Study 1: Simple Linear Regression Analysis

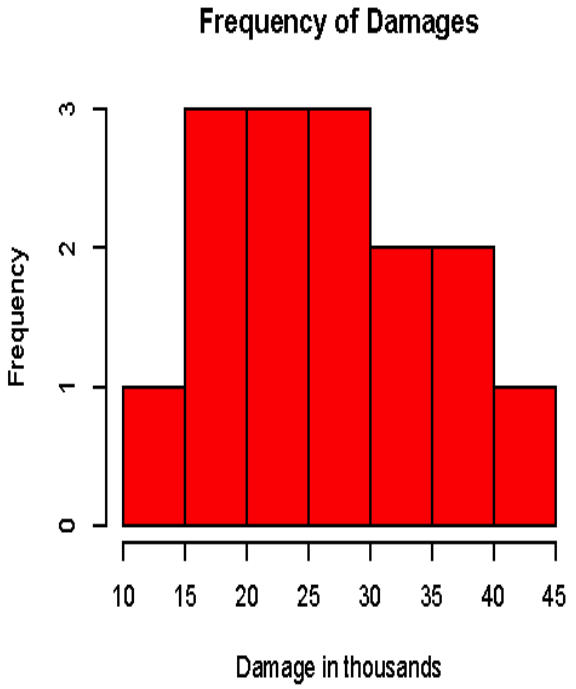
Introduction

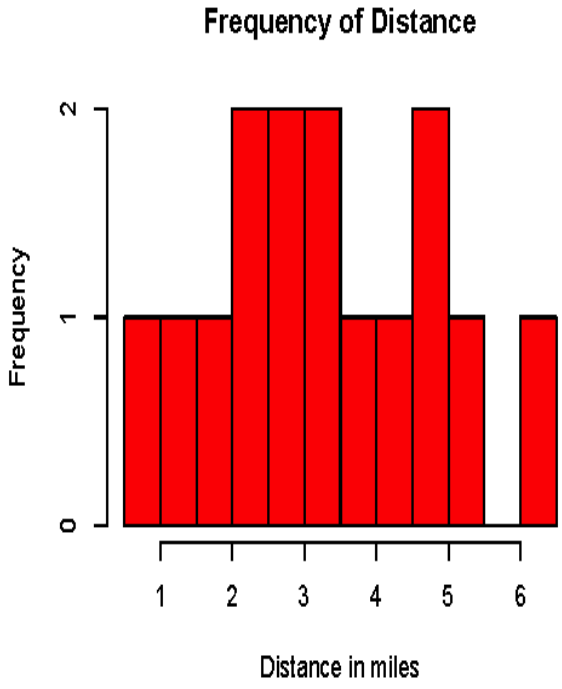
This report examines the relationship between the amount of fire damage sustained during major residential fires and the distance from the burning residence to the nearest fire station. The study is based on a dataset containing 15 cases of recent fires in a large suburb of a major city, where the amount of damage (in thousands of dollars) and the distance between the fire and the nearest fire station (in miles) were recorded. This analysis aims to determine whether there is a statistically significant association between the two variables. By conducting a simple linear regression analysis, we aim to assess the extent to which proximity to fire stations influences the severity of fire damage.

Exploratory Data Analysis

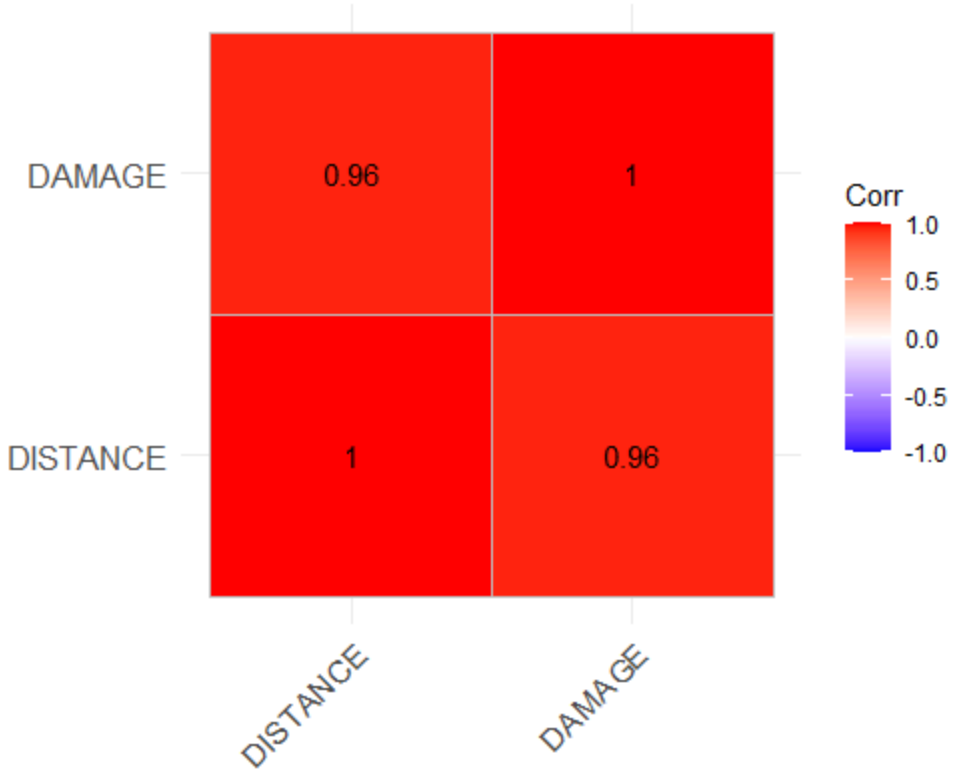
Firstly, we looked at the descriptive statistics for each variable. The average amount of damage was approximately $**26.41 thousand**, with a standard deviation of $**8.07 thousand**, which indicates a moderate spread of damage amounts around the mean. The average distance between the fire and the nearest fire station was approximately **3.28 miles**, with a standard deviation of **1.58 miles**, suggesting that the distances varied moderately among the 15 incidents. Still, most were within a few miles of a fire station.

Next, we utilized data visualizations to create and interpret a variety of exploratory plots. These visual tools allowed us to understand the underlying structure of the data. By inspecting these plots, we gained deeper insights into the trends within the data and the relationships between variables. The following plots are a result of these applied methodologies.





These histograms show the distributions of fire-related damages and distances to the nearest fire stations. The first graph shows that most damages fall in the range of $**22** and $**30 thousand**, with the highest frequencies in the $**24–30 thousand** range, indicating that mid-levels of damage are most common and that the data is slightly skewed to the right. There are fewer cases of very low or very high damages, but no obvious signs of outliers exist. The second graph shows that most locations are between **2** to **4 miles** from the nearest fire station, with the highest concentration at **2–3 miles**. This indicates that the majority of areas are within a reasonable response range for fire stations, with fewer outliers at the extremes. Taken together, these graphs show that while most properties are relatively close to fire stations, moderate levels of fire damage are still common.



To further explain the relationship between the variables, we created a simple correlation matrix between damage and distance. This correlation matrix is used to measure the strength and direction of the linear relationship between the two variables. The matrix shows a very strong positive linear relationship, with a correlation coefficient of **0.96**. This suggests that as the distance to the nearest fire station increases, the amount of fire damage also tends to increase significantly. In other words, properties located farther from fire stations are more likely to experience higher levels of fire damage.

Linear Regression Model #1

To obtain a better understanding of this relationship, we built a simple linear regression model of our two variables. Our predictor value is distance (in miles), and our response value is damage (in thousands). After generating our summary of the linear model, we find that we can use the equation:

$$\widehat{Y} = 10.2779 + 4.9193X$$

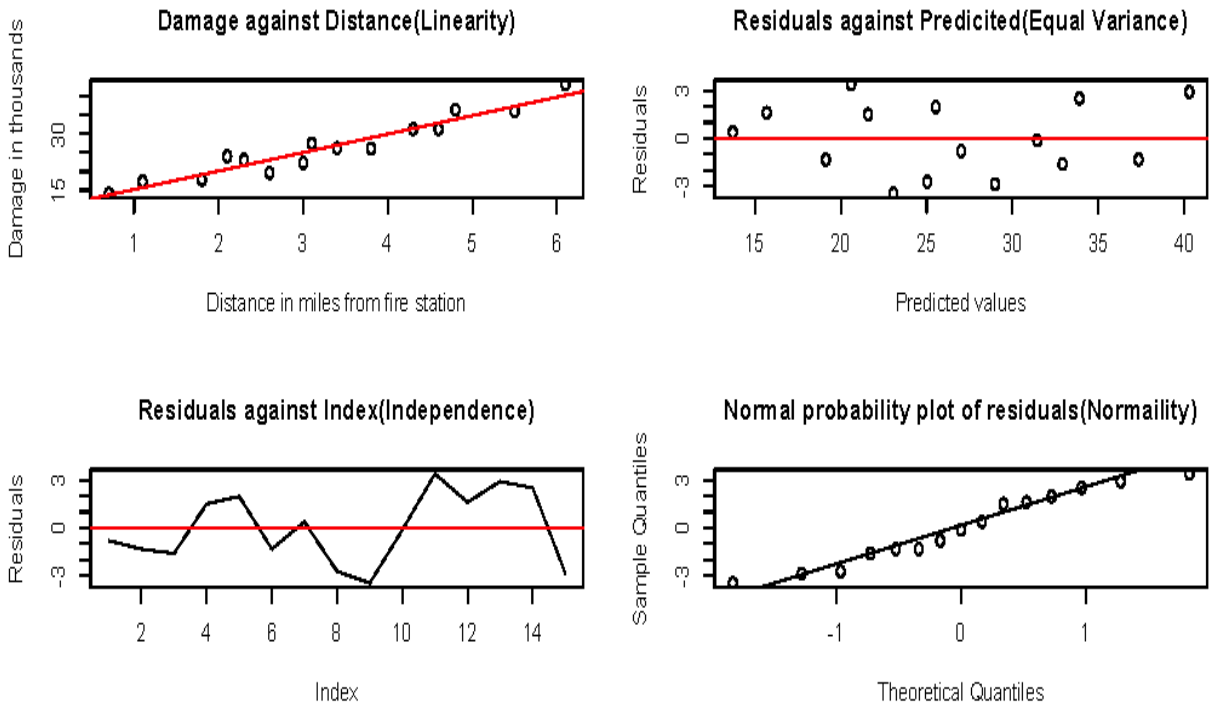
Where $\hat{\mathbf{Y}}$ is the predicted fire damage (in thousands of dollars), **X** is the distance to the nearest fire station (in miles), $\boldsymbol{\beta_0}$ **= 10.2779** is the intercept, and $\boldsymbol{\beta_1}$ **= 4.9193** is the slope (change in predicted damage per additional mile). It is found that this model has a **92.35**% R^2, which makes it a very strong and close to accurate model to use, as 92.35% of the variance in damages is explained by distance. Additionally, both the intercept and the distance are statistically significant as they have very small p-values (**6.59e-06** for the intercept and **1.25e-08** for the distance variable), which means that they contribute to the model. The following depicts all summary statistics of the model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4682 -1.4705 -0.1311  1.7915  3.3915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
DISTANCE      4.9193     0.3927  12.525 1.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared:  0.9235,    Adjusted R-squared:  0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

To evaluate the usefulness and reliability of a statistical model, particularly in the context of regression analysis, it's important to check several key assumptions: **linearity**, **homoscedasticity (equal variance)**, **independence**, and **normality**. Each plays a critical role in ensuring valid inferences and accurate predictions. We achieved this with the following plots depicted below.



To begin looking at the linear regression scatter plot, we can make some assumptions about the relationship between distance (x) and damages (y). The red line in the middle of the graph displays the linear regression line, which has a positive slope. As noted by the visual, there can be an assumption made that the relationship is strong and positive, as the points are very close to the regression line. Before we can assume that we can trust this model to give us accurate response values, we have to first check the key assumptions of linear regression using

different plots. The first assumption we can check for is the linearity assumption using the linear regression visual above. Our model fits the linearity assumption as the relationship of the plots follows a straight line.

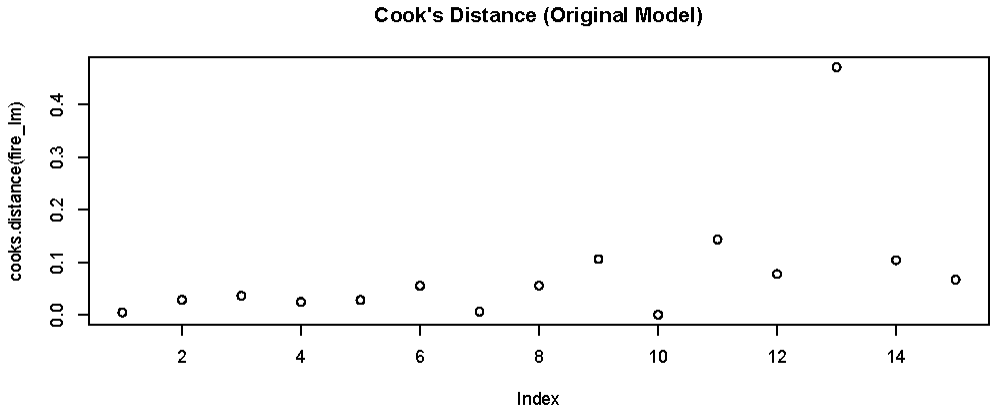
Next, we check a residual plot of the model to check for the equal variance assumption. The observation we can make from this visual is that there is no clear pattern between the plots. This indicates that the variance of the residuals is consistent, as the residuals are distributed around the red line with no signs of increasing or decreasing.

Following that, we check the independence assumption using a plot of residuals on the y-axis against the observation index on the x-axis. According to this plot, the independence assumption is not violated, as there is no clear trend in the line. This means that the observations are not dependent on each other, and the errors are independent.

Lastly, we look for the normality assumption with a normal probability plot of residuals. With this visual, it is seen that the normality assumption is not violated as the plots follow the straight 45° line. This means that the residuals are normally distributed. In conclusion, based on the fact that we have no assumptions violated and a 92.35% R^2, we can assume that this simple linear regression model will be useful in predicting damages from a given distance to the nearest fire station.

Linear Regression Model #2

While the study conducted used simple linear regression methodology, we felt that the inclusion of a second model would be beneficial. When examining the Cook's Distance plot of our original model, depicted below, there was a clear outlier.



After investigating further we found that this outlier came from the 13th instance within our dataset. To further our analysis we subset this outlier and once again built a model that regressed our predictor of damage (in thousands) against distance from nearest fire station (in miles). This new subsetted model presented with the following summary statistics.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4205 -1.3871 -0.2625  1.3800  3.2945

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.1019     1.4417    7.70 5.55e-06 ***
DISTANCE      4.5841     0.4279   10.71 1.69e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.193 on 12 degrees of freedom
Multiple R-squared:  0.9053,    Adjusted R-squared:  0.8975
F-statistic: 114.8 on 1 and 12 DF,  p-value: 1.692e-07
```

When comparing our original model to the subsetted version, we observed that the majority of the summary statistics and the significance levels of the coefficients remained relatively consistent. The second model, similarly to the first, met all assumptions needed in the context of linear regression, proving the significance of the second model. However, there were a few notable changes. The subsetted model exhibited a slight decrease in R^2, indicating a reduced explanatory power. Despite this, it showed an improvement in the overall fit of the residuals, suggesting better adherence to model assumptions such as homoscedasticity and normality of errors. Additionally, the influence of a previously identified outlier was significantly reduced, which likely contributed to the increased stability and robustness of the model estimates. Overall, while the model sacrifices a small amount of variance explanation, it gains in diagnostic and reliability.

Model Comparison

When looking at both linear regression models built, it is noticeable that they are both very strong in predicting damages based on distance. When looking closely at the resulting statistics, we have a stronger model when outliers are kept (first model), as it has a **92.35**% R^2, meanwhile when we removed the outlier, we had a **90.53**% R^2, which is a slight decrease of **1.82**%. This metric assists in drawing the conclusion that in this specific model, the outlier should be kept as it helps predict more extreme cases. Though we do have a lower R^2 in the model without the outliers, that model has increased stability as the influence of that point was reduced. See the graph below for a visualized comparison.