# California Fire Incidents

By: Oleksii K, Judith C, Raahim S, Tito A

# Introduction to Topic and Dataset

California is one of the places having the most deadliest and destructive wildfire seasons. The year of 2020 has been the largest wildfire season recorded in California history due to the particularly dry months from January to March. The LA Times found that wildfires and their compounding effects have intensified in recent years, likely due to climate change.

The dataset contains the list of **wildfires that has occurred in California between 2013 and 2019**.

The dataset contains the location where wildfires have occurred including the county name, latitude and longitude values and also details on when the wildfire has started.

# Learning Objectives

➔ Visualize the relationship between predictors in a visually appealing way

➔ Effectively apply statistical modelling techniques such as as regression to a large and complex dataset

➔ Perform geospatial analysis to observe the 2D-density of an event over longitude and latitude coordinates

➔ Connect geospatial analysis to predesignated hypothesis to arrive at a conclusion

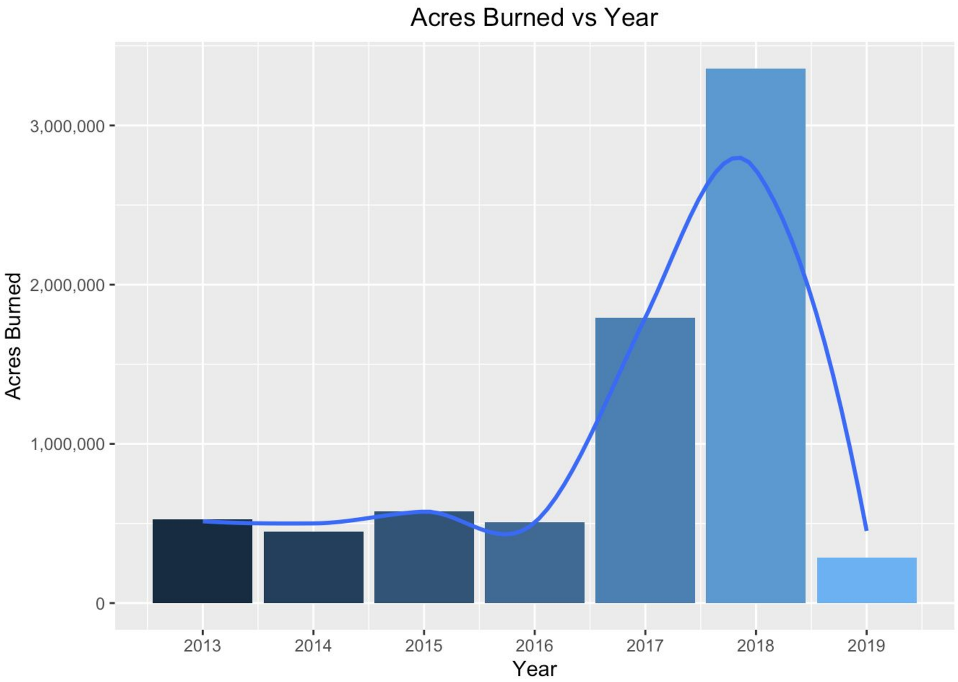
# Data Overview

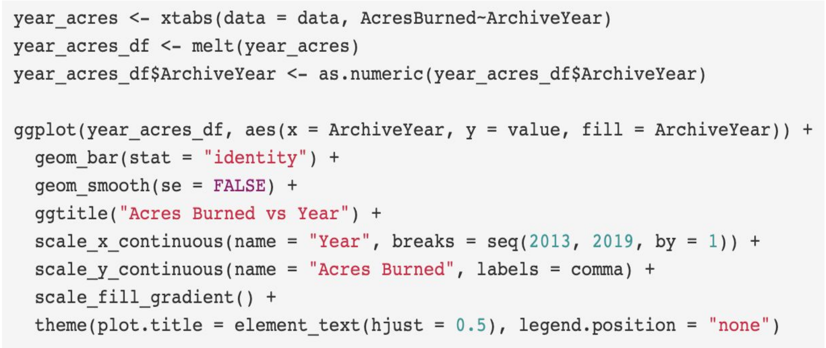
We used a dataset from Kaggle that contains the list of wildfires that has occurred in California between 2013 and 2019.

The dataset contains the **location** where the wildfires occurred including the **County name**, **latitude** and **longitude** values and also details on the **date** the wildfire started, how many **teams** and **helicopters** that were involved in fire suppression, number of **injured people** and much more.

The data consists of more than 1600 wildfires; however, it also has a lot of NA values, which complicates the reliability of the analysis.

# Exploratory Data Analysis

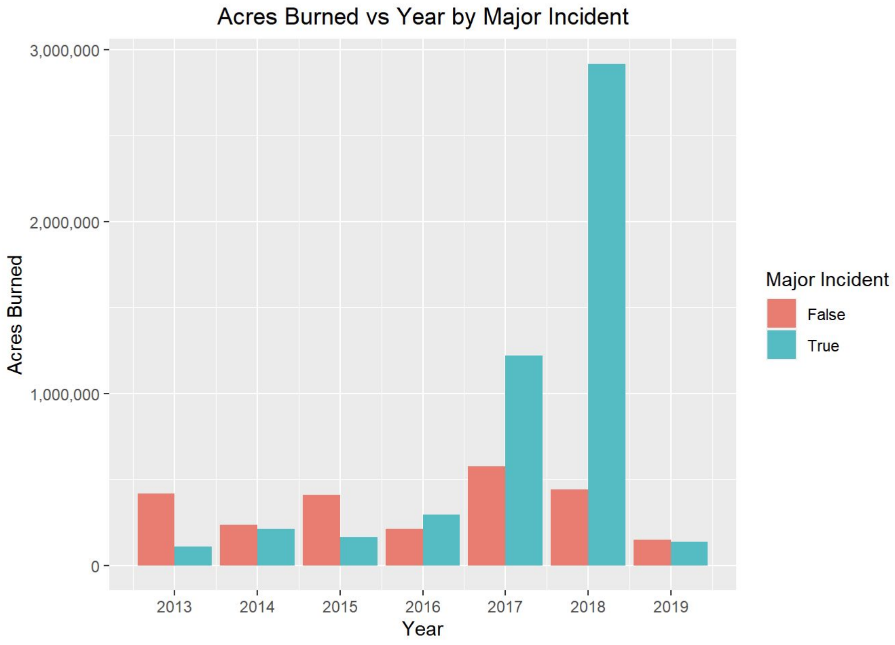


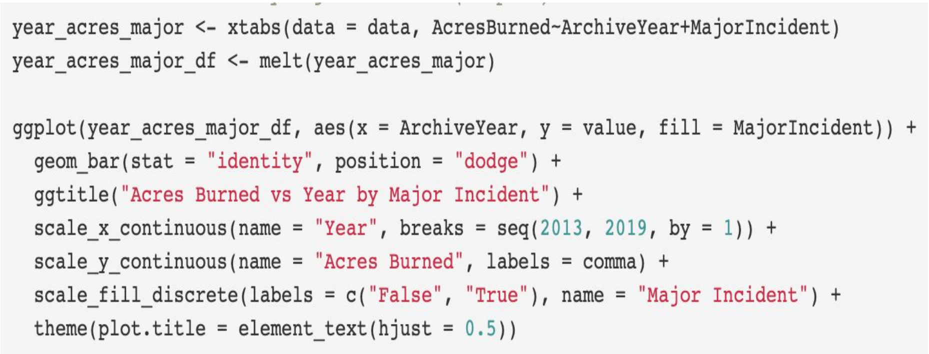


```
year_acres <- xtabs(data = data, AcresBurned~ArchiveYear)
year_acres_df <- melt(year_acres)
year_acres_df$ArchiveYear <- as.numeric(year_acres_df$ArchiveYear)

ggplot(year_acres_df, aes(x = ArchiveYear, y = value, fill = ArchiveYear)) +
  geom_bar(stat = "identity") +
  geom_smooth(se = FALSE) +
  ggtitle("Acres Burned vs Year") +
  scale_x_continuous(name = "Year", breaks = seq(2013, 2019, by = 1)) +
  scale_y_continuous(name = "Acres Burned", labels = comma) +
  scale_fill_gradient() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```

# Exploratory Data Analysis

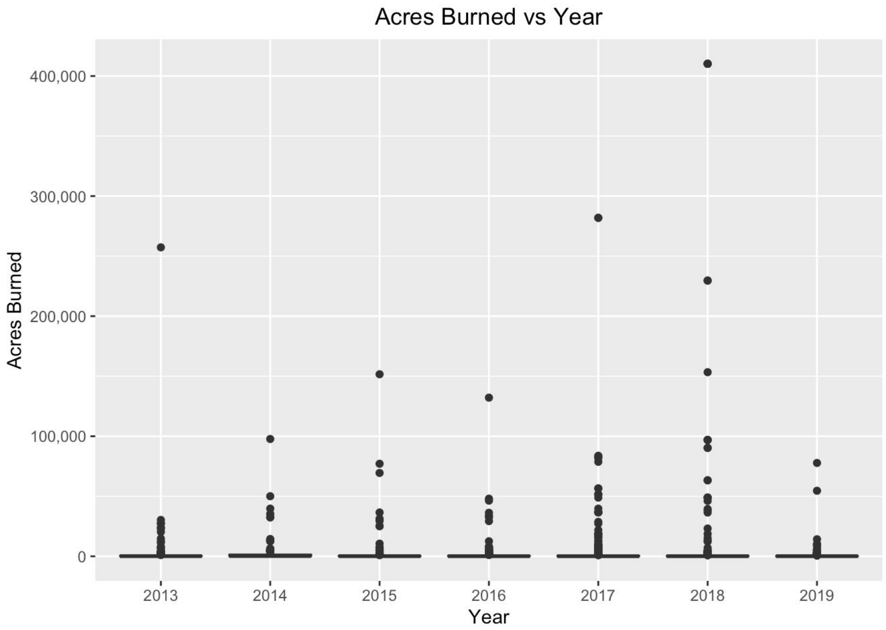


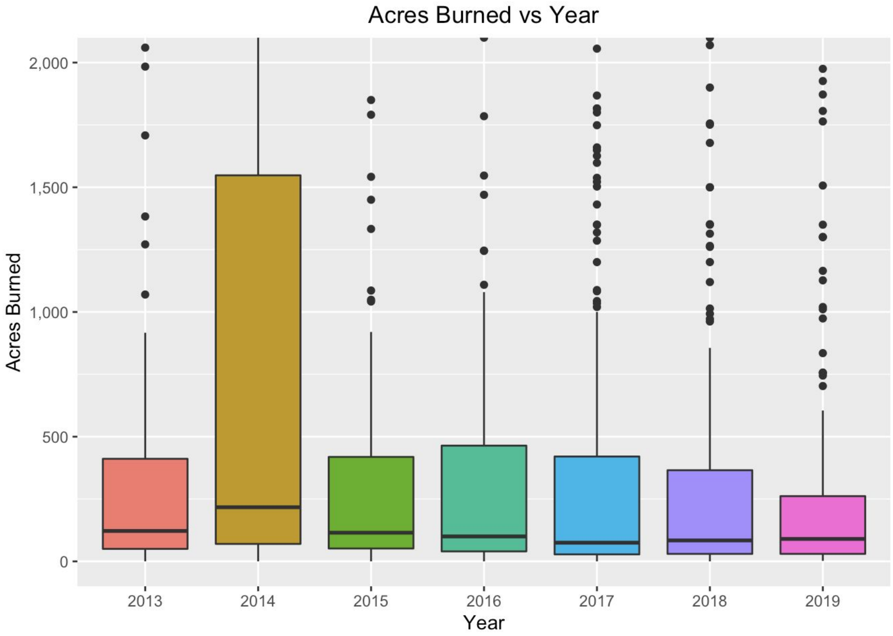


```
year_acres_major <- xtabs(data = data, AcresBurned~ArchiveYear+MajorIncident)
year_acres_major_df <- melt(year_acres_major)

ggplot(year_acres_major_df, aes(x = ArchiveYear, y = value, fill = MajorIncident)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Acres Burned vs Year by Major Incident") +
  scale_x_continuous(name = "Year", breaks = seq(2013, 2019, by = 1)) +
  scale_y_continuous(name = "Acres Burned", labels = comma) +
  scale_fill_discrete(labels = c("False", "True"), name = "Major Incident") +
  theme(plot.title = element_text(hjust = 0.5))
```

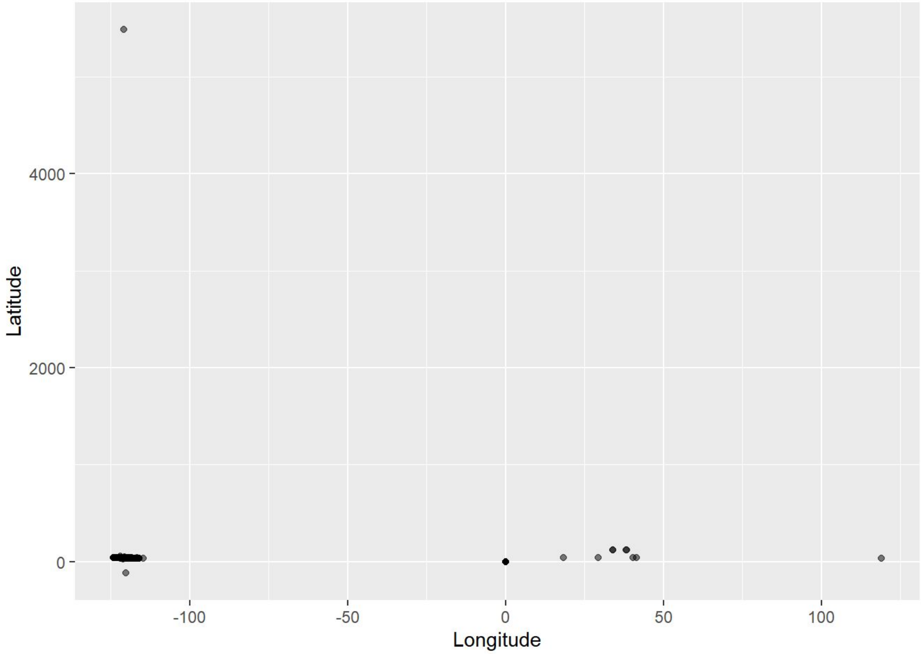
# Exploratory Data Analysis



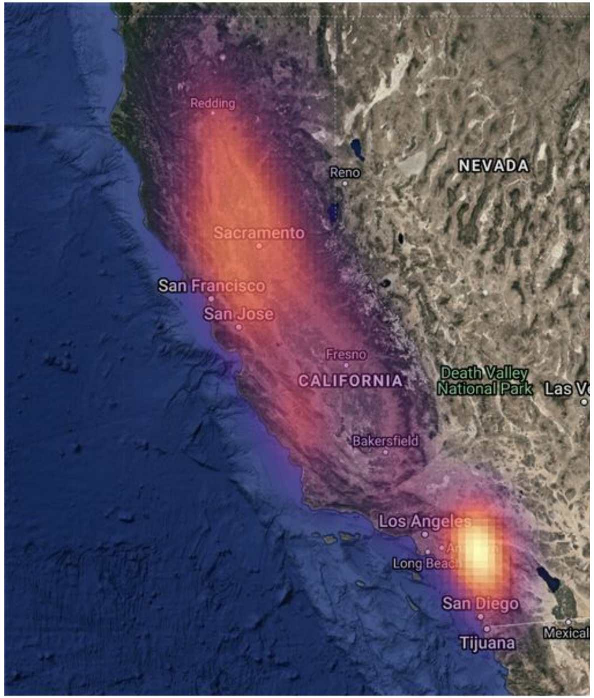


# Geospatial Analysis - Cleaning Data

- Clear Outliers in the dataset
- Removed outliers by selecting a range for geospatial coordinates

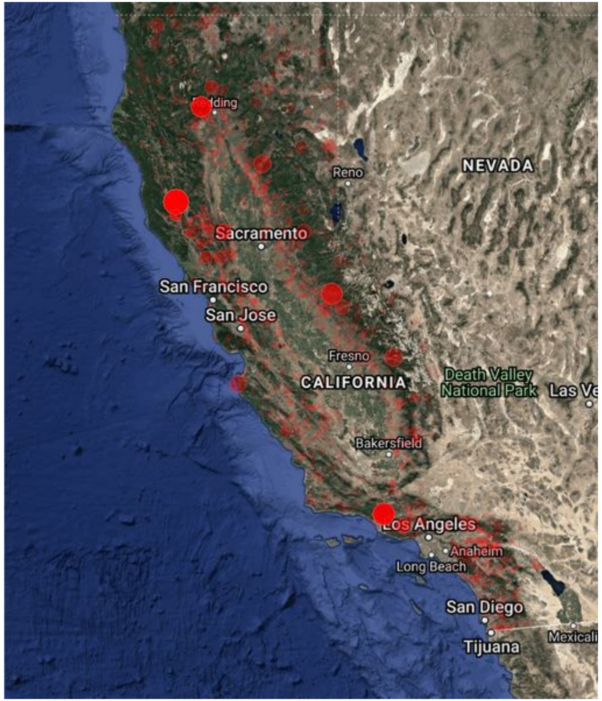


# Geospatial Analysis



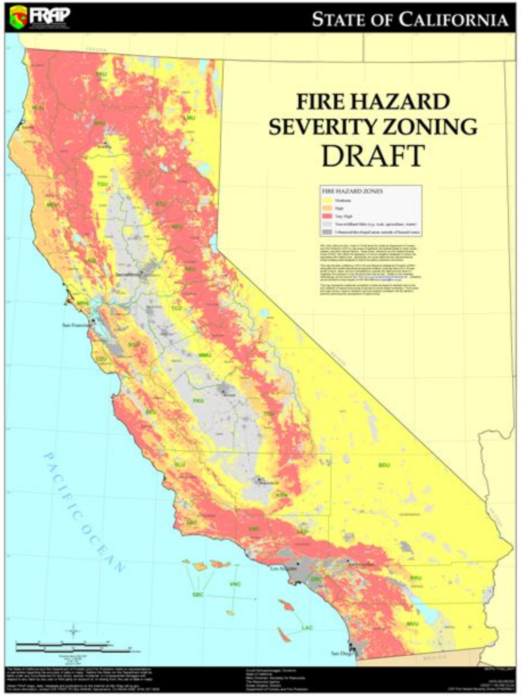
Frequency of Wildfires over map

(Brighter indicates larger frequency)

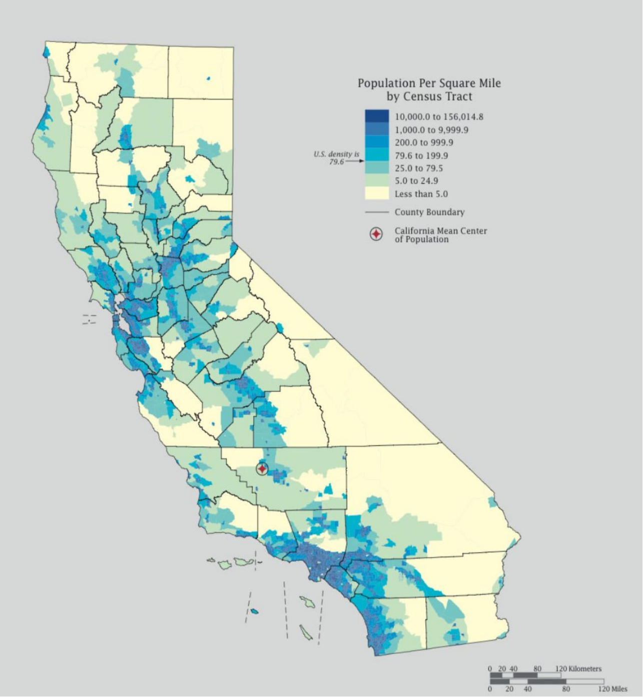


Acres Burned over Location

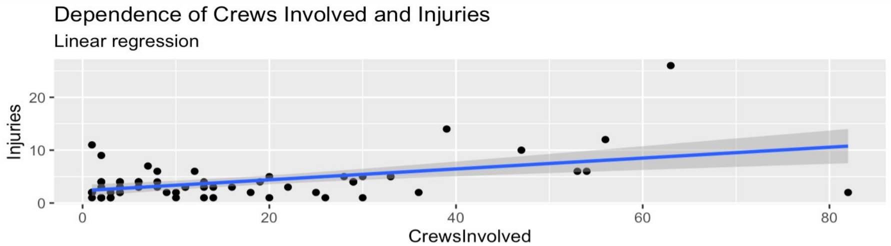
# Conclusions from Geospatial Analysis



Fire Hazard Severity Zoning

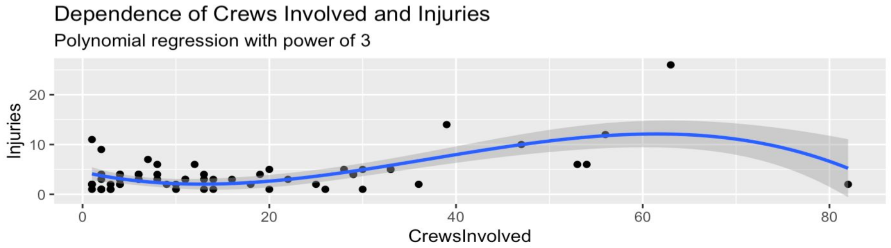


Population Density California



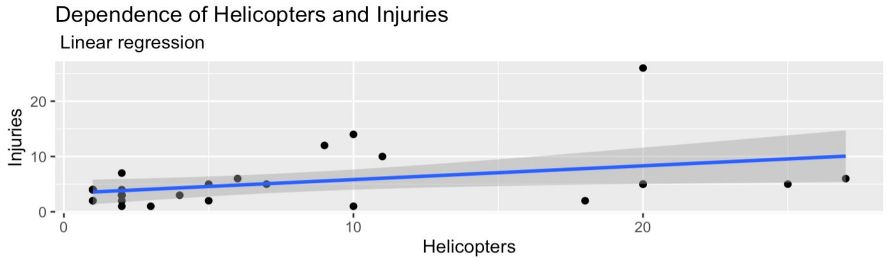
Dependence of Crews Involved and Injuries

Linear regression



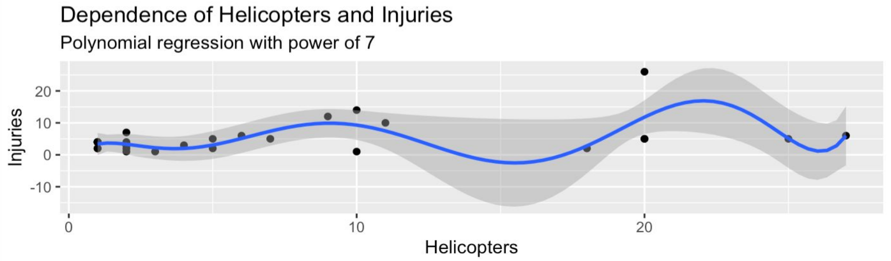
Dependence of Crews Involved and Injuries

Polynomial regression with power of 3



Dependence of Helicopters and Injuries

Linear regression



Dependence of Helicopters and Injuries

Polynomial regression with power of 7

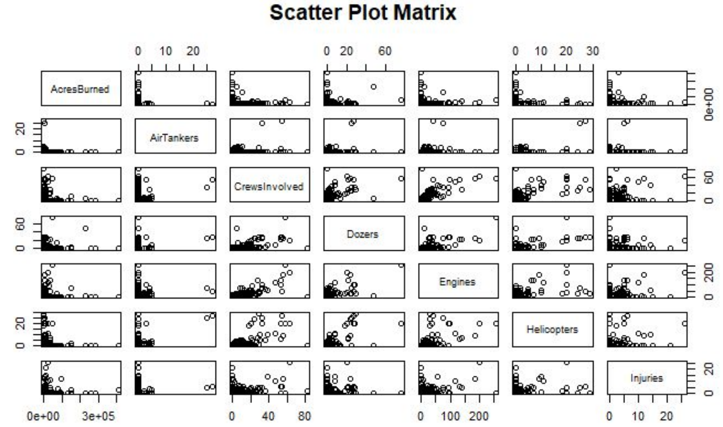
# Dependence of Injured with Helicopters and Crews Involved

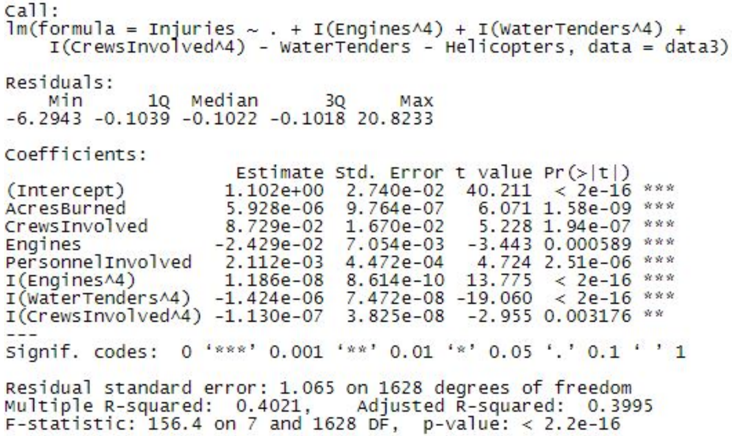
- Number of injuries per fire are positively correlated with both number of helicopters and crews involved

- To fit the current data, both linear and polynomial regression were used.

- However, it is visible that linear regression is more likely to be better prediction, since it is subject to less overfit.

# Multi-Linear Regression





```
Call:
lm(formula = Injuries ~ . + I(Engines^4) + I(WaterTenders^4) +
    I(CrewsInvolved^4) - WaterTenders - Helicopters, data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2943 -0.1039 -0.1022 -0.1018 20.8233

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.102e+00  2.740e-02  40.211  < 2e-16 ***
AcresBurned         5.928e-06  9.764e-07   6.071 1.58e-09 ***
CrewsInvolved       8.729e-02  1.670e-02   5.228 1.94e-07 ***
Engines            -2.429e-02  7.054e-03  -3.443 0.000589 ***
PersonnelInvolved   2.112e-03  4.472e-04   4.724 2.51e-06 ***
I(Engines^4)        1.186e-08  8.614e-10  13.775  < 2e-16 ***
I(WaterTenders^4)  -1.424e-06  7.472e-08 -19.060  < 2e-16 ***
I(CrewsInvolved^4) -1.130e-07  3.825e-08  -2.955 0.003176 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.065 on 1628 degrees of freedom
Multiple R-squared:  0.4021,    Adjusted R-squared:  0.3995
F-statistic: 156.4 on 7 and 1628 DF,  p-value: < 2.2e-16
```

**Process:**

1. **Analysis on Full Model**
   a. ANOVA test to see if new model improved outcome
2. **Collinearity**
   a. Calculated VIF
   b. ANOVA test for every predictor dropped
3. **Polynomial Terms**
   a. Didn’t change QQ-Plot or residual plot by a lot
4. **Backwards Selection using BIC**
   a. Output same model as in step 2

# Learning Outcomes

- ➔ Preprocessing and cleaning data
- ➔ Regression Analysis by Region
- ➔ Ensuring data fits regression assumptions (ie normality, equal variance,  linearity)
- ➔ Performed 2D Kernel Density Estimation in order to determine frequency of wildfires over a geospatial space
- ➔ Using results from Kernel Density Estimation in conjunction with other data to arrive to the conclusion that a large population in a high risk fire zone results in the largest fires in California

# Thank you!