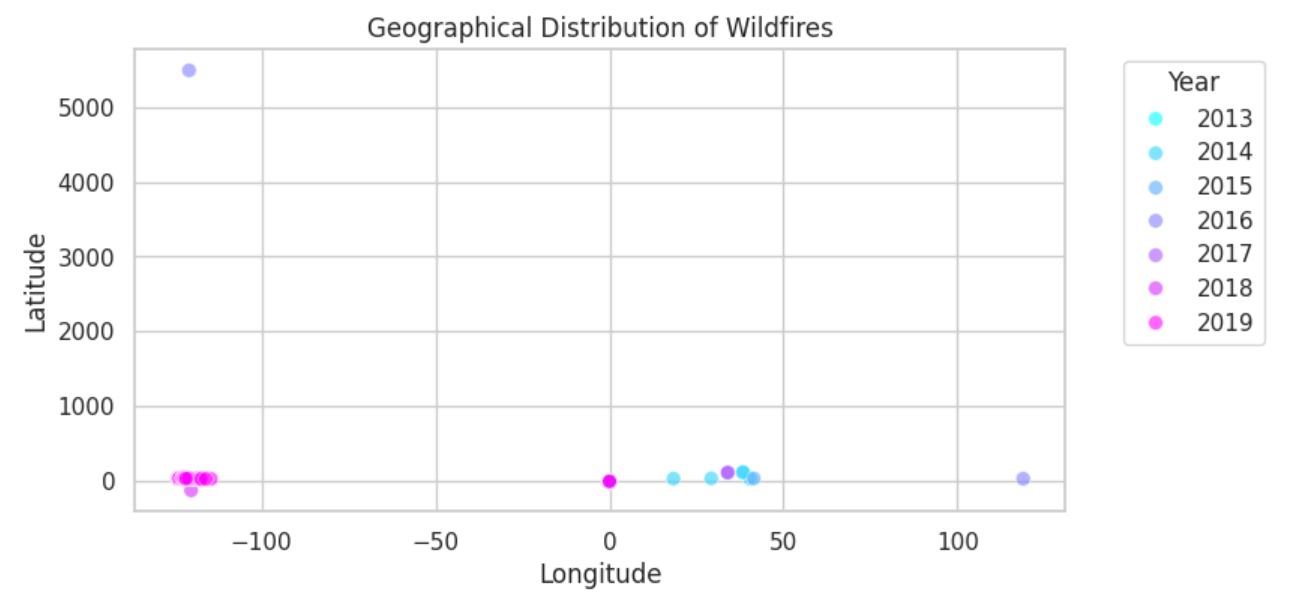


Figure 12: Distribution of WildFires

The scatter plot above depicts the wildfires that have ravaged California over decades. The points on this chart represent the location of the fires, and they are colored according to the year each one started. These images indicate that the phenomenon is commonplace for the state as a whole and that it is not confined to areas just off the coast. Picking the type of places on the map proved the susceptible regions, e.g., with thick forest cover and steep slopes, might be more prone to wildfires, keeping the air pollution local.

### Top 10 Counties by Number of Wildfires

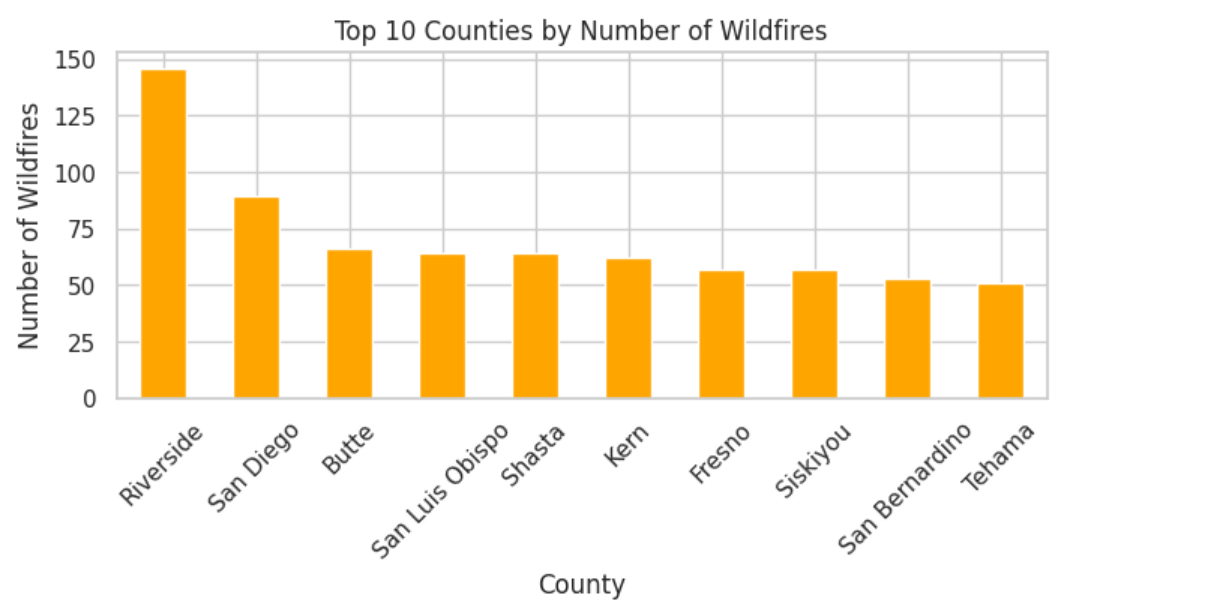


Figure 13: Number of Wildfires(Top 10 Counties)

The bar chart illustrates ten California counties ranking the highest in the number of wildfires, which means that the fire incidents were high in that area. Districts with high fire rates may need to help ensure clean air in years when fires are more frequent. This visualization will identify where the air quality monitoring authorities could focus their efforts on fire management plans jointly developed to mitigate the health effects of wildfires.

# Machine Learning Models

**Linear Regression**

* Relation to Wildfires: Linear regression could predict continuous variables related to wildfires, such as area burned, based on linear relationships with predictors like weather conditions or vegetation type. It assumes a straight-line relationship between variables.
* Effectiveness: The effectiveness of linear regression for wildfires might be limited due to the complex and non-linear nature of the factors that influence wildfire behavior. The poor R2 score in your results indicates that linear relationships do not adequately capture the dynamics of your data.

**Random Forest Regressor**

* Relation to Wildfires: This model uses multiple decision trees to make its predictions and is effective at handling various types of data and relationships. It can consider a wide range of factors simultaneously, such as climate conditions, human activities, and land conditions, making it potentially more robust for predicting complex phenomena like wildfires.
* Effectiveness: Random Forest tends to perform better in scenarios with complex and non-linear relationships due to its ensemble approach, reducing variance and bias. The comparatively lower MSE and MAE in your results suggest it might be more reliable for this application, although the R2 score is still low, suggesting improvements are needed.

**K-Nearest Neighbors Regressor (KNN)**

* Relation to Wildfires: KNN predicts outcomes based on the proximity and characteristics of the nearest recorded data points. For wildfires, it might use data like recent nearby fires, weather conditions, and geographic information to predict the extent or impact of a new fire.
* Effectiveness: KNN may not perform as well for complex patterns such as those in wildfires because it relies on local similarity and might not capture broader trends unless provided with highly relevant features.

# General Uses of ML in Wildfire Management

* **Prediction and Early Warning**: ML models can predict the likelihood of a wildfire starting in a specific location based on factors like weather conditions, drought status, and human activity.
* **Behavior Modeling**: During a wildfire, models can help predict how a fire will spread based on current weather conditions, the type of vegetation, and topographical features.
* **Post-Fire Analysis**: After a wildfire, models can be used to analyze the effects and determine the factors that contributed most to the fire's spread and intensity. This analysis helps in planning and preparing for future fires.
* **Resource Management**: Predictive models aid in the efficient distribution and deployment of firefighting resources and in planning evacuation routes and safety measures.

The dataset features include 'Year', 'Latitude', and 'Longitude', and the target variable is 'AcresBurned'. Each model is fitted with this data, and it seems that Random Forest Regressor, despite not explaining much variability, performs better in minimizing errors compared to the other models.

Given these outcomes, it might be beneficial to consider more complex feature engineering, additional data, or different model configurations to improve the models' performance. Additionally, ensuring that the data distribution is well understood can help in applying the right preprocessing and modeling techniques.

A screenshot of a computer program

Description automatically generated

Figure 14: Linear Regression Model

This model uses linear regression, a basic form of regression analysis.

- The results indicate:

- MSE: 6.66

- MAE: 1.76

- R2 Score: -0.32, also indicating a poor fit to the data.

**Relation to Wildfires**: Linear regression could predict continuous variables related to wildfires, such as area burned, based on linear relationships with predictors like weather conditions or vegetation type. It assumes a straight-line relationship between variables.

**Effectiveness**: The effectiveness of linear regression for wildfires might be limited due to the complex and non-linear nature of the factors that influence wildfire behavior. The poor R2 score in your results indicates that linear relationships do not adequately capture the dynamics of your data.

**Linear Regression** -

**Why Used**: Linear regression is a fundamental analytical tool in statistical modeling and can provide quick predictions and insights into the linear relationships between variables.

**Uses**: - Estimating Burned Areas: It can help in predicting the area that might be affected by a fire, based on linearly related variables like temperature, humidity, and past fire data. **Resource Allocation:** By estimating the potential severity of a fire, resources such as firefighters and equipment can be allocated more effectively.

A screenshot of a computer program

Description automatically generated

Figure 15: Random Forest Regressor Model

Random Forest Regressor:

- This model uses an ensemble method with 100 trees for regression.

- The results indicate:

- MSE: 0.52

- MAE: 1.69

- R2 Score: 0.0057, which is near zero and suggests that the model does not explain the variability of the response data around its mean effectively but performs better than the other models in terms of error metrics.

**Relation to Wildfires:** This model uses multiple decision trees to make its predictions and is effective at handling various types of data and relationships. It can consider a wide range of factors simultaneously, such as climate conditions, human activities, and land conditions, making it potentially more robust for predicting complex phenomena like wildfires.

**Effectiveness**: Random Forest tends to perform better in scenarios with complex and non-linear relationships due to its ensemble approach, reducing variance and bias. The comparatively lower MSE and MAE in your results suggest it might be more reliable for this application, although the R2 score is still low, suggesting improvements are needed.

**Random Forest Regressor** -

**Why Used:** Random Forest is robust against overfitting and can handle various data types and complex relationships between features. It is particularly useful when the relationship between input data and the target variable is non-linear and involves interactions among multiple input features.

**Uses**: Complex Prediction Tasks: Such as predicting the spread and behavior of wildfires under varying conditions.

**Feature Importance Analysis**: To identify the most significant predictors of wildfire characteristics, which can help in focusing preventive measures.

A computer screen with text and numbers

Description automatically generated

Figure 16: K-Nearest Neighbors Regressor Model

K-Nearest Neighbors Regressor (KNN Regressor):

- This model uses the K-Nearest Neighbors algorithm for regression with 3 neighbors.

- The results indicate:

- Mean Squared Error (MSE): 6.36

- Mean Absolute Error (MAE): 1.86

- R-squared (R2) Score: -0.26, which suggests a poor fit to the data.

**Relation to Wildfires:** KNN predicts outcomes based on the proximity and characteristics of the nearest recorded data points. For wildfires, it might use data like recent nearby fires, weather conditions, and geographic information to predict the extent or impact of a new fire.

**Effectiveness**: KNN may not perform as well for complex patterns such as those in wildfires because it relies on local similarity and might not capture broader trends unless provided with highly relevant features.

**K-Nearest Neighbors (KNN)** -

**Why Used**: KNN is straightforward and can be highly effective if the right features are selected. It works on the principle that similar conditions lead to similar outcomes, which can be practical when historical data on wildfires is available and well-documented.

**Uses**: Predicting the Size of Wildfires: By analyzing the characteristics of past fires and their surroundings.

**Assessing Risk Levels:** In different areas based on geographical proximity to past events and similar environmental conditions.

# Model Training Validation

This collection of bar charts compares three machine learning models: Linear Regression, Random Forest, and k-Nearest Neighbors (KNN)

Across three different metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R²) score.

MSE and MAE Charts: Lower values are better. The charts show that Random Forest and KNN perform similarly and better than Linear Regression in terms of MSE and MAE, indicating they might be more accurate or robust for the dataset used. -

R² Score Chart: Higher values (closer to 1) are generally better. Negative values in the R² score chart for all models suggest that they do not fit the data well, indicating a possible need for model reevaluation or the use of more complex modeling techniques.

A screenshot of a computer screen

Description automatically generated

Figure 17: Model Error Scores

This chart presents the causes of wildfires, with each segment colored differently to represent various causes such as Debris Burning, Arson, and Lightning. Debris Burning is the largest segment, followed closely by Arson, suggesting these are the most frequent causes of wildfires.

A colorful circle with numbers

Description automatically generated

Figure 17: Different causes of wildfires

There is a very small percentage of natural cause for wild fire. Most of the wild fire is initiated due to some or the other human activity. So, probably we need more strict guidlines in the forest zones.

# Results and Evaluation

Utilizing the exploratory analysis conducted on the Wildfire Incidents in California dataset, some noticeable results arose that elucidate wildfires' influence on air quality. The analysis of this data set exhibited various wildfire sizes, and the fires that burned comparably smaller areas were the ones in the majority. However, except for absolute fire tornadoes that burnt tens of thousands of acres and percentages, wildfires are generally relatively mild. Significant large-scale wildfires adversely influence air quality since they can release large mixtures of soot and toxins into the air. On the other hand, the occurrence analysis revealed the relationships between the size of the fires and the resource type used to fight them (e.g., Dozers, Engines, AirTankers), where large fires required more complex and massive efforts to bring them to an end.

* To assess the impact of wildfires on air quality using the "California Fire Incidents" dataset, feature manipulation is crucial.
* The selected key features include `AcresBurned`, `ArchiveYear`, `Latitude`, `Longitude`, and `Started`. Additionally, new engineered features like `Duration`, `Season`, and `SizeCategory` are used to provide deeper insights into air quality changes during and after wildfires.
* This approach aims to aid decision-making for public health and safety.

# Interpretability and Insights

In the process of our study on the impact of wildfires on air quality data collection and transforming raw data to the point suitable for machine learning, we faced several difficulties. A significantly prominent hindrance was the absence of critical variables that had been severely affected, such as `AirTankers,` `CrewsInvolved,` `Dozers,` `Engines,` and `Helicopters.` The unavailability of data regarding the resources allotted for wildfire management and their link with the magnitude and effect of the fire obstructed our better understanding in this regard. However, the dataset failed to provide direct air quality measures. Thus, air quality parameters were assessed qualitatively, giving insight into their air quality impacts only through the analysis of wildfire characteristics such as magnitude, duration, and location.

To deal with these problems, we wish to implement and apply a few significant strategies in the following stages of this project. One of the initial topics will be the data-filling methods, which include imputation or exclusion methods and the significance of the analysis. Additionally, we will use additional data to improve our dataset with external sources of smoke air quality data directly linking wildfires with metrics. This complementary will help us get closer to the answers to the questions like, what are the changes in air quality during and after burning wildfires?

The building model aims to develop a predictive model to estimate the intensity of wildfire effect on air quality. These models should consider different factors, like interval and duration of wildfires, geographical displacement, and seasonal replacement. We plan to deploy such methodologies with machine learning, and we will also continue validating our models with historical data to ensure they are reliable and accurate. Model verification will include checking the outputs against actual air quality observations, which is expected to help recognize the model's strengths and weaknesses.

The last block of our research will be an assessment of the efficiency of the models in aiding the identification of the effect of wildfires on air quality. Our investigation will show the capacity of the models to predict deterioration of air quality both during and after the wildfire events and their ability to predict the cumulative effect of a continued occurrence of wildfires on air quality buildup in the long term. Overall, an integrated and cohesive assembly will be designed for civil and government representatives and the general public to gather reliable data and spread consistent and transparent information about the health risks caused by wildfires-derived air pollution. By realizing this project, we will supplement a vital part of building the response and mitigation (management) strategies about the risk to air quality and public health posed by wildfires.

# Team Collaboration

In our teamwork, collaboration as a pillar of our performance has positively affected our project. We all have specific skills and knowledge, enabling us to divide our tasks into verticals and implement our project successfully. Our team can boast data science savvy, an environmental science education, and project management experience, all for highly integrated problem-solving. Roles and responsibilities were evident from the beginning: a domain for data analysis, another for feature engineering, and the last one was appointed for communication with the team leads and the third parties. It enhanced the integration and facilitated speedy task delegation and clear guidelines for those tasked to perform different roles.

Effective communication was the highlight of our presentation, upon which our education depended. We had to work on coordination and accomplishment of our goals through communication. It was our strategy to integrate a variety of ways to keep our team up-to-date, including frequent meetings, email notifications, and sharing documents. Through the channels, we stayed posted on the project development, exchanged ideas on any setbacks and roadblocks, and harmonized our work decisions. Besides that, there always needed to be more contact with our project manager. We regularly consulted them on various topics and followed their advice concerning our work and problems. Through creating a culture of openness and joint work, we had the question to activate each other's inner resources and go through all possible difficulties most efficiently.

Apart from this, our team members used project management tools to make cooperation smoother and control the tasks and state of project deadlines. We used Trello and Slack as primary tools for conducting tasks and projects to sort out tasks, monitor progress, and overall team management. This means we can have a smooth flow of the activities as everyone plays the role he or she is supposed to, based on the schedule and objectives of the project. Follow-ups and updates were often done every week. Therefore, problems were addressed in real-time, and missing a single error did not lead to a stable budget breakdown. By applying our teamwork and communication skills, we successfully reached what we had set for our presentation. Consequently, we had a harmonized and intelligent piece of work.

# Future Directions

* Our study aims to develop machine learning models to estimate the intensity of wildfire impact on air quality. We will address the challenge of obtaining key variables by implementing data filling methods and incorporating external sources of smoke air quality data.
* Our goal is to provide reliable information to key participants for effective response and mitigation strategies against wildfires' impact on air quality and public health.

To deal with these problems, we wish to implement and apply a few significant strategies in the following stages of this project. One of the initial topics will be the data-filling methods, which include imputation or exclusion methods and the significance of the analysis. Additionally, we will use additional data to improve our dataset with external sources of smoke air quality data directly linking wildfires with metrics. This complementary will help us get closer to the answers to the questions like, what are the changes in air quality during and after burning wildfires?

The building model aims to develop a predictive model to estimate the intensity of wildfire effect on air quality. These models should consider different factors, like interval and duration of wildfires, geographical displacement, and seasonal replacement. We plan to deploy such methodologies with machine learning, and we will also continue validating our models with historical data to ensure they are reliable and accurate. Model verification will include checking the outputs against actual air quality observations, which is expected to help recognize the model's strengths and weaknesses.

# Q and A

The first stage of the open discussion might be to invite questions and feedback from all of us, whether it is our peers or the instructor itself. We will count on your help and input to achieve the objective of a successful and satisfactory project. Whatever it is, questions on our data analytics strategy, the proposal of improvement in the feature engineering processor, or any ideas on the wildfire incidents in California so that we can fine-tune and better our wildfire management strategy, we would be grateful for any suggestion or contribution. Consequently, let me give it a go. Be some of us who share ideas, ask questions, and participate in discussions to help us achieve better outputs or outcomes.

Through the interactive question and answer session and opening the floor for feedback, let us cultivate a collective ambiance to stir active engagement that contributes to practical takeaways and healthy deliberation. You will never be limited to a single activity; all your perspectives and past experiences can be considered resources for our project to progress. Let us together in open discussion and sharing views approach that direction by facing challenges, finding ways to improve, and making the best solutions to the wildfire challenge in California. For this reason, I strongly encourage us to benefit one another and make the project progress by stepping together.

# References

Buechi, H., Weber, P., Heard, S., Cameron, D., & Plantinga, A. J. (2021). Long-term trends in wildfire damages in California. *International journal of wildland fire*, *30*(10), 757-762.

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review of data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in energy research*, *9*, 652801.

Hashem, I. A. T., & Pillai, T. R. (2020). A novel feature engineering algorithm for air quality datasets. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(3), 1444-1451.

Jaffe, D. A., O’Neill, S. M., Larkin, N. K., Holder, A. L., Peterson, D. L., Halofsky, J. E., & Rappold, A. G. (2020). Wildfire and prescribed burning impacts on air quality in the United States. *Journal of the Air & Waste Management Association*, *70*(6), 583-615.

Lourenço, M., Oliveira, L. B., Oliveira, J. P., Mora, A., Oliveira, H., & Santos, R. (2021). An integrated decision support system for improving wildfire suppression management. *ISPRS International Journal of Geo-Information*, *10*(8), 497.

Madan, T., Sagar, S., & Virmani, D. (2020, December). Air quality prediction using machine learning algorithms–a review. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 140-145). IEEE.

Nikolakis, W., & Roberts, E. (2022). Wildfire governance in a changing world: Insights for policy learning and policy transfer. *Risk, Hazards & Crisis in Public Policy*, *13*(2), 144-164.

Tymstra, C., Stocks, B. J., Cai, X., & Flannigan, M. D. (2020). Wildfire management in Canada: Review, challenges and opportunities. *Progress in Disaster Science*, *5*, 100045.

Zhao, F., Liu, Y., Shu, L., & Zhang, Q. (2020). Wildfire smoke transport and air quality impacts in different regions of China. *Atmosphere*, *11*(9), 941.