Predicting Fatalities from Tornado Data (Group 1)

Erika Delorme, Marcela Flaherty, Riddha Tuladhar, Edwin Yeung

Table of contents

Summary

In our project, we attempt to build a multilinear regression model that will predict the number of fatalities from tornadoes using the features width (yards) and length (miles) of the tornado. We tested our multilinear regression model with and without outliers and compared differences in coefficients and RMSPE scores. Both models had low positive coefficients, suggesting a minimal yet positive impact on the prediction of tornado fatalities, and both had low RMSPE scores, suggesting a low amount of error in its predictions. The model without outliers had a lower RMSPE score, which is partly explained by the lack of outliers and thus making predictions on a smaller range, which reduces the error. Despite the limitations of our model, we believe that it can still have some utility in predicting tornado fatalities with little error. However, the model should be improved in the future before being deployed to improve the size of the coefficients and its predictive power. In the future, we may consider exploring other features in predicting fatalities, predicting the number of injuries from the same features, or

even predicting the number of casualties (injuries and fatalities) from the same and additional features.

Introduction

Tornadoes are a common type of natural disaster in the United States; in fact, the United States gets more tornadoes than any other country at over 1,150 thousand recorded every year (Chinchar 2022). Furthermore, the United States has experienced many of the most violent tornadoes, with 59 of the 67 most violent tornadoes in recorded history taking place in the country (Storm Prediction Center 2023). As such, the ability to predict the number of fatalities caused by these tornadoes based on their physical characteristics is desirable in order to employ preventative measures and reduce the number of casualties caused by these disasters.

This project will be using a data set from the US NOAA's Storm Prediction Cente (Storm Prediction Center 2023), which contains information on all tornadoes recorded in the United States from 1950 to 2022. For each tornado, the data set records many of its features, including but not limited to its length, width, the state in the US, the date and time it occurred, the number of fatalities and the number of injuries it caused, and the financial losses it incurred. Using this information, this project will use a multivariable linear regression to answer the question "How does the length and width of a tornado affect the number of fatalities it causes?"

Methods

Data wrangling and cleaning

The R programming language (R Core Team 2022) and the following R packages were used to conduct our analysis: repr (Angerer 2023), tidyverse (Wickham and RStudio 2017), tidymodels (Kuhn et al. 2023), psych (Revelle 2019), and GGally (Schloerke 2020).

The code used to perform the analysis and create this report can be found here: https://github.com/DSCI-310-2024/DSCI-310-Group-1-Predict-Fatalities-From-Tornado-Data/blob/main/src/tornado_fatalities_predictor.ipynb.

1.First, we load the necessary packages.

```
Warning: package 'kableExtra' was built under R version 4.3.3
```

2.Then, we read in the data about tornadoes directly from the website.

om	yr	mo	dy	date	time	tz	datetime_utc	st	stf	mag
192	1950	10	1	1950-10-01	21:00:00	America/Chicago	1950-10-02T03:00:00Z	OK	40	1
193	1950	10	9	1950-10-09	02:15:00	America/Chicago	1950-10-09T08:15:00Z	NC	37	3
195	1950	11	20	1950-11-20	02:20:00	America/Chicago	1950-11-20T08:20:00Z	KY	21	2
196	1950	11	20	1950-11-20	04:00:00	America/Chicago	1950-11-20T10:00:00Z	KY	21	1
197	1950	11	20	1950-11-20	07:30:00	America/Chicago	1950-11-20T13:30:00Z	MS	28	1
194	1950	11	4	1950-11-04	17:00:00	America/Chicago	1950-11-04T23:00:00Z	PA	42	3

3.Next, we wrangle and clean the data. Firstly, we check for missing values in our cleaned data.

	x
om	0
yr	0
mo	0
dy	0
date	0
time	0
tz	0
datetime_utc	0
st	0
stf	0
mag	756
inj	0
fat	0
loss	27170
slat	0
slon	0
elat	0
elon	0
len	0
wid	0
ns	0
sn	0
f1	0
f2	0
f3	0

f4	0
fc	0

We can see that there are no missing values, except for the feature `loss`, with 27,170 missing values and 756 missing values for the feature mag. The feature `loss` refers to the financial loss of each tornado. For our regression problem, we did not deem this feature to be an important feature. Therefore, we decide to remove this column.

Because there are not so many rows missing for the feature `mag`, we decide to filter those rows out.

4.Then, we remove irrelevant or repetitive columns and then filter for missing values of the column `mag`, which stands for magnitude. Then, we change the feature names to make them more descriptive. We show the top 6 rows of our cleaned data.

ID	year	month	day	time	datetime_utc	state	mag	injuries	fatalities	start_lat
192	1950	10	1	21:00:00	1950-10-02T03:00:00Z	OK	1	0	0	36.73
193	1950	10	9	02:15:00	1950-10-09T08:15:00Z	NC	3	3	0	34.17
195	1950	11	20	02:20:00	1950-11-20T08:20:00Z	KY	2	0	0	37.37
196	1950	11	20	04:00:00	1950-11-20T10:00:00Z	KY	1	0	0	38.20
197	1950	11	20	07:30:00	1950-11-20T13:30:00Z	MS	1	3	0	32.42
194	1950	11	4	17:00:00	1950-11-04T23:00:00Z	PA	3	1	0	40.20

5.We split our data into two sets: `train_df` and `test_df`. `train_df` consists of 75% of our original data set and is used to train our regression model. The remaining 25% of our original data is `test_df`, which we use later to test the accuracy of our model at prediction.

EDA: Descriptive statistics and visualizations

6.Next, we create a summary table of features that could be useful to use in our regression model.

Table 1: Summary table of numerical features.

	vars	n	mean	sd	min	max	range	se
mag	1	50952	0.7814217	0.8959607	0.0000	5.0000	5.0000	0.0039692
injuries	2	50952	1.4762914	19.3663368	0.0000	1740.0000	1740.0000	0.0857960
fatalities	3	50952	0.0916941	1.5975758	0.0000	158.0000	158.0000	0.0070775
start_lat	4	50952	37.1099379	5.1003232	17.7212	61.0200	43.2988	0.0225952

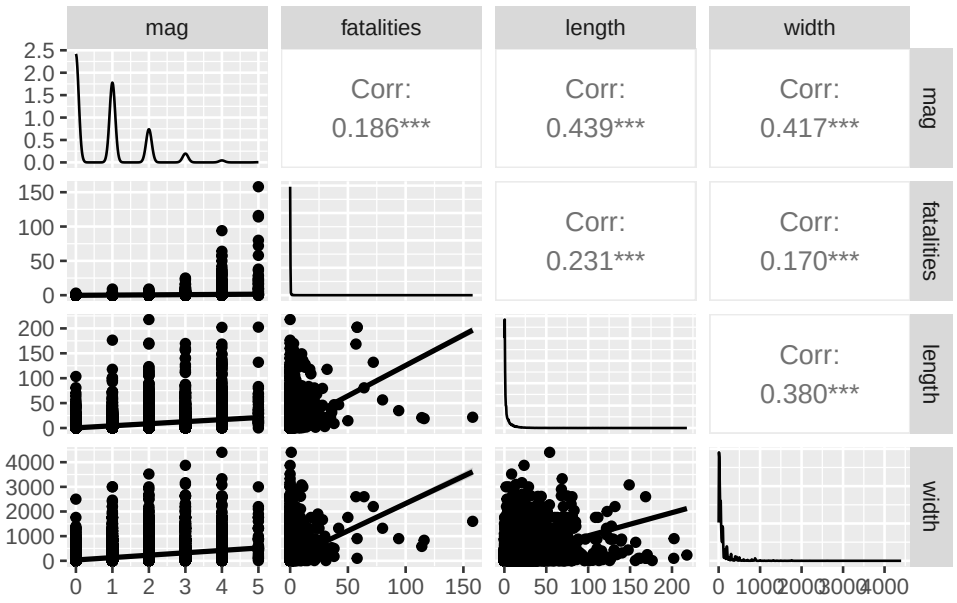


Figure 1: Correlation matrix of important numeral features and target.

Table 1: Summary table of numerical features.

	vars	n	mean	sd	min	max	range	se
start_lon	5	50952	-92.7131385	8.6782524	-163.5300	-64.7151	98.8149	0.0384460
end_lat	6	50952	22.8084368	18.5508885	0.0000	61.0200	61.0200	0.0821834
end_lon	7	50952	-56.4335822	45.3795249	-163.5300	0.0000	163.5300	0.2010385
length	8	50952	3.5363974	8.4491601	0.0000	217.8000	217.8000	0.0374311
width	9	50952	108.9442809	208.6282187	0.0000	4400.0000	4400.0000	0.9242563
ns	10	50952	1.0089692	0.0965440	1.0000	3.0000	2.0000	0.0004277

7.We create a correlation matrix to view the correlations between features related to `injuries` and `fatalities`.

From the correlation matrix, we observe that there are some small sized correlations between `fatalities` and other continuous features. For example, there is a correlation of `0.231` between `length` and `fatalities` and a correlation of `0.170` between `width` and `fatalities`.

There are also correlations between features that are not the target. For example, there is medium sized correlation of `0.439` between `length` and `mag` and a correlation of `0.417` between `width` and `mag`.

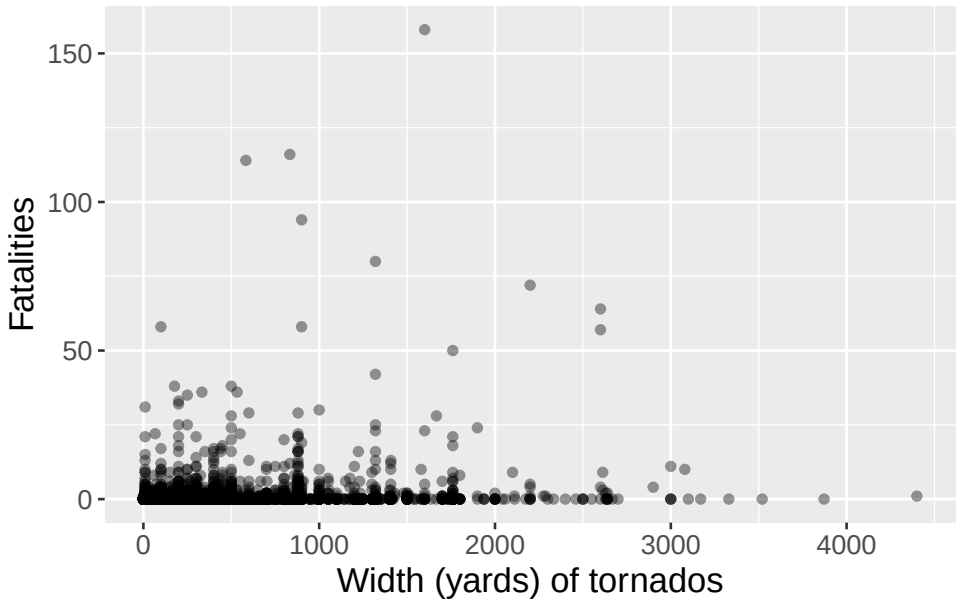


Figure 2: Scatterplot of width (yards) of tornado and fatalities.

These small to medium correlations may suggest that these features could be useful in predicting fatalities.

Based on the correlation matrix, we decide to use `length` and `width` as features in our model as they are numerical unlike `mag`.

Model selection and analysis

8.As mentioned previously, our group has decided that we will use a linear regression model to predict tornado casualties. We will be using the variables `width` and `length` as our predictors. The former is a measure of the width of a tornado, while the latter is a measure of length. The units of measurements are yards and miles, respectively.

Now that we have our training data `train_df`, we can fit our linear regression model. We will first specify our model, and then proceed to fit our model and obtain the regression coefficients.

```
== Workflow [trained] ==========================================================
Preprocessor: Recipe
Model: linear_reg()

-- Preprocessor ----------------------------------------------------------------
0 Recipe Steps
```

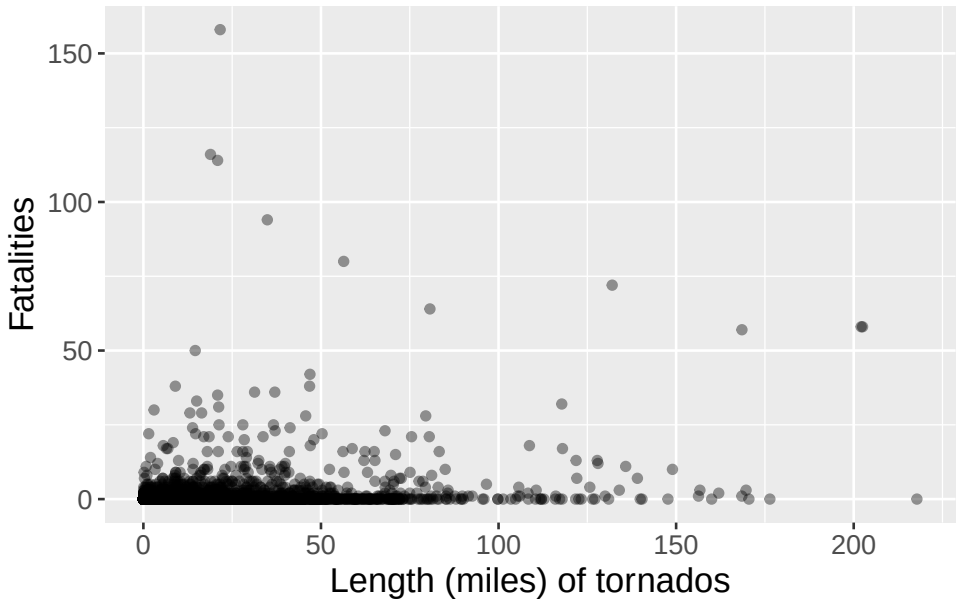


Figure 3: Scatterplot of length (miles) of tornado and fatalities.

```
-- Model ---------------------------------------------------------------------

Call:
stats::lm(formula = ..y ~ ., data = data)

Coefficients:
(Intercept)       length        width
 -0.1183680    0.0367960    0.0007337
```

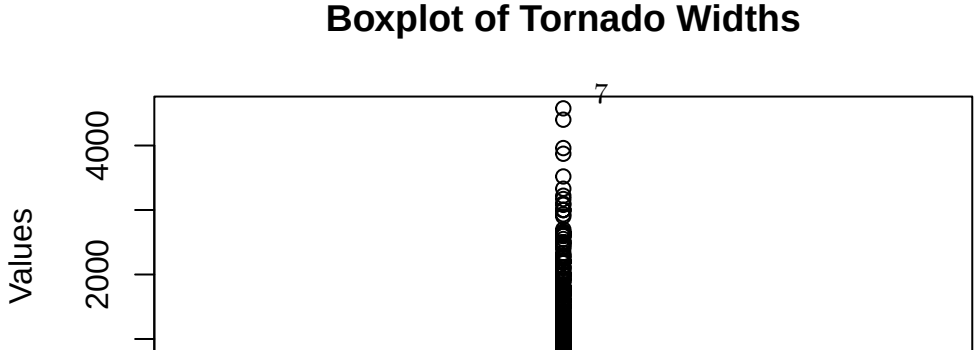
9.Now that we have our model, we can predict on the testing data `test_df` to assess how well it does.

Table 2: Results from model with outliers.

.metric	.estimator	.estimate
rmse	standard	1.0056770
rsq	standard	0.0898357
mae	standard	0.2298536

10.We can visualize our linear regression model to get a better idea of how well it performs.

Note there are clear outliers in the data, thus we will perform the same analysis after removing the outliers.



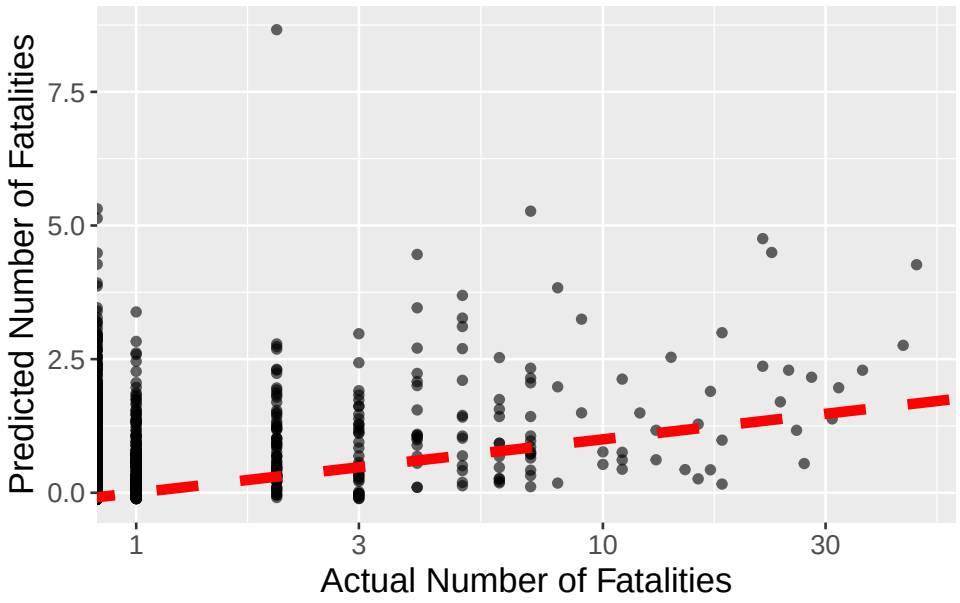
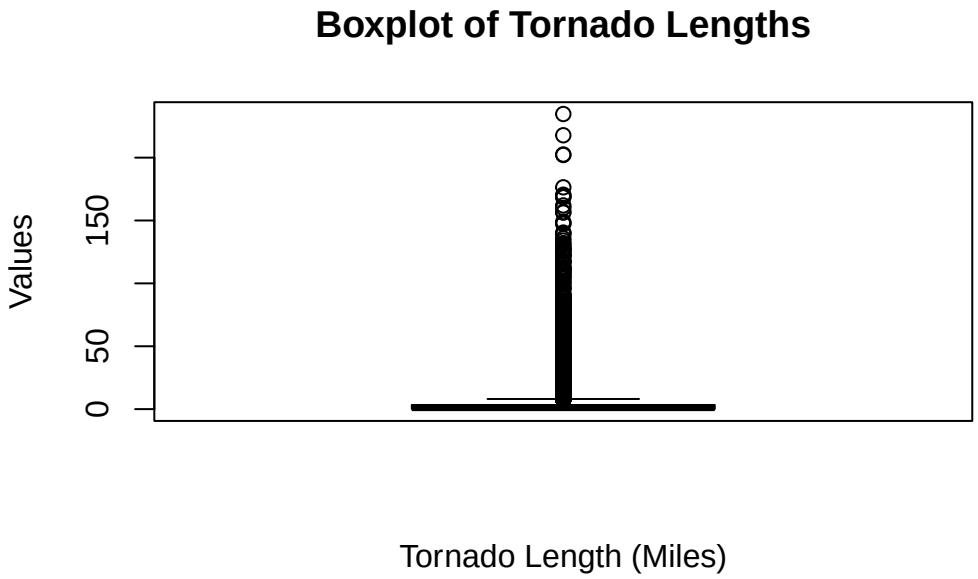


Figure 4: Actual Number of Fatalities vs Predicted Number of Fatalities.



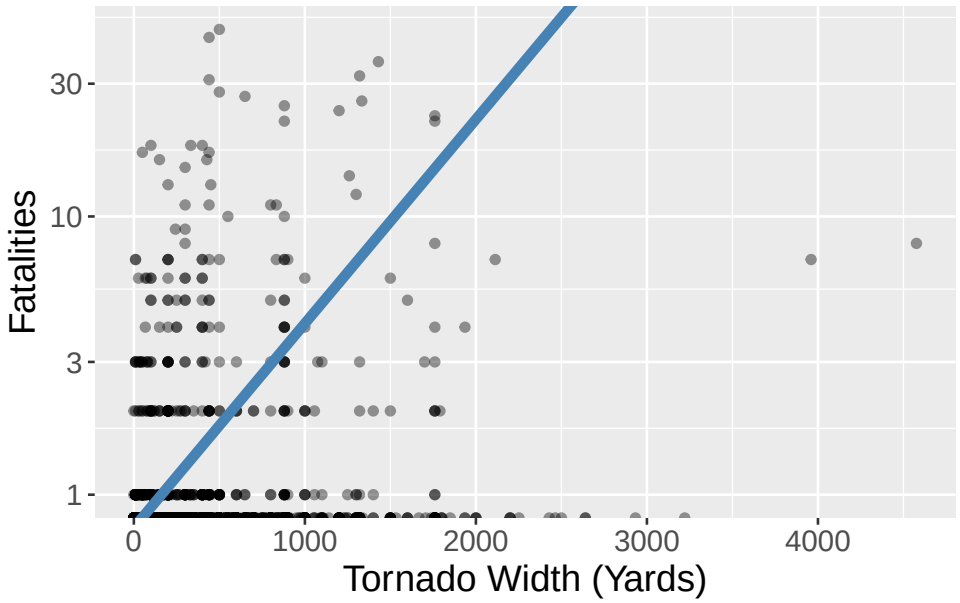
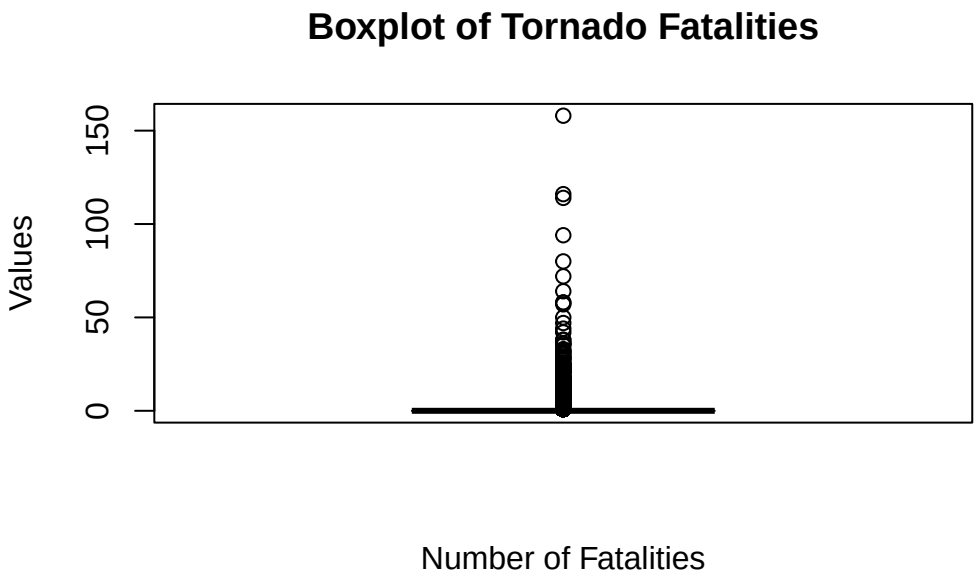


Figure 5: Fatalities vs Width Plot.



Note that, as the majority of tornadoes cause no fatalities and thus removing these outliers will leave us with data only containing tornadoes that caused no death, we will not be filtering outliers for fatalities.

```
== Workflow [trained] ==========================================================
Preprocessor: Recipe
Model: linear_reg()
```

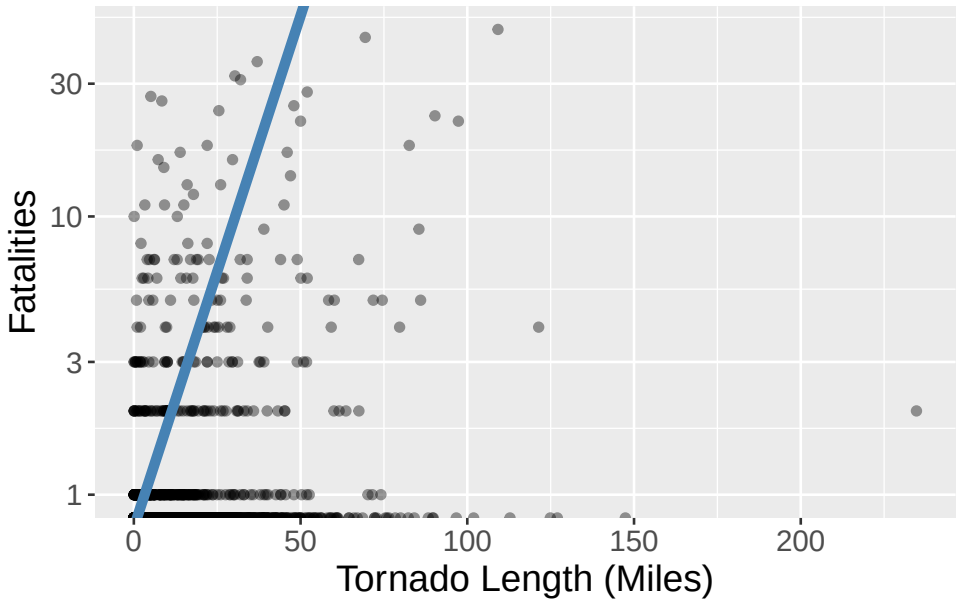


Figure 6: Fatalities vs Length Plot.

```
-- Preprocessor ----------------------------------------------------------------
0 Recipe Steps

-- Model -----------------------------------------------------------------------

Call:
stats::lm(formula = ..y ~ ., data = data)

Coefficients:
(Intercept)       length        width
 -0.0022491    0.0048939    0.0001736
```

Table 3: Results from model without outliers.

.metric	.estimator	.estimate
rmse	standard	0.1894697
rsq	standard	0.0046463
mae	standard	0.0242242

Discussion and Results

Our initial exploration of features width and length demonstrated a positive correlation with our target, fatalities. Length and fatalities show a correlation of `0.237`, whereas width and fatalities show a correlation of `0.174`. The values are small but illustrate a stronger positive relationship between length and fatalities.

Before constructing our model, we performed a 75/25 split on our data for reproducibility and validity.

From our multivariable linear regression model with the inclusion of outliers, we can write an equation of best fit:

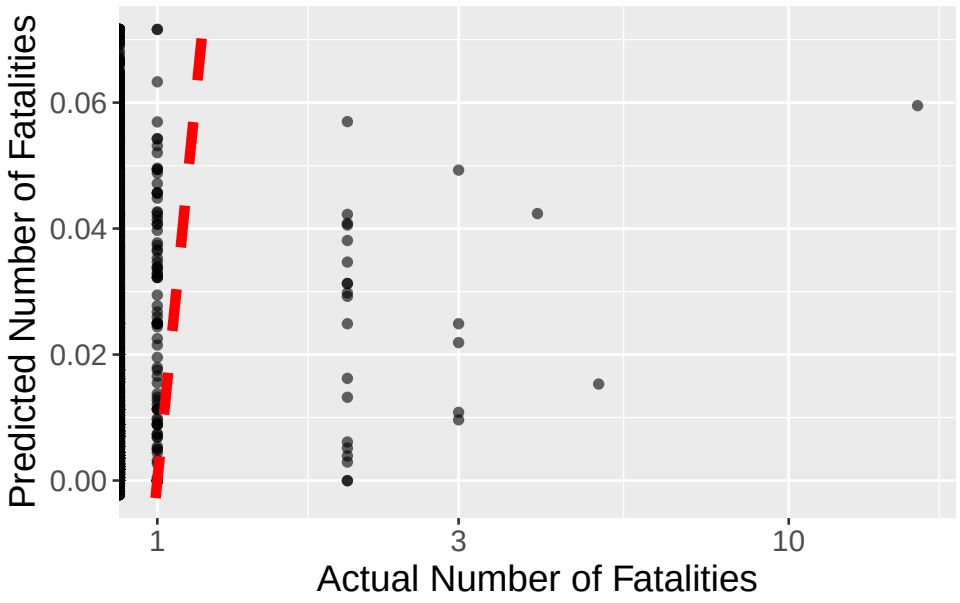


Figure 7: Actual Number of Fatalities vs Predicted Number of Fatalities.

On the other hand, we obtain an RMSPE score for this multivariable linear regression of `1.00567700` tornado fatalities (Table 2). This prediction error is very low, indicating that our model fits the data well and our predictions are more precise.

As noted in our analysis section, removing the outliers leaves us with tornadoes that cause no deaths. Nonetheless, we decided to keep the linear regression model without the outliers to observe if any differences were seen in the equation and RMSPE score.

The equation of best fit for our multilinear regression model without the outliers is:

$$\text{Tornado fatalities} = 0.0048939 \times (\text{length of tornado}) + 0.0001736 \times (\text{width of tornado}) - 0.0022491$$

Compared to the model including outliers, the coefficients and the intercept are much smaller, implying that these features have less of an impact on predicting tornado fatalities and are more likely to lead to less precise predictions.

The RMSPE score here is `0.189469674` tornado fatalities (Table 3), suggesting that it makes fewer errors than the model with the outliers.

From Figure 5, Figure 6, Figure 8, and Figure 9, the plots from the model with the outliers seem more interpretable than the ones from the model without the outliers, as the slope is an almost flat one. On the other hand, by observing Figure 4 and Figure 7, the plots showing the actual number of fatalities vs the predicted number of fatalities of both models, we can conclude that the reason why the model without the outliers has a lower RMSPE score is

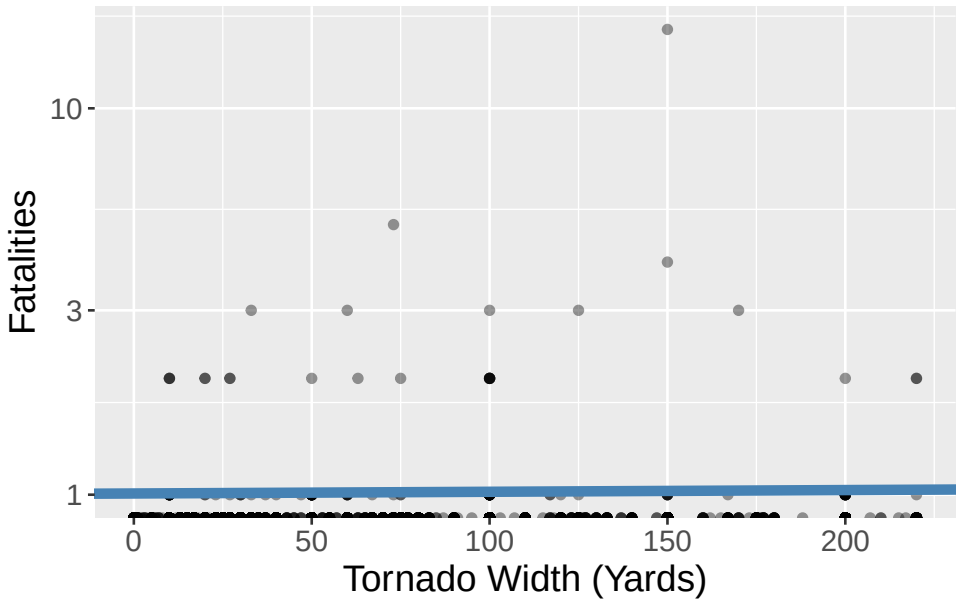


Figure 8: Fatalities vs Width Plot.

because the outliers were removed, thus making it easier to predict the fatalities that are more closely located than fatalities that are along a longer range.

Therefore, we can conclude that our model has a lower RMSPE score, which is preferable in a multilinear regression model. However, it also does not have very strong coefficients for predicting the number of fatalities.

In terms of expectations, given the small correlations that length and width have with the number of tornado fatalities, it is expected that our model does not have very large coefficients. In terms of the RMSPE score, we did not have an expectation as to how our model would perform. However, we did not expect the RMSPE score to be so low as typically if the coefficients are smaller, then they have less impact on the predictions and thus may lead to more erroneous predictions.

Despite the low coefficients of our model equations, the low RMSPE scores suggest that if this model were to be deployed to predict the number of fatalities from tornadoes in the U.S., then it may have the ability to predict them without so much error. Therefore, our model could have some utility in such aspects.

However, we believe that because of the limitations of our model, it would be interesting to observe the impact of using other features to predict the number of fatalities from tornadoes, such as `mag`. Furthermore, it could also be valuable to explore how well we could predict the number of injuries using the same features that we did, namely `length` and `width`. The total number of casualties (fatalities and injuries) could also be predicted using `length` and `width` and perhaps other features.

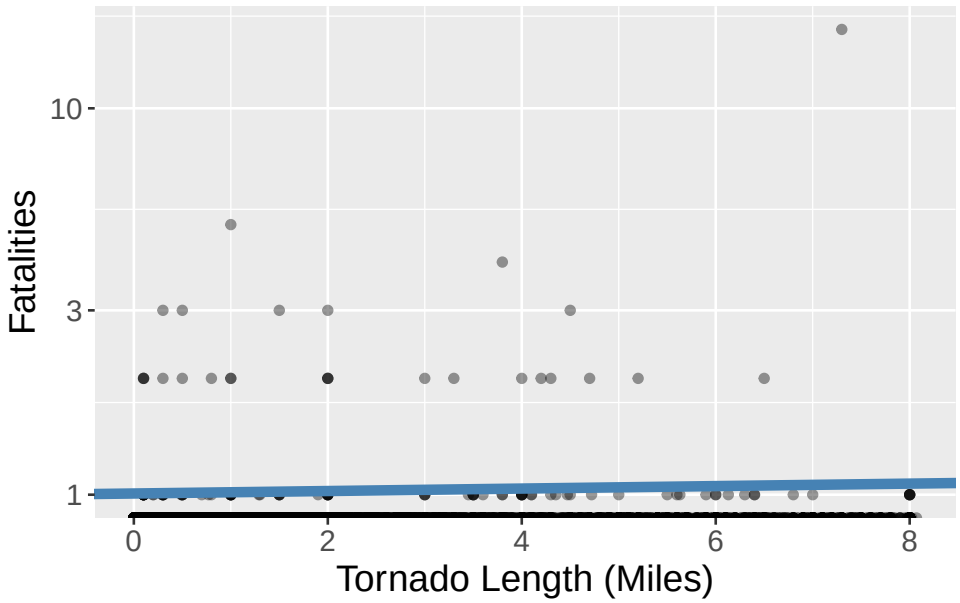


Figure 9: Fatalities vs Length Plot.

References

Angerer, Kluyver, P. 2023. "Repr (1.1.6): Serializable Representations." CRAN. https://cran.r-project.org/package=repr.

Chinchar, Allison. 2022. "Here's Why the US Has More Tornadoes Than Any Other Country." CNN. November 28, 2022. https://www.cnn.com/2022/11/28/weather/us-leads-tornado-numbers-tornado-alley-xpn/index.html.

Kuhn, Max, Hadley Wickham, P. Software, and PBC. 2023. *Tidymodels (1.1.0): Easily Install and Load the "Tidymodels" Packages*. https://cran.r-project.org/package=tidymodels.

R Core Team. 2022. *R (4.3.2): A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Revelle, William. 2019. *Psych (2.3.3): Procedures for Psychological, Psychometric, and Personality Research*. https://cran.r-project.org/package=psych.

Schloerke, Cook, B. 2020. "GGally (2.1.2): Extension to "Ggplot2"." CRAN. https://cran.r-project.org/package=GGally.

Storm Prediction Center. 2023. "F5 and EF5 Tornadoes of the United States - 1950-Present (SPC)." Noaa.gov; NOAA's National Weather Service. February 19, 2023. https://www.spc.noaa.gov/faq/tornado/f5torns.html.

Wickham, Hadley, and RStudio. 2017. *Tidyverse (2.0.0): Easily Install and Load the "Tidyverse"*. https://cran.r-project.org/package=tidyverse.