

---

---

# California Wildfire Predictor

By Aanvi Goel

---

# Overview

- In 2021 there were 5267 wildfires started across the state of California, burning 204,921 acres of land, displacing locals from their homes, causing property destruction, poor air quality and other consequences
- Predicting the size of wildfire would help the fire departments and the Forest Services in deploying appropriate resources right at the time of fire discovery, leading to quicker containment times and reduced environmental and infrastructural damage

# Project Objective:

Build a Linear Regression model to predict the size of wildfire for occurrences in the state of California based on weather conditions, location and cause factors



# 1. Methodology

- Web Scraping and Data Collection
- EDA and Feature Engineering
- Building a Baseline Model
- Applying Transformations
- Scaling Features and Reducing Model Complexity
- Regularization
- Finalize Model and Test

# – Web Scrapping

visualcrossing.com/weather/weather-data-services/39.0225,-120.3036111/us/2018-7-1/2018-7-1

visualcrossing Weather Data Weather API Query Builder Pricing More Search docs... Sign in Sign up

**Weather Query Builder** Legacy version

Guided Data Download Manual Explore

Locations View data Download My Datasets

You aren't signed into an account. Some features will be restricted. Sign in.

39.0225,-120.3036111 Date range 07/01/2018 → 07/01/2018 US (°F, miles)

Addresses, partial addresses or lat,lon History or forecast data Data units

Query options

Data sections Weather elements Degree days Wind & solar Agriculture Weather stations

API Grid Chart JSON CSV

Weather Data Additional Data Data Details

Daily Hourly Current Events Alerts Info Stations

Download

Available weather data for 39.0225,-120.3036111. These results are filtered by your query options.

tempmin	temp	feelslikemax	feelslikemin	feelslike	dew	humidity	precip	precipprob	precipcover	precipctype	snow	snowdepth	windgust	windspeed	winddir	pressure	cloudcover
47.2	67.1	81.2	44.4	66.4	37.1	36.23	0		0		0	0	21.9	9.9	154.3	1017	0

# – Final Combined Data

## Fire Occurrence + Corresponding Weather Data

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3657 entries, 0 to 3811
Data columns (total 19 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   DISCOVERY_DATE                       3657 non-null   object
1   NWCG_CAUSE_CLASSIFICATION            3657 non-null   object
2   CONT_DATE                           3657 non-null   object
3   FIRE_SIZE                           3657 non-null   float64
4   LATITUDE                            3657 non-null   float64
5   LONGITUDE                           3657 non-null   float64
6   LAT_LONG                            3657 non-null   object
7   WEATHER_DATE                        3657 non-null   object
8   TEMP_MAX                           3657 non-null   float64
9   TEMP_MIN                           3657 non-null   float64
10  TEMP_AVG                           3657 non-null   float64
11  DEW                                3657 non-null   float64
12  HUMIDITY                           3657 non-null   float64
13  PRECIP                            3657 non-null   float64
14  WIND_SPEED                         3657 non-null   float64
15  WIND_DIR                          3657 non-null   float64
16  SEA_LEVEL_PRESSURE                 3657 non-null   float64
17  CLOUD_COVER                       3657 non-null   float64
18  VISIBILITY                        3657 non-null   float64
dtypes: float64(14), object(5)
memory usage: 571.4+ KB
```

# — Baseline Model

Dep. Variable:	FIRE_SIZE	R-squared:	0.004
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.9907
Date:	Tue, 09 Aug 2022	Prob (F-statistic):	0.460
Time:	19:23:29	Log-Likelihood:	-38395.
No. Observations:	3657	AIC:	7.682e+04
Df Residuals:	3642	BIC:	7.691e+04
Df Model:	14		
Covariance Type:	nonrobust		

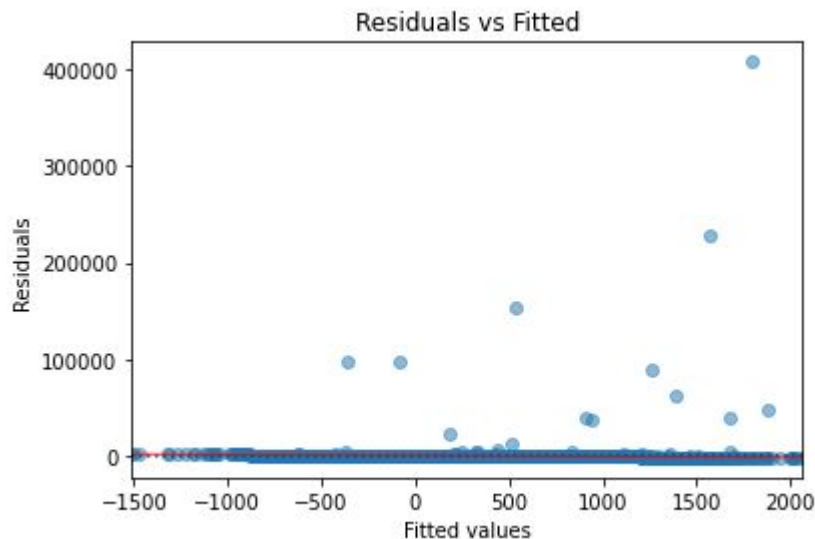
	coef	std err	t	P> t	[0.025	0.975]
const	-3.029e+04	4.45e+04	-0.681	0.496	-1.17e+05	5.69e+04
LATITUDE	-152.8464	146.685	-1.042	0.297	-440.440	134.747
LONGITUDE	-362.1123	175.902	-2.059	0.040	-706.988	-17.237
TEMP_MAX	29.7761	61.357	0.485	0.627	-90.521	150.073
TEMP_MIN	-18.5941	61.980	-0.300	0.764	-140.113	102.925
TEMP_AVG	17.8475	118.482	0.151	0.880	-214.451	250.146
DEW	-13.1248	56.464	-0.232	0.816	-123.829	97.579
HUMIDITY	-3.8237	37.098	-0.103	0.918	-76.558	68.910
PRECIP	428.6397	2456.670	0.174	0.861	-4387.946	5245.226
WIND_SPEED	-13.4792	37.380	-0.361	0.718	-86.767	59.808
WIND_DIR	-1.8674	2.620	-0.713	0.476	-7.005	3.270
SEA_LEVEL_PRESSURE	-7.8254	41.228	-0.190	0.849	-88.657	73.006
CLOUD_COVER	2.1035	9.296	0.226	0.821	-16.123	20.330
VISIBILITY	-32.2027	132.957	-0.242	0.809	-292.880	228.475
CAUSE	-227.6666	557.723	-0.408	0.683	-1321.147	865.814

For 80/20 test and validation data split:

**Train score:** 0.0066973648088055615

**Val score:** -0.0022044748479403964

## Overfit Model



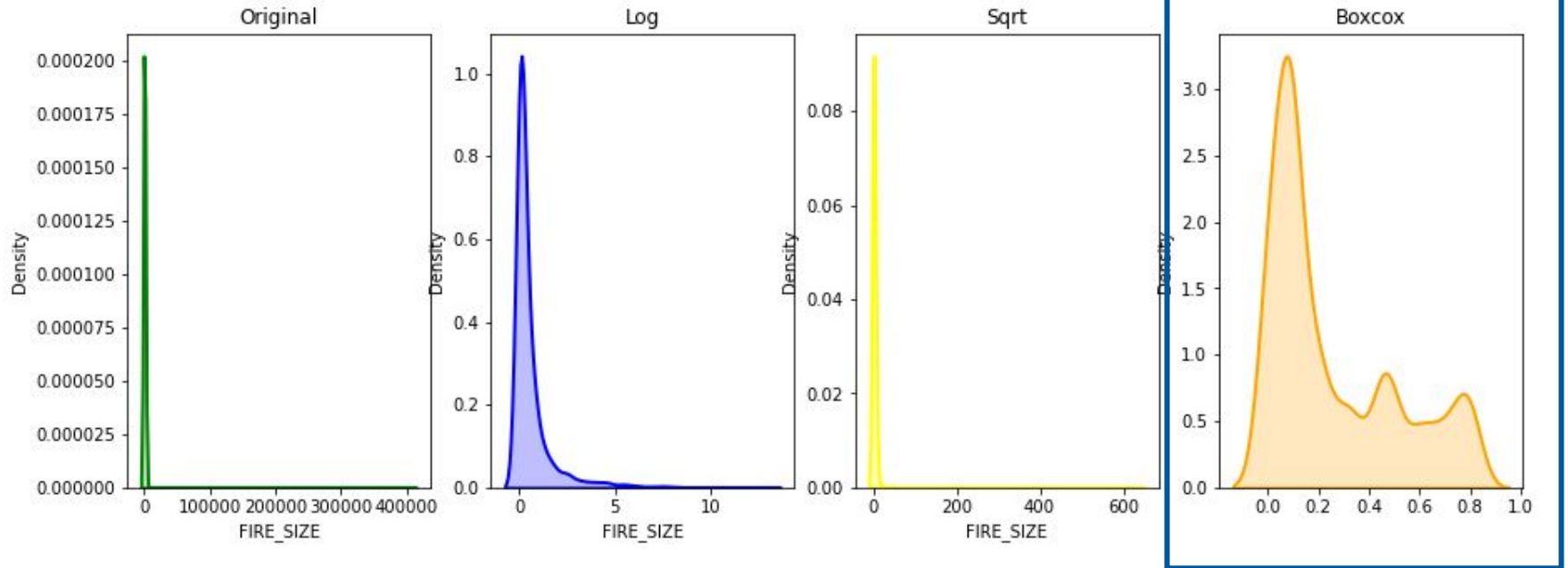
### Observations:

1. Residuals are unbalanced along the Y-axis -> The data is not normally distributed
2. Target (y) is heavily right skewed -> Fire size values have outliers beyond the std deviation

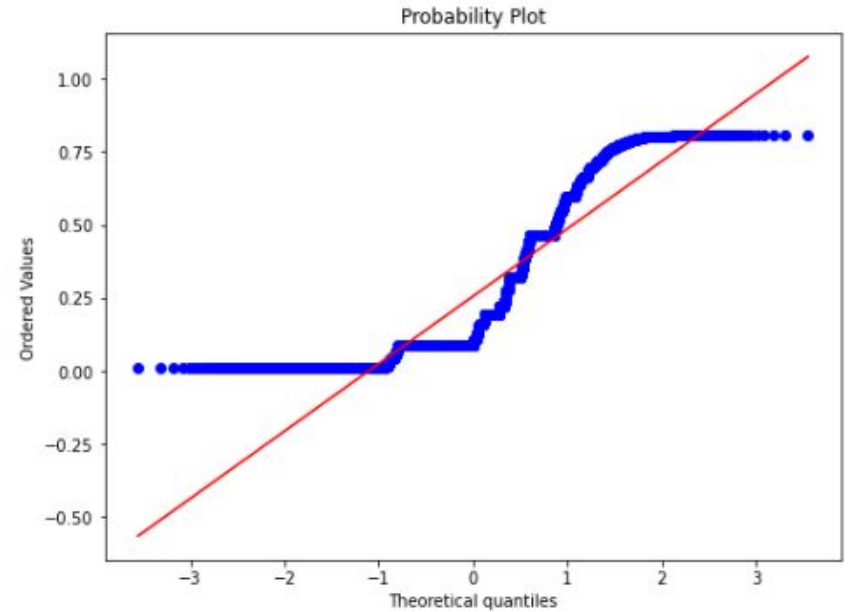
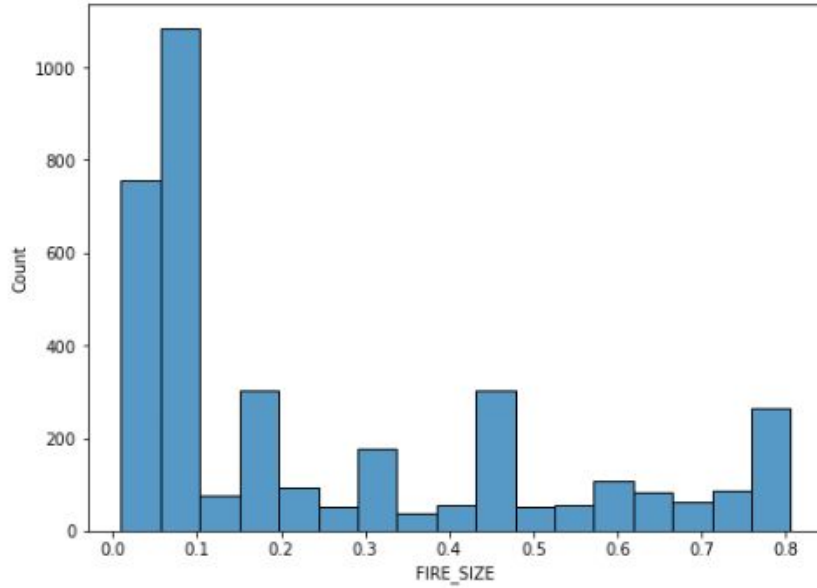
	FIRE_SIZE	LATITUDE	LONGITUDE	TEMP_MAX	TEMP_MIN	TEMP_AVG	DEW	HUMIDITY	PRECIP	WIND_SPEED	WIND_DIR
skew	35.204903	-0.190024	0.43162	-0.298382	-0.251725	-0.160159	-1.206296	0.314177	12.194452	0.856114	-0.012204



# Applying Transformation to Y

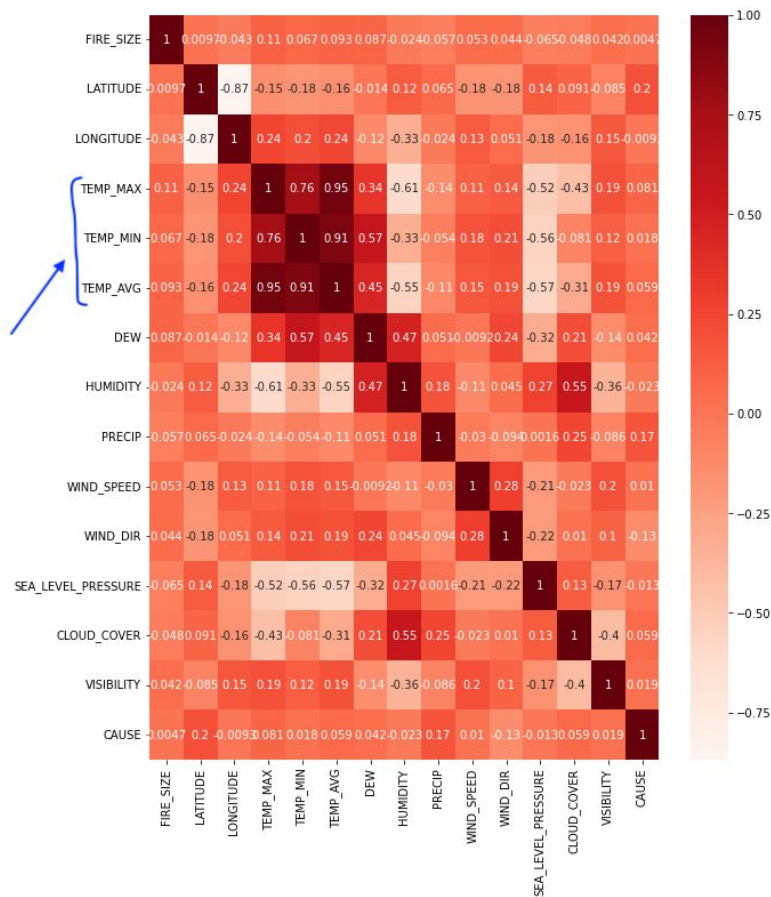


Skew: 0.914



**Y data is still not ideal. Fire Size distribution has LIGHT TAIL**

# – Handling MultiCollinearity



## – Scaling the Data

	<b>Model</b>	<b>Train_Score</b>	<b>Val_Score</b>	<b>Test_Score</b>
<b>1</b>	Baseline Model	0.006697	-0.002204	-
<b>2</b>	Y Transformed Model	0.026182	0.042815	-
<b>3</b>	Simplified Model	0.024159	0.029803	-
<b>4</b>	Scaled Model	0.029192	0.014666	0.0257

## – Regularization

	Model	Test_Score
1	Lasso Model	0.026562
2	Ridge Model	0.025944
3	Elastic Net Model	0.026485
4	Polynomial + Lasso Model	0.03436
5	Polynomial + Ridge Model	0.038441
6	Polynomial + Elastic Net Model	0.032941

# – Final Model

## Evaluating Performance of Final Model:

R2 score: 0.04

Mean absolute error: 0.21

Mean squared error: 0.06

Root mean squared error: 0.24

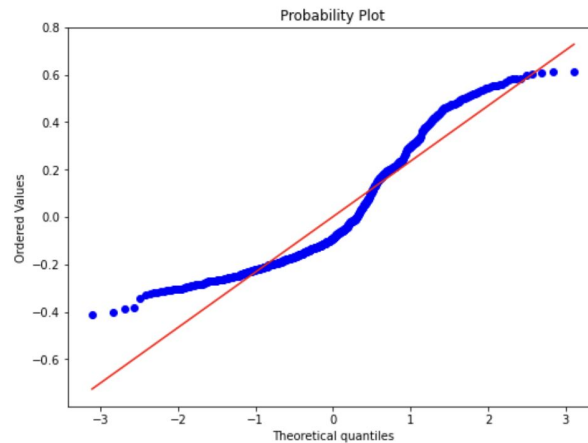
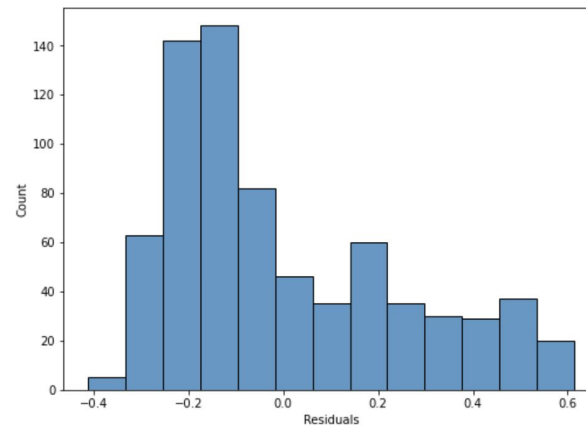
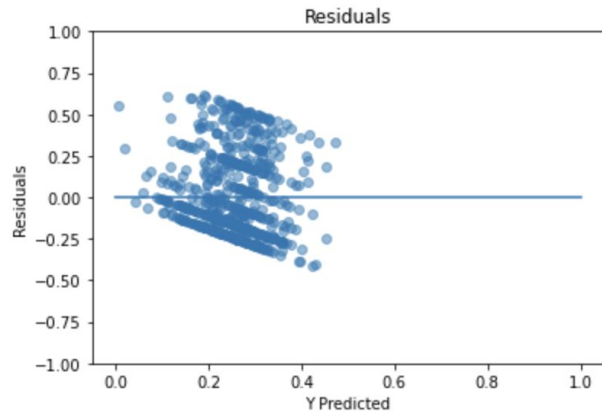
Comparing R2 Scores:

**Baseline model R2 score: 0.007**



**Final model R2 score: 0.04**

— There is still scope for improvement!



# Future Improvements

Getting average  
weather data  
across multiple  
days

Get data points  
across different  
years

Combine Fire  
data points from  
same date and  
neighbourhood