CEN 6940 – Computing Practicum

Deliverable 5 – Data Analytics – Final Report

Due date: December 12

The final report for the data analytics option should include the following items:

Team name: MAD

Team members: Dafni Dimitriadi, Morgan Stephens, Ahmed Sayed

Title for the project: Wildfire Prevention and Prediction in the US

Executive summary:

The executive summary should briefly describe the problem, data, data analytic solution, issues faced, key findings, and other relevant items.

The problem is wildfires have increased over the past years. We found a dataset from data.gov that has over 2,000,000 rows of data regarding wildfires and where they were discovered. We created Tableau dashboards to look at patterns within the data. The data analytic solution is that we are using linear regression to predict wildfire locations. An issue faced was making sure the dates were correctly formatted and getting rid of unwanted attributes. One key finding is that Alaska and California are high-risk locations for wildfires. Another key finding is that humans are the main cause of wildfires.

Introduction:

Describe the problem, why it is vital to address the problem and your motivations for addressing the problem, and choose the approach used to develop the solution. What can be done to address or reduce the problem? Describe your solution to the problem. What kinds of research/project outcomes could come out of this idea? List multiple goals in a brief, itemized format. Why will the project outcomes matter? How will the project outcomes help relevant stakeholders? Briefly describe your motivation for conducting this project. Briefly describe the significance of the project work and how it will contribute to the existing body of knowledge.

Wildfires have been significantly increasing over the years, in terms of frequency and intensity. This issue is multifaceted and arises from the complex interaction of human, natural, and environmental factors. Data analysis allows us to gain more insight on the problem and explore patterns of fires such as location and intensity. This project holds significant importance due to its potential to understand wildfire dynamics and eliminate such incidents. Identifying wildfire patterns can result in a more timely and efficient response preventing wildfires from becoming uncontrollable.

Some things we explored are regression model, finding patterns in humans or nature-based attributes, frequency of fires relative to state, what states have the most wildfires, what time periods had the most fires, and many more. The solution is a forecasting model that can predict and prevent a fire from occurring.

We began by defining the issue of increasing wildfires. Our problem was investigating what caused wildfires to increase. Then, we found a suitable dataset. Next, we explored the data to determine the most suitable analytical approach. Utilizing a modeling tool like Tableau, we visually identified any correlations and patterns within the data. Our Tableau exploration resulted in the creation of two interactive dashboards. One dashboard will show the relationship of fire sizes between the states and the time of year. The second dashboard will show the distinct fire causes along with containment and discovery times. Each dashboard contains 3 to 4 insightful visualizations. Our last step was to interpret the results of our data science process.

Given that our findings highlight human activities as a significant cause of wildfires, our society needs to raise awareness. Encouraging communities to embrace eco-conscious behaviors like proper disposal of cigarette butts, campfire safety, and using caution during outdoor activities, can mitigate the risk of human-caused fires. Moreover, collaborative efforts involving the US government, environmental organizations, and local communities are vital in preserving the environment. Potential outcomes of this project: - **Goals:**

- Use data analytics to uncover underling patterns from historic wildfire data:
  - Wildfire starting factors.
  - Analysis of natural wildfire behaviors
  - Analysis of artificial wildfire behaviors
  - Analysis of wildfire locations, in general and per state.
  - Analysis of discovery vs contain time of wildfires.
  - Uncovering underlying seasonal trends in wildfires.
  - Analysis of wildfire severity.
  - Uncover underlying patterns between fires and other attributes collected.

- Create Machine Learning models to:
  - Predict the location of the wildfire.
  - Evaluate wildfire risk factors and behavior.

By achieving these goals, the project outcomes will provide crucial insights and tools for relevant stakeholders. Firefighters can benefit from more accurate predictions, assisting in proactive planning and resource allocation based on risk level of each area. Policymakers gain essential data-driven insights to formulate more targeted preventive measures and regulations. Additionally, communities benefit from increased awareness, reducing the risks caused by wildfires. The significance of the project's outcome lies in its proactive approach, aiming to reduce the destructive consequences of wildfires and protect both human lives and ecosystems.

Original data description:

Provide details of procedures followed and methods used to capture the data. What is the file format of the dataset obtained from the source? Did you need to process the original data to get it into a more accessible, more compressed format? How large is the dataset (in numbers of rows and columns)? What are data types of presents (symbolic, numeric, etc.)?

The dataset was in a CSV file that we got from the government website which is data.gov. We did not process the original file while exploring the data at first. But we got rid of columns that we were not interested in exploring. The dataset is 40 columns and 2,303,566 rows of data. The data types that are present are strings, numbers, integers, and dates.

Data cleaning:

Are there attributes with missing values and blank fields? If so, is there a meaning behind such missing values? Are there spelling inconsistencies that may cause problems in later mergers or transformations? Are there deviations in data distribution that can be considered as noise? Have you explored deviations to determine whether they are worth analyzing further? Was a plausibility check conducted for data values? Were any data attributes being excluded that have no impact on your hypotheses?

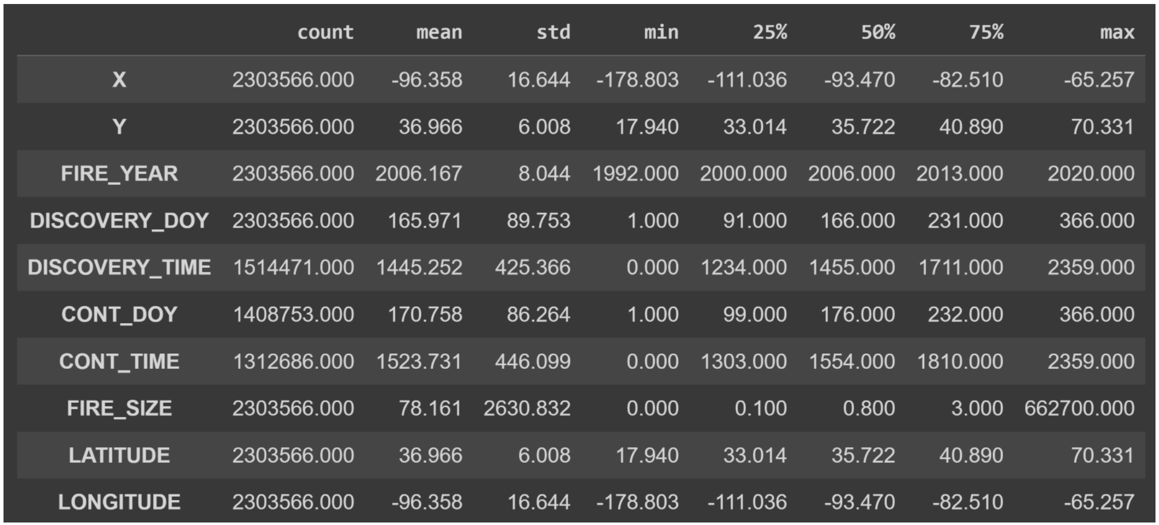
There are attributes with missing values and blank fields. The meaning behind it is that they did not track the fire for that year or the fire station. We found no spelling inconsistencies that may cause problems later. There weren't any big deviations in data distribution to be considered as noise when exploring our data. We made sure to check all the data attributes that we are going to use it plausible. We had some data attributes excluded from our hypotheses that had no impact, such as incident id numbers from a local station or source system type.

Descriptive analysis:

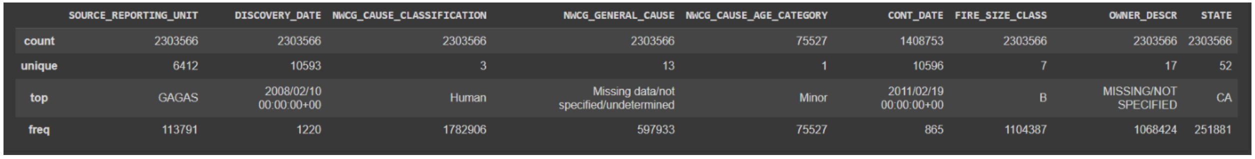
Describe basic descriptive statistics for the key attributes. What insight did this provide into the problem? Include appropriate visualizations to present the descriptive analytics results. Describe the prioritization of relevant attributes. If not, identify and describe relevant literature to provide further insight. What are the insights gained about crucial attributes from using exploratory visuals? Have data explorations revealed new characteristics of the data? Are any interesting interactions between data attributes? Were subsets of data identified for analysis?

Our key attributes are fire size, fire class size, longitude, latitude, fire discovery time, fire containment time, fire general cause and if the fire was done by a person.

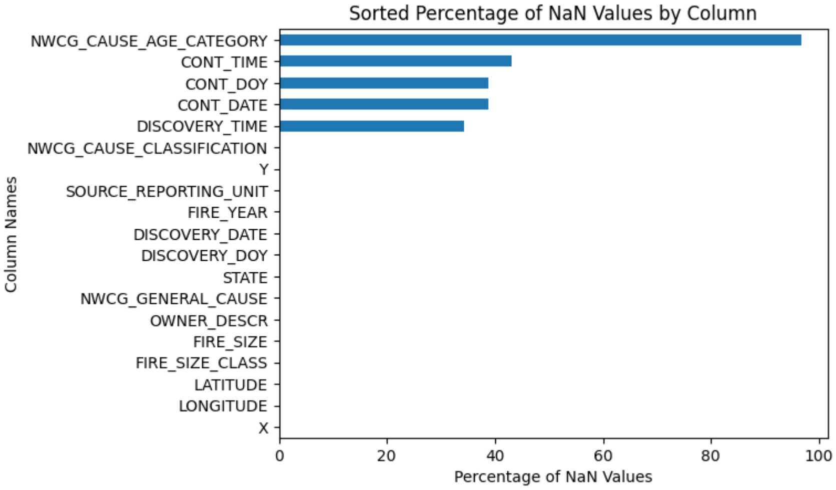
The range of Fire Size is between 0 and 6,627,000 acres (about the area of New Jersey). The average size of a wildfire is 78 acres (about the area of a large shopping mall). The most common fire size class is B which is 47% of the classes. The number of distinct values for the fire size class is 7 which means that the classes are from A to G. Another attribute we look at is states. The state that is in there the most is California, and it is 10% of that column. There are 52 unique values in the states column which means every state was affected at least once. The most common discovery date in our data was October 2nd, 2008. Most of the fires are discovered at three o'clock in the afternoon and 166 days (about 5 and a half months) into the year which would be around the middle of June and July. The data contains several missing values across the (NWCG_Cause_Age_category, CONT_TIME, CONT_DATE, CONT_DOY, and DISCOVERY_TIME) after further analysis of those attributes, a null value in NWCG_Cause_Age_category is not an indicator of missing data, it just means the fire was not caused by a minor.



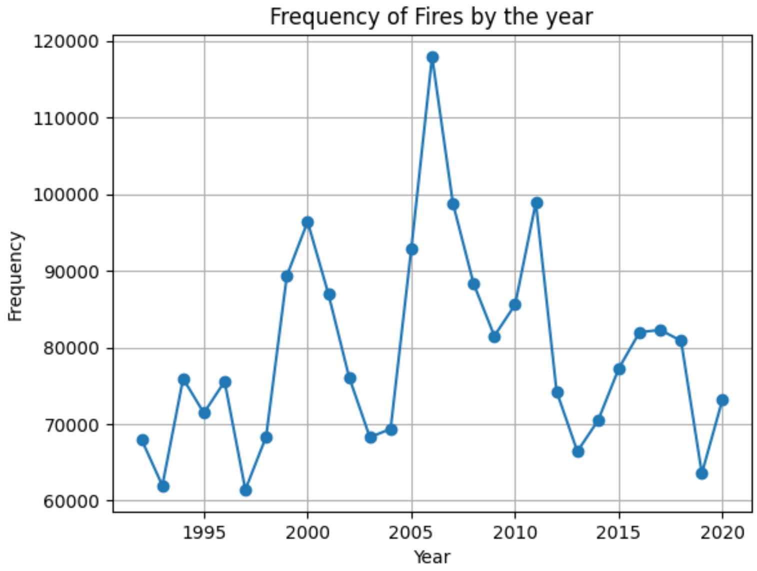
	count	mean	std	min	25%	50%	75%	max
X	2303566.000	-96.358	16.644	-178.803	-111.036	-93.470	-82.510	-65.257
Y	2303566.000	36.966	6.008	17.940	33.014	35.722	40.890	70.331
FIRE_YEAR	2303566.000	2006.167	8.044	1992.000	2000.000	2006.000	2013.000	2020.000
DISCOVERY_DOY	2303566.000	165.971	89.753	1.000	91.000	166.000	231.000	366.000
DISCOVERY_TIME	1514471.000	1445.252	425.366	0.000	1234.000	1455.000	1711.000	2359.000
CONT_DOY	1408753.000	170.758	86.264	1.000	99.000	176.000	232.000	366.000
CONT_TIME	1312686.000	1523.731	446.099	0.000	1303.000	1554.000	1810.000	2359.000
FIRE_SIZE	2303566.000	78.161	2630.832	0.000	0.100	0.800	3.000	662700.000
LATITUDE	2303566.000	36.966	6.008	17.940	33.014	35.722	40.890	70.331
LONGITUDE	2303566.000	-96.358	16.644	-178.803	-111.036	-93.470	-82.510	-65.257



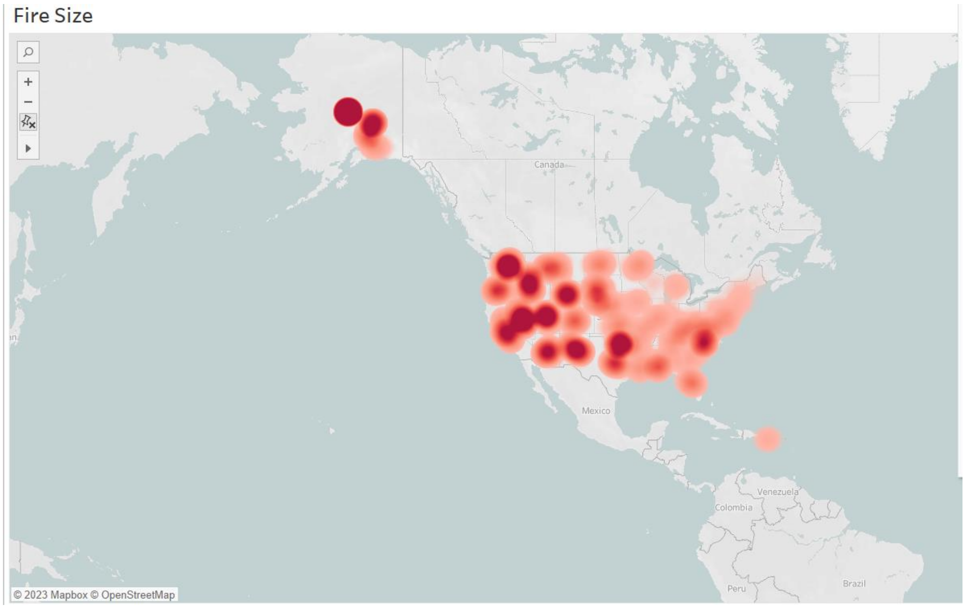
	SOURCE_REPORTING_UNIT	DISCOVERY_DATE	NWCG_CAUSE_CLASSIFICATION	NWCG_GENERAL_CAUSE	NWCG_CAUSE_AGE_CATEGORY	CONT_DATE	FIRE_SIZE_CLASS	OWNER_DESCR	STATE
count	2303566	2303566	2303566	2303566	75527	1408753	2303566	2303566	2303566
unique	6412	10593	3	13	1	10596	7	17	52
top	GAGAS	2008/02/10 00:00:00+00	Human	Missing data/not specified/undetermined	Minor	2011/02/19 00:00:00+00	B	MISSING/NOT SPECIFIED	CA
freq	113791	1220	1782906	597933	75527	865	1104387	1068424	251881



The insights this problem provided were that wildfires are at a constant rate. It increased from 1992 to 2006, then it declined from 2006 to 2019. After that, it started to rise from 2019 to 2020.

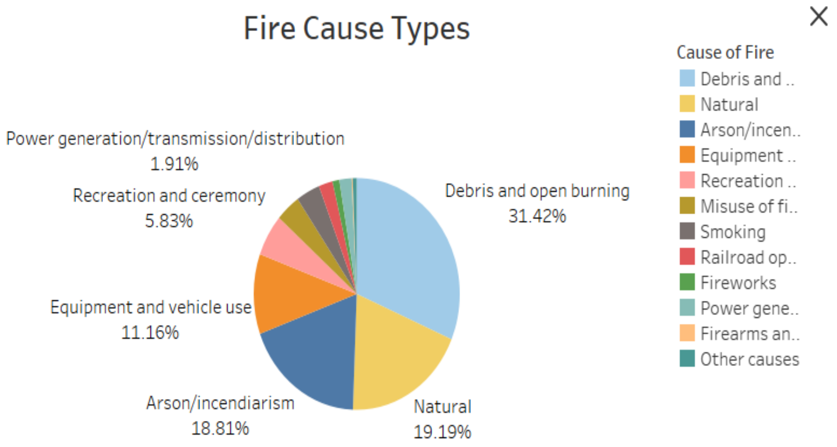


Frequency of Fires by the year

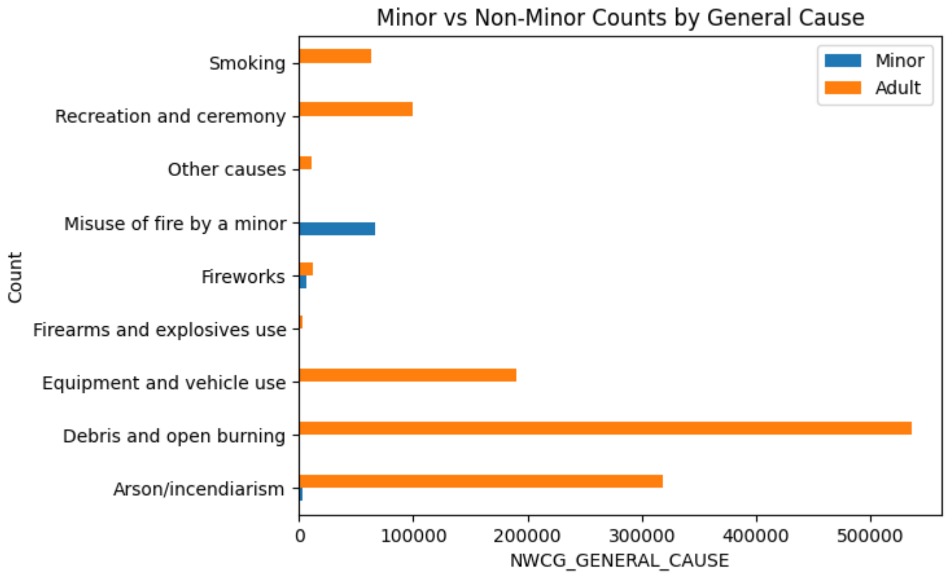


Fire Size

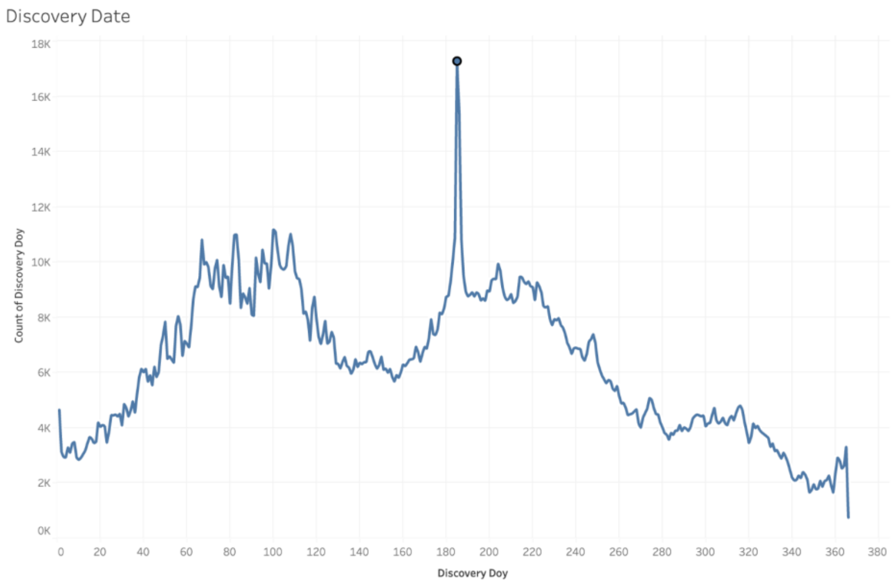
As illustrated in the above plot, we can see that Alaska and parts of the West Coast of the United States are experiencing the most significant wildfires in terms of size and severity.



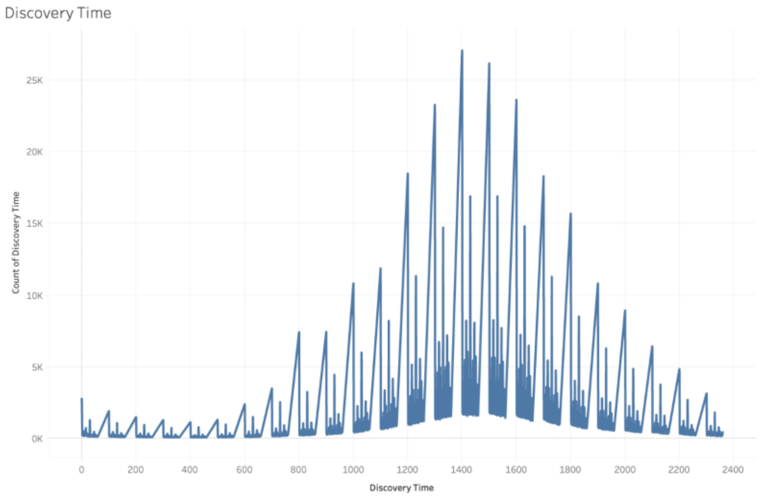
In this visualization, it is evident that the primary causes of fire are debris & open burning, natural, and arson. Human activities can contribute significantly to wildfires.



This visualization shows us that most wildfires are caused by adults, offering us insight into the responsible parties and the methods by which these fires occur.

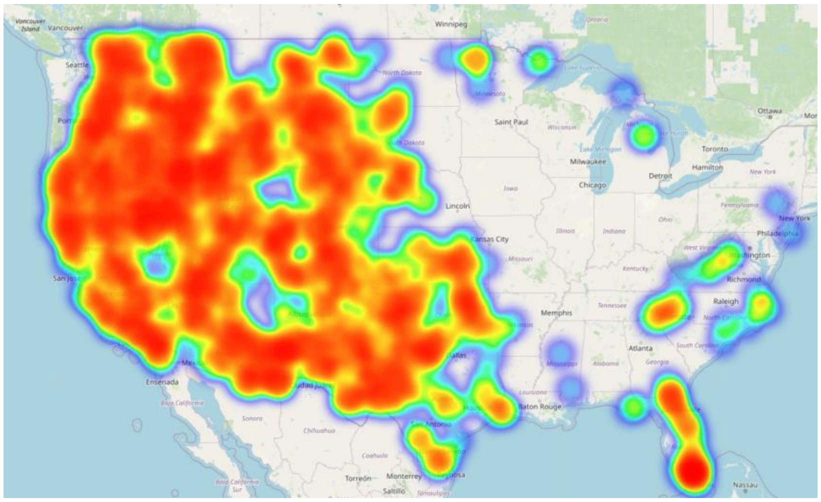


This insight shows us what time of year wildfires are most prevalent. As the visualization shows, most of the discoveries of wildfires happen in the summer months which are June and July.



The insight we get from this visualization is that wildfires tend to happen at 1400, which is 2:00 pm. In the afternoon is when the sun is at its peak time during the day.

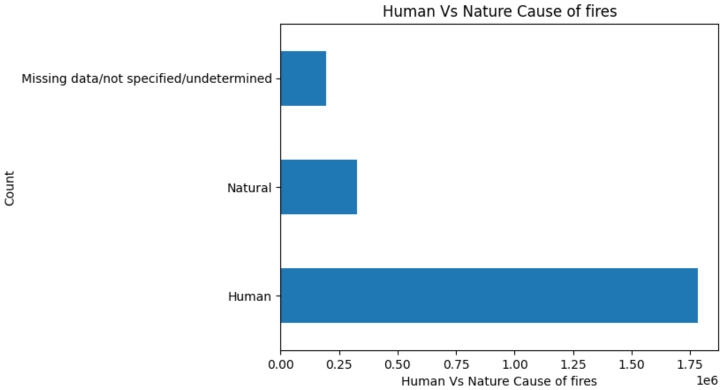
Using python folium, we created a heatmap of fires with size of more than 5000 Acres. The map shows that most of those fires are in the west states of the United States except for Florida where the biggest fires were inside the national parks.

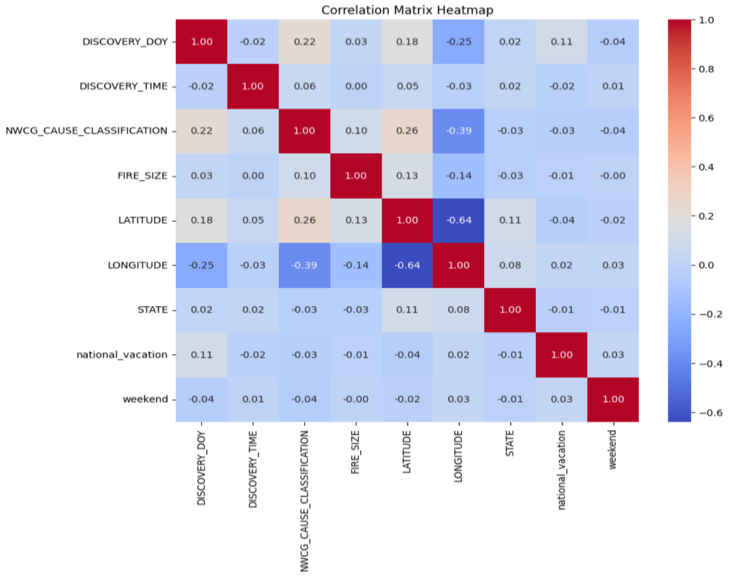


The most important relevant attributes are fire size, longitude, latitude, discovery data, discovery time, and whether human activities caused the fire. Analyzing these features enables the identification of patterns or correlations within the data. These features are the leading factors for a wildfire to occur based on our dataset.

The insights we gained from the use of exploratory graphics are we found different patterns and correlations. For example, heat maps enabled us to identify high-risk areas effectively. We used applications such as Tableau and Python to create these insights. In Tableau, we created visualizations to look at how fire size was affected over the years and examine the variations across different states. Additionally, our analysis identified the major factors influencing wildfires, along with distinguishing between the sources of fire origin, categorizing them as human-induced or natural occurrences.

During our data exploration, we differentiated between fires caused by minors and adults. We also explored whether fires occurred more frequently on holidays or weekends. Using a correlation heat map, we examined potential relationships between variables where we found an interesting correlation between the cause of the fire and the day of year the fire was discovered. Our exploration also aimed to determine if natural factors outweighed human-caused wildfires. Surprisingly, we discovered that humans were responsible for initiating more wildfires than nature. Furthermore, we analyzed the mean and median of fire size and time to extinguish. An interesting discovery appeared: There is a meaningful relationship between the duration required to extinguish a fire, and the subsequent rise in the average fire size.





We are focusing on wildfires that are uncontrollable and cause the most significant losses in human life and in costs. We used NWCG classification to use a subset of the data. NWCG classification for the fires are as follows:

- **A:** 0 < Fire < ¼ Acres
- **B:** ¼ < Fire < 10 Acres
- **C:** 10 < Fire < 100 Acres
- **D:** 100 < Fire < 300 Acres
- **E:** 300 < Fire < 1000 Acres
- **F:** 1000 < Fire < 5000 Acres
- **G:** 5000 < Fire < max Acres

We determined that a fire of class C or higher is the fire that causes the most damage, hard to control, and is the type of fire this project should focus on.

Data preprocessing:

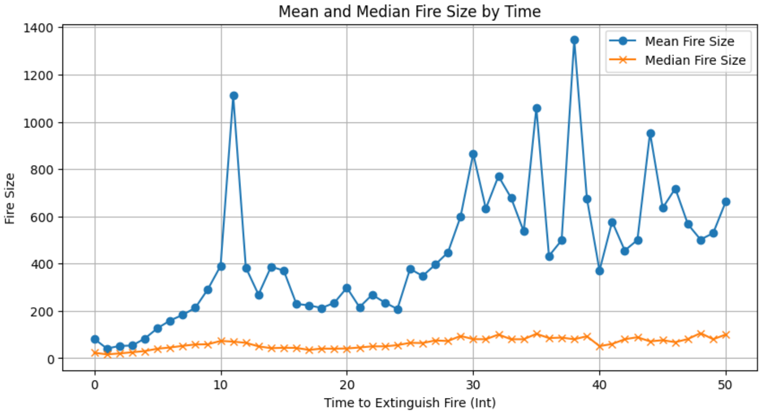
Have you simplified the data set? Removed some rows or columns of data? Any feature selection done? Did you enrich your dataset with other columns or more information? Imputing/removing missing values? Simplification of values? Normalization? Remember that you should describe all procedures performed on your raw dataset. Are multiple data sets appropriately integrated? Were there any merging problems that should be documented? Have you researched the modeling tools and data mining algorithms you plan to use? Are there any data formatting issues you need to address before modeling?

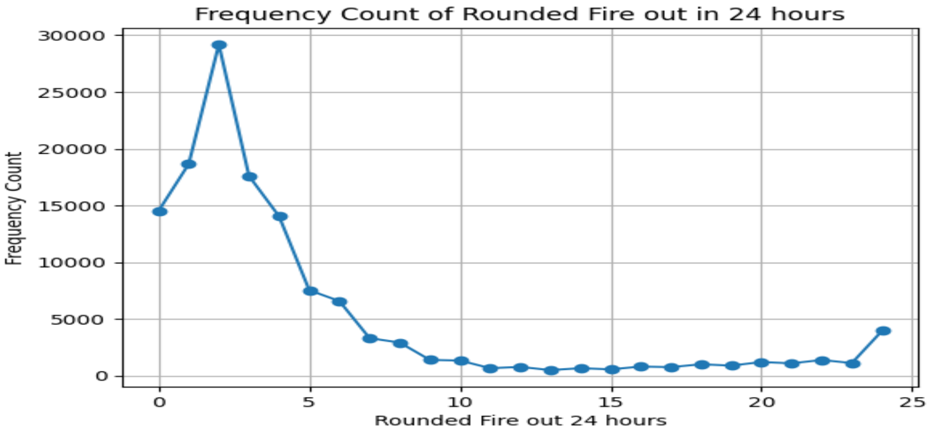
We have simplified the data set by removing several columns that were duplicates and the id columns. As the dataset is a merge between several wildfire systems it contained several id columns and duplicate columns which were removed. Also, the rows that did not contain discovery date and containment date were removed as we theorized that they would be especially important attributes later when making the machine learning models.

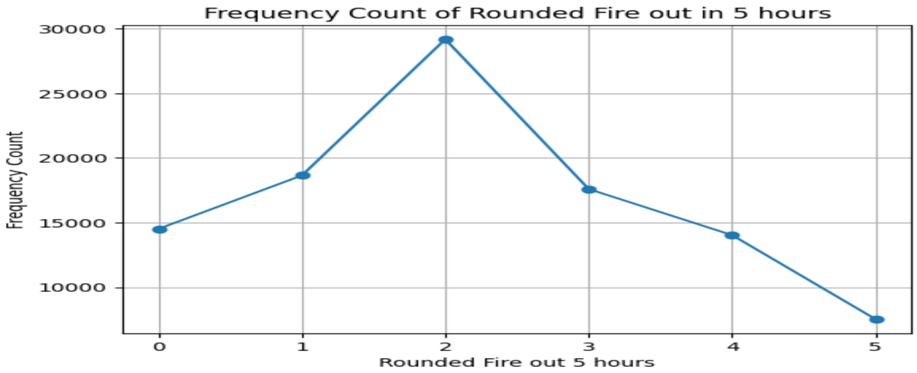
The features selected that were used for processing are the states, Lat, Long, Year, Discovery Day of the year, Discovery time, Containment Day of the year, Containment time, fire size, cause of fire, is the fire done by a minor or not. We also added if it is a vacation, and if it is a weekend.

We removed null rows from our raw dataset. The rows that were removed were rows that contained null values in the discovery date & time, and containment date & time.

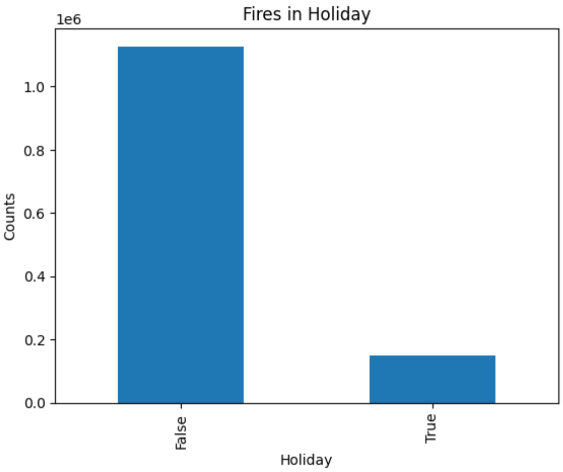
Based on the attributes left we created a time to containment column which calculates the time it took to contain the fire. As seen in the diagrams below for fires of size 10 Acres or more the time it takes from discovering the fire to containing it is usually around 2 hours. If the time taken to extinguish is more than 2 hours, the fire size might not increase but the risk of a single fire going completely out of control increases greatly.

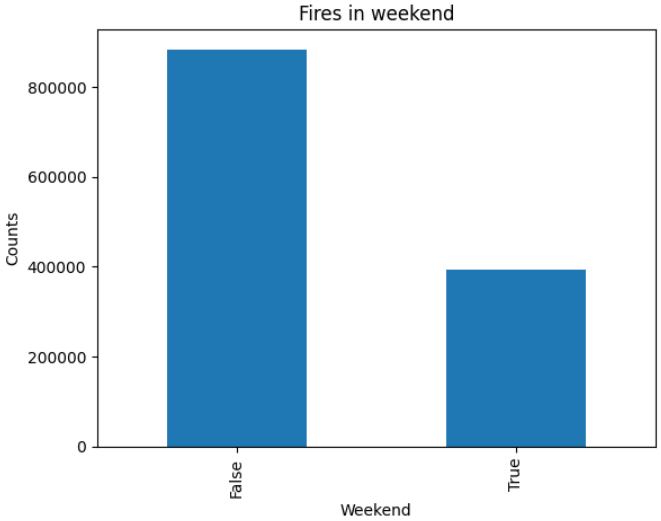






We also created national holiday and weekend columns where we determine the day of the fire was either on a holiday or 2 days before and after a national holiday and the results show that the number of wildfires on national holidays amounts to almost 15% of all fires and 1/3 of all fires are on weekends.





We are using one dataset that contains 2,303,566 rows of data. So, we did not have any problems merging our dataset.

We have researched the requirements for Tableau and data mining algorithms such as a regression algorithm. For Tableau, you must ensure it is formatted appropriately so you can show it the correct way in your visualizations. Like, for example, you need to make sure that the variable you are using the correct type.

We changed the discovery date format to meet our requirements. It was changed to just show the month, day, and year without the time stamp next to it. Also, we used a label encoder on the cause of the fire and the states to feed it forward to the machine learning model.

Data science process:

Describe the data science process adopted to implement your analytics approach. Provide an overview of data mining and machine algorithms employed as part of your analytics approach. Explain how the selection of mining and machine learning algorithms helps address the problem and associated issues. Discuss approaches and algorithms used by others to solve the problem. Discuss what aspects of your approaches are different from others.

The data science process we adopted to implement our analytics approach is OSMEN. This process starts off with obtaining the right data for our topic. The next step is scrub which is looking at the quality of our data and converting the raw data into usable data. We are making sure that we are removing duplicate data and missing data. The third step is explored which means exploring our dataset. At this stage, you start looking at descriptive statistics and we use both Tableau and Python to get the stats. The fourth step is to model, and this is where you figure out what is the best algorithm for your data. We did this by selecting the features that we wanted for our model and creating training data with it. The last step is to interpret, which means you look at the results from the model. This is also where we share our results with people who are interested in the topic.

For our analysis, we used data mining techniques such as regression analysis, K-means clustering, and time series analysis. The time series analysis we used allowed us to examine the progression of fire sizes over the years, visualized through a heat map of the US. Tableau was the modeling tool used to present this visualization. We use Python notebooks to perform regression analysis and k-means clustering. K-means clustering is examining how we can cluster the fire sizes based on longitude and latitude.

The selection of mining and machine learning algorithms helps us address the problem by giving us evidence of what is causing the problem. For the prediction of latitude and longitude, we will employ a Dense neural network using Pytorch. The model criterion will be Mean Squared Error (MSE) as the Lat and Long are presented in a float form. The optimizer we will use is Adam optimizer with a learning rate of 0.001. We will perform the regression using Python, enabling us to predict the areas in the US most likely to experience high fire rates, considering both location (state-wise) and fire frequency.

We found two papers that were trying to solve our problem of wildfires. The first paper was about wildfire monitoring in Canadian forests. They used remote sensing to create their dataset and machine learning to predict wildfires. They used a big data platform called Databricks to run their machine learning algorithms. They used two supervised learning techniques, neural networks, and

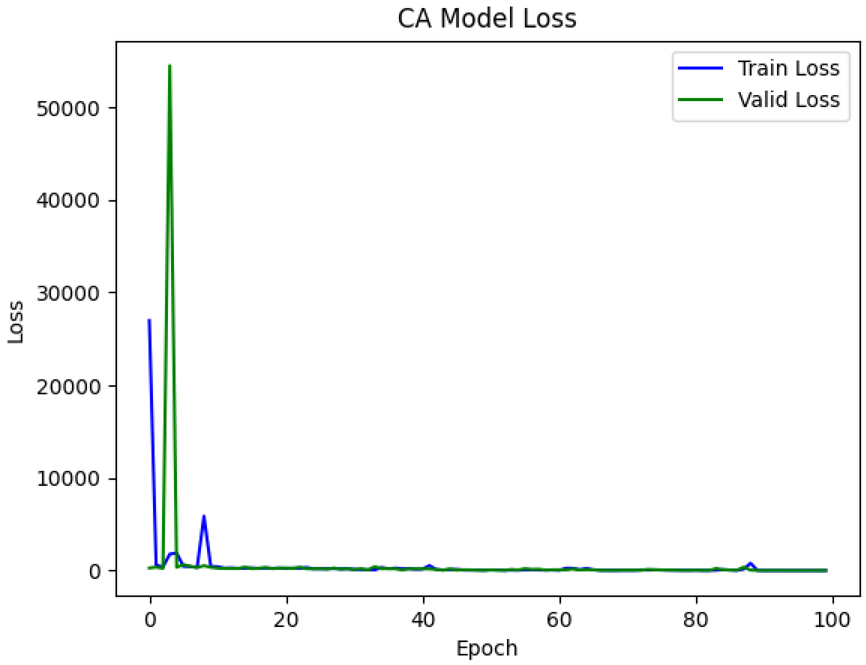
SVM, to predict wildfires. The simulation accuracy of the neural networks was 98% and the SVM was 97%. They used two cross-validation techniques which are k-fold and shuffle split to evaluate and validate the machine learning techniques. The second paper was about making a wildfire susceptibility prediction model in Australia. They used Explainable artificial intelligence (XAI) to make a deep learning model to look at the wildfire. They feed the model several contributing factors that were part of three distinct categories (meteorological, topographic, and vegetation). Used an explanation model called SHAP, which analyzes each of the feature's importance and chooses the best one.

This is different from our approach because we are using regression analysis, k-means clustering, and visualizations to solve the problem. Regression analysis is used for predicting and forecasting the data values and to predict the potential latitude and longitude of the fire. We use k-means clustering and visualizations to look at the patterns within our data. These aspects are different because we are using one supervised and one unsupervised learning algorithm to look at the data. We are also using different software from the other papers to run our machine learning algorithms.

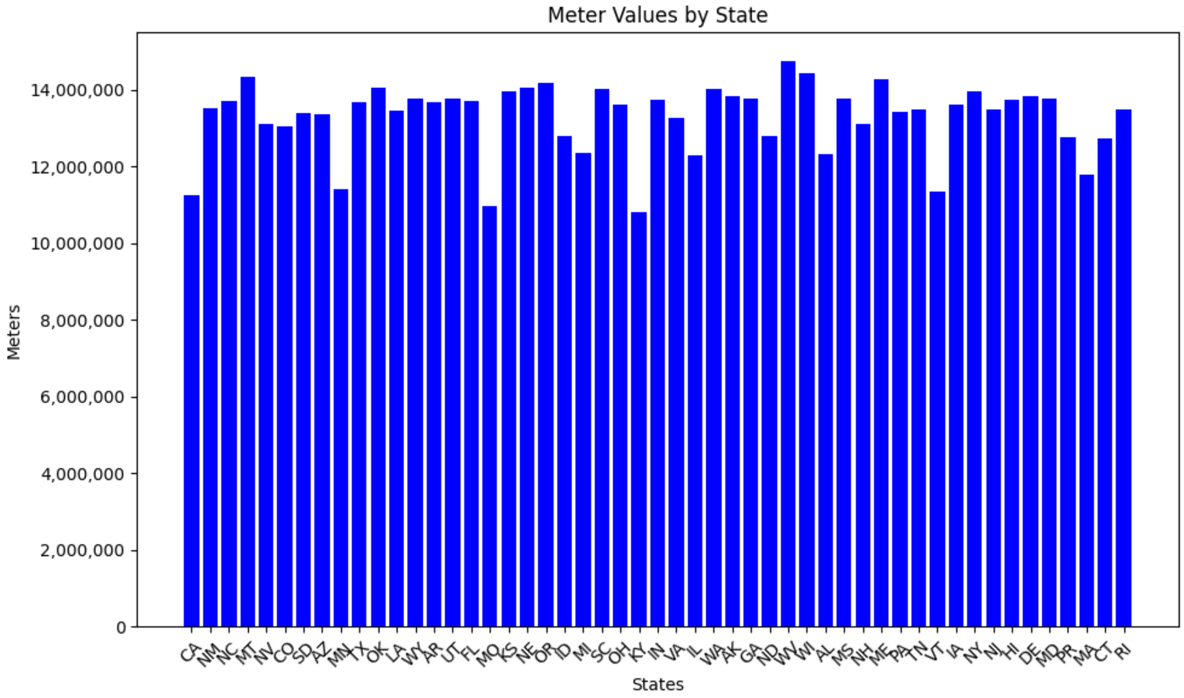
Mining and Machine Learning Model Results:

What data was used to test the models? How was the data partitioned into train/test sets? How long does it take to train the model? How long does it take to generate predictions using the model? Were there any execution problems? Describe how you adjusted the parameters for the algorithm. Were there any calculation inconsistencies that should be noted? How did you measure the success of the models? How many times did you rerun the data analytic technique with adjusted settings before attempting another type of model? Describe the charts and other visualizations created to compare and evaluate the accuracy of the results. How did you evaluate the models? Report the performance of the model on the validation dataset.

The data we used in the prediction of the models are data that contains fire type of G, since we are concerned with uncontrollable wildfires that go out of control. Since this is a time series data, we used wildfires of all the years till 2019 as the training dataset, and wildfires after 2019 as the testing dataset. After training a singular model to predict fires in the entire USA, the model score was unbelievably bad. To improve on that score a model was developed for each state which increased the prediction score by a big margin. The total time to create a model for each state was exceptionally long. Even though we are using a simple ANN, the size of the data is too large, and each model also had 100 epochs to train. For all models, it took more than 8 hours to train. With the exceptionally long training time also came the resource limitation, since the work was done on Google Colab the code was usually interrupted due to being idle or reaching resource limits. This limitation has not allowed us to experiment much with the architecture of the ANN. Adam optimizer was used with a learning rate of 0.001, the criterion was Mean Squared Error (MSE) since our prediction is latitude and longitude which is a continuous number. The loss was plotted for each of the state models, both the training and validation loss. Also, the distance between the actual location and the prediction was calculated using a written function to get a sense of how far off the prediction was. The Mean Squared Error was the metric used to evaluate our model's performance. All models exhibited an MSE of 20-30 at the 100th epoch, which is not a satisfactory result given how a single digit change in either latitude or longitude equated to Kilometers on the earth's scale.



California model training and loss scores, there are other plots for each of the states in the file.



Each State model and the average errors in meters per epoch

What were the results of the data analytic technique in relevance to the project goals? Do the results make sense to you from a purely logical perspective? Are there apparent inconsistencies that need further exploration? Do the results address the problem? Was more than one technique explored? If so, provide a discussion comparing the results? Describe how the results of the data analytic techniques will be deployed. In case multiple data analytic techniques were used, rank order the techniques and findings in order of their applicability to the project goals and problem contexts? What meaningful conclusions were drawn from the selected data analytic technique? Are there new insights or unusual patterns revealed by the application of the technique?

The data analytics techniques helped us uncover underlying information for the cause of wildfires. Starting by analyzing the states with the most wildfires, and the states with the largest wildfires. Also, by using an interactive heatmap and selecting fires of class C or higher, we understood the areas with extreme risk of uncontrollable wildfires. We understood using frequency analysis that humans are the cause of most wildfires and analyzed the reasons, and seasonal information as to when the fire started. From that information we hypothesized a substantial portion of fires happen during weekend and holidays which turned out to be true. By using that information, and how it increases the risk, we added it to our data as it might be able to determine if a fire is going to happen and where it will happen. We also used multivariate analysis where we calculated the fire size given the time it took to take out the fire from the discovery time, which confirmed our hypothesis that the longer it takes to put out a fire the bigger the fire is. But we were surprised by the huge gap between the mean and median of the fire sizes, which means usually when duration to put out a fire increases, it is rare for its size to go out of control, and it very slowly increases in size, but once it goes out of control fire size increases exponentially.

Where possible, results should be summarized using clearly labeled tables or figures and supplemented with written explanations of the significance of the results

Conclusion: Provide concluding remarks for the project and outline potential future work directions. Are the things tried to solve the problem reasonable? Are the employed algorithms or applications developed to solve the problem and produce interesting results? What novel insights about the problem, mining algorithms, and data sciences do your project convey? What are key challenges you faced while working with the data, and how did you address them? Describe what you did to evaluate the validity of your data analytics solution. Discuss any important limitations or caveats to your results and solutions with respect to the problem.

In conclusion, this project marks a significant step toward understanding the complex issue of wildfires through data analytics and machine learning. The efforts to recognize patterns in wildfire behaviors, predicting their locations, and evaluate risk factors have produced valuable insights. In the future, we wish to look at more factors that can affect wildfires, such as temperature, drought level, and air quality.

Looking ahead, potential future directions could be refining the machine learning model for more accurate predictions, as well as integrating real-time data sources to enhance the response time. Additionally, exploring the impact of wildfires in our society and creating community- specific preventive measures could be crucial for mitigating the issue.

The novel insights our project conveys about the problem, mining algorithms, and data sciences is that this idea can help with more information and time. A lot of people lose things during wildfires, and that can happen at any time. It can be caused by most humans and natural causes.

One key challenge we faced while working with the data was making sure that it was set for our dashboards and models. For example, one problem we faced making sure the date in the data was correctly formatted so that it just should be the month, day, and year without the time in the same column. We addressed this by splitting the column between the year and time. Another key challenge we faced was figuring out what columns to drop and which contained multiple rows of null values.

To evaluate the validity of our solution, we utilized the Mean Squared Error (MSE) as our primary metric to analyze the performance of our models. This allowed us to quantify the difference between predicted and actual latitude and longitude, which is vital for the accuracy of our predictions.

Our results have several limitations. Resource limitations and time constraints prevented us from exploring more complex parameters, potentially decreasing the model's predictive capabilities. Moreover, focusing on wildfires of a specific type might limit the model's adaptability to different wildfire variations in environmental conditions.

Despite current limitations, this work lays a foundational framework for future advancements in wildfire prevention and prediction, crucial for reducing the destructive consequences of these natural disasters.

References: Cite all the references used in the report. You can cite using ACM or APA styles.

Sayad, Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal*, 104, 130–146. https://doi.org/10.1016/j.firesaf.2019.01.006

Abdollahi, & Pradhan, B. (2023). Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. *The Science of the Total Environment*, 879, 163004–163004. https://doi.org/10.1016/j.scitotenv.2023.163004

Publisher U.S. Forest Service. (2023, September 29). *National Interagency Fire Occurrence Sixth Edition 1992-2020* (feature layer). Data.gov. https://catalog.data.gov/dataset/national-interagency-fire-occurrence-sixth-edition-1992-2020-feature-layer

Appendix: Give an appendix that contains fully documented code segments, programs, and scripts used as part of your analysis, actual screenshots of the results, the descriptions, and references to any third-party tools used in the analysis.

Appendix A

Tableau used as a tool for visualization and dashboards

Appendix B

Python Notebook for models