2023 Data Science Competition Report

Team : Aggieland DataDunkers

Members : Ram Pangaluri, Shane Olson, Vardaan Kola, and Josiah Kuncheriah

Executive Summary

The primary goal of our team was to find trends in number of fires, acres burned, and fire location. Another main goal of our project was to see if there were any relationships between demographic features and wildfire risk. We limited our focus to the United States and made special focus on Texas. Most of the time and energy of this project was dedicated to finding good data and changing it into a usable form. This data was then transferred to a data visualization program to be analyzed and presented in a series of interactive dashboards. From these dashboards our team was able to learn new insights and we can now recommend new strategies for fire departments and the general public to better prepare and fight wildfires and even predict future fire hotspots.

Problem Statement

Our team chose to focus on three aspects related to wildfires. These pillars are Area damaged by Wildfires, People Affected by wildfires, and Economic Costs of wildfires. We also chose to limit the location of analysis to the United States of America. We tried to be as detailed as possible given the data at hand. For example, we had more detailed information on Texas so we had a dashboard that solely focused on Texas analysis.

Datasets

All Datasets used can be found in the github provided in this report. All sources of this data can be found in the references.txt file that can also be found in the github provided. While all sources listed in references.txt were not used in creating datasets or dashboards, the team nevertheless found them important in understanding the problem area and scope.

Competition Datasets

In terms of the competition datasets our team heavily relied on data from the National Interagency Fire Center (NIFC) and the Fire Information for Resource Management Systems(FIRMS). These data sources were very important. The NIFC was useful for getting data about the number of fires and acres burned in the United States. The FIRMS source was good for getting satellite images, as well as detailed fire location information with longitude and latitude. This allowed the team to accurately map fire histories across different states and time periods. Other competition datasets were explored but were found to be hard to access data from by our team, so we chose to ignore them given the time frame of our project and schoolwork and other activities.

Additional Data

Numerous additional data sources were used to supplement the competition datasets. This was especially true in order to fulfill our team's goal to estimate economic costs of wildfires and wildfire impact on different demographic groups. For example, to get data for our vulnerable population dashboard we leveraged data from the US census bureau, American community survey, and headwaters economics. For economics data we procured this from headwaters economics and value penguin, an insurance company that insures against wildfires. Another major dataset was a kaggle dataset that contained United States wildfire records for the past 20 years.

Preprocessing

For the most part, the datasets were pretty complete. Some cleaning was required especially with filling in missing data. For example some states/counties were missing data, this was filled in by consulting data from other resources to complete empty cells. Other times the team chose to leave values empty to show that the data was missing. Outliers were not treated, while outliers could affect a model performance, outliers are not so problematic in a visualization sense so the team felt we could leave outliers in the datasets. Sometimes we had to add new information like FIPS codes to outline county lines which adds additional granularity and detail to our data..

Data Exploration

Initial data analysis included finding excel and csv files to use. This was done by consulting the competition data sources as well as by independent research by the team. Based on the data sources gathered our team narrowed down our problem formulation to focus on the three aspects discussed before. Our team also decided that our solution would consist mainly of data visualization and analysis with a minor focus on future prediction. The team was also able to get many quick ideas by doing exploratory data visualization using tableau.

Methodology

Our team used a myriad of data analysis methods to reach our solution. Some methods were qualitative while others were quantitative in nature. Methods used include cluster analysis, regression analysis, time series analysis and data mining.

Cluster analysis was used primarily to group states, counties, and locations by a certain metric like wildfire risk, exposed housing units, etc. We found this analysis to be useful to provide overall groups to compare against each other. For example it was interesting to see the differences between different US regions (West, South, East, Midwest).

Regression analysis was used to predict and forecast future number of fires, and future number of acres burned for the United States. This was done by making a linear regression model. We found an increasing number of acres being burned with this model, even though the number of fires wasn’t increasing which was a definite surprise.

Time series analysis was used to track wildfire history and metrics over time. The team used Gradations of time that were as big as years to as small as month and day. Generally our datasets already had a good enough spread of data to cover multiple time points, so this analysis was relatively simple to complete.

Data mining was also used especially in our bigger datasets. This consisted of sorting datasets by different factors to see relationships between variables/columns that were not readily apparent at first glance. As an extension to this our dashboards and visualizations also have the ability to be sorted by different metrics. We also made our dashboards as interactive as possible so users can filter data to make it specific for their own purpose.

Modeling and Analysis

While the main bulk of our work did not focus on quantitative models and predictions, we did make a linear regression model and a wildfire risk prediction model based on pattern recognition over ten years of data. These models will be explained in more detail now.

Linear regression model

This model can be seen here:
https://public.tableau.com/app/profile/ram.pangaluri/viz/FireAcrePrediction/FireAcreForecasts

We can draw some conclusions from this visualization and model. First we can see that time is not related to the number of fires. We see the r-squared value for the fire actual regression line is a very small 9.8 * 10^-5 also the p value is 0.952295. These values indicate that the time/year doesn't explain the variation of number of fires in the USA and that time effectively has no effect on the number of fires (large p value). However on the Acres actual regression line we can see a clear increasing trend over time. As a result the r squared value for this line was 0.469 and the p value was less than 0.0001. These values show that almost half of the variation in acres burned can be explained by the time/year, and that time had a strong effect on the number of acres burned(very small p value). From this model we can conclude that while the number of fires is plateauing the number of acres burned will increase. This is an indication of the intensity and strength of modern fires as they are burning more acres than previous fires. Our fire actual linear regression model predicts that the number of fires for 2023-2028 will fall in the range of 20,000 to 100,000 fires per year. The acres actual linear regression model predicts that the number of fires for 2023-2028 will fall in the range of 4 million to 13 million acres per year.

Wildfire Risk Prediction Pattern Recognition

This model/animation can be seen here:
https://public.tableau.com/app/profile/ram.pangaluri/viz/FireAcreGridAnimation/Bubble

We can draw conclusions from this visualization and model. The first is that there are 4 clear zones that are made that divide states based on average number of fires and average number of acres burned (the purple vertical and horizontal lines). These zones can be used to predict potential wildfire risk and damage. Moreover if states can proactively use such a tool then they would be able to request and prepare firefighting resources to fight fires during the fire season. The four zones are summarized below in a picture.

We see that most states tend to be in the green zone, meaning that they have less than the average number of fires and less than the average number of acres burnt in their state. This green zone can be considered a low risk area and less resources are needed for states in these areas. A step above this are the yellow zones where either fires or acres are above average. These zones represent medium risk and these states should be prepared for firefighting activities. Finally we have the red zone. States in this zone are at high risk and probably actively fighting fires. If a state falls in this zone, it is advisable for this state to declare an emergency and sanction help from other state fire departments preferably the green zone states. Active monitoring of this grid could help state fire departments better understand fire risks for their states and be able to send firefighting aid to states at a higher risk level.

Limitations

Some limitations for these models is the lack of time data. The team would have liked to include more years but there was not enough time to search for data from previous years. Obviously adding more data would have made the models even more accurate and useful in predictive scenarios.

Areas for Further Analysis

Some areas for further analysis in terms of predictive modeling that were identified by our team include predictions for monetary loss by wildfires, prediction for casualties and fatalities by wildfires, and further analysis for firefighting efficiency. Having models that predict monetary loss and casualties and fatalities caused by wildfires could help various organizations from state governments to insurance companies. Analyzing which states or fire departments are the most efficient at limiting fire damage(this could be done by tracking metrics like average acres burned per fire), could help other fire departments learn methods and techniques from these high-performing departments leading to better fire damage reduction capability across the board.

Visualization and Interpretation

Our team's main focus for this project was in making visualizations and interpreting them. As a result this section will be lengthy, as we have many visualizations to talk about. With this point in mind we will skip over talking about the model visualizations already discussed in the modeling and analysis sections. While there is some information overlap in some visualizations we believe that the visualizations we made

are all sufficiently unique and tackle our three main goals of determining area damaged, people affected, and economic cost.

https://public.tableau.com/app/profile/ram.pangaluri/viz/FireStateRaceAnimation/USFireRace

This visualization is an animation that shows how different states rank to each other in terms of number of fires from 2010-2021. States are colored based on the region they belong to, so it is easy to see which regions experience a higher number of fires. From this visualization it's clear to see that states belonging to the South and West are responsible for a majority of the fires in this time period. Our team included it in the project because it is a fun animation but useful as well.

https://public.tableau.com/app/profile/ram.pangaluri/viz/FireEconomics/FireEconDash

This visualization is very important to our project because it tackles analysis of the economic costs of wildfires which was one of the three main pillars of the team problem statement. The visualization has a map that contains a percentage of housing units with risk to be exposed to a fire as a percentage at the county level. It also includes a treemap with breakdown of housing units by state, and another treemap showing value at risk(total monetary loss possible) by state. Finally the visualization includes a dual axis line and barchart at the bottom. The bars represent firefighting costs per fire, while the line represents total firefighting costs for a particular year. The visualization also features a filter at the top to focus on a particular state in detail. From the visualization its clear to see that housing units are at threat pretty much everywhere, the safest spots seem to be in the midwest however. As for top states with exposed housing units Texas, Pennsylvania, Florida, North Carolina and California round out the top 5, with New York also in the mix. It is interesting to see that many residential areas are at high fire risk, possibly because of the close packing of fire starters and general crowding of hazards and people. For monetary costs California by far leads all of the states accounting for 58.81% of possible monetary loss due to fires at over 343 billion dollars estimated. This is probably due to California's high property values and the existence of many wealthy companies in the state. We can also see that in general total firefighting costs and costs per fire have been going up each year based on data from 1985-2020.

https://public.tableau.com/app/profile/ram.pangaluri/viz/VulnerablePopulations_16799314425620/Dashboard1

This visualization is also very important to our project because it answers the question of who are the people impacted by wildfires, which is one of the three pillars of our problem statement. The visualization allows users to select a state, select values for 8 demographic filters, and see the effect of their selections on the map and counties below. The visualization also includes a demographic summary below the map summarizing population numbers for different minority and marginalized groups. As with other natural disasters it is often these groups who struggle the most in recovering from wildfire disasters.The map is stratified to the county level, and wildfire risk of a county is calculated based on that county's performance to its own states counties. This is more accurate than comparing to some sort of national average because the United States has a varied climate across state borders. Hovering over counties gives information on county population, name, and vulnerable populations. We believe that if state departments can use this information they can better help marginalized communities.

https://public.tableau.com/app/profile/ram.pangaluri/viz/Fire_16776980457310/Summary

https://public.tableau.com/app/profile/ram.pangaluri/viz/doi/Dashboard1

https://public.tableau.com/app/profile/ram.pangaluri/viz/TexasCountyWildfireFocus/Dashboard1

All of these visualizations are similar in that they both answer the team problem statement of Area/Location damaged by wildfires. All show wildfire locations, causes, size, and year. All dashboards are also filterable by this information and allow users to click on bars and labels to explore further. The main difference is that the first shows a national fire trend while the second focuses on the state and county level. The second also has more time granularity by allowing users to filter by month and day and not just year. Meanwhile the third focuses solely on Texas fire trends and features an animation. From these visualizations we can see that most fires are very small (less than a quarter of an acre), caused by lightning, and most occur in the west. We also see that fire frequency is high during the summer months(June, July, August) nationally, although this is not always the case at the state level. Another interesting finding is that some particular holidays like July 4th and New Years day saw an uptick in the number of fires, due to the nature of celebrating with fireworks.

https://public.tableau.com/app/profile/ram.pangaluri/viz/WildfiresinUSA2000-2018_16791902070870/NamedWildfiresUSA2000-2018

This visualization also helped answer the question of where fires occur, but with a special emphasis on acres burned and larger named fires (some fires are so small that they fizzle out by themselves thus requiring no name by the fire department to identify them). Filters for year, and specifying an acres burned range are available.

Conclusions and Recommendations

Overall our team enjoyed taking part in this opportunity. It allowed us to strengthen our interdisciplinary research skills, as well as storytelling, visualization, graphics, and data mining skills. We also believe that our work does have a lot of potential and focuses on aspects that other research teams could have overlooked. Research is a vast field but unfortunately its impact on public discourse and societal challenges can fly under the radar. Our team recognizes the impact of research and the products of research. We recommend that organizations show research or products of research competition(like this one) to the public at large. Utilizing social media is an effective way to drive enthusiasm and curiosity towards research products and goals while also tackling societal challenges. To that end all our dashboards will be made public and the shareable links to these dashboards are compiled below in the Supplementary materials section. We also recommend contacting various parties that could be helped by our research (in our case this could be fire departments, nonprofits for marginalized groups, insurance companies). Often use cases of research products can be applied to industries not necessarily thought of as intuitive as well. For example Henry Ford got the idea for car assembly production lines, from watching a slaughterhouse packing process. Ford took the same principles used in the slaughterhouse of division of labor and repeatability to produce cheaper cars. We see that the same can be said of research. It often takes an outside force to find a use case for the research. This outside force can be an individual like Ford, or an organization. To that end, our team would like to thank the organizers and sponsors of this event, as well as the faculty available to guide and advise our project development. Thank you.

Executable code

Tableau Workbooks can be downloaded and viewed along with their respective datasets from this github.

There are also html reports included in this github that can be downloaded and viewed in chrome.

Some files were too big to upload to github, so the team did their best to publish them and make them publicly available for viewing and downloading.

https://github.com/rampang/AggielandDataDunkers2023

Supplementary materials

All Dashboards and Animations

https://public.tableau.com/views/VulnerablePopulations_16799314425620/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link - wildfire impact on vulnerable populations

https://public.tableau.com/views/WildfiresinUSA2000-2018_16791902070870/NamedWildfiresUSA2000-2018?:language=en-US&:display_count=n&:origin=viz_share_link - named wildfires in USA

https://public.tableau.com/views/TexasCountyWildfireFocus/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link - texas county wildfire focus

https://public.tableau.com/views/doi/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link - fire year, class, cause, by month and day and county map and state filter

https://public.tableau.com/views/Fire_16776980457310/Summary?:language=en-US&:display_count=n&:origin=viz_share_link - fire cause, size, name and national density map 2002-2012

https://public.tableau.com/views/FireEconomics/FireEconDash?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link - wildifre economic impact and housing under threat

https://public.tableau.com/app/profile/ram.pangaluri/viz/FireAcrePrediction/FireAcreForecasts?publish=yes - fire and acres regression prediction

https://public.tableau.com/app/profile/ram.pangaluri/viz/FireStateRaceAnimation/USFireRace?publish=yes - fire state bar race animation

https://public.tableau.com/app/profile/ram.pangaluri/viz/FireAcreGridAnimation/Bubble?publish=yes - fire acre grid animation

References

These references can also be found in the references.txt file in the github.

https://www.nifc.gov/fire-information/statistics/wildfires

https://www.iii.org/table-archive/23284

https://www.worldatlas.com/articles/the-officially-recognized-four-regions-and-nine-divisions-of-the-united-states.html

https://firms.modaps.eosdis.nasa.gov/download/

https://www.fs.usda.gov/rds/archive/Catalog/RDS-2020-0016

https://www.kaggle.com/datasets/behroozsohrabi/us-wildfire-records-6th-edition?select=data.csv

Short, Karen C. 2022. Spatial wildfire occurrence data for the United States, 1992-2020 [FPA_FOD_20221014]. 6th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.6

https://www.census.gov/

https://imis.county.org/iMIS/CountyInformationProgram/QueriesCIP.aspx?QueryMenuSelectedKeyctl01_TemplateBody_WebPartManager1_gwpciNewQueryMenuCommon_ciNewQueryMenuCommon=01024814-3279-4a72-ae86-91c667f8d2af

https://headwaterseconomics.org/apps/economic-profile-system/48000

https://www.census.gov/programs-surveys/acs

https://www.valuepenguin.com/homeowners-insurance/wildfire-statistics