**Homework 1: Project Background, Data Background, Problem Formulation, and Data Strategy Plan**

By David Jean

**Project Background**

I’ve been hired by the National Interagency Fire Center on behalf of the United States Department of Agriculture to acquire, analyze and predict wildfire trends. As temperatures continue to rise across the United States, so does the number of wildfires annually. On average, the US has around 60,000 wildfires per year which destroy over 6.8 million acres of land. Although some of these wildfires are started by naturally occurring lightning strikes, most are started by humans, either accidentally or on purpose. We know that the most fires occur in the western united states and that this surplus of fuel is a result of the arid climate, high winds, and history of fire suppression. In 2021, wildfires cost the United States a total of 2.7 billion dollars. This aggregated figure is comprised of the cost of property damage, firefighting expenses, medical responders, personnel, trucks, equipment, and aircrafts. We will use all available information at our disposable to provide critical information to decision makers in hopes of that they can better allocate resources and improve wildfire management practices.

**Problem Formation**

My approach is to use state of the art machine learning techniques to derive meaningful insights from the following questions to better allocate federal resources for high-risk areas:

What locations have the most fires? What are the primary causes of the fires? What time of day do most of the fires occur? How much area is burned per fire? How long are the fires active? How long does it take to put out a fire based on class size? What agencies respond to the fires? What are the costs associated? Are there diminishing returns for dollars spent vs the amount of land destroyed per fire? What agency reported the fires?

I believe these questions are critical for us to better understand the patterns and behaviors behind this destructive force. We need to find better ways to predict, prevent, and respond to wildfires, and data can help us do that. By analyzing the available data on wildfire activity, we can identify patterns and trends in fire behavior, determine the most effective strategies for fire management and suppression, and develop better tools for predicting and preventing wildfires before they occur.

**Data Strategy Plan**

The strategy for obtaining and using this data will come from a variety of sources such as the National Interagency Fire Center, the National Wildfire Coordinating Group, the US Forest Service, Nasa, and local and state agencies.

The dataset we will be investigating contains a mixture of at geospatial data, financial data, environmental data, weather data, police report data, federal data, socio-economic data, and historical data for every single datapoint. The data we have obtained is also found on Kaggle in SQL format. We will be examining date ranges from 1992-2020 to see how wildfire behavior has changed over time. We will use a combination of data collection methods and data processing techniques to ensure that the data we have obtained is accurate, complete, and relevant to our research questions. This will involve working closely with data sources to ensure that we are applying appropriate methods to clean, analyze, visualize, and model the data.

**Homework 2: Summary of Data Cleansing and Exploratory Data Analysis**

**Preliminary Investigation and Exploration:**

Our dataset for this project consisted of combining several datasets from federal and state agencies showing a total number of 2.3 million fires ranging from the dates 1992 -2020. Every observation had 30 unique columns with information such as date, time, unique identifiers, source information, agency and unit identifiers, fire location and size, cause classification, and owner/manager information. Quite a few observations had missing data, primarily on the local government level pertaining to the fire management details. We chose to either fill some in with the averages for some columns, and for columns that didn’t have a start or end time/date, we filled them in so that it would not make the total duration of a fire negative. To make predictions in the future, we will make a slimmed down version of the dataset that removes a few of the lesser important variables.

**Initial Insights:**

We first inspected the number of fires taking place year over year and saw that 2006 had the most fires at 117,944, with 1997 having the least at 61,442. The number of fires fluctuates drastically each year with the average being around 95,000 and since 2018 we have seen an upward trend in the number of fires per year. Over time, class B fires have accounted for 1,104,387 or 47.9% of the total fires. Class B is defined as a fire that destroys .25-9.9 acres of land. The fire size code is determined by the acreage within the final fire perimeter, where A represents greater than 0 but less than or equal to 0.25 acres, B represents 0.26-9.9 acres, C represents 10.0-99.9 acres, D represents 100-299 acres, E represents 300 to 999 acres, F represents 1000 to 4999 acres, and G represents 5000+ acres.

Next, we wanted to determine what was causing most fires and where they were taking place. According to our data, 600,000 or 25% of our data indicates that the cause of the fire was unknown. A few other key contributors were lightning, debris burning, arson and equipment malfunctioning. In order to dive deeper into our data, we wanted to see which states had the most fires. California tops our list with over 250,000, followed by Georgia and Texas with over 175,000 each. However, these fires were not the most destructive due to effective fire management efforts already in place. Alaska had the longest average burning fires at over 12 days per fire and destroyed over 35 million acres of land over 25 years. This can be attributed to sparse population, dense wilderness, and lack of fire responders in the area. In our dataset, there were over 20 fires that last 300 days or more. Over half of the fires in our dataset, or 1,174,725 to be precise, were contained on the day they were discovered. The next highest duration was contained in 1 day and that occurred 129,708 times. The number of occurrences continues to decrease by half for every day after the first. They are a few outliers that we noticed such as one fire that destroyed over 309,200 acres over the span of 237 days before it was contained.

With most fires being contained on the same day and mostly being relatively small (Class B) we can determine that our agencies resources are being well spent, with some areas that could be improved upon. One area for improvement is in the accuracy and completeness of the data. As mentioned earlier, quite a few observations had missing data, primarily on the local government level pertaining to the fire management details. Improving data collection and reporting practices could help provide a more comprehensive picture of fire management efforts and their outcomes. Additionally, it may be worth exploring how fire management resources could be allocated more effectively to prevent fires from growing to larger sizes and causing more damage. This could involve targeted outreach and education campaigns, improved detection and response systems, or other strategies. Overall, our analysis suggests that while there is room for improvement, current fire management efforts are generally effective and should continue to be supported.

**Homework 3: Summary of Dashboard and Visualizations**

For this section, I’m including graphs that highlight some of our analysis that I’ve created in python using the Plotly library. Each of these graphs is interactive and you can click the legend on the sides to change what you want to see and it will update automatically.

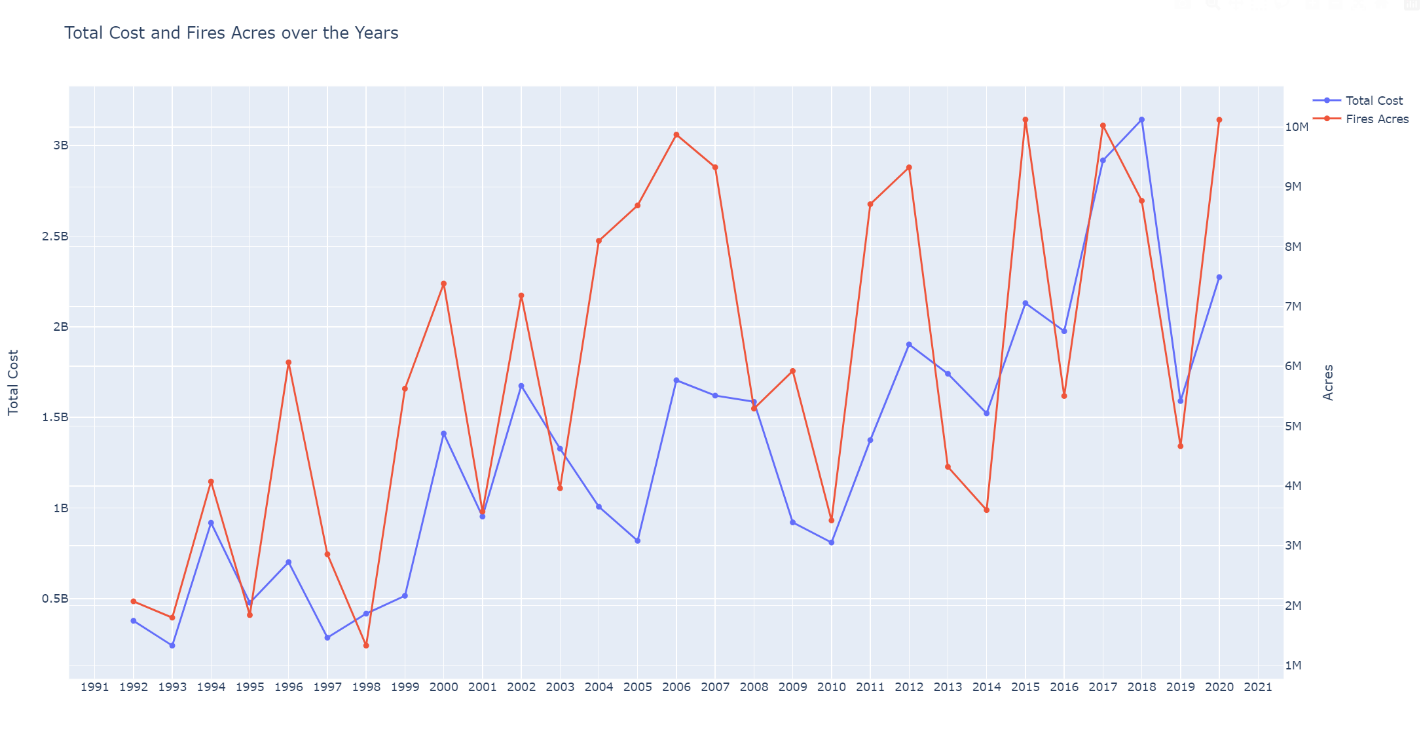
To begin with, I created a dual Y axis line chart that shows the number of acres destroyed for the time frame of 1992-2022 as well as the total associated costs. To provide clarity, I included descriptions on the left and right sides. 

Figure1

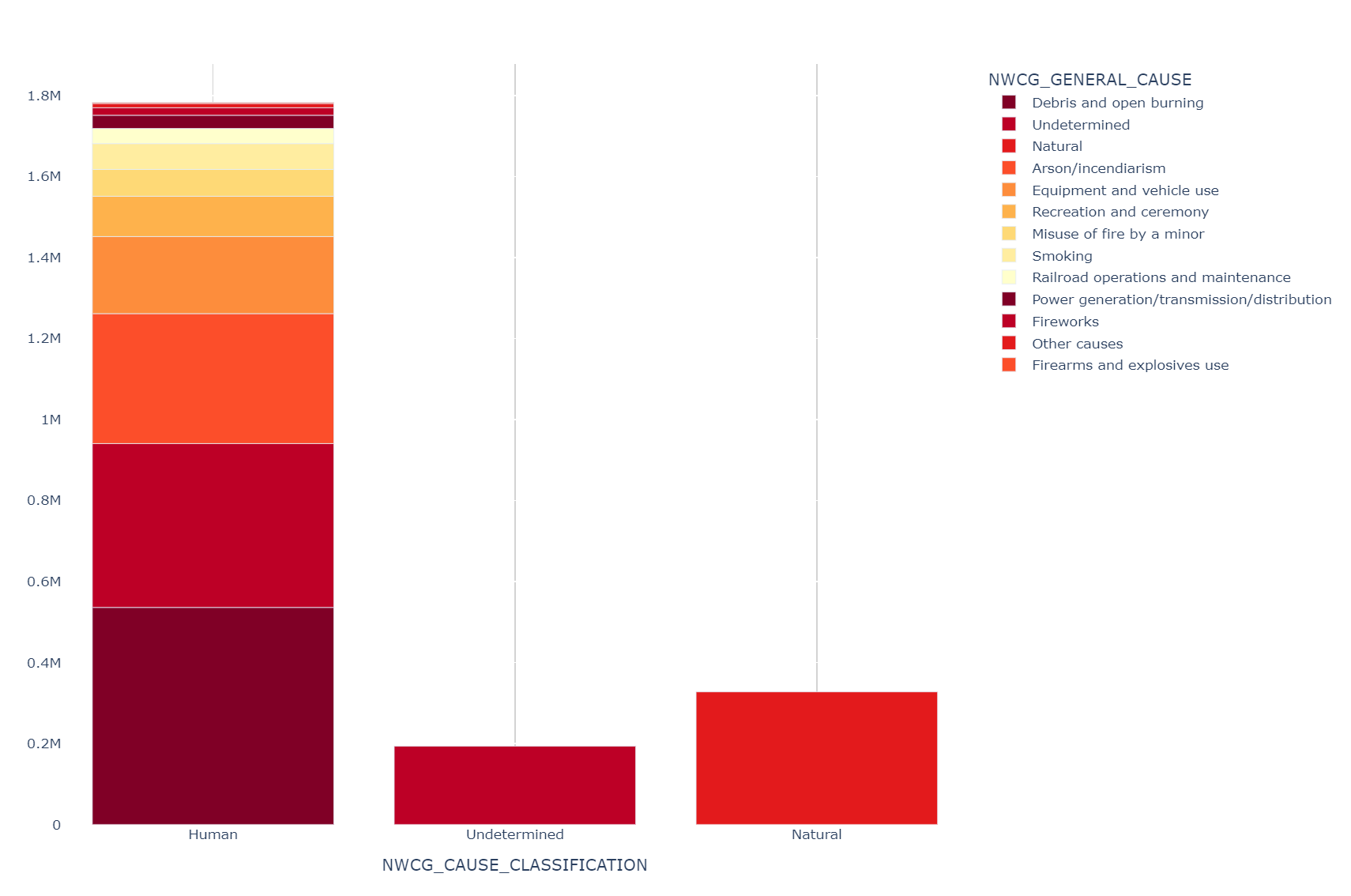
Each datapoint included the general cause of the fire, and this stacked bar chart groups them by whether they were of human cause, natural, or unidentified and shows the proportions for each. I think I am going to go back and split this one into two separate graphs with one of them showing the general cause in a little more detail but overall I think this one looks great.

Figure2

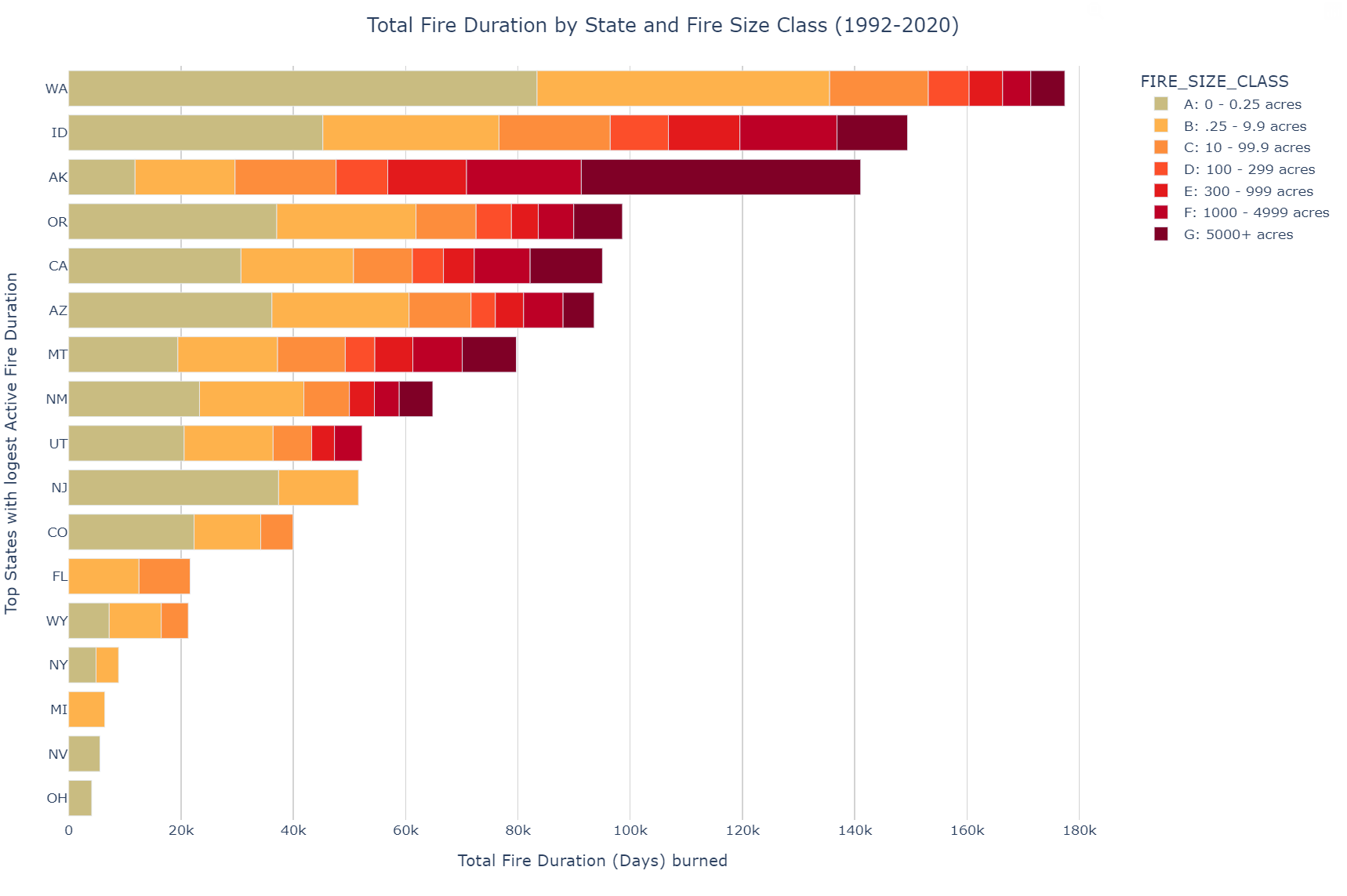
To investigate even further, we wanted to look at the duration of each fire size class by state. Each bar represents the sum of a state over time and the subsections of each bar are the proportion of each fire size class’s burn time(days).

Figure3

After applying our machine learning algorithm, we drilled down on identifying the root cause in order to best determine if it was caused by humans or not. This donut chart delivers the information in a highly effective manner.

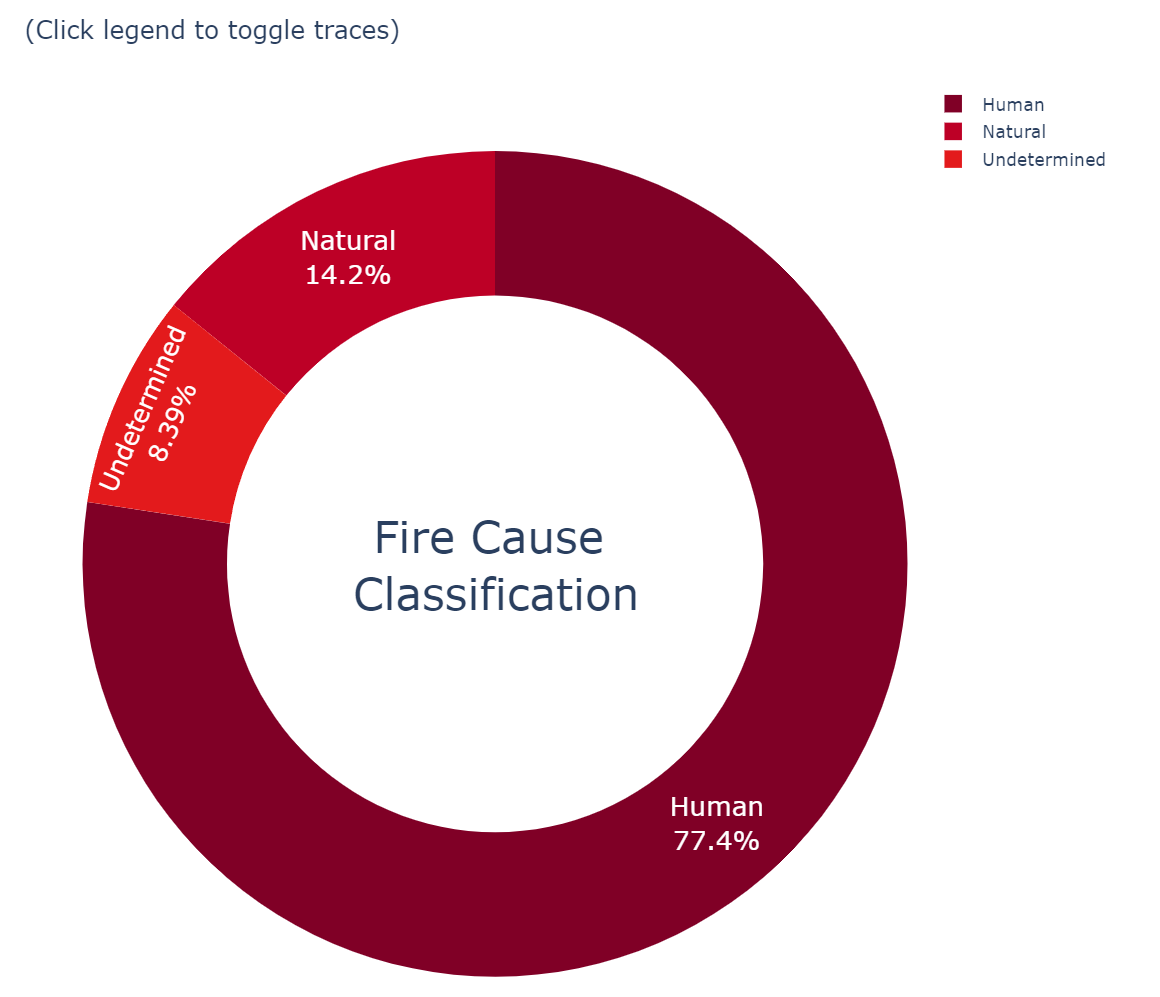


Figure4

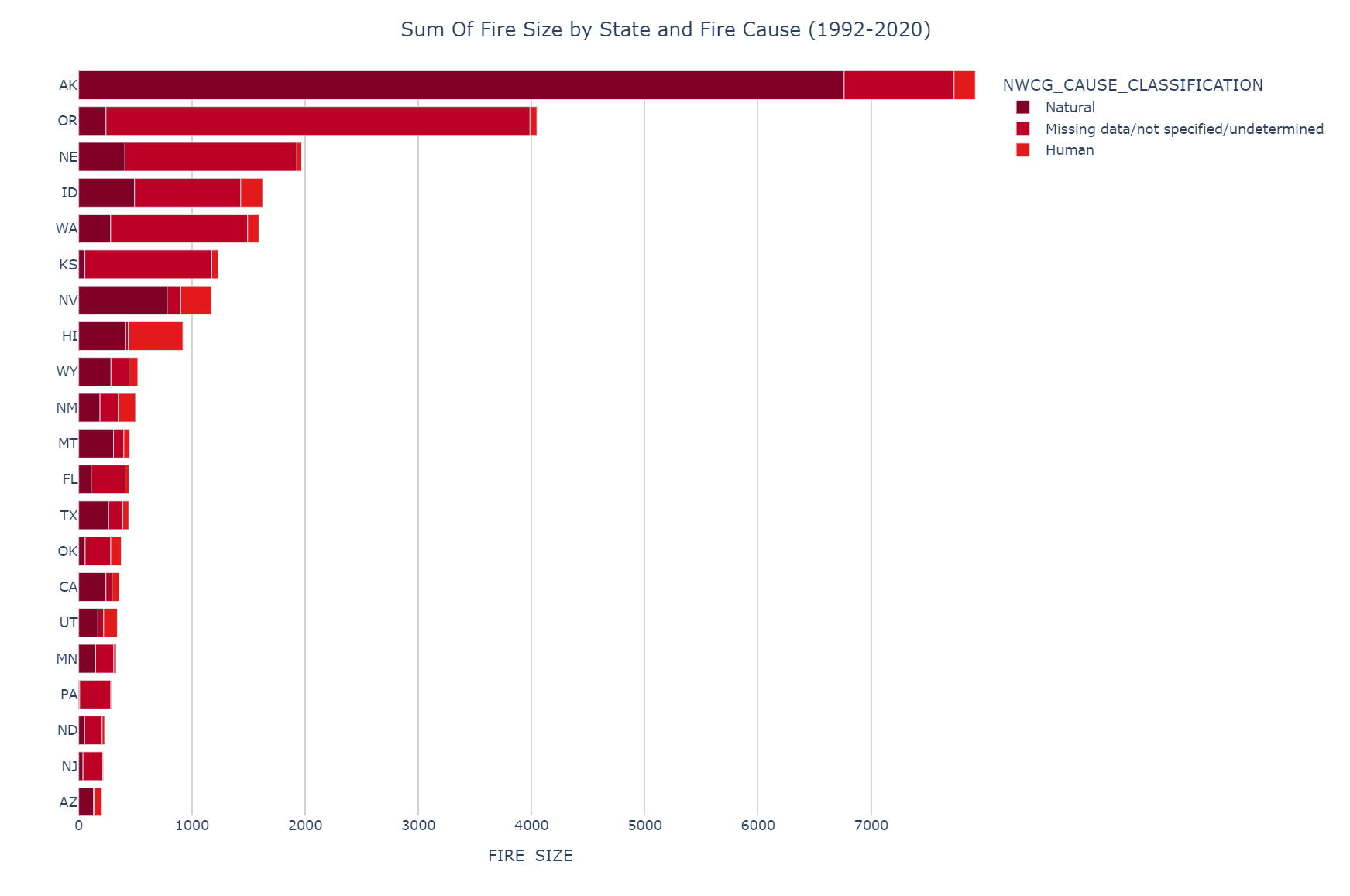
We then decided to dive into state level and see which states had the largest sum of of fire sizes by type. 

Figure 5

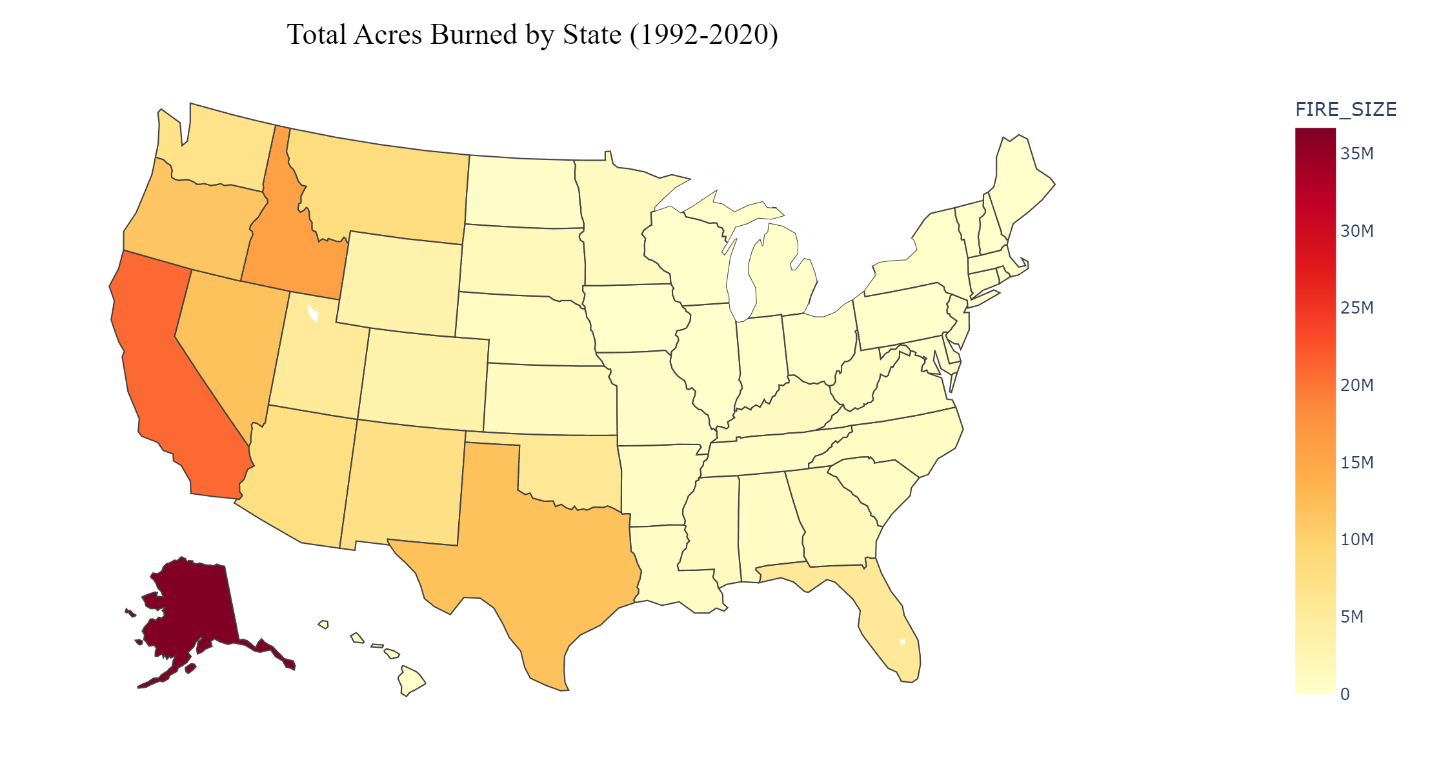
Lastly, we wanted to visualize the number of acres destroyed in a different way. This choropleth map shows the number of acres destroyed over time(millions).

Figure 6

This Sunburst chart shows the total reports by unit type.

Chart, sunburst chart

Description automatically generated

Figure 7

This stacked bar chart shows the specific reporting agency reports by Month.

Chart, bar chart

Description automatically generated

Figure 8

This chart visualizes the fire count by region for the last 10 years.  
Chart, bar chart

Description automatically generated

Figure 9

This chart visualizes the average fire size by region over time for the last 10 years.

Chart, bar chart

Description automatically generated

Figure 10

This treemap visualizes the unique human caused fire types and their proportions.

Chart, treemap chart

Description automatically generated

Figure 11

This map shows a sample of 50,000 fires from last year, their size, and their cause.  
Map

Description automatically generated with low confidence

Figure 12

**Homework 4: Analysis, Business Insights, and Final Summary**

For our analysis, we chose to use 2 different models to answer questions. We used a regression model to predict the total cost of fire suppression for next year and a classification model to predict whether a fire will be contained on the same day it was discovered or not. This can be particularly useful, but the cost of a single fire can easily escalate to tens or even hundreds of millions of dollars if not contained quickly and efficiently. More importantly, whenever a new fire is discovered in real time, our model can predict if it is able to be contained on the same day based on such features as latitude, longitude, fire size class, state, and several other factors.

Secondly, our Regression model can be used to predict the total cost of fires for the next year based on historical data. Since the number of fires, size of fires, and cost can vary drastically each year due to weather conditions and human involvement, having an accurate model will be particularly useful for government officials lobbying for additional resources from congress. Out of next years 1.5 trillion dollar budget for the United States Federal Government, our model predicts that in 2021, fire suppression costs will only 1.5 billion or 0.1% of that total budget. With NWCG receiving funding negatively proportional to the damage of last years fires, receiving more funding at the federal level would be more beneficial to allocate it to states with a high amount of class B+ fires. Since the US averages 60,000 fires a year, being able to put them out before they grow out of control is an upmost priority.

**Where it could be the most useful:**

* We were able to determine that California, Alaska, Georgia, Texas and North Carolina had the most fires overall through the time period. Fortunately, because of a existing presence of suppression ready forces, California, Georgia, Texas, and North Carolina have adequate measures in place.
* Notably, the areas with the most fires did not necessarily have the longest burning or largest with Alaska being the outlier. This is due to having adequate suppression and prevention measures already in place.
* One key area for discussion is Alaska which had the largest number of class G fires and over 20 fires that burned for 300+ days each. Because of the remoteness, and low population, this key state is not a high priority in terms of suppression, and we should not allocate any additional resources.

Having accurate figures on fire suppression costs and containment probabilities is crucial not only for effective resource allocation and decision-making but also for the safety and well-being of individuals and communities. When wildfires are not contained quickly and efficiently, they can spread rapidly, resulting in significant damage to property and natural resources, as well as threatening the lives of residents in the affected areas. Moreover, the cost of wildfires can be substantial, ranging from millions to billions of dollars, which can place a significant financial burden on communities and governments. Therefore, having reliable models that can predict the total cost of fire suppression and the likelihood of containment can help prevent and mitigate the impact of wildfires, ultimately protecting lives, property, and the environment.

**Model 1:**

This model used XGBoost Classifier to determine whether a fire would be put out on the same day it was discovered. We had to take significant cleaning measures to ensure accuracy including imputing missing values in the discovery time and contained time to make sure there would not be negative values.

We used binary classification for this model with XGBoost which is a very popular and state of the art model which predicts whether something belongs to one of two categories. It works by creating a series of decision trees that work together to make a final prediction. Each tree is trained on a different subset of the data, and the final prediction is based on the combined results of all the trees. XGBoost Classifier is also able to handle missing data and outliers, making it a robust choice for many real-world problems.

We then created a pipeline is being created using the Pipeline class from the Scikit-learn library. The 'clf' estimator is defined to use the XGBClassifier algorithm with 'binary:logistic' as the objective function, which is a common setting for binary classification problems. The 'random\_state' parameter is set to 21 to ensure that the results are reproducible. The pipeline is then created by passing the 'estimators' list to the Pipeline constructor. This pipeline can be used to process and transform data, fit the model to the data, and make predictions on new data in a single step. The resulting pipeline will first apply any data preprocessing steps, such as scaling or feature selection, and then fit the XGBClassifier model to the transformed data.

We then will be performing a hyperparameter tuning for a machine learning model using Bayesian optimization. Hyperparameters are settings that need to be specified before training a machine learning model, and they can greatly affect the model's performance. The search\_space dictionary specifies the range of values that the hyperparameters should take. The opt variable is created using the BayesSearchCV class from the Scikit-learn library, which performs Bayesian optimization to find the best combination of hyperparameters for the given model and data. It uses a cross-validation strategy to evaluate the performance of each combination of hyperparameters, and it tries different combinations based on the search space dictionary. The 'cv' parameter specifies the number of cross-validation folds, and the 'scoring' parameter specifies the evaluation metric used to compare the performance of different models. Finally, the 'verbose' parameter determines the level of output messages during the optimization process.

After our training time of 4 hours, it is time to measure our results. The outcome of our efforts for this model’s ROC/AUC score score of 0.739 means that the machine learning model has an overall moderate ability to distinguish between the positive and negative classes in the data. The score ranges from 0 to 1, with 1 indicating a perfect classifier and 0.5 indicating a random classifier. A score of 0.739 suggests that the model is performing better than a random classifier but may still have room for improvement. This score can be used to compare the performance of different models and to evaluate whether the model is suitable for the intended use case.

To further analyze our model’s results, we must look at the Accuracy, Recall, F1-score, and Precision. The results above show the performance of a machine learning model on a binary classification task. The accuracy of 0.8818 indicates that the model correctly classified 88.18% of the data. The recall of 0.9622 indicates that the model correctly identified 96.22% of the positive cases, which means it has a low false negative rate. The precision of 0.9026 indicates that when the model predicted a positive case, it was correct 90.26% of the time, which means it has a low false positive rate. The F1-score of 0.9314 is a harmonic mean of precision and recall, and it provides a balanced measure of the model's overall performance. These results suggest that the model is performing well on the given task, but it's important to evaluate its performance on new, unseen data as well.

When used in live production, the importance of this model lies in its ability to predict whether a newly reported wildfire is likely to be contained on the same day or not. This prediction has the potential to save tens of millions of dollars per fire, and enables a swift decision on resource allocation. The most impactful features to determine whether a fire would be contained on the same day it was discovered or not were the latitude, longitude, fire size, and state. Western states tend to have the highest amount of destructive fires, and this model confirms that all financial resources should continue to be supported.

**Model 2:**

Our 2nd model uses XGBOOST’s linear regression on the Suppression costs dataset to predict the total costs of firefighting for next year. The provided dataset has six variables and 34 observations, each of which represents a year from 1985 to 2018. Year, Fires, Acres, ForestService, DOIAgencies, and Total are the variables in the dataset. The Fires variable shows how many forest fires there were each year. The total number of acres burned as a result of forest fires is represented by the Acres variable. The ForestService and DOIAgencies variables show the corresponding sums of money spent on fighting forest fires by the Forest Service and the Department of the Interior. The Total variable shows the overall sum of funds used by both organizations to put out forest fires. The relationship between the number of forest fires, the damaged region, and this dataset can be studied.

A dependent variable (also known as the response variable) and one or more independent variables (also known as the predictor variables) are compared using a statistical technique called linear regression. The change in the dependent variable is assumed to be proportionate to the change in the independent variable(s) because it is based on the linear relationship between the variables. When making predictions for fresh data, linear regression calculates the values of the coefficients that best suit the observed data. It can be used to assess the importance of the calculated coefficients, detect outliers or significant data points, and study the direction and strength of the association between variables.

A potent machine learning technique called the XGBoost Regressor can be used to forecast future values using historical data. The XGBoost Regressor will use the numbers entered by the user for 2021 to produce a virtually accurate price prediction. This is possible because the algorithm can examine a lot of data and spot patterns that aren't immediately obvious to the naked eye. The computer can develop a model that precisely forecasts future prices by using these patterns. Overall, the XGBoost Regressor is a useful tool for anyone trying to predict future pricing accurately.

**In Conclusion:**

Wildfires have become an increasing threat to lives, property, and the environment, particularly in regions prone to dry, hot weather conditions. The cost of suppressing these fires can reach tens or even hundreds of millions of dollars if not contained quickly and efficiently. Therefore, there is a pressing need for reliable models that can predict the total cost of fire suppression and the likelihood of containment, which can help prevent and mitigate the impact of wildfires.

To address this need, our team used two different models to answer important questions related to wildfires. First, we used a regression model to predict the total cost of fire suppression for the next year based on historical data. This model takes into account various factors such as the number and size of fires, as well as weather conditions and human involvement. Having an accurate model can be particularly useful for government officials in lobbying for additional resources from congress. Our model predicts that in 2021, fire suppression costs will only amount to 1.5 billion or 0.1% of the United States Federal Government's total budget.

Secondly, we used a classification model to predict whether a fire will be contained on the same day it was discovered. This model takes into account features such as latitude, longitude, fire size class, and state, among others. The ability to predict containment probability in real-time can help authorities make critical decisions regarding resource allocation and public safety.

Our classification model uses XGBoost, which is a state-of-the-art machine learning algorithm that creates a series of decision trees to make predictions. We created a pipeline using the Scikit-learn library to process and transform data, fit the model, and make predictions on new data in a single step. We also performed hyperparameter tuning using Bayesian optimization to find the best combination of hyperparameters for the given model and data.

One key finding from our analysis is that areas with the most fires did not necessarily have the longest burning or largest fires. This is due to having adequate suppression and prevention measures already in place. We also found that Alaska had the largest number of class G fires, and over 20 fires burned for 300+ days each. Because of the remoteness and low population, this state is not a high priority in terms of suppression, and we should not allocate any additional resources.

In conclusion, having accurate models to predict the total cost of fire suppression and the likelihood of containment can help prevent and mitigate the impact of wildfires, ultimately protecting lives, property, and the environment. Our analysis highlights the importance of adequate suppression and prevention measures, as well as the need for effective resource allocation and decision-making to manage the growing threat of wildfires.