Tackling the HOT Topic of Wildfires

By: Data Nerds

~ Azaan Barlas, LaKeisha Bridges,
Khushi Gandhi, Carol Gong, Rishab Shah

Project Purpose

- Changing weather patterns resulting in increased risk of natural disasters
- Today's effects of climate change

Research Questions:

- Is the data correlated? How? What is significant about this data at first glance?

- Can we find out how correlated the data are to each other? Can we find out which columns are best to predict wildfires? Can we determine that the weather increase caused the increase in fires?

- Can we accurately predict wildfires with this data? After analyzing the data, is there anything we learned that might help solve this problem?

Project Overview

Hypothesis

Climate change is linked to wildfires through increase in temperatures

Approach

Looking at details on global temperatures and extreme weather

Project Focus

Analysis is focused on California due to an abundance of wildfire data

Data Sets

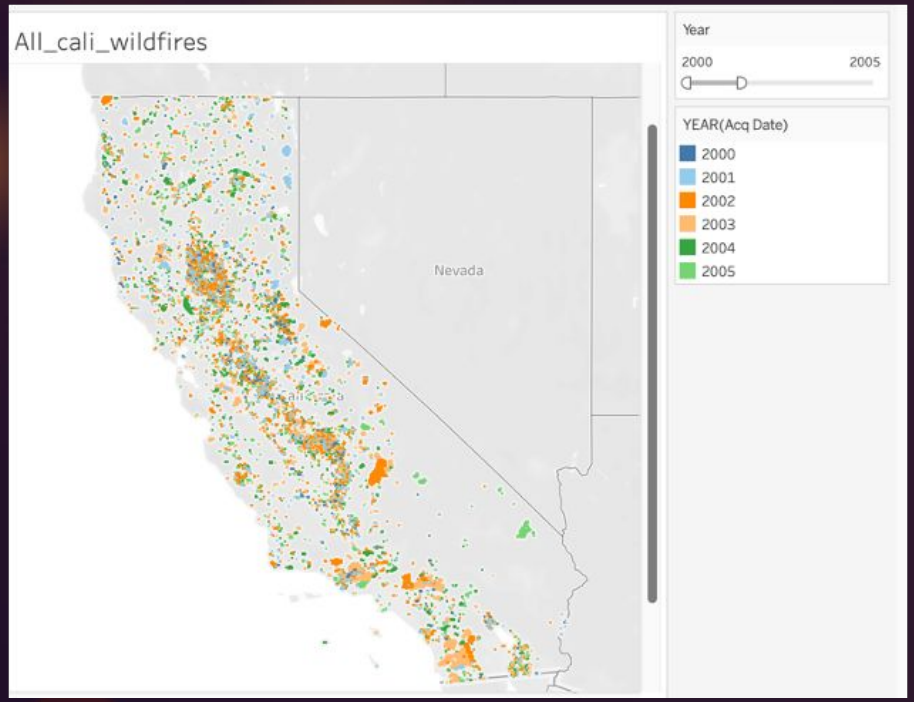
7 columns, ~30,000 records

Table	Attribute Name
Weather	Date
Weather	MaxTemperature
Weather	MinTemperature
Weather	AvgTemperature
Weather	AtObsTemperature
Weather	Precipitation
Weather	Snowfall
Weather	SnowDepth

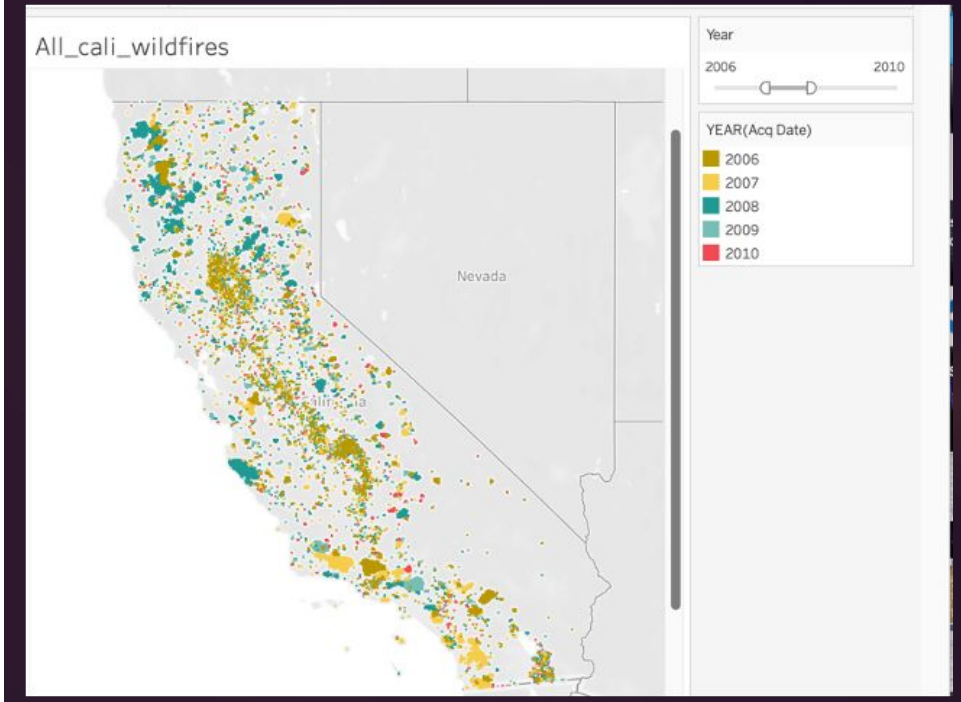
11 columns, ~220,000 records

Table	Attribute Name
Wildfire	date
Wildfire	year
Wildfire	month
Wildfire	latitude
Wildfire	longitude
Wildfire	acq_date
Wildfire	satellite
Wildfire	instrument
Wildfire	frp
Wildfire	type
Wildfire	bright_t31

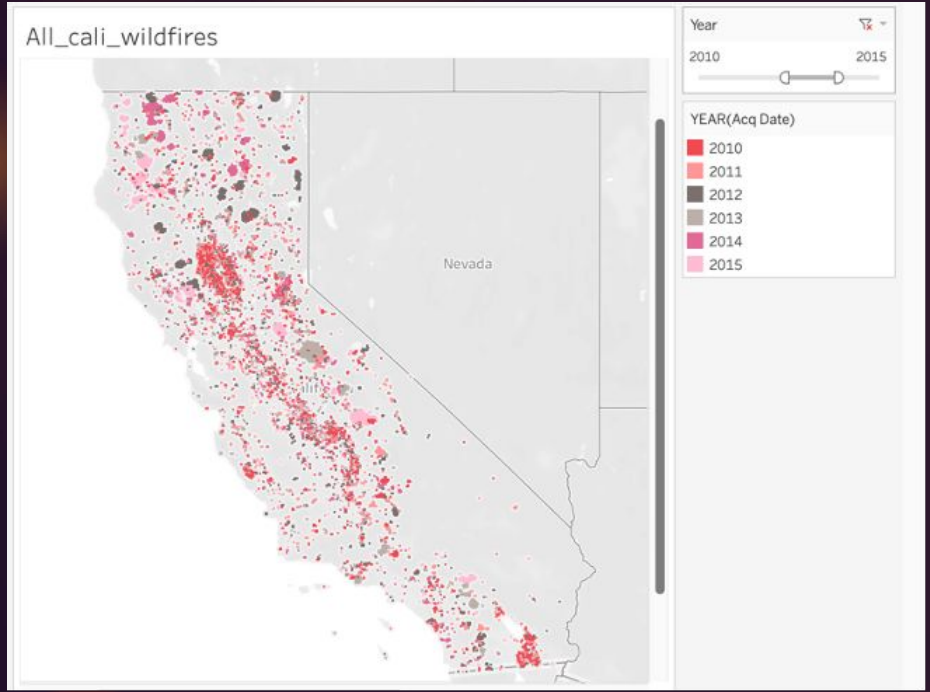
2000 - 2005



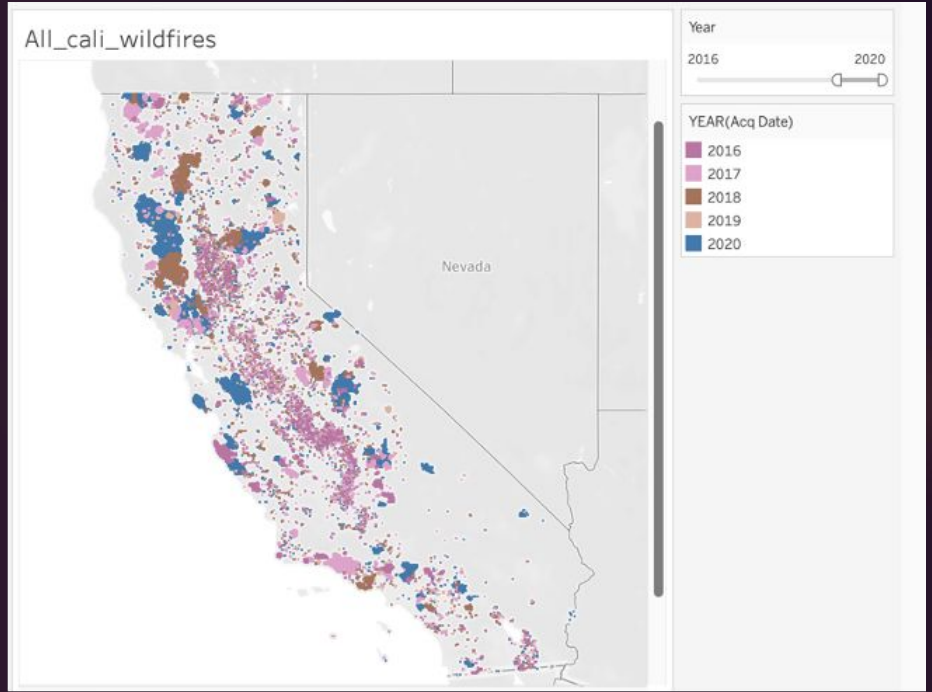
2006 - 2010



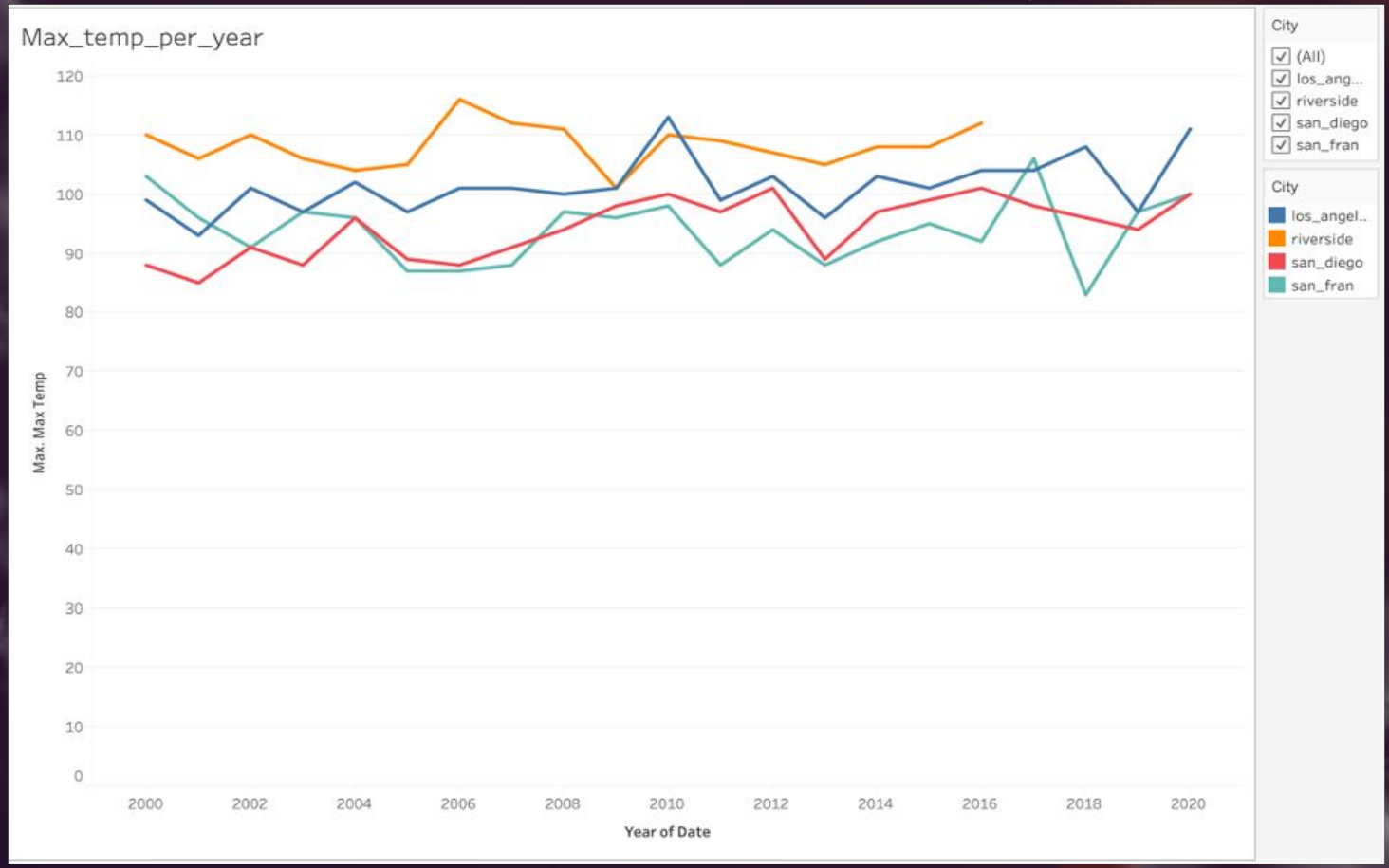
2011 - 2015



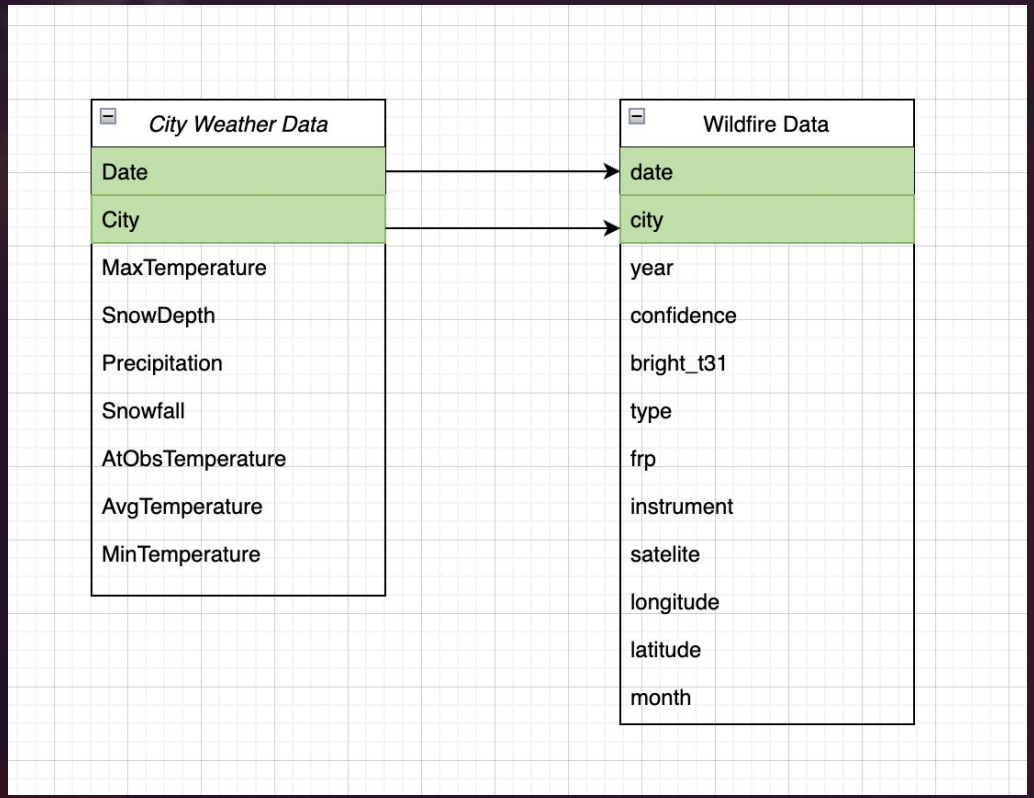
2016 - 2020



Extreme temperature insights



Data Dictionary & Schema

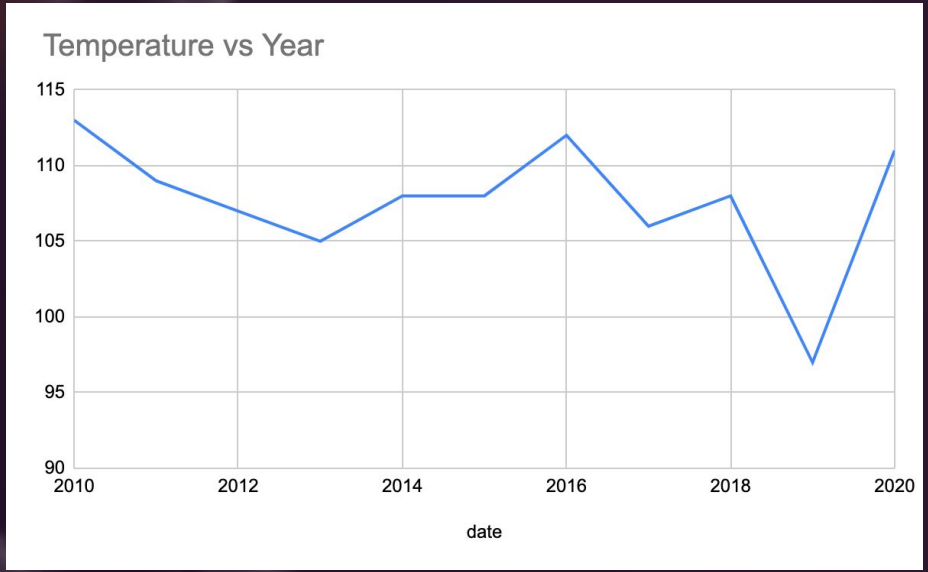


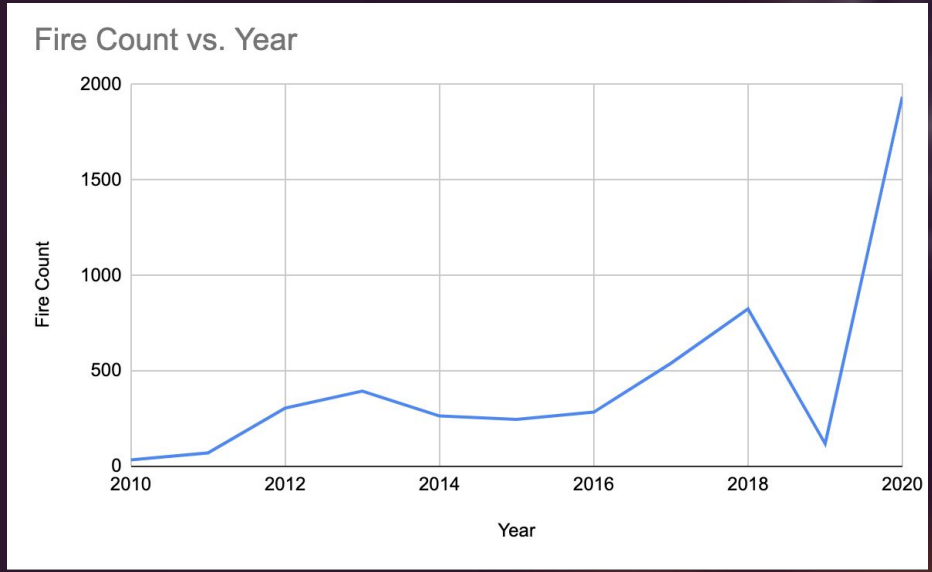
Data Dictionary

Table	Attribute Name	Description	Datatype
	record_#	Record number	Numeric
	date	Date data was recorded	Date
Wildfire	latittude	Latittude of location where satellite data was acquired	Numeric
Wildfire	longitude	Longitude of location where satellite data was acquired	Numeric
Wildfire	satellite	Name of the satelite that acquired data	Text
Wildfire	avg_frp	Avereage Fire Radiative Power; measurement of heat energy released from fire, or strength, in Megawatts (MW)	Numeric
	surface_temp	Surface temperature of location where satellite data was acquired	Numeric
Wildfire	confidence	This attribute represents the algorithmic confidence level that a pixel is a fire pixel. The confidence range is between 0 and 100, with 100 being the highest confidence. A higher confidence level indicates that the pixel is more likely to be a fire pixel.	Numeric
Wildfire	fire_count	Number of fires in a day	Numeric
Weather	max_temp	Maximum temperature recorded for a day	Numeric
Weather	rain	Rain levels of a day measured in inches	Numeric
Weather	city	Name of city in which data was collected	Text

Date

Interesting Initial Takeaways





Data Manipulation & Preparation

```
data1 = data1[data1['confidence'] > 80]
data1.count()
```

Limited data to high confidence for relevance. Accumulated 100 wildfires with highest average fire power.

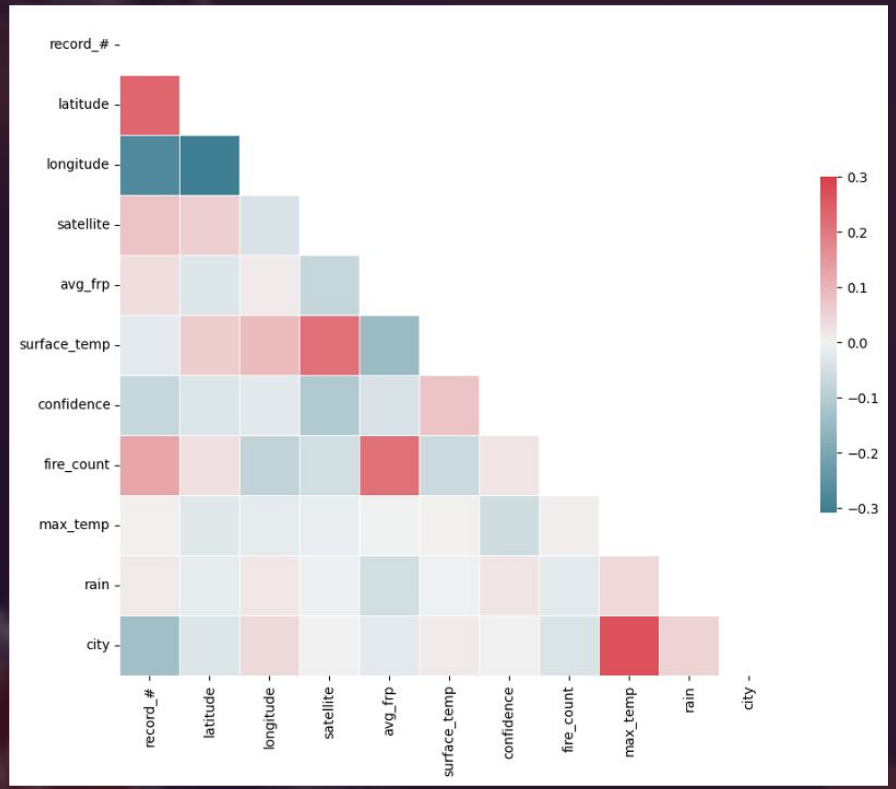
acq_date	
2000-11-02	4
2000-11-03	7
2000-11-05	2
2000-11-06	1
2000-11-07	18
	..
2020-12-27	1
2020-12-28	3
2020-12-29	2
2020-12-30	5
2020-12-31	1

```
# scratch
print(data1['frp'].nlargest(100))
# print(data1['instrument'].value_counts())
# data1.plot.hist(by=['frp'], bins=12)
data1['confidence'].mean()
data1['frp'].mean()
data1['frp'].median()
data1.groupby(data1['acq_date'].dt.year)['bright_t31'].mean()
# data1.groupby['bright_t31']
# data1
print(data1.count())
(data1['confidence'] > 70).sum()
# (data1['confidence'] == 60).sum()

# the mean is higher than the median meaning positively skewed
# meaning that there are some HUGE wildfires!
```

109928	11944.2
368	11800.8
449	11528.3
27983	11488.0
27804	11486.3
	...
9595	5553.4
30996	5552.7
30304	5531.4
101812	5527.5
142352	5498.6

Descriptive Statistics



ANOVA	P-Value < 0.05 F-Value > 1000 Significant

	PCA
1	0.26465734
2	0.1664484
3	0.114626
4	0.09644437
5	0.09270553

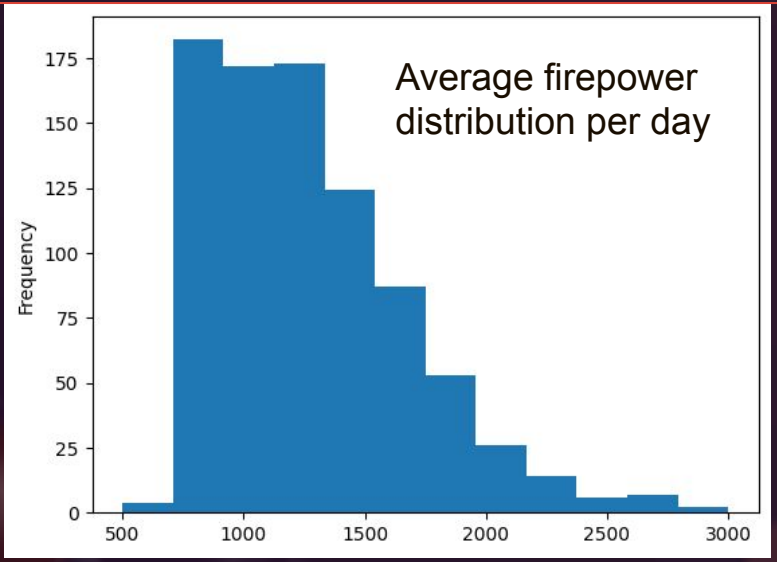
Date

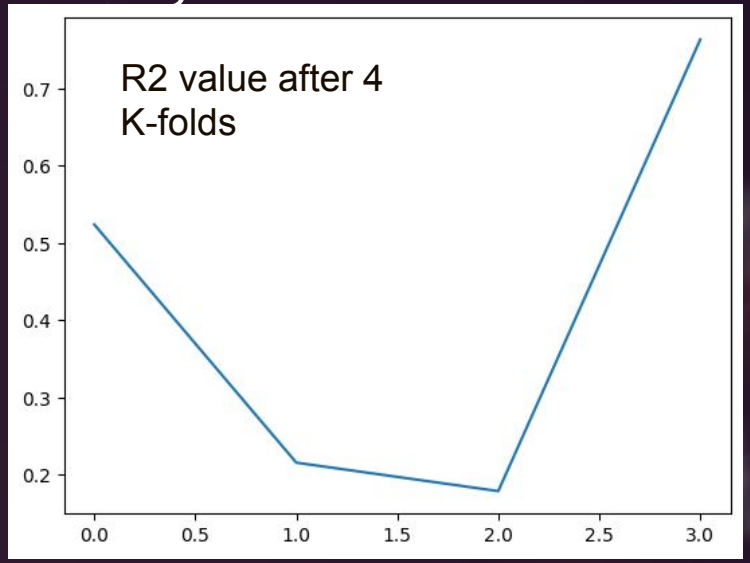
Regression Analysis

Distribution of firepower showed left-skewed; we used this as dependant variable for logistic regression

Binned these into outputs into 3 labels based on a threshold, from “low, medium, and high” intensities

Mean square error was low (good) for this around 0.5, but r2 was also around 0.5 so semi-accurate model





Used a Random Forest model, to predict wildfire counts per day. Some regression metrics from our Random Forest:

r2: 0.7634908613408767
mse: 12.389888853189625
mse: 153.50934579439252

We are accurately able to predict the count of wildfires with our existing model

Regression Analysis

column	importance
avg_frp'	0.22305946
surface_temp'	0.13883825
satellite'	0.12918465
latitude'	0.11165353
longitude'	0.10985316
max_temp'	0.09461917
record_#'	0.07824742
city'	0.07551562
confidence'	0.03625028
rain'	0.00277845

Forecasted Year	Strong fire count per year
2021	1941
2022	2017
2023	1817
2024	2003
2025	2350