Analyzing US Wildfire incidents

by

Radhika Vijayaraghavan
netID# zg4894

Instructor: Dr.Eric Fox

STAT 650, California State University East Bay

Fall 2022

Contents

Abstract

This dataset includes Wildfires incident reports from period 2011 to 2015. These reports have been filed by various federal, state, and local fire organizations.

Evidently, wildfires create huge ecological and economic damage to any country while endangering animal lives and human lives. Fast detection is a key element for controlling such events. In California alone, we observe some of the largest wildfires in state history.

The objective of this project is to apply the learnings/techniques taught in STAT 650 course to use R for data analysis and visualization to provide useful insights on wildfires, its causes and behavior. Effort has been made to use concepts of Data Wrangling, Data Transformation, Model building, Spatial visualization. This helps to make useful recommendations to increase public safety, minimize economic damage from future wildfires. My findings show the regions/sub regions that are most affected and the impact of seasons on wildfires.

To clean the raw data set obtained from the SQLite database, data wrangling techniques were applied. `discovery_date, cont_date` were in Julian format(float) and `discovery_time, cont_time` were in military format. Measures have been taken to change them to R Date format utilizing `Lubridate` package. The time taken to contain the fire was added using the contained_date, discovery date variables. `season` was also added as a new column by utilizing functions in `forcats` package. Special characters rows and NA rows have been filtered out. Majority of the Tidy data functions and techniques learnt in this course have been useful to apply in this exercise. Data transformation have been performed wherever necessary for model building and visualization.

Data Description

a. Data Source

The source of this data set is Kaggle(1.88 Million US Wildfires). This dataset is an SQlite spatial database of wildfire incidents that occurred in the United States from 2011 to 2015. Additionally, utilized SQLite database to extract related data sets of interest to investigate further in this exercise.

b. Data Description

The cleaned data set consist of 188,017 observations on the following 16 variables. 4 new variables have been added through data wrangling. Belo are the description of all the variables. Apart from this data set, `states` data set from R was used to perform data transformation.

- **fire_year** = Calendar year in which the fire was discovered or confirmed to exist.
- **discovery_date** = Date on which the fire was discovered or confirmed to exist.
- **discovery_time** = Time of day that the fire was discovered or confirmed to exist.
- **Stat_cause_descr** = Description of the (statistical) cause of the fire.
- **fire_size** = Estimate of Acres of land burnt within the final perimeter of the fire.
- **fire_size_class** = Code for fire size based on the number of acres within the final fire perimeter expenditures (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres)
- **latitude** = Latitude (NAD83) for point location of the fire (decimal degrees).
- **longitude** = Longitude (NAD83) for point location of the fire (decimal degrees).
- **owner_descr** = Name of primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
- **cont_date** = Date on which the fire was declared contained or otherwise controlled (mm/dd/yyyy where mm=month, dd=day, and yyyy=year).
- **cont_time** = Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, mm=minutes).
- **county** = County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.
- **state_abb** = Two-letter code for the state in which the unit is located (or primarily affiliated).
- Additionally below variables were added through EDA:-
  - **resolved_days** - #Days taken to resolve the wildfire
  - **season** - Season of the year - Summer, Spring, Fall, Winter
  - **state_name** - Full name of the state
  - **disc_month** - Month of the year

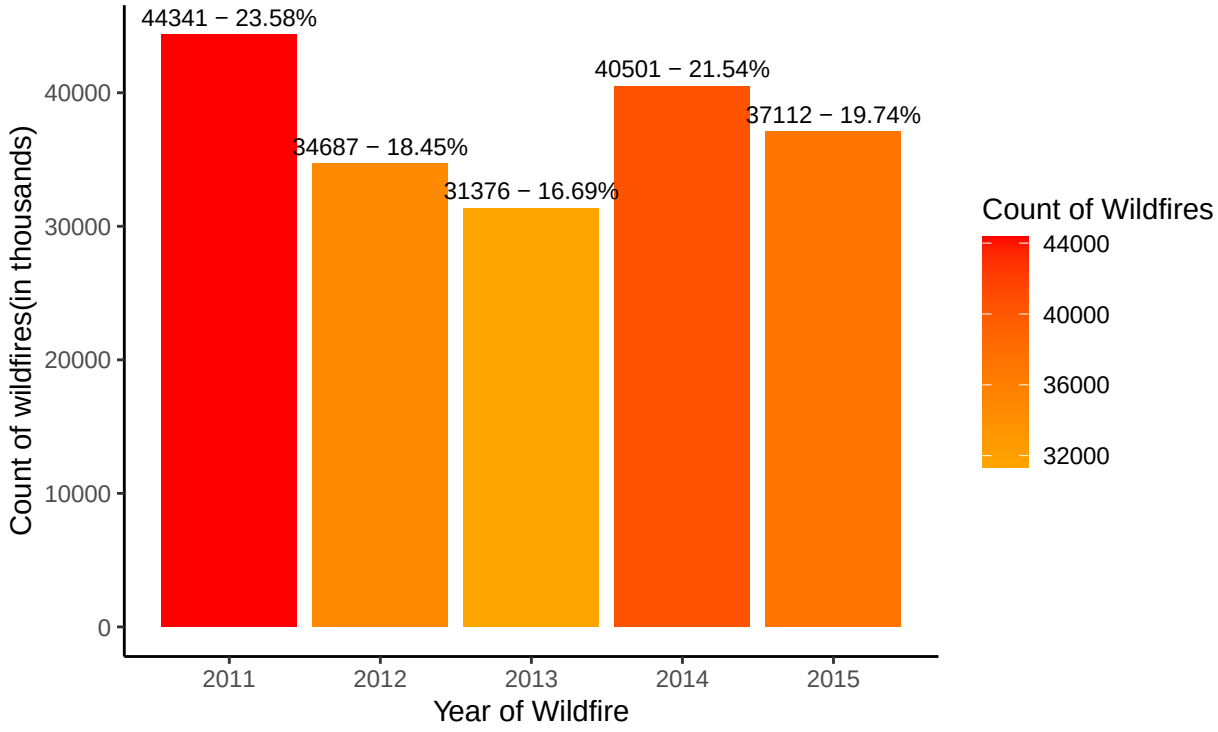
Questions of Interest

Research Question 1

"Does Global warming affect the number of fires? Has the number of fires increased over the period 2011-2015?"

Potential Use Case:

- With this information, we can support stakeholders in decision making - such as taking preventive measures and restrict the wildfire size with the help of modern equipment in highly destructive areas.
- From this barplot, although we see an upward trend at certain years, there has been no constant increase or decrease in trend over the years 2011 to 2015. But how do we know what causes these incidents? Our next research question answers this.



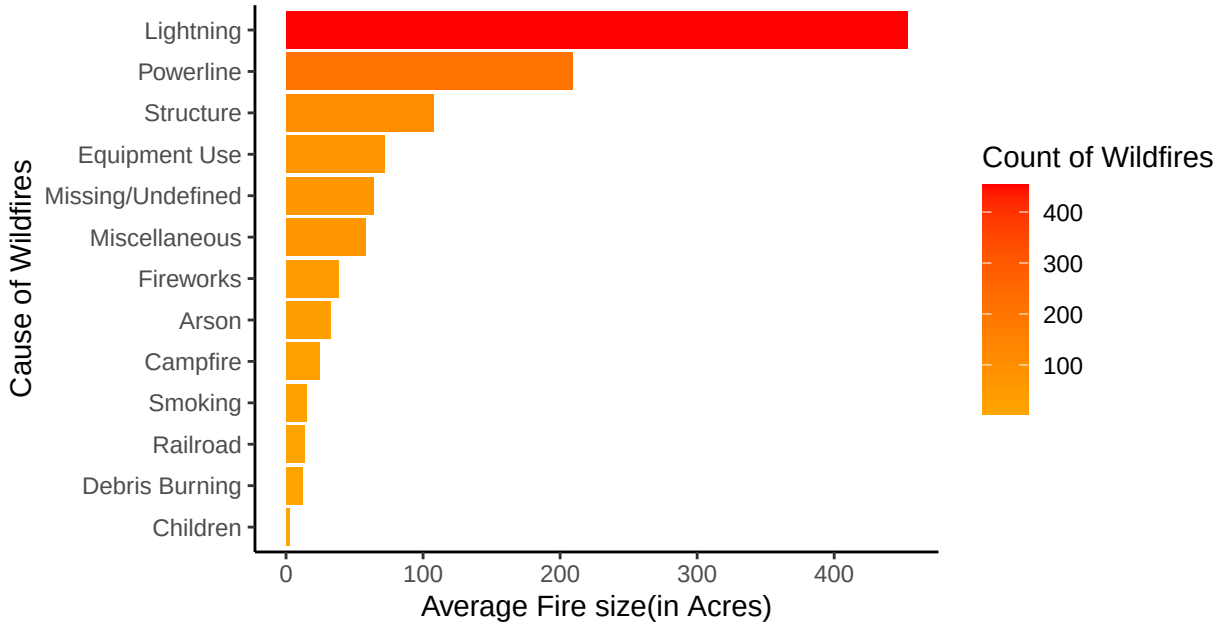
Research Question 2

"What causes the most wildfires? Which causes are associated with it?"

Potential Use Case:

- With this information, the concerned area owners(private/state) could work towards intelligent monitoring of power lines to mitigate substantial associated risks.

- In broad terms, intelligent monitoring can provide heightened awareness of power line health and events, enabling utility companies to act more quickly.



- Although `Debris burning` seemed to cause the highest number of wildfires in this period, the average size of fire caused by it is low.

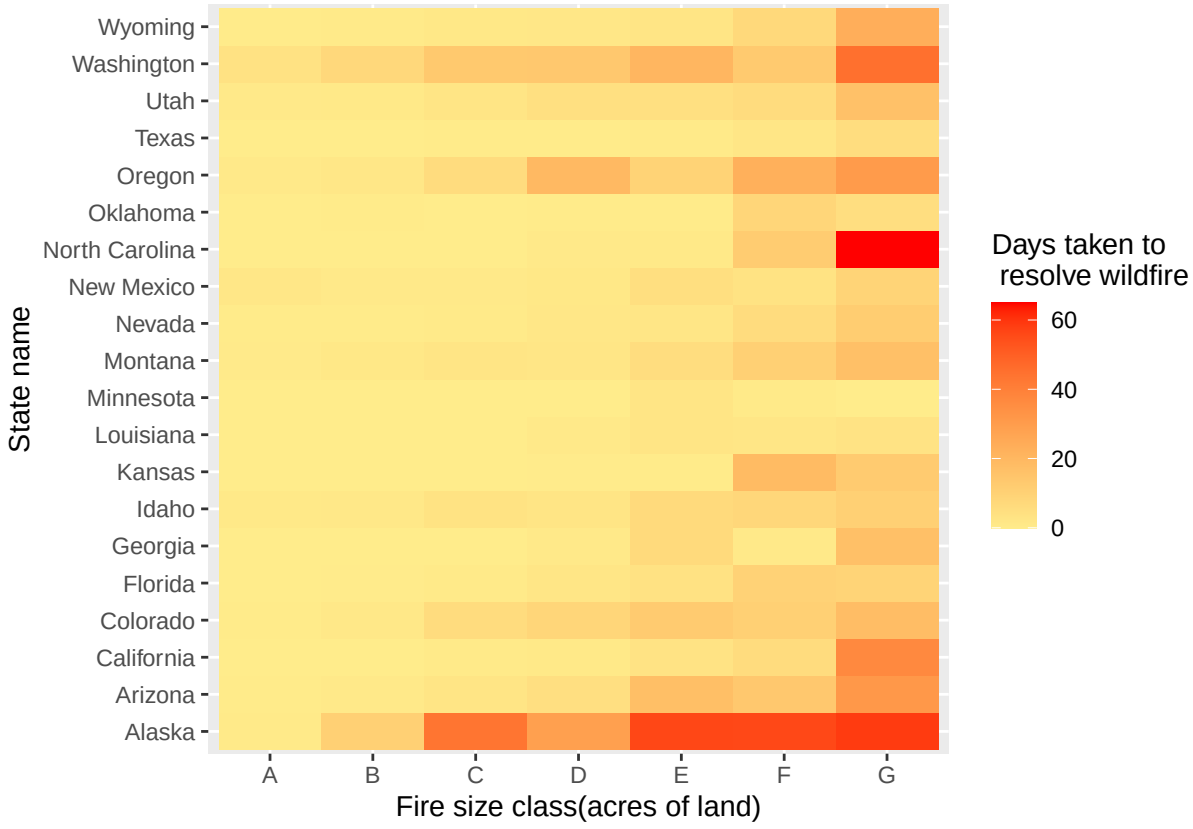
- On the other hand, `Lightning` and `Powerlines` i.e., fire started by electricity causes the most destruction wildfires in terms of fire size(acres of land burnt).

Research Question 3

"Does the time taken to resolve the fire vary by region and acres burnt? Is the resolution time related to season?"

Potential Use Case:

- From the Data set, we know that the following fire sizes A to G represents the respective acres of land burnt i.e., (*A= greater than 0 acres but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, G=5000+ acres*)
- From the below plot, it is interesting to note that some states such as Alaska, Orgeon, Washington have taken more days to resolve the wildfire although less area was burnt. Alarmingly, states like "Alaska" seems to have struggle resolving wildfires of any size. This should be investigated further.
- It is observed that certain states have been very efficient in handling dangerous fire sizes such as size "G" in states such as Nevada, Montana. The states with the dangerous fire size "G" can learn the best practices and readiness implemented by the states that took less number of days to resolve fire type "G".
- It is also possible that the wildfires of type G was extinguished soon due to temperature changes. Let's discuss this further.



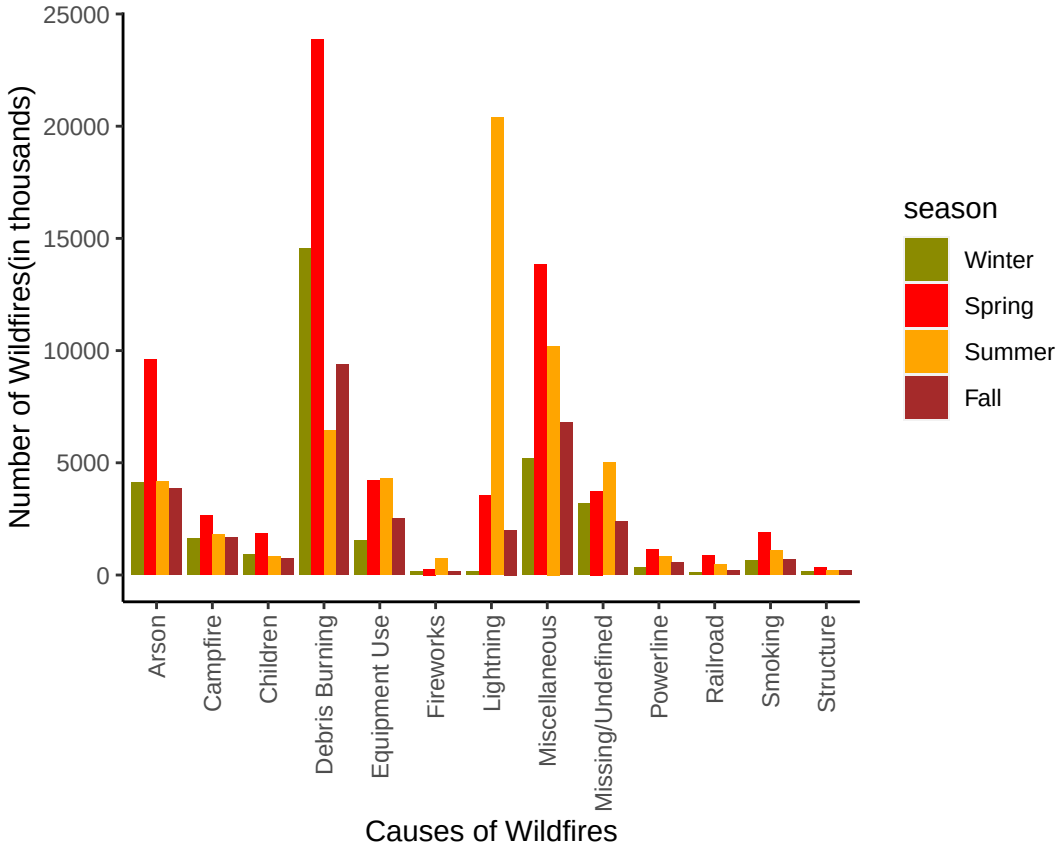
- From the above result, it is also interesting to note that more destructive wildfires are reported in Southeast states of US such as Texas, Kansas, Alaska, Georgia.
- Though we see the fire size per state and days taken to contain the wildfire, it is not clear what is causing them. Let's study this further.

Research Question 4

"Is season related to the cause of wildfires?"

Potential Use Case:

- By analyzing the below plot, we observe that Debris burning during Spring season and Lightning during Summer season causes the most wildfires.
- It is interesting to note how `seasons` have a significant impact on the `cause of fire`.
- By this, the forest and fire agencies can plan the schedule such that there are optimal number of officers to handle the expected load of fire incidents at peak seasons/times such as below.

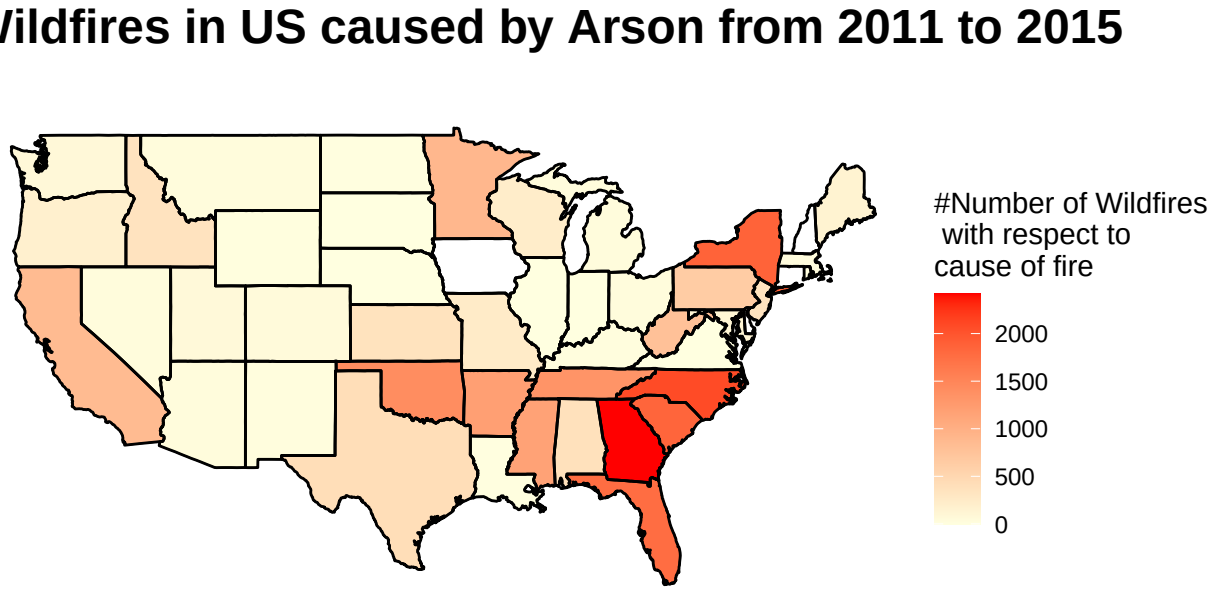


Research Question 5

"Geospatial visualization - Which states have the highest count of wildfires against the cause of fire"

Potential Use Case:

- This geospatial map depicts the count of wildfire incidents with respect to the cause of the fire by "Arson".
- It is interesting to note that South region has the highest concentration of fires induced by "Arson". Georgia seems to have a concerning number of Arson induced wildfires.
- This makes it noteworthy for the concerned regional agencies of this region to monitor the area and deploy preventive measures against such unlawful practices.



* With respect to states filtered by the cause of fire as "Lightning & Power lines", the West region and Florida have a substantially high number of wildfire incidents induced due to electricity. Florida seems to have high wildfire incidents caused by "Electricity" & "Arson"

Conclusion

Summary of Results

- Most number of wildfire incidents across the US are caused by Debris Burning, Arson. Although the average incidents are due to Lightning(especially during Summer) causes the most damage.
- As per this study, smaller southern US states like Georgia, Florida and Alaska must be more prepared with the right equipment and safety measures to fight wildfires.
- Our Geo spatial map will be helpful to figure out which regions of the country to focus more, with regards to the cause of fire.
- Local government departments in regions such as Alaska need to be vigilant as its land has been recklessly burnt by these wildfires.
- The MLR model of this study shows that a 12% of variation in the response variable can be explained by the predictors that were considered, but with further tuning and testing with other variables and by backward AIC we would be able to find a better r-squared and p-value.
- Finally, by educating people about local regulations/guidelines regarding Arson, trash burning, being cautious during campfire, and by implementing strict regulations against people committing Arson, we would be able to contain the prevent such human caused fires.

References

- 650 Lecture slides and notes by Dr.Eric Fox, Fall 2022
- R for Data Science by Hardley Wickham
- R Markdown LATEX Cookbook by bookdown.org
- Geospatial visualization by towardsdatascience.com

Appendix

For the whole R script, visit https://github.com/vijayaraghavan-radhika/650_final_project