California Wildfires Database

Aneesha Sreerama, Elena Dobryn, Ethan Reres, and George Bidgood

Northeastern University, Boston, MA, USA

Abstract

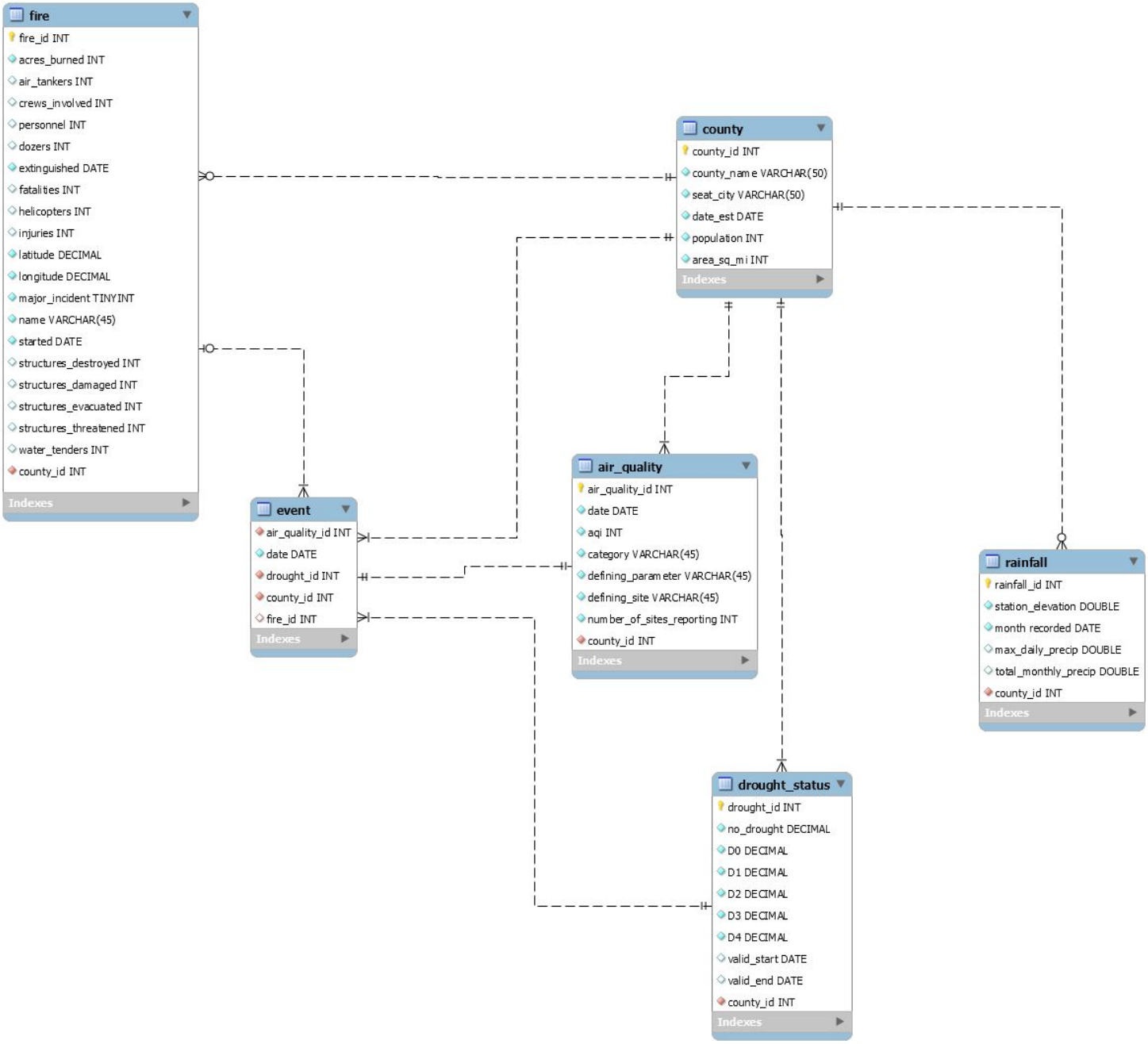
Wildfires on the West-coast of the United States have become increasingly more prevalent. Many scientists cite climate change, irresponsible civilian practices, and insufficient public policy to be responsible for the increase of frequency and severity of fires. This database was created in effort to shed light on the causation of depreciating conditions in the state of California in particular. The main objective of this project is to join together several tables based on fire data and relevant climate data in order to gain relevant and important insight on what factors allow the rapid escalation of wildfires to occur. The data may also be used as a tool for predictive modeling, to assist in scouting and proactively preparing for potential wildfires in coming years. Significant findings were successfully found with this database, such as frequency of fires, average Air Quality Indices in counties, worst places to live in-state, and risk-levels in terms of correlating Air Quality Indices with national guidelines.

Introduction

California leads the United States in number and severity of wildfires by far; in 2018, California had 800,000 more acres burned by wildfires than the second-most affected state, Nevada, for a total of 1,823,153 acres burned [1]. Furthermore, an estimated 90% of wildfires in the United states are started from human activity such as unattended campfires or downed power lines, while the remaining fires are caused by lightning strikes [2]. Given that an estimated 2 million homes in California are at high risk of wildfire damage [2] and the effect mass amounts of smoke can have on human and environmental health, this is an extremely current and pertinent issue for research.

The motivation of this project stems from the horrendous wildfire season that the state of California faced in the summer of 2020. We wanted to investigate where wildfires were occuring, quantify the escalation of the situation over the past couple years, and identify which factors may have contributed or been affected by the presence of wildfires. In order to do so we joined wildfire, drought, rainfall, and air quality data in order to gain more insight into the rise of wildfires in California. These entities may not have been put together to be examined in this way before and could lead to novel observations when analyzed. The goal of this endeavor was to determine if we can find patterns that can be used to understand how to control the fires, where to increase the number of fire personnel and to identify specific factors that may be responsible for their spread. In addition, this dataset could be used to help the average user understand the environmental risks of the county they live in and raise more awareness about the climate crisis that the state of California is facing.

Database Design



The three main tables of our database correspond to these entities: fire, drought_status, and air_quality. The county table and its attributes such as population and area was formed separately, and all other table records include a county_id foreign key to indicate location.

The fire table contains entries representing specific fires in California along with supporting details. The attributes crews_involved, air_tankers, dozers, fatalities, helicopters, injuries, water_tenders, and structures_ are all INTs as they report the number (not name) of each of these categories involved, and therefore cannot be extracted out to another table. Each fire was started in one and only one county, and each county may have zero or many fires started in.

The drought table includes the county_id and categories no_drought, D0, D1, D2, D3, and D4. Each category represents different worsening levels of drought and the record gives percent area of the county that experienced each of these categories over the specified week-long range indicated by the valid_start and valid_end date attributes. Details on the data provided for the drought, air quality, and precipitation tables are further discussed in the Data Sources and Methods section. A drought record is for one and only one specific county, and one county will have many drought records as they span one record per week for each county over the years of 2013-2020.

The air quality table follows a similar structure to the drought table, providing reports of the air quality index daily by county. It also provides the category the measurement falls in (from ‘good’ to ‘hazardous’), measurement type (defining_parameter) used to find the standardized AQI level, and the site code at which the measurement was taken. An air quality record is for one and only one specific county, and each county will also have many associated records as the AQI is reported every ~4 days for each county over 2013-2019.

The rainfall table includes the month the record was taken, county_id foreign key referencing the county table for the location the record was taken, and two precipitation recordings for that month, including maximum daily precipitation and total monthly precipitation, in inches. One of these attributes may have an occasional null instance as a result of the source we took it from, which is why they are nullable. Each precipitation record was taken from a station in one and only one county, and since data was not available for all counties, a county may have zero or many precipitation records associated.

The three main tables are linked together in the event table. A row in the event table represents the state of a particular county on a particular date. Every event has one and only one associated air quality and drought status. Since drought statuses are weekly, one drought status applies to a range of days, so it will appear one or more times in the table for its indicated county. Since there is not always a fire taking place during a specific date in a county in California, the fire field is nullable. A fire may last for one or more daily records depending on how long it took to be extinguished. The event table was created by joining the three main tables together by date and county. By creating the event table, queries requiring joins between two or more of the main tables became much faster. For instance, a three way join between fire, drought, and air quality took over 30 seconds. With the addition of the event table, the same query could be performed in under 5 seconds. For all the data, we created the tables in an sql script then ran the import wizard to load the data in from csv.

Data Sources and Methods

Wildfire data was obtained from a Kaggle dataset which scraped data on around 1,600 distinct wildfires in California from 2013-2019 from the California Department of Forestry and Fire Protection [3]. The file initially included several columns that were not needed such as text description and other administrative information. In excel we selected only the information that we determined relevant, useful, and directly attributable to the fire itself, such as name, location, acres burnt, date started and extinguished, number of crews, personnel and equipment needed, and damage caused (ex. number of affected structures), before importing the data into our database. We also extracted the county name so as just to include the county_id foreign key referencing the table of county information.

Drought data was taken from the US Drought Monitor website, with data provided from the National Drought Mitigation Center at the University of Nebraska-Lincoln [4]. The site allows selection by specific filters such as date, location, and type of statistics, so it was straightforward to select data by county in California for the date range of 2013-2020. Each weekly record details the county, the county's unique Federal Information Processing Standards (FIPS) code, and the 6 drought categories reporting the percent area of the county land that was classified in each particular state of drought during the indicated time period. The categories increase in severity from None, meaning no drought present, to D0, indicating abnormally dry (on the brink of going into a drought, coming out of a drought), to D1 (moderate drought), to D2 (severe drought), to D3 (extreme drought), to D4, indicating exceptional drought with significant crop loss, water emergencies, and fire risk.

Air quality data was found on the United States Environmental Protection Agency website from their available pre-generated data files [5]. The files provided daily records by county per year, so we had to download each year from 2013-2020 separately. Each record includes an Air Quality Index (AQI) measurement and category in which the measurement falls, from "Good" to "Hazardous". AQI is determined by the amount of fine particulate matter in the air. Higher AQI levels indicate more pollution and lesser overall air quality. Multiple different measurements of air particles and pollutants are taken about every three days, and the highest average measurement out of the categories is selected to represent the daily reading. The record may be obtained from measuring PM2.5 (concentration of particles less than 2.5 microns in diameter) or PM10 (concentration of particles less than 10 microns in diameter) in micrograms per cubic meter, or concentrations of pollutants such as ozone (in parts per million), carbon monoxide, sulfur dioxide, or nitrogen dioxide (in parts per billion) [6]. The concentration of each pollutant is normalized to a standard AQI index scale and corresponding AQI category that are used as the records in the table.

Precipitation data was collected from the National Centers for Environmental Information climate data online search tool [7]. The site has limited automatic filtering tools to get the data needed so manual selection of stations in each county in California with data from the time range of 2013-2020 was necessary, and unfortunately not all counties had available data. We chose monthly data as daily precipitation records would likely yield a large amount of 0 total daily

rainfall due to the dry conditions. The stations collect a large amount of climatological data, and we initially selected both temperature and precipitation data but upon examination of the csv file, very few of the stations had non-null records for the temperature readings. In addition, not all stations had records for every month within the time frame indicated. Each record includes precipitation data for the highest daily precipitation for the month, and total monthly precipitation, both in inches. The FIPS code or county name was not provided for the records, so manual search by station location and insertion of the FIPS code for the records was done in excel before importing the data.

Lastly, a list of California counties and information about the counties including established date, seat city, population, and area was manually compiled in excel from information provided by Wikipedia [8], the California State Association of Counties [9], and the US Census Bureau [10].

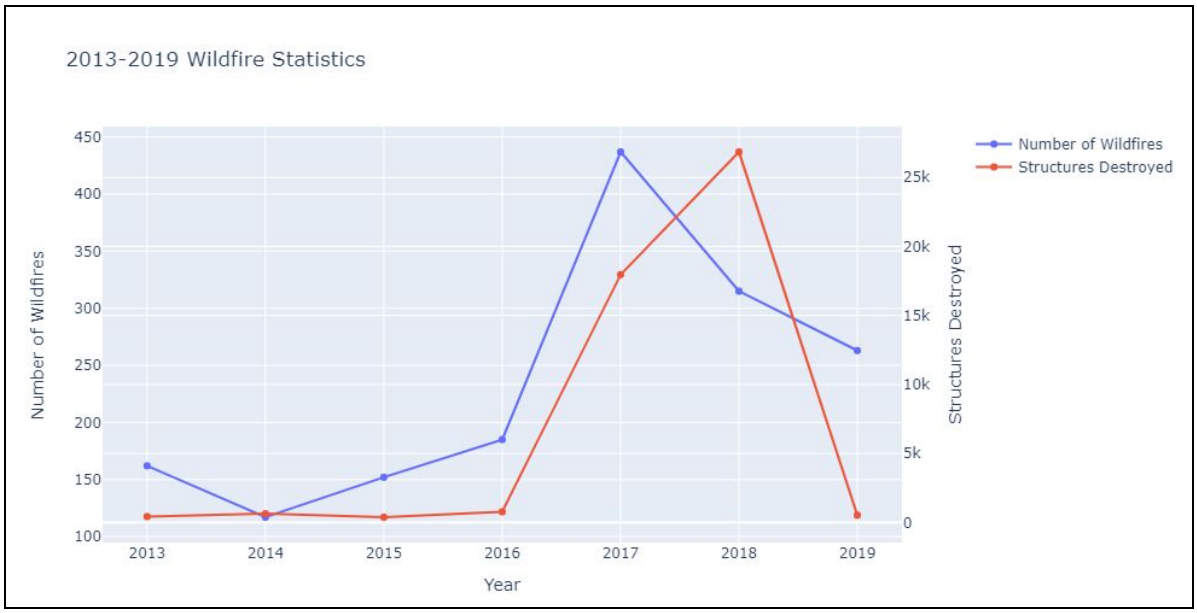
Use Cases

Non-trivial questions that users could ask of our database include the temperature, air quality, drought, and wildfire statistics broken down by each individual county, date, or year.

Queries that Summarize the Dataset:

Number of Structures Destroyed by Year

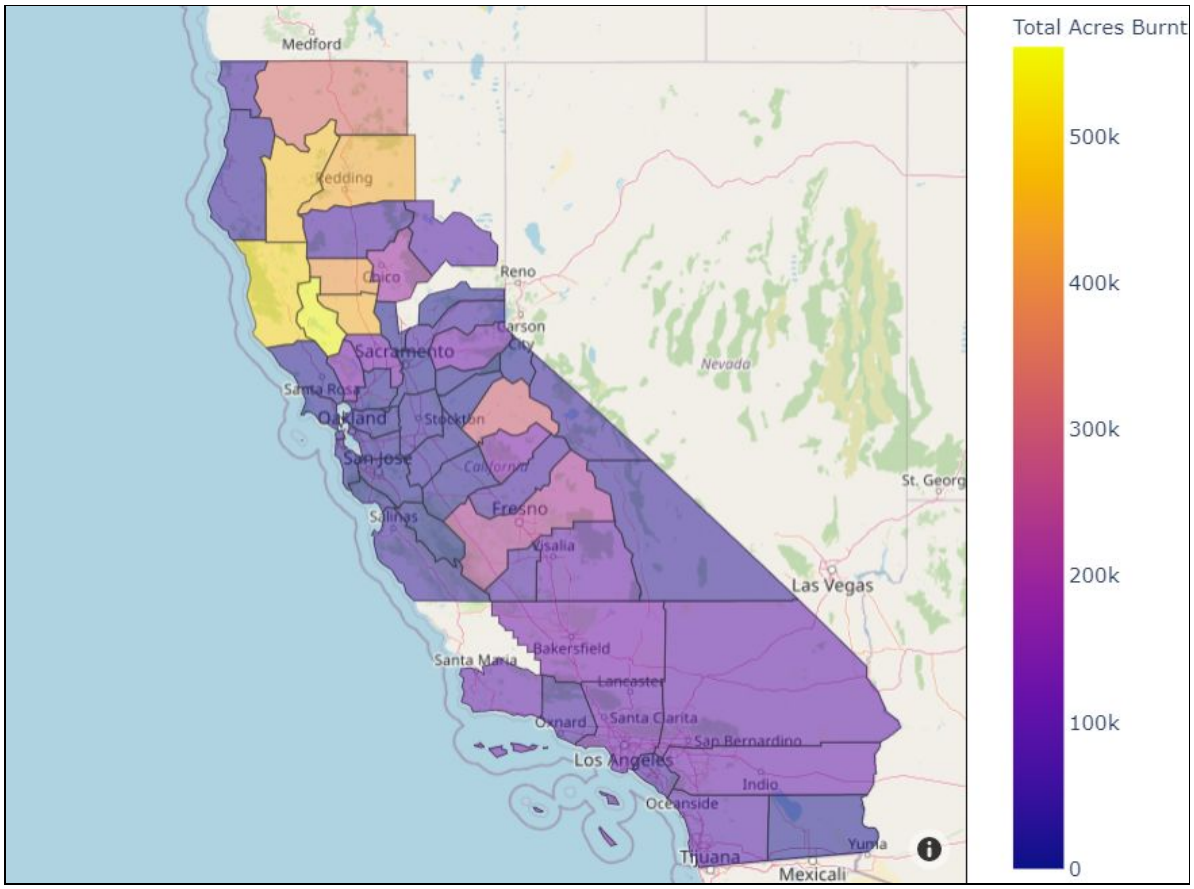
```
Select year(started), sum(structures_destroyed), count(*)
From fire
Group by year(started);
```



Summary: The line graph shown above displays the number of wildfires and structures destroyed for every year between 2013-2019. Our original hypothesis was that the number of wildfires were steadily increasing by the year. However, the data showed that the number of wildfires peaked in 2017. Moreover, when plotting the number of structures destroyed, we found the peak to be in 2018, as opposed to 2017. This realization prompted us to look deeper into what caused this difference and what may have influenced it.

Total Acres Burnt between 2013-19 by County

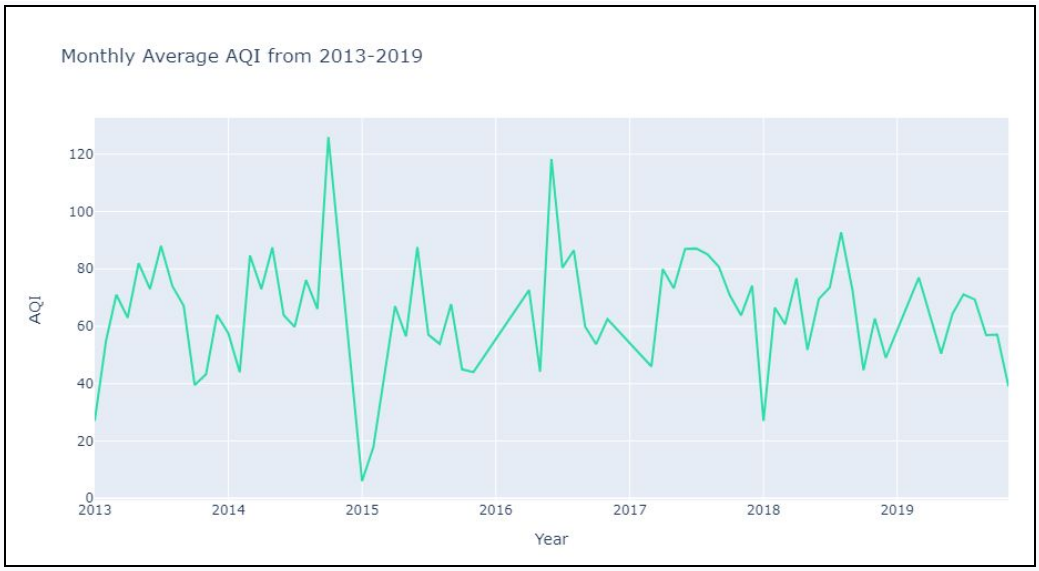
Select county_name, ifnull(sum(acres_burned), 0) as ‘total acres burnt’
From fire join county using (county_id)
Group by county_name;

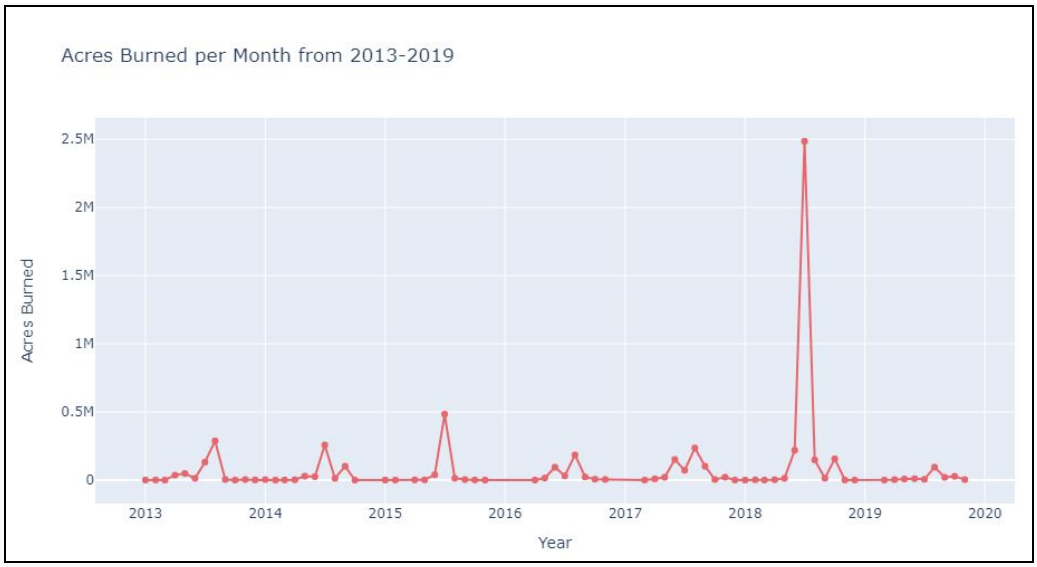


Summary: The image above is a map of the state of California broken down by county. The counties in Northern California had significantly more fires than any other part of the state. In addition, we noticed that the counties with high populations were the source of a significant number of fires and that this could be responsible for the numbers of structures destroyed in 2018. (Entries with null values for acres_burnt were removed for the purposes of producing the visualization)

Time Series Analysis of monthly average AQI for california vs. Total acres burnt from 2013-2019.

Select year(date), month(date), avg(aqi), ifnull(sum(acres_burned), 0) as 'acres burned' from air_quality a
left join fire f on (a.county_id = f.county_id and a.date < f.started and adddate(a.date, interval 4 day) >= f.started)
Group by year(date), month(date);



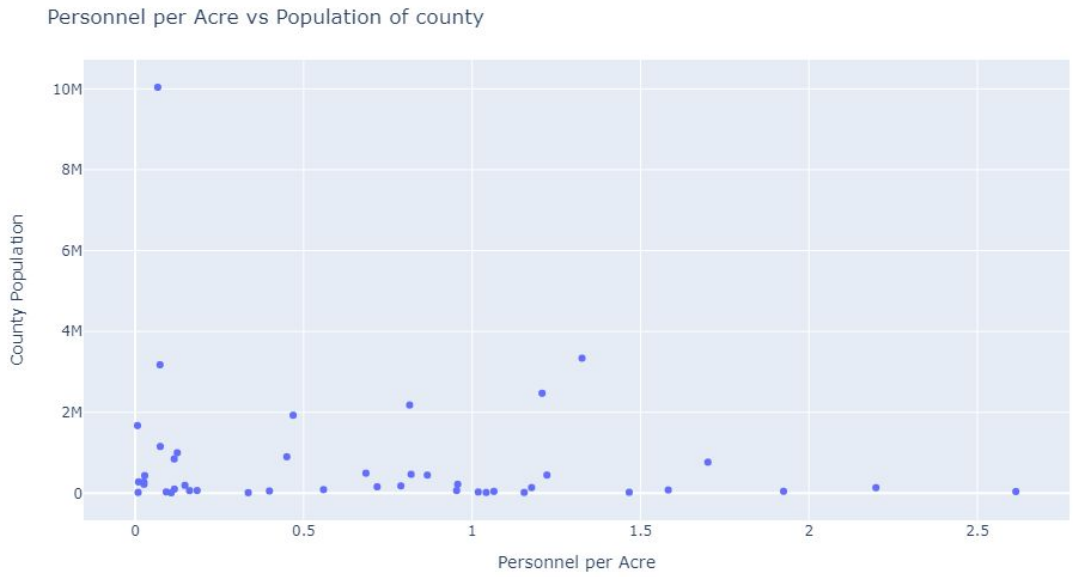


Summary: The spike in acres burned per in 2018 corresponds to the time when the Camp fire burned. It was the deadliest and most damaging fire that occured in California and 6th deadliest US wildfire in history [11]. There is also a spike in AQI levels around the same time, perhaps attributable to the damage caused and smoke levels from the fire. It is interesting to note the

clear spikes acres burned, and less obviously (but still present) in AQI, in the middle of each year, during the summer months.

Personnel per Acre vs Population Size

select counties, avg(personel/acres_burned) as 'personnel_per_acre', population
from fire
join county using (county_id)
where personnel is not null
group by counties
order by personnel_per_acre desc;



Summary: As can be observed from this plot, some sparsely populated counties tended to have more personnel involved in containing and extinguishing the fires. Less populated and less developed land with larger areas of forest and grassland causes fires to spread faster [12]. This fast spread and the land possibly being harder to access for containment measures, may contribute to the requirement for bigger crews for successful fire management. Los Angeles county with the highest population of about 10 million, actually devotes less than 0.5 personnel per acre burned. This is interesting as LA is highly affected by fires every year and due to the population density many structures are often damaged, so a higher number of devoted personnel would be expected. It is possible that the crews involved in fire containment in LA county are highly strategized and efficient at containing fires, so as many personnel are not needed. However, given that an estimated 12 homes and 21 buildings were destroyed in just one fire in LA this year [13], perhaps more crews or better strategies are needed.

Investigative Queries:

When is the worst time of year to be living in california? At what time of year do the most fires and highest AQIs happen? When do the least fires happen?

```
select month(started) as 'month', count(distinct name) as 'num_fires'
From air_quality a
left join fire f on (a.county_id = f.county_id and a.date < f.started and adddate(a.date, interval 4 day) >= f.started)
group by month
order by month;
```

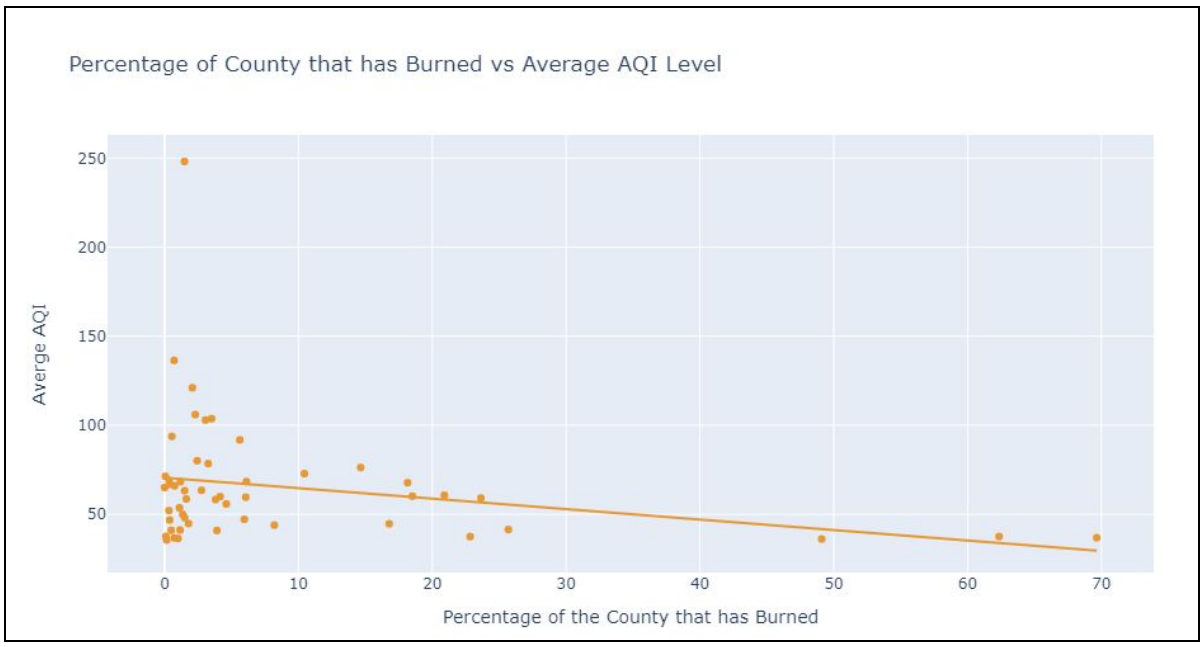
Output:

Month	# of Fires	Average AQI
1	10	53.2792
2	7	65.4655
3	6	52.234
4	33	89.4036
5	128	81.3723
6	260	76.8047
7	317	72.0915
8	223	66.8564
9	163	61.2899
10	108	60.4402
11	36	54.7545
12	17	55.5327

Summary: The summers in California seem to experience the worst AQIs as the number of fires generally increased towards the middle of the year. June and July particularly see an increase in about 100 fires from the months prior. This reflects increasing temperatures and dryness during the summer. January - March have the least number of fires, October-March have the lowest AQI among the other months of the year. So, if you want to avoid any fires it would be best not to visit California in July, and March is the safest bet if you want to be breathing in the healthiest air *and* see the fewest fires.

Where is the worst place to live in California? In terms of worst average air quality, drought, and % area burned (1 acre = 0.0015625 sq. miles)

Select county_name, ((sum(acres_burned)*0.0015625) / area_sq_miles) as 'percent_area_burned', avg(aqi)
from fire f join county c on (f.county_id = c.county_id)
join air_quality a on (a.county_id = f.count_id and a.date < f.started and adddate(a.date, interval 4 day) >= f.started)
Group by county_name
Order by 'percent_area_burned', avg(aqi);

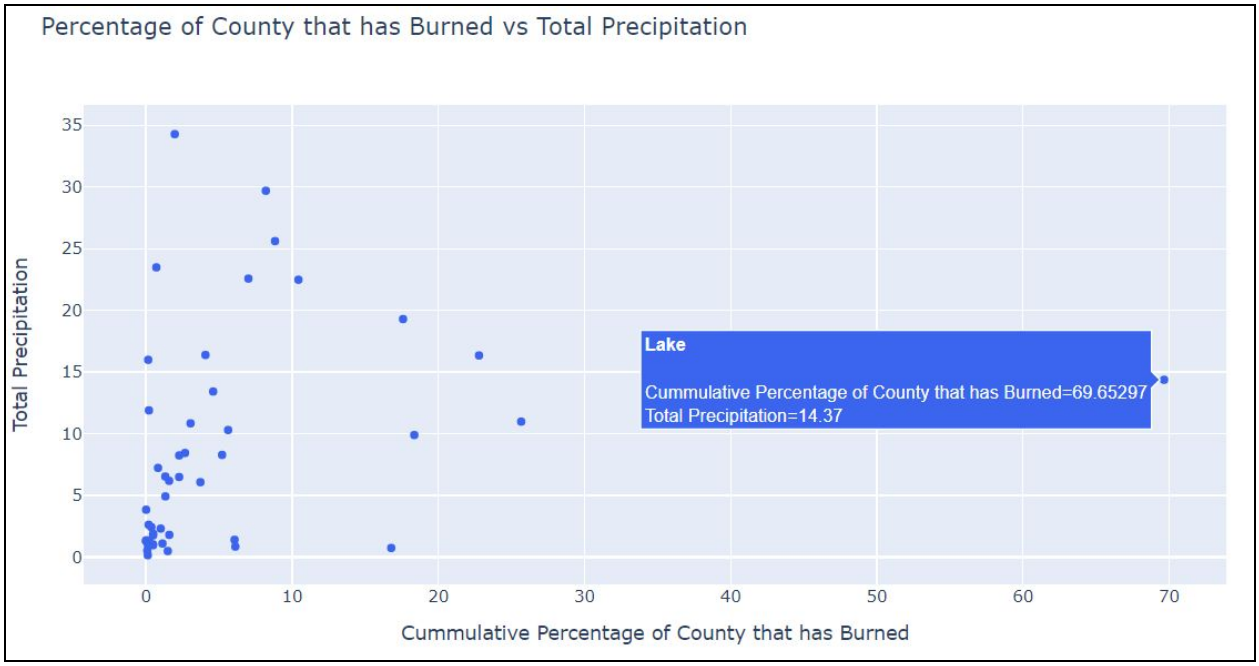


Summary: The number of acres burned over the county size is the percent of the county area that was burned. Each dot represents a county. Over this time frame there were likely different fires that burned over the same area, and also fires may spread over multiple counties. It is shocking that a few counties had around 50% or higher of their land area burned over this time. 6 counties also have an average overall AQI that is above 100, in the category of 'unhealthy for sensitive groups', which is concerning. The trend line included does not indicate a high correlation between percent area burned and average AQI. There is a large distribution of counties that had between 0 and 10% of their area burned over the 7 years, and they show a wide range of average AQIs. AQI is likely subject to many factors and the number of acres burnt may net be directly related to the Average AQI. For example, where the fire burns and what materials are being burned likely contribute to what is being released into the atmosphere as a result. Many of the counties with larger fires tend to be sparsely populated, with denser forestry which may contribute to the lower AQI. Fires burning in more populated areas with commercial

buildings and houses, and other made man structures will likely release more hazardous particles into the air when burned.

Do the counties with the highest rainfall have less impactful fires? Can we compare Acres Burned per Square Mile and total Precipitation?

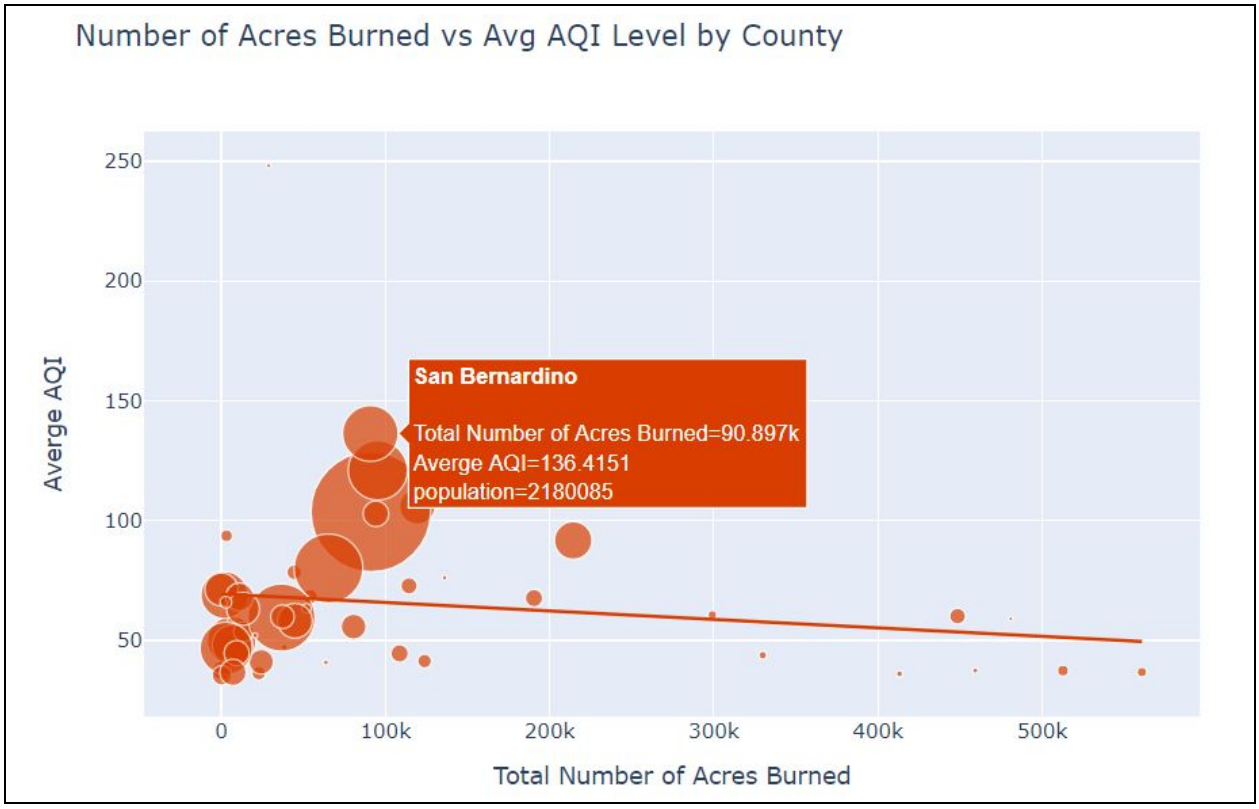
```
select county_name as 'County Name', sum(acres_burned)/area_sq_mi as 'Acres Burned per Sq Mile', sum(total_monthly_precip) as 'Precipitation'
from event e
join county c using (county_id)
left join fire f using (fire_id)
join rainfall r on (r.county_id = e.county_id and e.date > month_recorded and e.date <= date_add(month_recorded, interval 4 day))
group by county_name;
```



Summary: As shown in this visualization, the precipitation amount has minimal correlation with the amount of acres burnt per square mile. Lake county is the most extreme outlier, as it has a slightly higher than average total precipitation (in inches) in the state, yet has the most acres burned per sq. mile out of any county on the graph. Precipitation likely has little actual effect on the breadth, let alone severity of an instance of wildfire.

What is the average AQI vs Acres burnt in each county? Is there any correlation between the two variables?

```
Select county_name, sum(acres_burned), avg(aqi), population
from event e
join county c using (county_id)
left join fire f using (fire_id)
join air_quality a using (air_quality_id)
group by name, county_name;
```



Summary: This shows that there is actually very little correlation between AQI and a fire that occurs within the county. Densely populated areas are bound to have significantly higher AQIs, and areas with close proximity to fires, regardless of county boundaries, are bound to get affected by a wildfire if it is in trajectory of an area's wind pathing.

Do the Top 15 counties with the highest average percent area in exceptional drought (d4) have more fires over this time period?

```
Select county_name, (sum(D4) / count(drought_id)) as 'avg_percent_area_in_D4',
count(distinct fire_id)
From county join drought using (county_id)
Join fire using (county_id)
Group by county_name
Order by avg_percent_area_in_D4 desc
Limit 15;
```

County	Avg % Area in D4	# Fires
Kings	36.6562	5
Santa Barbara	36.0605	26
Tulare	35.9758	35
Ventura	34.2397	27
Fresno	33.109	54
San Benito	31.4044	17
Madera	30.3462	35
Mariposa	29.54	32
Los Angeles	27.9927	41
Merced	27.6416	14
Kern	27.5981	58
Monterey	24.845	39
Stanislaus	23.9322	17
Orange	23.6295	9
Tuolumne	22.7627	19

Summary: This question was formed with the idea that drier conditions may lead to more fires starting in that location. Once again, there appears to be little correlation between drought levels and the number of fires started in each county. The average number of fires started is 28.53 with a standard deviation of 15.56 so the data is not falling within a consistent range. Furthermore, these 15 counties have a comparable average number of fires started as the

whole set of 58 counties (average 28.22 fires started per county). There are likely other more specific factors that contribute to how many fires get started in a county.

Conclusions

The goal of our project was to create a database to support queries about the wildfires in california and write queries that provide insight into those fires.

We constructed a relational database using mySQL to model the links between entities such as droughts, air quality, precipitation, and wildfires. We modeled an event as a particular date in a particular county associated with an air quality reading and a drought status. Additionally, an event has an optional connection to a fire if there was a fire in the county at the time. The event table was populated by joining data from the fire, air quality, and drought status tables on the county and date values. This table made queries requiring two or more joins significantly faster and allowed for more coherent queries. A county table was also added to the model to contain general information regarding a county such as population and area, which were used to provide context to other queries, for example, calculating the percentage of a county that burned over a certain timeframe.

We collected a significant amount of data for our database, including fires that occurred in California from 2013-2019, daily air quality readings, weekly drought conditions, and monthly precipitation levels The collection of records by date and location allowed us to examine trends relating to fires over time and by location. One challenge we faced when using this data was the range of dates; the daily and weekly readings, plus start and end dates in which fires occurred, did not always line up to an exact date so we had to make sure we were selecting in a range of dates and eliminating possible duplicates.

The main questions that we wanted to address were how to control the fires, where to increase the number of fire personnel, and to identify specific factors that may be responsible for their spread, and we think that while our project did not exactly solve these issues, we took a step towards it, which in itself is significant.

We were able to successfully create queries that utilized our tables and attempted to find relevant information about wildfires and their causation. These findings were able to be modeled into visualizations such as graphs, tables, and scatterplots for easy interpretation. However, we were surprised at the results of some measurements: many of our preconceived hypotheses were disproven with our findings, and many queries even represented a lack of correlation in data comparisons. The scope of fires and their patterns and predictability proved to be more complex than we thought, and likely is reflected in many other factors that we did not incorporate in our database. This is not entirely disheartening because if it was that easy to find attributable patterns and causes of fires, this would be reflected in the better overall management and control of them. In the future, we may want to tighten the scope of our data even further by increasing the specificity of some sets; focusing on more in-depth environmental conditions such as topography, climate zone, wind patterns, and others may lend more towards

the characteristics of a fire and its severity. Focusing on the impact of fires on towns and cities is recommended as well; after a fire a tree can grow back but houses and buildings cannot. Building houses out of fire-resistance materials is already a surfacing strategy in California [14], so it would be interesting to look at the change in damage of structures relating to the material the buildings are made of. We can also decide to extend data towards socioeconomic factors, such as rent, average income in a county over time, and demographics, to see if wildfires have a significant impact on the overall quality of life for the human population in certain areas. Perhaps it can assist in public policy advocacy in the future, to help crack down on unhealthy and destructive habits that corroborate the influx of fires.

Author Contributions

Aneesha: Aneesha conceptualized the initial idea of the project and guided its goal. She sought out and found relevant data sets to incorporate into the database. She played a large role in cleaning the datasets and transforming the data. She created ideas for insightful queries and then visualized them using plotly, mapbox, geojson, and pandas.

Elena: Elena also helped to incorporate data from various sources into the project, mainly the drought table, counties table, and precipitation table, which required a lot of manual searching and clean up. She proposed interesting queries and wrote the required SQL statements to realize them, and also analyzed their results and significance.

Ethan: Ethan cleaned a large amount of the data used in the project: from eliminating inconsistent values to removing logical inconsistencies that would have created issues down the line. He also assisted in some ideas for potential SQL queries that could be significant to the overall objective of the assignment. General formatting of reports and presentations were handled by him.

George: George had a major role in the design of the database, both in determining the requirements and in implementing the DDL to satisfy those requirements. George also handled the population of the database with values from the various datasets including adding primary and foreign keys to the original datasets.

References

[1] “Fire Information.” *National Interagency Fire Center*, www.nifc.gov/fireInfo/fireInfo_statistics.html.

[2] “Facts + Statistics: Wildfires.” *III*, www.iii.org/fact-statistic/facts-statistics-wildfires.

[3] Ares. “California WildFire Incidents (2013-2020).” *Kaggle,* 9 Feb. 2020, www.kaggle.com/ananthu017/california-wildfire-incidents-20132020.

[4] “Comprehensive Statistics.” *U.S. Drought Monitor*, *National Drought Mitigation Center.* https://droughtmonitor.unl.edu/Data/DataDownload/ComprehensiveStatistics.aspx.

[5] “AirData Website File Download”. *Environmental Protection Agency*. aqs.epa.gov/aqsweb/airdata/download_files.html.

[6] Cardinal, David. “How Air Quality and the AQI Are Measured.” *ExtremeTech*, Ziff Davis, 21 Nov. 2018, www.extremetech.com/electronics/280956-how-air-quality-and-the-aqi-are-measured.

[7] “Climate Data Online Search.” *National Centers for Environmental Information (NCEI).* www.ncdc.noaa.gov/cdo-web/search?datasetid=GSOM.

[8] “List of Counties in California.” *Wikipedia*, Wikimedia Foundation, 15 Nov. 2020, en.wikipedia.org/wiki/List_of_counties_in_California.

[9] “California County History.” *California State Association of Counties*, www.counties.org/county-history.

[10] “California Counties by Population.” *California Outline*, www.california-demographics.com/counties_by_population.

[11] “Camp Fire - 2018 California Wildfires.” *The United States Census Bureau*, 9 July 2019, www.census.gov/topics/preparedness/events/wildfires/camp.html.

[12] Siegel, Ethan. “The Terrifying Physics Of How Wildfires Spread So Fast.” *Forbes*, Forbes Magazine, 6 Sept. 2017, www.forbes.com/sites/startswithabang/2017/09/06/the-terrifying-physics-of-how-wildfires-spread-so-fast/?sh=134e70727791.

[13] Garrova, Robert, et al. “What We Know About The Wildfires Burning In LA County.” *LAist*, laist.com/latest/post/20200817/la-county-wildfires-updates-august-17.

[14] Simon, Matt. “California Fires: We Know How to Keep Cities From Burning.” *Wired*, Conde Nast, www.wired.com/story/cities-have-turned-into-fire-bait-but-we-can-fix-them/.