Gabrielle Beinars
https://github.com/gbeinars

Forest Fires in the United States

Executive Summary

Background

The ability to predict what locations are more prone to fires, the most common causes of fires, or the size of a fire, can provide endless benefits. By understanding more about forest fires, preventative measures can be taken and possibly save lives, homes, forests and wildlife. Fires can be violently destructive and difficult to manage. Fires are known to destroy millions of acres of forest every year and can spread and grow quickly. Some causes of forest fires are controlled, accidental, intentional or natural. Three types of forest fires are surface, ground and crown.. Surface only impacts a small area of the forest floor and are easiest to stop. Ground fire, also referred to as underground fire, burns where there is dead and dry vegetation, spread slow, but difficult to put out because of its unpredictable pattern. Crown fires are serious, destructive and deadly.[4] This data will provide insight into forest fires with the purpose of developing a better understanding, and determine whether any factors of forest fires can be predicted, thus allowing for preventative measures to be taken.

Problem Statement

Analyzing data collected on forest wildfires can serve a variety of purposes. For example, if the cause of fires could be predicted, then preventative measures can be put in place. This could save thousands of lives, the forests, and the wildlife. Predictions of fire size can also allow for prior action to be taken.

- Are wildfires increasing over the years?
- Do certain geographic areas have more fires than others?
- Are there any features that show correlation with fires?
- Is it possible to predict the cause of a fire based on certain features?
- What is the most common cause of forest fires?

Gabrielle Beinars
https://github.com/gbeinars

Scope

This project involves various exploratory analysis techniques and predictive modeling, to gain a better understanding about forest fires, and predict the fire size and location.

Method

Data Source

The dataset that will be analyzed for this project can be found at https://www.kaggle.com/rtatman/188-million-us-wildfire. This dataset includes forest fires in the US from 1992 to 2015. There are 1.88 million records, with 130 columns, where each observation is a different fire, and total of all fire sizes span 140 million acres. The primary features to be analyzed are the discovery date, final fire size, and location (precise within 1 square mile). The data has already undergone basic error checks and duplicate records were removed. [1][6]

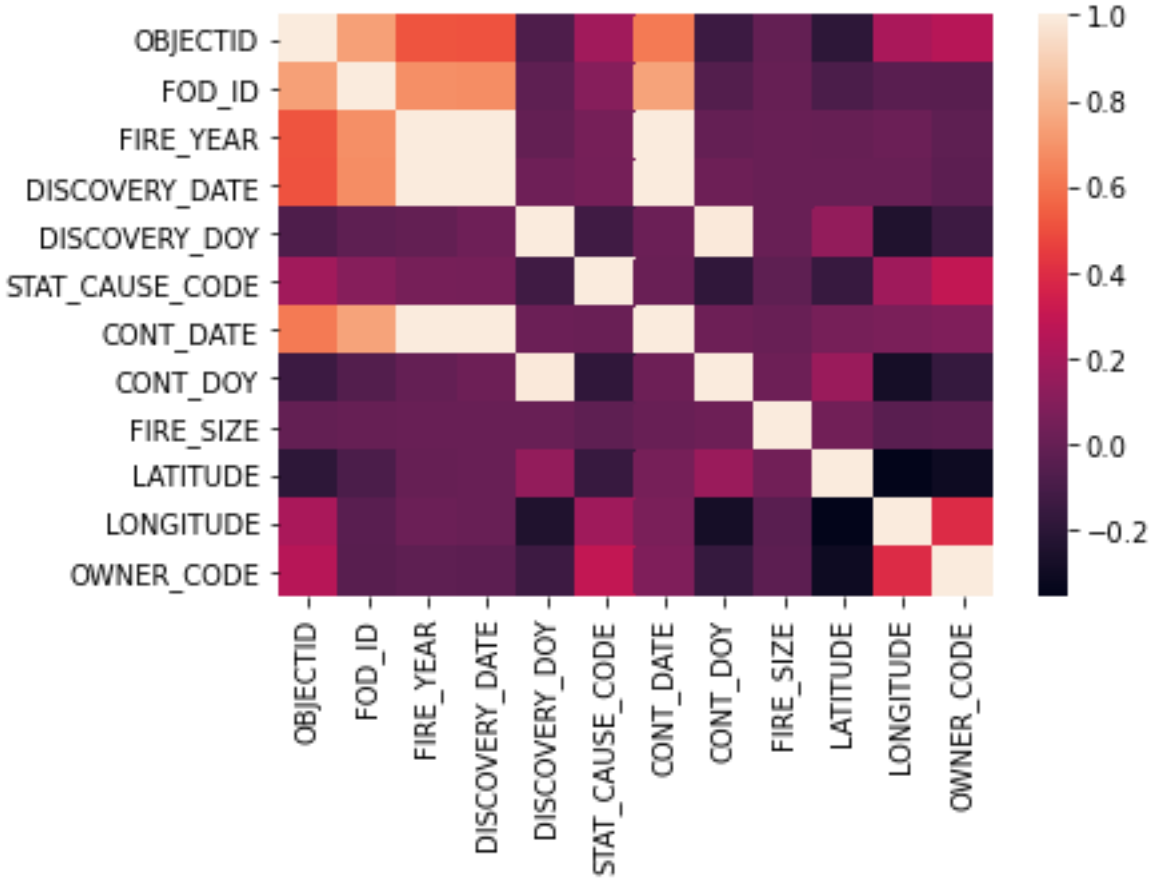
The publication of the data was in support of the national Fire Program Analysis (FPA) system where the records were originally collected from systems of federal, state and local level.[2] The FPA is meant to replace older systems including National Fire Management Analysis System (NFMAS), FirePro and FireBase. The FPA supports the wildland fire program and is considered important for the Forest Service and the Department of the Interior.[3]

Data Import/Cleaning

The data set was imported using Python's sqlite3 library as the data is stored in an SQL database. The table was loaded to a pandas DataFrame where it was further analyzed. Cleaning the data involved checking for missing values and removing any features necessary. Three variables involved dates that which the fire was declared contained or controlled were broken up into day of year, full date, and time. These were removed due to an abundance of missing observations, and multicollinearity displayed with the discovery variables (Figure 1), including date and day of the year the variable was discovered or confirmed. The following variables were included in a subset of the original DataFrame to be used for further analysis: fire year, discovery date, discover day of year, statistical cause code, statistical cause description, fire size, fire size class,

latitude, longitude, state and county. The only variable with missing values was county, however, this is not a concern because we can look at location through state or the coordinates.

Figure 1: Correlation Heat Map



Exploratory Data Analysis (EDA)

After the data was cleaned, exploratory analysis began, where various visuals were created to gain a better understanding about location and cause of forest fires. The top six causes of forest fires from 1992 to 2015 were lighting, miscellaneous, missing/undefined, arson, equipment use and debris burning (Figure 2). With missing/undefined and miscellaneous being the second and third top categories, it raises the question, what more can be done to gain further insight into these categories?

Figure 2: Horizontal Bar Chart

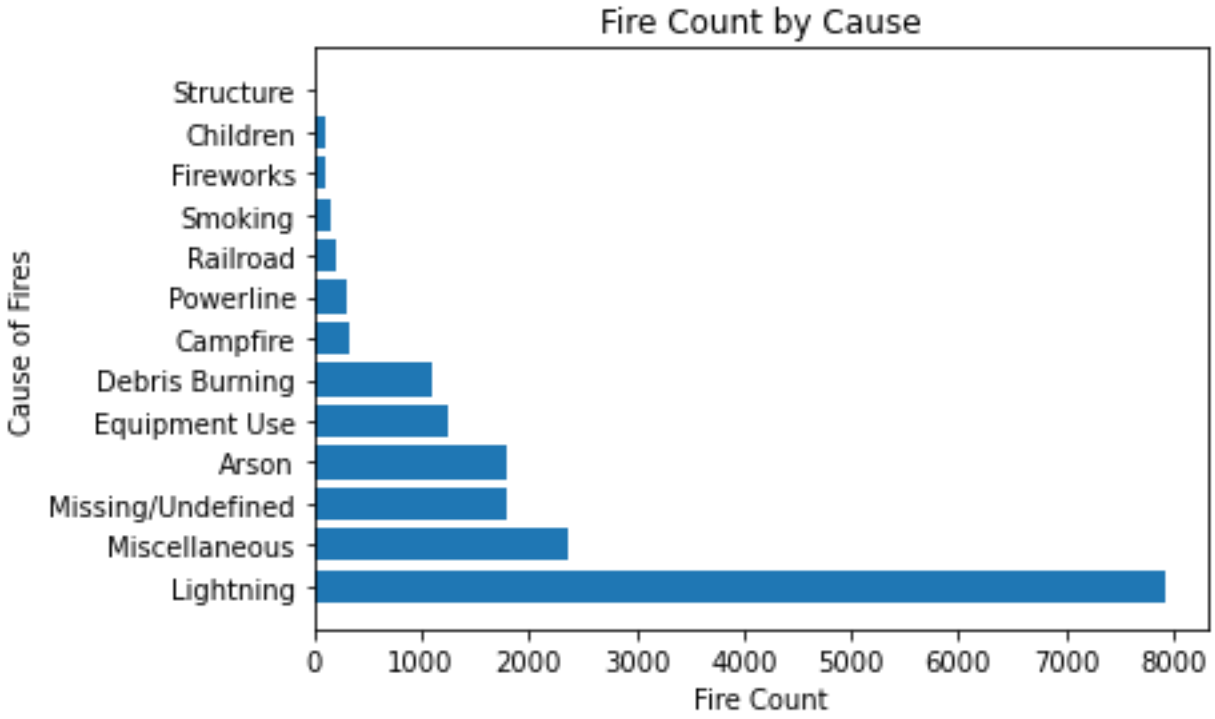
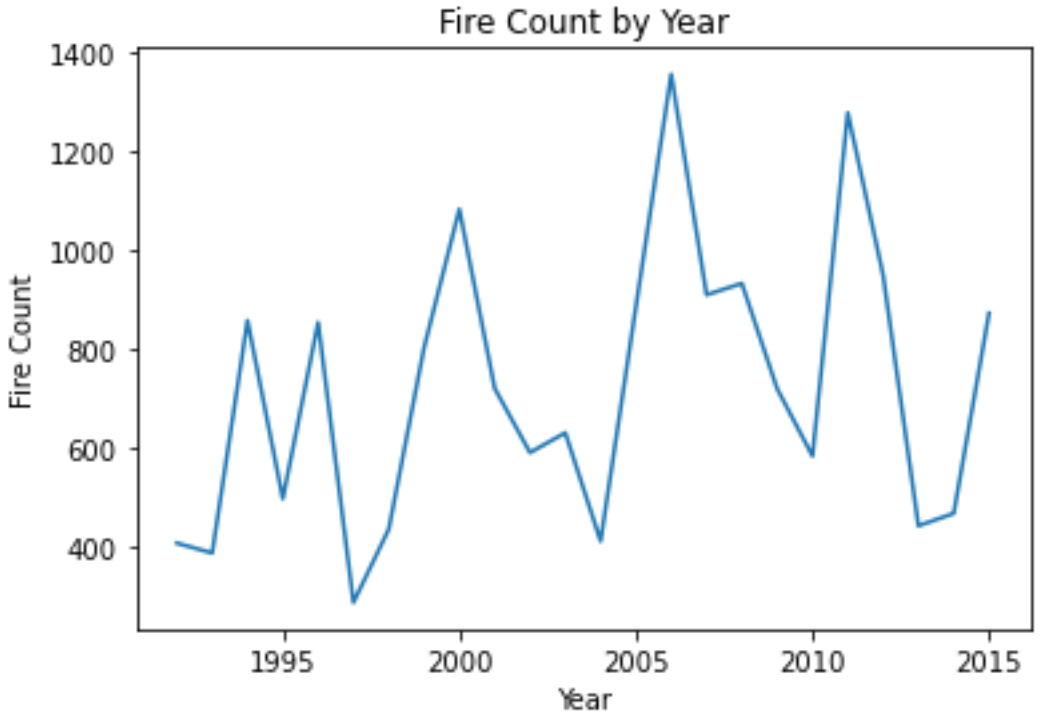


Figure 3: Line Chart



Gabrielle Beinars
https://github.com/gbeinars

Modeling

The approach for modeling utilized decision tree classifier from the scikit learn package. The purpose of decision trees is to predict the value of a target variable by learning decision rules inferred from the other features. A decision tree classifier was used as it can handle multi-output problems, allows for validation, and will still perform well when assumptions are violated. It is also important that the classes used as target variable are balanced to avoid biased.[5] The classes used for modeling are the three fire size classes, E, F and G, which correspond to sizes, 300 to 999 acres, 1000 to 4999 acres, and 5000+ acres (Figure 4). The features used for modeling are location, year and cause.

Figure 4: Bar Chart

Additional models that may be explored are random forests, k-nearest neighbors, gradient boosting and multi-layer perceptron.

Gabrielle Beinars
https://github.com/gbeinars

Results

Discussion

Assumptions

Limitations/Challenges

It is important to be aware when using decision trees that they can be overfitted, and pruning may need to be used to avoid this problem.[5]

Next Steps

Gaining more insight into the fire cause descriptions of miscellaneous and missing/undefined, could provide additional benefits and better the understanding of forest fires.

References

[1] *1.88 million US wildfires*. Kaggle. (n.d.). Retrieved October 16, 2021, from https://www.kaggle.com/rtatman/188-million-us-wildfires.

[2] Short, K. C. (n.d.). *Spatial wildfire occurrence data for the United States, 1992-2018 [fpa_fod_20210617] (5th edition)*. Home. Retrieved October 16, 2021, from https://www.fs.usda.gov/rds/archive/catalog/RDS-2013-0009.5.

[3] *Fire program analysis application*. Bureau of Land Management. (n.d.). Retrieved October 16, 2021, from https://www.blm.gov/policy/ib-2009-040.

[4] Rodgers, B. (2021, August 18). *Understanding forest fire safety, preparation and survival*. Prepper's Will. Retrieved October 16, 2021, from https://prepperswill.com/understanding-forest-fire-safety-preparation-and-survival/.

[5] *1.10. decision trees*. scikit. (n.d.). Retrieved October 16, 2021, from https://scikit-learn.org/stable/modules/tree.html.

[6] Short, Karen C. (2017). *Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508] (4th Edition)*. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

[7] Coffield, S. R., Graff, C. A., Chen, Y., Smyth, P., Foufoula-Georgiou, E., & Randerson, J. T. (2019, September 17). *Machine learning to predict final fire size at the time of ignition*. CSIRO PUBLISHING. Retrieved October 16, 2021, from https://www.publish.csiro.au/wf/WF19023.

[8] Koehrsen, W. (2017, August 12). *Data wrangling with python and SQLite*. Medium. Retrieved October 17, 2021, from https://williamkoehrsen.medium.com/data-wrangling-with-python-and-sqlite-900d21bc5a53.

[9] *Public fire information websites*. US Forest Service. (n.d.). Retrieved October 17, 2021, from https://www.fs.usda.gov/science-technology/fire/information.

[10] *National fire news*. National Fire News | National Interagency Fire Center. (n.d.). Retrieved October 17, 2021, from https://www.nifc.gov/fire-information/nfn.

[11] Dailydatablog. (2017, November 13). *SQLite tutorial: Bike Share Company Data Analysis*. Daily Data Blog. Retrieved October 17, 2021, from https://dailydatablog.wordpress.com/2017/10/01/sqlite-tutorial-bike-share-company-data-analysis/#:~:text=SQLite%20is%20a%20Python%20package%20which%20allows%20the,Hubway%20is%20a%20bike-share%20company%20based%20in%20Boston.

[12] *Data and Tools*. Wildfire - Real-time Data. (n.d.). Retrieved October 17, 2021, from https://www.usgs.gov/products/data-and-tools/real-time-data/wildfire.

Gabrielle Beinars
https://github.com/gbeinars

Appendix

Below is a screenshot of an interactive map provided by Geospatial Multi-Agency Coordination and the wildfire community. It shows active incidents (shown as purple and black squares, and red and pink flame symbols) and satellite fire detections (shown as red, orange and yellow circles).

https://maps.nwcg.gov/sa/#/%3F/%3F/45.3869/-132.9872/4