Shubhangi Singhal

DWBI Final Project Report

Wildfire Trends in the US

Abstract - I intend to work on the 2018-07-wildfire-trends dataset obtained from Buzz Feed News' GitHub repository. I aim to develop 2-3 dashboards depicting the relationship between the various variables present in the dataset. My dataset is of 502 MB size. It has 18,80,465 rows and 38 columns. The columns are objectid, fod_id, fpa_id, source_system_type, source_system, nwcg_reporting_agency, nwcg_reporting_unit_id, nwcg_reporting_unit_name, source_reporting_unit, source_reporting_unit_name, local_fire_report_id, local_incident_id, fire_code, fire_name, ics_209_incident_number, ics_209_name, mtbs_id, mtbs_fire_name, complex_name, fire_year, discovery_date, discovery_doy, discovery_time, stat_cause_code, stat_cause_descr, cont_date, cont_doy, cont_time, fire_size, fire_size_class, latitude, longitude, owner_code, owner_descr, state, county, fips_code, fips_name.
I intend to show case the trend in the wildfire occurrences over the years. Moreover, I aim to create data distributions using columns like fire_year, discovery_time, fire_size, state, county. Other than this I plan on to showing the relationship between different variables through my dashboards.

Columns of Interest – Stat Cause Desc, Fire Year, Fire Size, State, County/ Fips Name
URL/ Source of the Dataset - https://github.com/BuzzFeedNews/2018-07-wildfire-trends
File Type – .csv
Size of the Dataset – 502 MB

Introduction –
This dataset represents manifestation of wildfires in the United States from 1992 to 2015. This is the third revise of the publication originally produced to back the national Fire Program Analysis (FPA) system. These wildfire accounts were developed from the reporting systems of federal, state, and local fire organizations. The data was altered to adapt, when possible, to the data standards of the National Wildfire Coordinating Group (NWCG). Rudimentary error-checking was achieved, and redundant records were recognized and removed, to some extent. The resultant is mentioned as the Fire Program Analysis fire-occurrence database (FPA FOD), which includes 1.88 million geo-referenced wildfire records, demonstrating a total of 140 million acres burned throughout the 24-year period.

Data Dictionary

I have collated a metadata sheet for the dataset which has been attached within the metadata.xlsx file.

Shubhangi Singhal

Objective of the project

The objective of the project is to develop a series of Tableau dashboards using Tableau Desktop to analyze and visualize the "Wildfire Trends in the US". I used the 2018-07 Wildfire Trends Dataset, obtained from Buzzfeed News to conduct an analysis of some of the attributes which I believe help us answer the following questions – (*The tableau workbook is a set of 13 sheets and 4 interactive dashboards.)*

- Have the fires become more frequent over the years?
- Which counties are more vulnerable to the wildfires?
- What are the causes behind these wildfires; Are they Natural or Human?
- Which States are more prone to these fires?
- What are days in a year (wildfire season) when these fires occur?

Challenges encountered and how they were resolved -

While extracting, transforming and loading of the data and while developing the dashboards I encountered multiple challenges. Some of these were-

- **Row limit per excel sheet** – Since my dataset has about 18,00,000 rows and 38 columns, I faced with a limitation on the number of rows that could be inserted within a single worksheet of an excel workbook. Due to this, I had to distribute my data across different worksheets and compile them into one final excel workbook for easy data transfer to Tableau.
- **Null values and random values** – there were many null values and random irrelevant values present in the dataset which could have distorted any significant results erroneously. To curb this, I preferred to exclude those null values during the loading process in Tableau and worked on removing any irrelevant random values manually.
- **Choice of correct attributes for the visualization(s)** – While working on one of the visualizations in Tableau, I was required to use the *County* variable from the dataset, based on a manual analysis, I understood that this variable had values of alpha numeric type which would have been difficult to visualize at a later stage. As an alternative I used the Fips_Name attribute, which was also the County names, just with the string data type. I used an attribute with a consistent data type as my choice for the visualizations.
- **Merging the excel sheets in Tableau** – There were some excel sheets that were added to the data source at a later stage which required me to replace, and refresh the data source and then take a union of the new sheets with the already existing sheets in Tableau. Exclusion of these sheets would have distorted any significant outputs to the project.
- **Alternative to Pie Charts** – During the exploratory stage of the analysis, especially for Dashboard 1, I worked with a few pie charts to visualize categorical data (*fire size class*) because of which I later on faced a challenge of changing my choice of chart into a more effective alternative. I had more than 3 categories for the fire size attribute due to which the pie chart fell short of being an effective choice of chart. Finally, I chose the heat map over the pie chart to accommodate for the drawback.

- **Redundant columns** – One of the columns in the original dataset (*Source Reporting Unit*) has is repeated twice. The redundant column has not been used during the analysis.

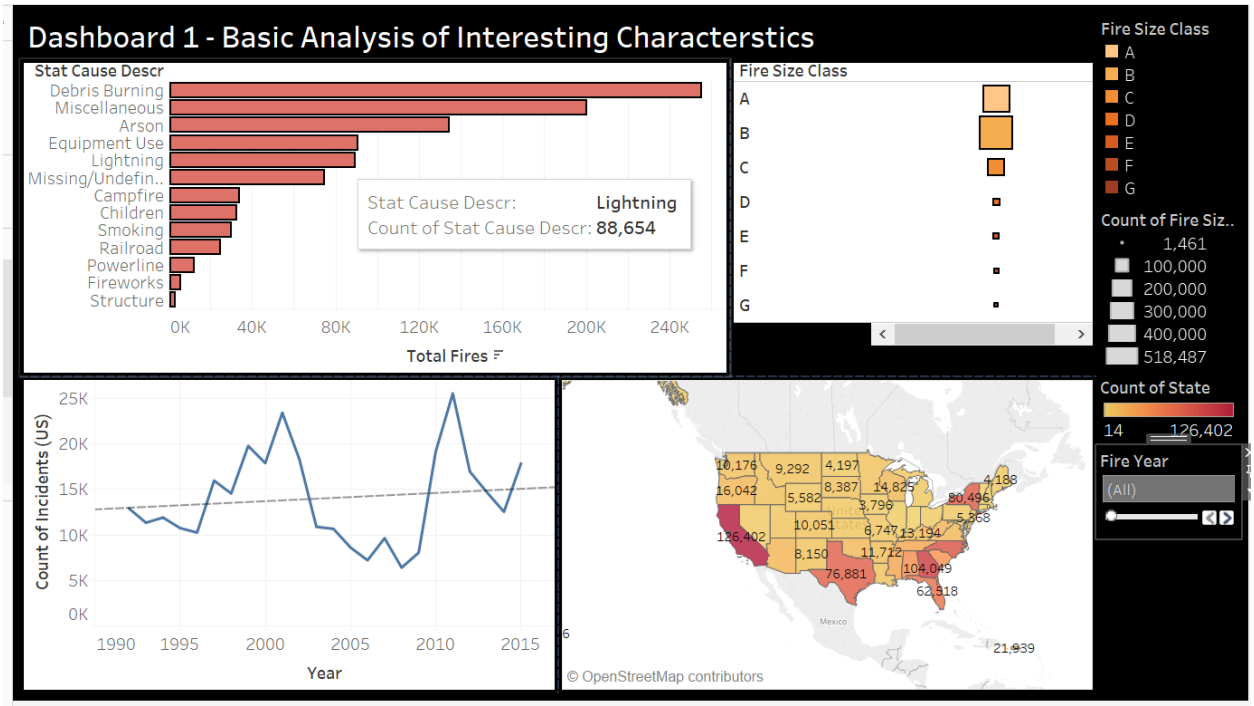
Result

I started with conducting a basic exploratory analysis of some key dimensions and measures which I believe are relevant to the objective of the project, these attributes were - fire_year, discovery_doy, stat_cause_code, stat_cause_descr, fire_size, fire_size_class, state. This is where Dashboard 1 comes into the picture. I let the audience play with these key attributes and let them understand the impact on the cause of the wildfires, the fire size class, etc.

From this dashboard I try to come up with some basic data distributions for the above-mentioned attributes. I finally conclude that the top five causes of these forest fires are Debris Burning, Miscellaneous reasons, Arson, Equipment use and lightning, this suggests that overall, it is the human factors dominating the increase in the forest fires over these years. We can also see in the adjacent heat map, that there are 7 different classes of the fire size (acres burnt) where the lighter shades of orange denote the lower number of acres and darker shades denote higher number of acres burnt. We see that most of the fires that occur are of smaller size. Looking at the 3rd graph, which is a line chart, it talks about the number of fire incidents over the span of 20 years, we understand that the count has incremented as per the trend line. We assume that this may be due to globalization, more human interference, and climate change. Lastly the geo map portrays the count of wildfire incidents per state in the US. We again use the darker shades for the higher counts and vice versa. We find California and Texas among other highly impacted states which still are prone to these fires in the year 2019-2020.

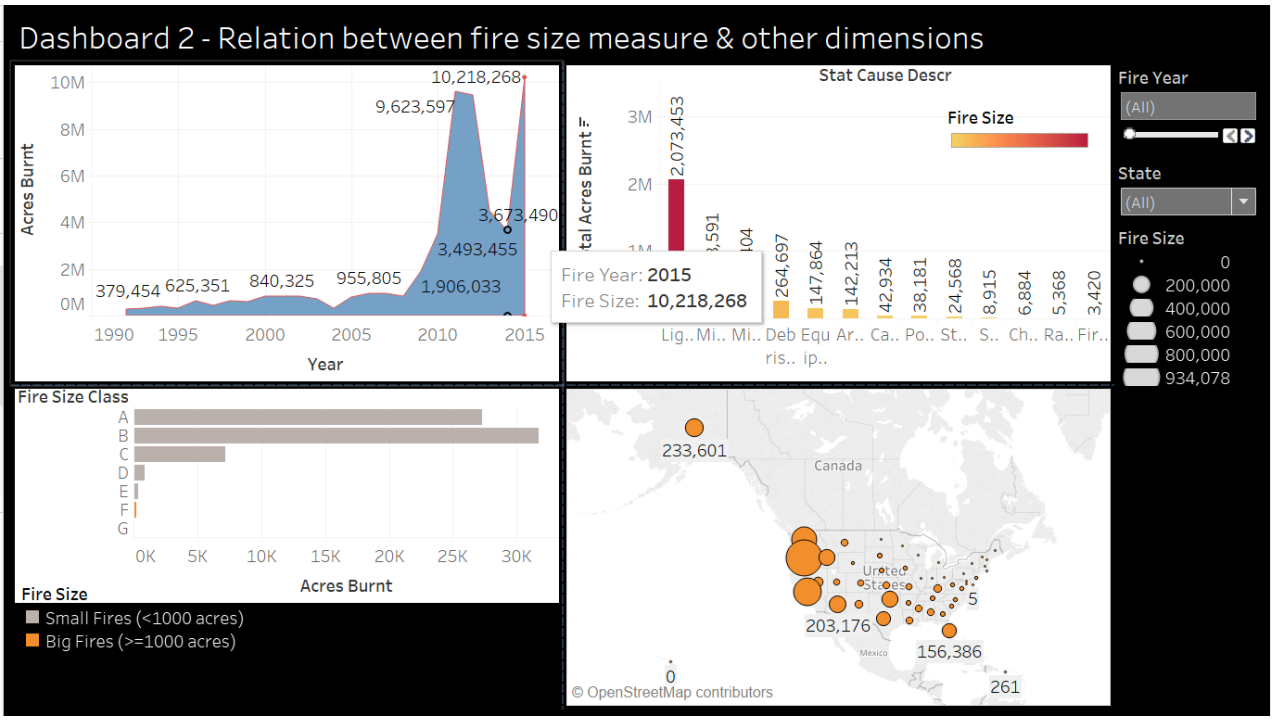
Screenshot of the Dashboard 1

Shubhangi Singhal



Secondly, I decided to further my analysis on the fire size (number of acres burnt) attribute as I believe that this is the measure of the impact of the fire, hence becomes a very important characteristic to check. The first graph on the top left is an area graph showing the number of acres burnt over these years from 1992 to 2015. We see there has been a significant increase in the fires and the sum of acres burnt. Second graph to the right is a vertical bar graph depicting the cause of the fire vs the total area burnt in acres. We see that lightening has been causing the most acres to burn even though majority of fires have been caused by human factors. The horizontal bar graph on the lower left showcases the fire size classes A, B, C, D and E which represent the classes with fires resulting in less than 1000 acres of land burning and classes F and G with fires burning more than or equal to 1000 acres. Lastly, the geo map shows the relation between the US State and the number of acres burnt in that state for that particular year. We see that most of these states lie on the west coast of the country.

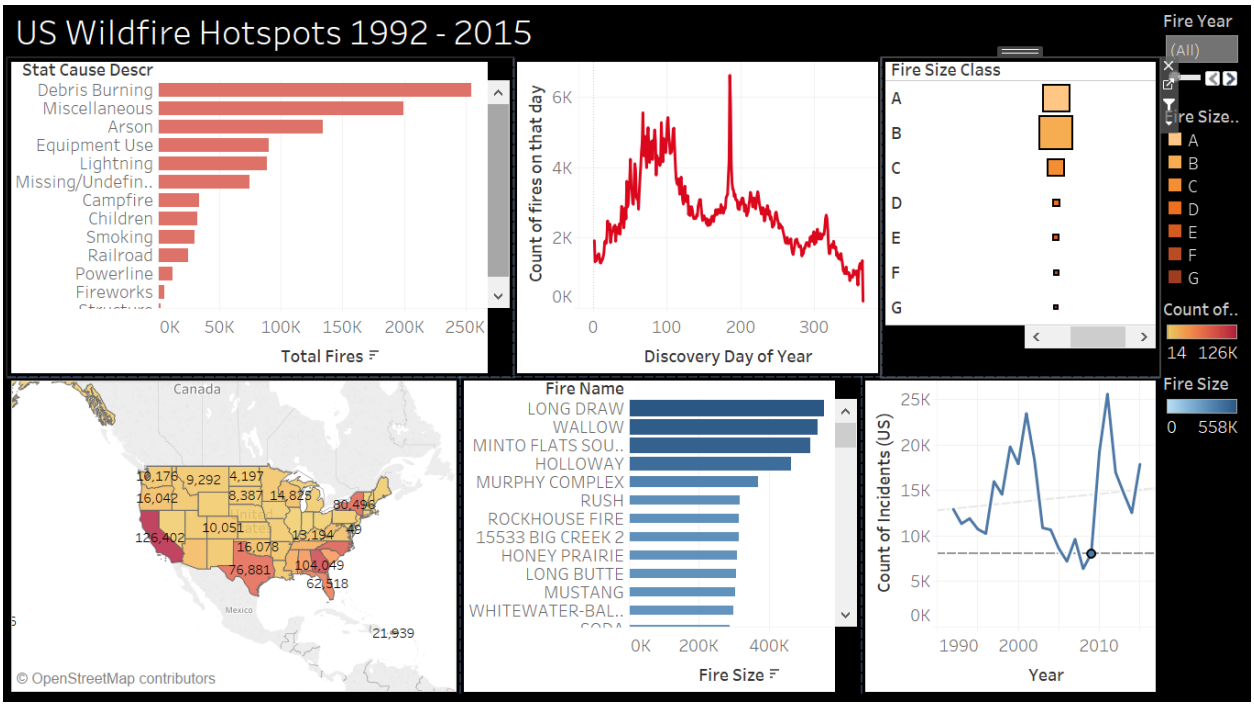
Screenshot of Dashboard 2



Dashboard 3

This dashboard is for the audience to use the Year filter located towards the right and make usage of these visualizations to understand the overall impact on the land area affected, the top causes, and the days of the year which are more prone to these onsets of forest fires. We conclude from the newly added line chart and the bar graph in the lower section of the dashboard in addition to the previous results that Long Draw, Wallow, Minto Flats and Holloway as the top 4 fires from 1992-2015. We also see that California has been the most affected by the onset of these forest fires. We also see from the second graph which is a line chart (red) that there is a peak in the fires between 100th and 200th day of the year. These days are represent the summer season and maybe because of the increased heat and dryness, these fire onsets tend to increase during this period.

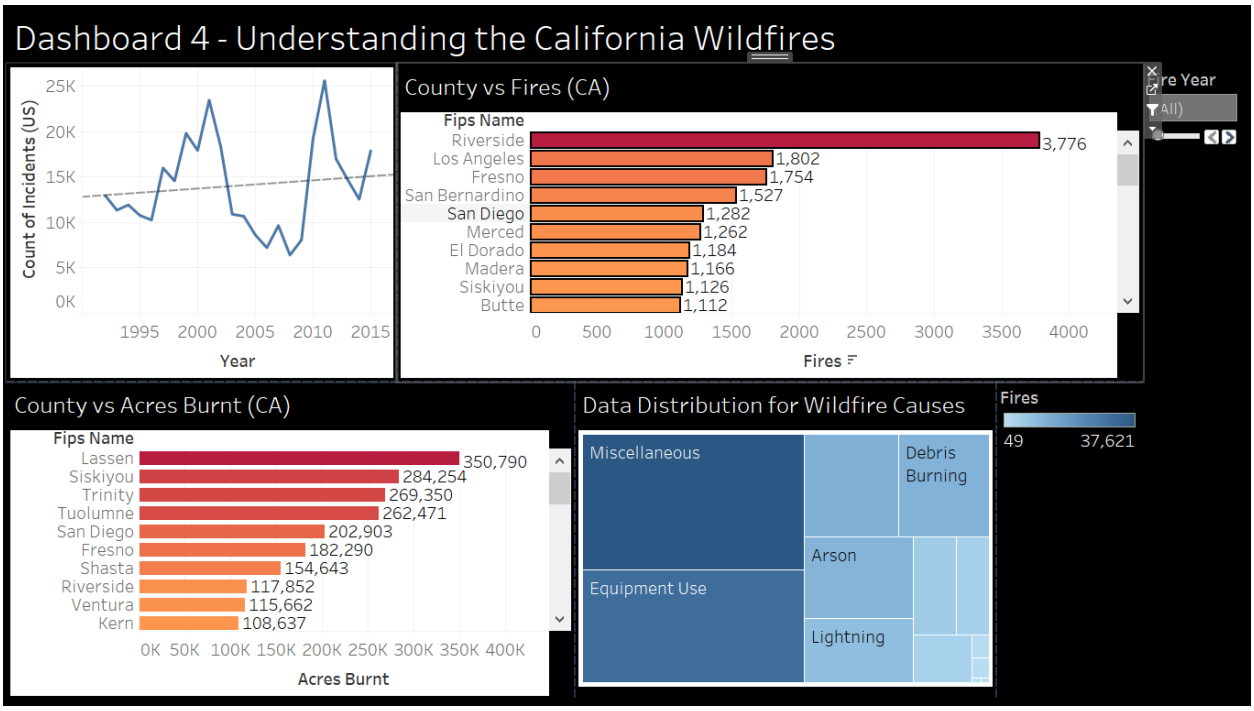
Shubhangi Singhal



Dashboard 4

This dashboard is specifically for county level detailed analysis of California Wildfires which recently had been a trending issue in the news. The user can use the slider towards the right to interact with the bar graphs and the tree map mentioned alongside. We see that Riverside and Los Angeles would the counties with the highest number of fires in the US over these years. We also find that Lassen, Siskiyou, and Trinity counties faced with the highest number of acres affected because of these wildfires irrespective of the count of fires that ensued there. One can also check the top causes for a county using the tree map below.

Shubhangi Singhal



Finally, to conclude we can say that human involvement would be the major cause behind the increment in the number of fires over these years. We also see that these fires take place mostly during the summer season but if we individually see the trend for each year there have been instances of this fire seasons expanding beyond the summer season. We can also conclude that California has been consistently the epicenter for these fires every year and a couple of other states towards the south west and south regions of the US are the secondary impact points. Hence, this analysis is a reasonable proof that human involvement and climate change has adversely impacted the nature including the increment of these wildfires.

References-

1. https://www.fs.usda.gov/rds/archive/products/RDS-2013-0009.4/_metadata_RDS-2013-0009.4.html (Metadata)
2. https://hub.arcgis.com/datasets/e4d020cb51304d5194860d4464da7ba7_0

Glossary-

ICS – Incident Status Summary
MTBS- Monitoring Trends in Burns Severity