**USA Wildfires Project**

I have chosen this data set as it contains both geospatial and time series data. With over 2 million records, I am interested to see what patters and trends could be uncovered. It is also something which directly impacts communities, wildlife and the environment.

The data set contains information on wildfires which occurred in the USA over 27 years (1992 – 2018). It was the fourth update of a publications with the original purpose of supporting the US Fire Program Analysis (FPA) system. Although the program has since been retired, the purpose was to support the process for strategic fire management planning and budgeting.

The database includes over 2 million records representing a total of 165 million acres burned during the period.

**Data Source**

The wildfire records were collated from federal, state and local fire organisations making it a more reliable source. Basic error-checking was performed and redundant records were removed, making the data set easier to work with.

It is an open-source data set without additional permissions or fees:  
<https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0009.5>

Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA\_FOD\_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.5

**Data Collection**

In some states, reporting wildfires is voluntary. It has been estimated that only two-thirds of fires of any type, are reported to the system of record for U.S. fire departments, NFIRS. Subsequently, this dataset is likely to be an incomplete picture of wildfires in the US.

**Data Contents**

The database contains both geospatial and time series data as well as key data points on individual wildfires including the size.

**Data Relevance**

*Hypotheses*

The frequency and geographical range of wildfires have increased over time.

Due to higher temperature climates, southern states have a higher risk of wildfires.

Historical wildfire data can be used to determine trends over time and also help identify high risk states.

**Data Profile**

The data set contains 2,166,753 rows and 37 columns (listed below). The 14 columns kept for data analysis are highlighted.  
  
FOD\_ID = Unique numeric record identifier.  
  
FPA\_ID = Unique identifier that contains information necessary to track back to the original record in the source dataset.  
  
SOURCE\_SYSTEM\_TYPE = Type of source database or system that the record was drawn from (federal, nonfederal, or interagency).   
  
SOURCE\_SYSTEM = Name of or other identifier for source database or system that the record was drawn from. See Table 1 in Short (2014), or \Supplements\FPA\_FOD\_source\_list.pdf, for a list of sources and their identifier.  
  
NWCG\_REPORTING\_AGENCY = Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report (BIA = Bureau of Indian Affairs, BLM = Bureau of Land Management, BOR = Bureau of Reclamation, DOD = Department of Defense, DOE = Department of Energy, FS = Forest Service, FWS = Fish and Wildlife Service, IA = Interagency Organization, NPS = National Park Service, ST/C&L = State, County, or Local Organization, and TRIBE = Tribal Organization).  
  
NWCG\_REPORTING\_UNIT\_ID = Active NWCG Unit Identifier for the unit preparing the fire report.  
  
NWCG\_REPORTING\_UNIT\_NAME = Active NWCG Unit Name for the unit preparing the fire report.  
  
SOURCE\_REPORTING\_UNIT = Code for the agency unit preparing the fire report, based on code/name in the source dataset.  
  
SOURCE\_REPORTING\_UNIT\_NAME = Name of reporting agency unit preparing the fire report, based on code/name in the source dataset.  
  
LOCAL\_FIRE\_REPORT\_ID = Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year.  
  
LOCAL\_INCIDENT\_ID = Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year.  
  
FIRE\_CODE = Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (https://www.firecode.gov/).  
  
FIRE\_NAME = Name of the incident, from the fire report (primary) or ICS-209 report (secondary).  
  
ICS\_209\_PLUS\_INCIDENT\_JOIN\_ID = Primary identifier needed to join into operational situation reporting data for the incident in the ICS-209-PLUS dataset.  
  
ICS\_209\_PLUS\_COMPLEX\_JOIN\_ID = If part of a complex, secondary identifier potentially needed to join to operational situation reporting data for the incident in the ICS-209-PLUS dataset (2014 and later only).  
  
MTBS\_ID = Incident identifier, from the MTBS perimeter dataset.  
  
MTBS\_FIRE\_NAME = Name of the incident, from the MTBS perimeter dataset.  
  
COMPLEX\_NAME = Name of the complex under which the fire was ultimately managed, when discernible.  
  
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\_DOY = Day of year on which the fire was discovered or confirmed to exist.  
  
DISCOVERY\_TIME = Time of day that the fire was discovered or confirmed to exist.  
  
NWCG\_CAUSE\_CLASSIFICATION = Broad classification of the reason the fire occurred (Human, Natural, Missing data/not specified/undetermined).  
  
NWCG\_GENERAL\_CAUSE = Event or circumstance that started a fire or set the stage for its occurrence (Arson/incendiarism, Debris and open burning, Equipment and vehicle use, Firearms and explosives use, Fireworks, Misuse of fire by a minor, Natural, Power generation/transmission/distribution, Railroad operations and maintenance, Recreation and ceremony, Smoking, Other causes, Missing data/not specified/undetermined).  
  
NWCG\_CAUSE\_AGE\_CATEGORY = If cause attributed to children (ages 0-12) or adolescents (13-17), the value for this data element is set to Minor; otherwise null.   
  
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\_DOY = Day of year on which the fire was declared contained or otherwise controlled.  
  
CONT\_TIME = Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, mm=minutes).  
  
FIRE\_SIZE = The estimate of acres 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 (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.  
  
STATE = Two-letter alphabetic code for the state in which the fire burned (or originated), based on the nominal designation in the fire report.  
  
COUNTY = County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.   
  
FIPS\_CODE = Five-digit code from the Federal Information Process Standards (FIPS) publication 6-4 for representation of counties and equivalent entities, based on the nominal designation in the fire report.  
  
FIPS\_NAME = County name from the FIPS publication 6-4 for representation of counties and equivalent entities, based on the nominal designation in the fire report.

**Data Wrangling**

The data set was downloaded as an SQLite database and uploaded in Python as a data frame. Out of the 27 columns in the table, only 14 were imported into Python (as highlighted above).

After checking the data frame column data types, the “DISCOVERY\_DATE” column was changed from an “object” to “datetime64[ns]” type.

**Consistency Checks**

Upon investigation, the ‘DISCOVERY\_TIME’ column was found to have a statistically significant number of missing values and was dropped.

The “COUNTY” code and “FIPS\_NAME” county columns also had missing values but they were not dropped as it could be worth exploring to replace the missing values later on using the “LATITUDE” and “LONGITUDE” coordinates.

Although on the face of it, the “OWNER\_DESCR” did not have any NaN values, there are over one million rows with “MISSING/NOT SPECIFIED” owner description. Although

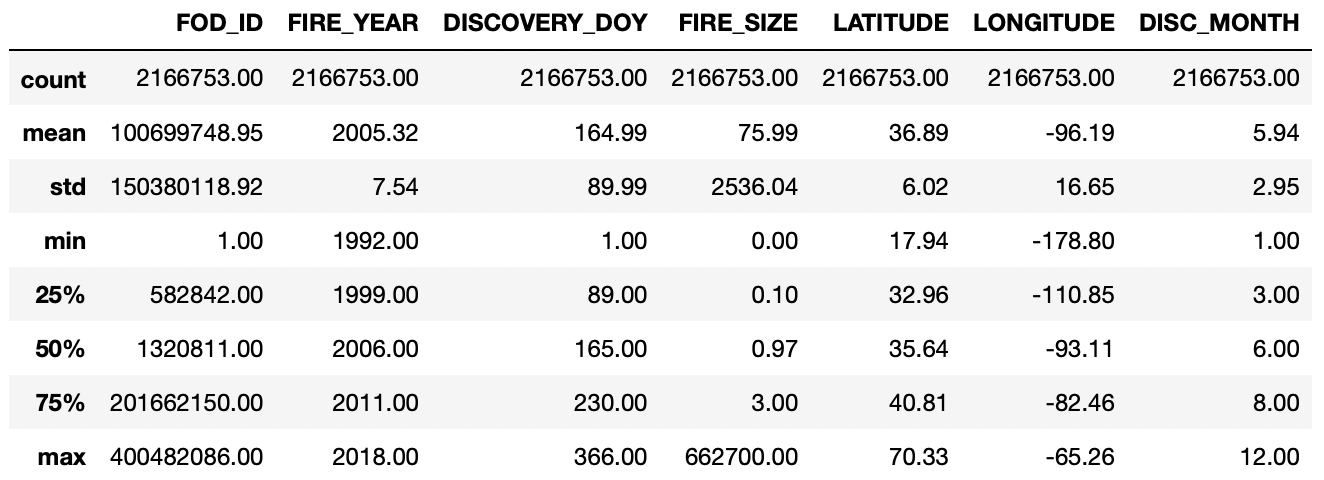
statistically significant, I am not dropping the column at this stage without further analysis as it could still be useful.

The were no columns with mixed data types and there were no duplicate records found.

**Data Profile**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Time-variant / invariant | Structured / Unstructured | Quantitative / Qualitative | Nominal / Ordinal | Discrete / Continuous |
| FOD\_ID | Invariant | Structured | Qualitative | Ordinal |  |
| FIRE\_YEAR |
| DISCOVERY\_DATE |
| DISCOVERY\_DOY |
| DISCOVERY\_TIME |
| NWCG\_GENERAL\_CAUSE | Nominal |
| FIRE\_SIZE | Quantitative | Ordinal | Continous |
| FIRE\_SIZE\_CLASS | Qualitative |  |
| LATITUDE | Nominal |
| LONGITUDE |
| OWNER\_DESCR |
| STATE |
| COUNTY | Unstructured |

**Basic Statistical Descriptive Analysis**



**Limitations and Ethics**

As mentioned above, although the data set may be the most complete and official of wildfires in the US, due to the fact that in some states reporting wildfires is voluntary, this is likely not the complete picture. The implication is that incorrect fire management planning and budgeting decisions could be made if used for that purpose.

There is no personally identifiable information (PII) in the data set, meaning there are no ethical concerns from that perspective.

**Questions to Explore**

* What are the leading causes of wildfires?
* What percentage of wildfires are caused by humans?
* When do wildfires happen? i.e., in which month and day of the week
* Have incidents been increasing or decreasing over time?
* Where do fires occur? i.e., in which states and counties?
* On which type of land do wildfires occur e.g., government owned etc.

1. **Who:** The client is a US government agency tasked with identifying the leading causes of wildfires.
2. **Why:** The agency’s ultimate goal is to devise initiatives to reduce the impact of wildfires to be piloted in key states before being rolled out nationally.
3. **What:** The storyboard will provide more information about the main causes of wildfires, focusing of the top 5 states affected by wildfires.
4. **When:** It will be published on various national websites
5. **Where:** Tableau Public

Government agencies to allocate resources

General public awareness on climate change or human cause wildfires

Wildlife conservation agencies

Compare impact vs. football fields etc.

Dashboard plan

1. Why are wildfires a problem? What is the extent of wildfires?
   1. Definition
   2. Trend by region (1992 – 2018)
   3. Number of largest size category wildfires
   4. Impact on people, wildlife and environment size example of 2018 impact
   5. Resources used in tackling wildfires
   6. What can be done to reduce the number of and impact of wildfires?
2. Largest wildfires
   1. Regression analysis
   2. Cluster analysis
3. Which states have been impacted the most in the last x years?
   1. Number of wildfires vs. area impacted
4. What are the main causes?
5. Natural causes trend and possible climate change link
   1. 15-year trend and forecast of acres burned using average US temperature, Max temperature, average precipitation
   2. 15-year trend of average size of fire
6. Human causes, focus short and mid-term resources?

Analyse below causes and identify top 5 states by number of wildfires and area burned. Spatial maps

* 1. Arson
  2. Equipment & Vehicle use
  3. Debris & open burning
  4. Recreation and ceremony
  5. Other/Unknown

1. Suggest to use and introduce drone and satellite technology, new laws and equipment regulations prioritising and piloting based on above results

### This script contains the following:

#### 1. Importing Visualization Libraries and Data

#### 2. Data Checks

- Missing values

- Duplicate records

- Data types

#### 3. Data Wrangling

- Deriving new columns

- Year & month

- Fire size description

- Fire cause category

- Import US Regions

- Drop Puerto Rico data (not a US State)

- Import Monthly Average Temperatures

- Import Annual Average Temperatures

- Convert average monthly temperatures from Fahrenheit to Celsius

- Rename average monthly temperature column

#### 4. Exploratory Data Analysis

- Time series trend

- Export data to .csv for Tableau use

- Create recent data subset based on last 10 years

- Bar graphs analysis

- Create various subsets

#### 5. Explore Data Correlations

- Correlation matrix

- Scatterplots

- Pair Plots

- Categorical Plots

#### 6. Regression Analysis

- Data preparation

- Regression analysis

- Performance after removing outliers

- Export data to .csv for Tableau use

#### 7. Machine Learning

- Derive fire discovered day of week number column

- Change derived column type to int64

- Apply the elbow technique

- K-means clustering

- Scatterplot & Spatial plot analysis

- Export data to .csv for Tableau use