STAT423 Final Report

California Wildfires - Predicting Damage

Maia Czerwonka	**Research Question 4**
Eliana Dietrich	**Research Question 5**
Elyn Franson	**Research Question 3**
Carson Lindholm	**Research Question 1**
Adriana Vassek	**Research Question 2**

1 Introduction

Wildfires wreak devastating havoc on Californians annually, with tens of billions of dollars in damage and economic loss and dozens of people injured or killed. We wish to explore the relationship between a variety of predictive variables and the destruction done in order to better understand the factors that play into this tragedy on a yearly basis.

1.1 Research Questions

We propose the following questions to guide our exploration of the data:

i. What is the minimum number of response variables needed to reliably predict average estimated financial loss?

ii. What is the minimum number of response variables needed to reliably predict the number of fatalities?

iii. What is the minimum number of response variables needed to reliably predict the number of injuries?

iv. Do specific locations predict significantly different numbers of fatalities and injuries than others?

v. Do certain times of year predict significantly different numbers of fatalities and injuries than others?

1.2 Dataset Description

Our data comes from the California Wildfire Dataset (2014-2025). This dataset is randomly generated from data observed by Cal Fire, (California Department of Forestry and Fire Protection), FEMA (Federal Emergency Management Agency), and USGS (United States Geological Survey). With the following variables:

Variable Name	Variable Type	Description
Estimated_Financial_Loss	Quantitative	Estimated loss (in millions) from particular fire
Injuries	Quantitative	Number of injuries seen in particular fire
Fatalities	Quantitative	Number of fatalities seen in particular fire
Location	Nominal	County location of the fire
Date	Ordinal	Date of fire outbreak
Area_Burned	Quantitative	Acres of land burned by the fire
Homes_Destroyed	Quantitative	Number of homes destroyed by the fire
Businesses_Destroyed	Quantitative	Number of businesses destroyed by the fire
Vehicles_Damaged	Quantitative	Number of vehicles damaged by the fire

We will focus on the variables ’Estimated_Financial_Loss’, ’Fatalities’, and ’Injuries’ as our response to the remaining predictors.

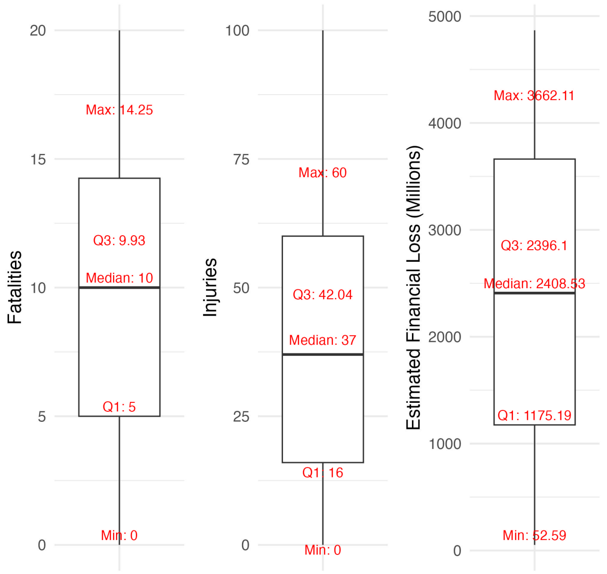
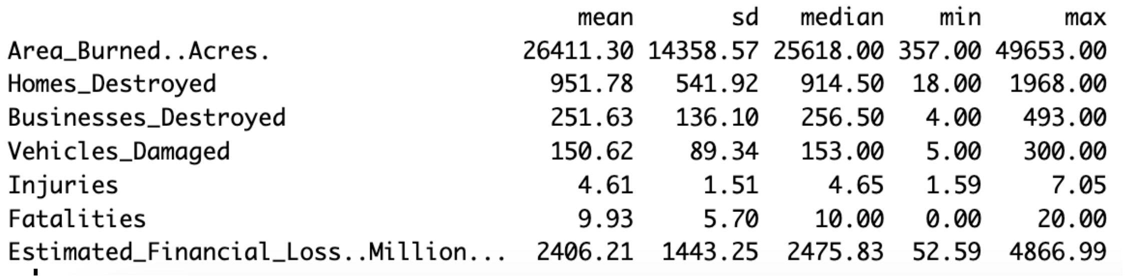


Figure 1: Distributions of Outcome Variables



```
                                      mean       sd   median    min      max
Area_Burned..Acres.               26411.30 14358.57 25618.00 357.00 49653.00
Homes_Destroyed                     951.78   541.92   914.50  18.00  1968.00
Businesses_Destroyed                251.63   136.10   256.50   4.00   493.00
Vehicles_Damaged                    150.62    89.34   153.00   5.00   300.00
Injuries                              4.61     1.51     4.65   1.59     7.05
Fatalities                            9.93     5.70    10.00   0.00    20.00
Estimated_Financial_Loss..Million... 2406.21 1443.25  2475.83  52.59  4866.99
```

Figure 2: Descriptive statistics of all quantitative variables.

2 Method

2.1 Data Processing

There were no missing values in this dataset so no modifications were needed. The response variables were not skewed, and there was no evidence of a non-linear relationship so no other regular data transformations were applied. In order to correctly deal with 'Date', the variable was first transformed into a factor and binned into twelve levels, each one a month of the year. An additional factor variable was created to represent the four seasons of the year to allow for further exploration of the effects of time of year on prediction. Additional subsets of the data were also made to explore some confounding effects - these took the form of subsetting by cause (to fires caused specifically by human activity) and/or location, including removing counties such as "Los Angeles County", "Orange County", "Riverside County", and "Santa Barbara".

2.2 Research Question 1, 2, 3

To predict financial losses, fatalities, and injuries caused by wildfires, we began by analyzing the distribution of each response variable and examining key predictor variables, including location, acres burned, and the number of buildings destroyed. Covariance matrices and visualizations such as boxplots and scatterplots were used to explore relationships between these predictors and wildfire fatalities. A comprehensive linear regression model was initially constructed to evaluate their combined influence. Model refinement was conducted using ANOVA and the F test to assess the necessity of each predictor, aiming for an efficient model with the most minimized amount of predictors. To meet statistical assumptions, we examined normality and variance conditions using TA and QQ Plots, applying Box-Cox transformations where necessary. Various transformations, including logarithmic, square root, and fractional adjustments, were tested to improve model performance. The final evaluation involved comparing transformed and reduced models against the original through ANOVA, while diagnostic checks, including QQ plots and residual analysis, ensured normality and homoscedasticity of errors.

2.3 Research Question 4

Linear regressions were fitted to predict fatalities and injuries according to the location of the fire as a factor:

a. lm(Injuries $\sim$ Location)

b. lm(Fatalities $\sim$ Location)

Full models for both injuries and fatalities were fitted as well. An ANOVA was used to assess the differences between the two models for both injuries and fatalities.
A Shapiro-Wilk test was used to test the normality of residuals for both injuries and fatalities. A subsequent Box-Cox transformation was attempted in the injury model.

California locations were grouped into three categories: NorCal, BayArea, and SoCal.
A linear regression was fitted on these new location groupings for both injuries and fatalities to check if it made a difference in the results.

2.4 Research Question 5

Linear regression models were fit to predict both fatalities and injuries. A variety of beginning models were explored for the potential, but the foundation of the question revolved around the following four:

a. lm(Fatalities ∼ Month)

b. lm(Fatalities ∼ Season)

c. lm(Injuries ∼ Month)

d. lm(Injuries ∼ Season)

Model fit was analyzed visually using the Tukey-Anscombe and QQ-plots to assess validity of model assumptions. Additionally, the Shapiro-Wilk test was used to confirm the normality of the residuals computed.

3 Results

3.1 Research Question 1

The Covariance matrix (Figure 3) shows a low correlation between the predictors and our response variable, financial loss. This makes it initially difficult to find a linear regression model to predict financial loss. This issue was compounded by a bimodal distribution of financial loss (figure 4). This indicates that many insignificant fires caused relatively little damage and many expensive fires that caused lots of damage. Bimodal distributions are difficult to fit to a linear regression because they typically have non linear relationships and cause non normal residuals and non-constant variance.

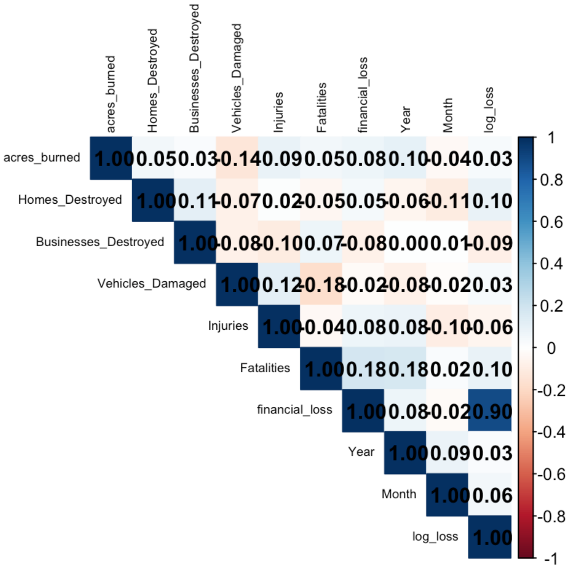


Figure 3: Covariance matrix

We used multiple methods to find a linear model that has the greatest adjusted r^2. Initially, we used a stepwise forward algorithm to compute a linear model. This resulted in $financial\ loss \sim fatalities$ with $r^2 = 0.03$. The full untransformed model found significant relationships between financial loss and fires in Mendicino, San Diego, and Sonoma County, but produced an adjusted r^2 of -0.04. To further investigate the problem, we added interaction terms to the linear model, using the model, $financial\ loss \sim Location(buildings\ destroyed + acres\ burned + Fatalities + Season + Year)$ This linear model was quite complex, which may be an indicator that it is overfit, but produced an adjusted r^2 of 0.42. This model was the best in terms of the adjusted r^2 and indicated significant relationships for the variables like: Napa Valley, Napa Valley:buildings destroyed, Santa Barbara County:acres burned, Los Angeles County:SeasonSummer, Mendocino County:SeasonSummer among others.

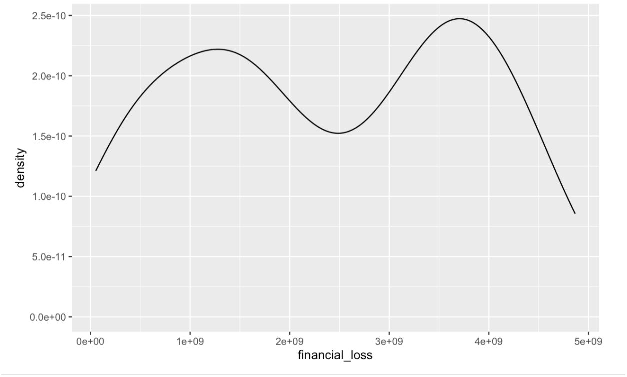


Figure 4: PDF of financial loss

3.2 Research Question 2

From the scatterplots, we observed that there was no apparent linear relationship between the predictors and the number of fatalities. A full linear model was created with all the predictors against the response variable (fatality). The model produced high residuals, ranging from -10 to 10, but appeared to be homoscedastic (exhibiting constant variance). However, the QQ plot indicated that the residuals deviated from normality, both at the tails and in the middle of the distribution. Furthermore, there were no significant p-values, and the adjusted R-squared value was -0.01947, indicating a poor model fit.

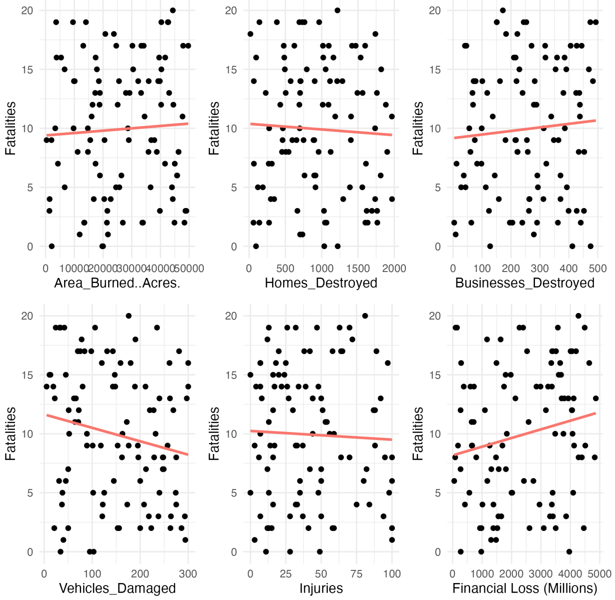


Figure 5: Fatality vs Predictor Plots

The use of transformations (inverse, logarithmic, square, and square root) also proved ineffective in linearizing the relationship between the predictors and fatalities. ANOVA and F-tests conducted on the transformed and reduced models showed no significant relationship between any combination of the predictors and fatalities.

3.3 Research Question 3

The scatter plots first plotted without transformations showed no linear relationship between the dependent variables and the injuries. The points are scattered across all plots.

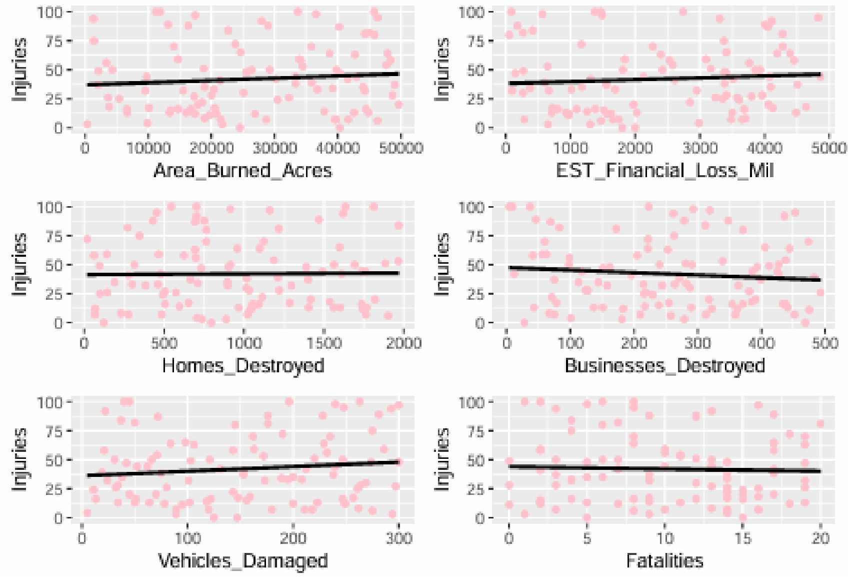


Figure 6: Variables Plotted Against Injuries

A full linear model was made with all the dependent variables against the response variable injuries. The residuals of this model ranged from -40 to 60. These high residuals suggested a poor model fit. The residuals were mostly normal based on the QQ plot, but there were some deviations on the tail ends. Based on the TA plot, we saw a spread of residuals that seem homoscedastic.

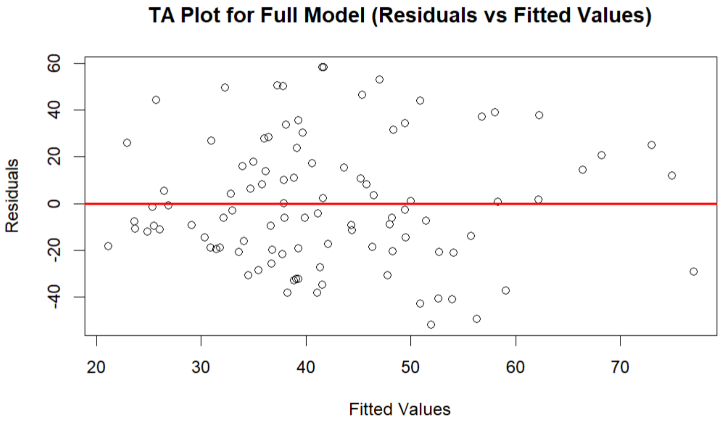
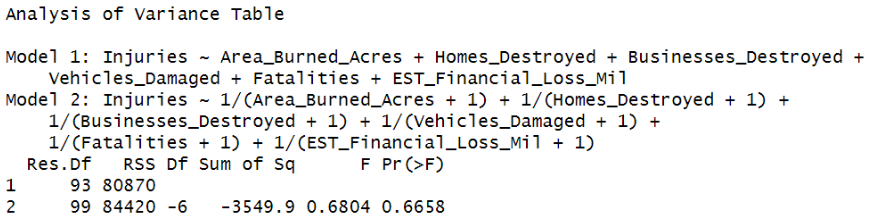


Figure 7: TA Plot for Full Model Relating to Injuries

Based on the full model one location, Mendocino County, was statistically significant. All other variables had very high p-values and were deemed insignificant. Since this one variable was deemed significant, we wanted to explore this subset of data. Just Mendocino County was selected from the data, and scatterplots and the same full linear model were made. The same results were found, with no variables found significant for just this county. Additionally, when the insignificant variables were removed from the full model and it was just Injuries and Location in the model none of the locations were significant at the 0.05 level.

Transformations were performed to see if the models would become more linear and significant using square root, log, and fractions. None of the transformations made a significance in the model using ANOVA and the F statistic and scatterplots of the transformations. No variables were found to be statistically significant using the transformations. This ANOVA test has a high p-value showing that this fraction transformation was not significant.



```
Analysis of Variance Table

Model 1: Injuries ~ Area_Burned_Acres + Homes_Destroyed + Businesses_Destroyed +
    Vehicles_Damaged + Fatalities + EST_Financial_Loss_Mil
Model 2: Injuries ~ 1/(Area_Burned_Acres + 1) + 1/(Homes_Destroyed + 1) +
    1/(Businesses_Destroyed + 1) + 1/(Vehicles_Damaged + 1) +
    1/(Fatalities + 1) + 1/(EST_Financial_Loss_Mil + 1)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     93 80870
2     99 84420 -6   -3549.9 0.6804 0.6658
```

Figure 8: ANOVA Test

3.4 Research Question 4

Looking at the distribution of injuries and fatalities by location, we can see that Shasta County has the highest number of fatalities, while Los Angeles County has the least. Mendocino County has the highest number of injuries, while Riverside County has the lowest.

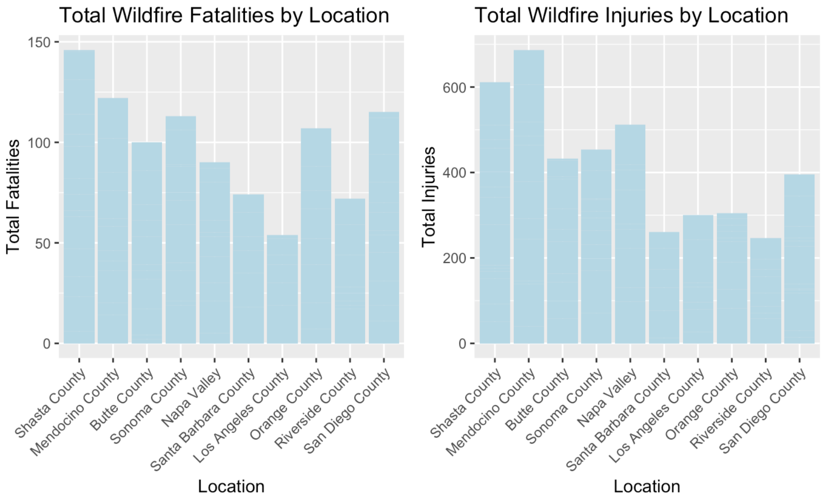


Figure 9: Distributions of total injuries and fatalities by wildfire location.

A full model was fit for both injuries and fatalities, resulting in one significant predictor (Location Mendocino County) for the full injury model. All other injury predictors and fatality predictors had high p-values deemed insignificant. A reduced model was fitted for both injuries and fatalities with only location as the explanatory variable. None of the location levels were significant for either of the models.

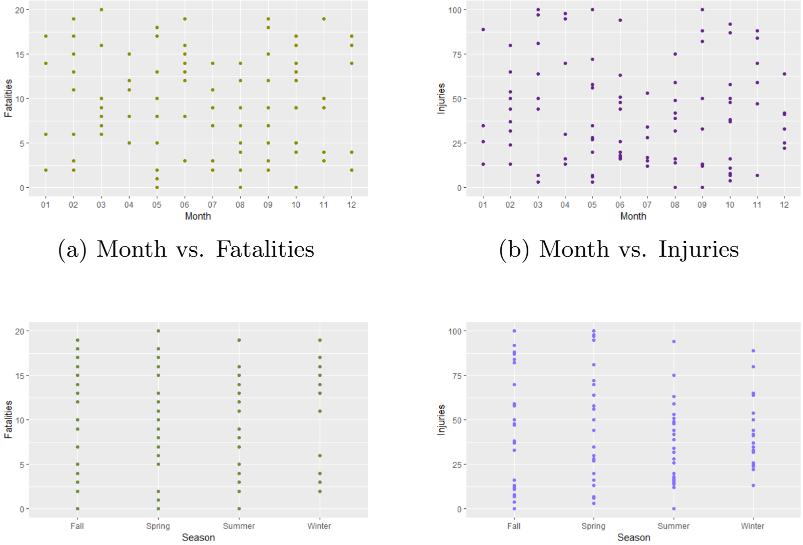
An ANOVA was done for both fatalities and injuries to compare the full models to the reduced models. Both came back insignificant.

A Shapiro-Wilk test and Box-Cox transformation function was applied to both full models and the reduced models. None of the models came back significant after transformations.

A linear model was fit with locations grouped by geographical proximity to other locations to see if fewer levels would reveal a clearer pattern. Both models show no significance with this grouping.

3.5 Research Question 5

From the original exploration first performed on the variables, there was no visual correlation immediately obvious to the viewer.



(a) Month vs. Fatalities

(b) Month vs. Injuries

(c) Season vs. Fatalities

(d) Season vs. Injuries

Figure 10: Predictor vs Dependent Variable - 'season' and 'Month'

Despite that, simple linear regressions were first fitted to equations (a) through (d) from the methodology. Further models were built around subsets of the data to exclude potential confounding locations, however no further significance was found. Additionally full models for both Fatalities and Injuries were built using either 'Month' or 'season'. ANOVA tables and F-tests were used to compare the full and transformed model to the original SLRs, however this was largely ineffective given that all predictors were insignificant. For all models, every month and season was determined to be an insignificant factor level (using $\alpha = 0.05$), with no R^2 value ever reaching higher than 0.2. Models were also built to include interactions between the 'Area_Burned' and 'Month' (or 'Area_Burned' and 'season'). The assumption was that certain months/seasons may be drier, leading to higher acreage burn rates. Even if this assumption holds in sense, it did not have a tangible impact on the model, with all predictors remaining insignificant and the R^2 value only reaching 0.31 when interacting 'Month' and 'Area_Burned'.

4 Discussion

4.1 Research Question 1

Our analysis of financial loss faced two main challenges: low correlations and nonlinear relationships between predictors and the response variable. To overcome these challenges, we found that a interaction heavy linear model performed the best compared to models with less or no interaction terms. Key predictors in the final models tended to be location based variables.

To further analyze the financial loss caused by fires, we may consider decomposing the bimodal distribution of financial loss into lower- and higher-cost fires, which may lead to better prediction. In addition, we may search for better predictors with higher correlations with financial loss to strengthen the linear modeling.

4.2 Research Question 2 and 3

We found that none of the variables in the data set help predict or are linearly correlated with the number of fatalities and number of injuries in a wildfire. We tried transformations, and selecting certain counties but none of them improved the model.

Logically, we would have expected to see variables such as acres burned or financial loss to help predict fatalities or injuries (i.e. the more damage done would result in more fatalities or injuries). However, this was not the case. Based on our sample, the size, damage, or even location of the fire did not have correlation with the number of fatalities or injuries. This could be explained by the well established evacuation systems in California for wildfires, in combination with the prevalence of predictive modeling for areas that will be hit, allowing people to evacuate early.

4.3 Research Question 4

Although none of the locations came back as significant predictors of injury or fatality in California wildfires, we can still analyze data trends. Although there was no significant difference between Northern California, Bay Area and Southern California counties, we can see in Figure 5 that, unexpectedly, Northern California and the Bay Area see more fatalities and injuries overall than Southern California. Although Southern California is generally more prone to wildfires, it is likely more prepared for such events, perhaps to the point that there are less injuries and deaths than in less prepared Northern California. Another reason this may be is that Northern California is more rural than Southern California, making it easier for wildfires to spread, thus injuring and killing more people.

4.4 Research Question 5

There seemed to be no relationship between time of year and either injuries or fatalities. Even interactions between time of year and a logical other term of 'Area_Burned' showed no predictive power. This is surprising given that the intuitive understanding would say that certain times of year are more likely to experience conditions that contribute to larger fires, thus making them more destructive and dangerous and causing more injuries and fatalities. It is possible that the areas where the fires are occurring are so far removed from populations, and that warning is given far enough in advance, that the human toll is independent from that of damage caused. In other words, even if the fires are dangerous, they may not cause a large enough number of casualties for us to analyze at a statistically significant level.

4.5 Implications

It is interesting to note that with regard to both injuries and fatalities, all variables were found to be insignificant predictors. As mentioned in the previous discussions particular to specific research questions, this could be due to the evacuation and warning systems in place to allow people to escape fires quickly, keeping casualties low despite the area burned and property damage done. Another possibility is that the location where the fires occur in the counties themselves is more removed from larger population centers, thus causing damage but not injuries. Similarly, it was found that very few predictors were significant with regard to financial damage. It is important to note that this data was based on real data but randomly generated. This may be one reason we don't see significant results as it is possible the data could be skewed because it is randomly generated. Future exploration in different datasets would be needed to see if relationships between our variables do exist, and if the lack of significance is because of the actual data relationship or the generation of our dataset.

References

Vivek Attri. (2025, February). California Wildfire Damage (2014 - (feb)2025), Version 1. Retrieved February 19, 2025 from https://www.kaggle.com/datasets/vivekattri/california-wildfire-damage-2014-feb2025/data