Descriptive Data Analysis of Harmful Tornadoes in the United States in 2017

Data Analytics I – ADS522

James Sears

Applied Data Science Program, Graduate School, Longmeadow, MA

Submitted To: Professor Aja Shabana

Spring 2021

Jim Sears
ADS522Z2 – Data Analytics I
R Project 1

Objective

The objective of this project is to describe a dataset consisting of qualitative and quantitative variables by means of organizing and plotting them with R and, where applicable, determine the summary statistics and describe the shapes of their distributions.

Data Source & Description

The data for this project is from the Tornadoes_2017.csv data file provided with the electronic textbook Statistics: Informed Decisions Using Data. It includes various attributes for 1,472 tornadoes that occurred in the United States in 2017 such as date, location, width, path length, property loss, as well as human injuries and fatalities. The variables of interest in the data set exhibit a wide range of values, are greatly skewed, and contain many outliers. The resulting plots and summaries[1] are not ideal for the objective of this project. Therefore, this project is focused on the 88 tornadoes from the data set that are termed "harmful", those that resulted in human injuries and/or fatalities. Those variables are explored along with the length of the tornado path to see if there is any correlation. The location of each tornado is also explored to determine which states are the most prone to harmful tornadoes.

Jim Sears
ADS522Z2 – Data Analytics I
R Project 1

Qualitative Variable Analysis

State. The location of the tornadoes is available by state; this is a qualitative variable because it is descriptive and enables classification and does not enable meaningful arithmetic operations as it is non-numeric (Sullivan, page 7).

There are 26 states that experienced at least one harmful tornado in 2017. Table 1 displays the frequency (and relative frequency) of tornadoes by state. Tornadoes are typically associated with the central plains states like Texas, Oklahoma, and Nebraska, but there are harmful tornadoes in states that may be surprising, such as Massachusetts, Maine, Wisconsin and Wyoming. It is notable that most states, 14 of the 26, experienced just one or two tornadoes.

Table 1. Frequency tables of harmful tornadoes by state: alphabetical and by greatest number.

State	Frequency	Relative Frequency
AL	4	0.045
AR	5	0.057
DC	1	0.011
FL	1	0.011
GA	6	0.068
IA	3	0.034
IL	5	0.057
IN	5	0.057
KS	1	0.011
KY	2	0.023
LA	11	0.125
MA	1	0.011
MD	1	0.011
ME	1	0.011
MO	10	0.114
MS	5	0.057
NC	3	0.034
NE	1	0.011
OH	2	0.023
OK	2	0.023
PA	1	0.011
SC	5	0.057
TN	1	0.011
TX	8	0.091
WI	1	0.011
WY	2	0.023
Total	**88**	**1.000**

State	Frequency	Relative Frequency
LA	11	0.125
MO	10	0.114
TX	8	0.091
GA	6	0.068
AR	5	0.057
IL	5	0.057
IN	5	0.057
MS	5	0.057
SC	5	0.057
AL	4	0.045
IA	3	0.034
NC	3	0.034
KY	2	0.023
OH	2	0.023
OK	2	0.023
WY	2	0.023
DC	1	0.011
FL	1	0.011
KS	1	0.011
MA	1	0.011
MD	1	0.011
ME	1	0.011
NE	1	0.011
PA	1	0.011
TN	1	0.011
WI	1	0.011
Total	**88**	**1.000**

It is also notable that the three states – Louisiana (LA), Missouri (MO), and Texas (TX) – that experienced the most tornadoes accounted for approximately a third of all tornadoes. These states that are most prone to dangerous tornadoes are demonstrated by the largest areas of the pie chart in Figure 1 and the leftmost bars of Figure 2.

Figure 1. Pie chart of harmful tornadoes by state.

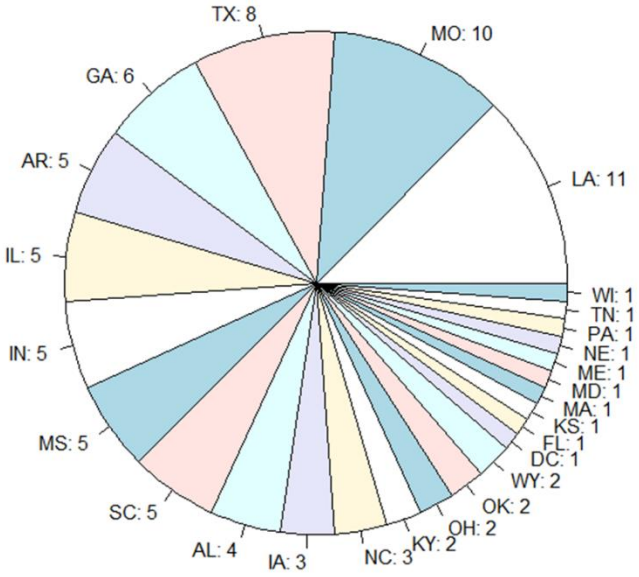
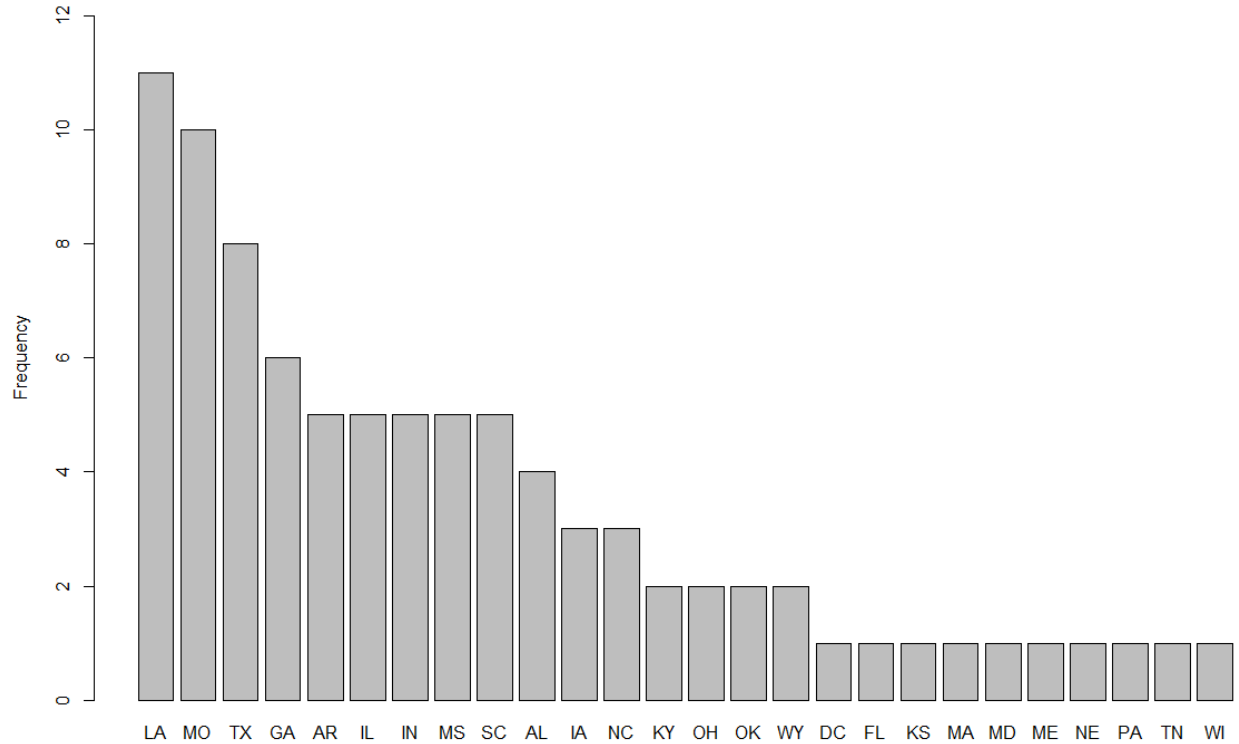


Figure 2. Bar chart of harmful tornadoes by state.



Although it is interesting to see which states have more or fewer harmful tornadoes, there are a few considerations that should be made when making comparisons. One is the size of each state that varies greatly. Some states may be more prone to tornadoes simply because they are larger, and the square mileage of each state should be factored in to normalize the data. For example, Texas experienced eight times as many harmful tornadoes than Massachusetts, but it is more than 25 times larger (268,597 square miles versus 10,565 square miles). As a result, Massachusetts has more than three times more tornadoes per square mile (9.47×10^{-5}) than Texas (2.98×10^{-5}).

Another detail to consider is the awareness and responsiveness of the states that are more prone to tornadoes. Those states are likely to have more robust warning systems and residents that may be more accustomed to tornado protocols, both of which can reduce the number of injuries and fatalities caused by a tornado.

Quantitative Variable Analysis

Length, in miles. The length of the path of each harmful tornado is a quantitative variable because it is a numeric measurement that can be subjected to arithmetic operations to provide meaningful results (Sullivan, page 7). Further, it is a continuous variable because there are an infinite number of possible values that are not countable (Sullivan, page 8).

The summary statistics of the Length data (Table 2) of the 88 harmful tornadoes reveal that the shortest tornado was just 0.25 miles long, and the longest was 82.53 miles long. Within that 82.28 mile range, the mean length is 12.05 miles, while the median is just 7.24 miles. A median that is less than the mean indicates that most of the harmful tornadoes are shorter than the average, but that there are several longer ones to draw the average of the data up. This is exhibited by the frequency table (Table 3) that shows that at least 50 of the 88 tornadoes were between 0 and 10 miles long (below the average length) and the remaining observations are spread out over a range of 75 miles.

Table 2. Summary Statistics of the length of the path of harmful tornadoes.

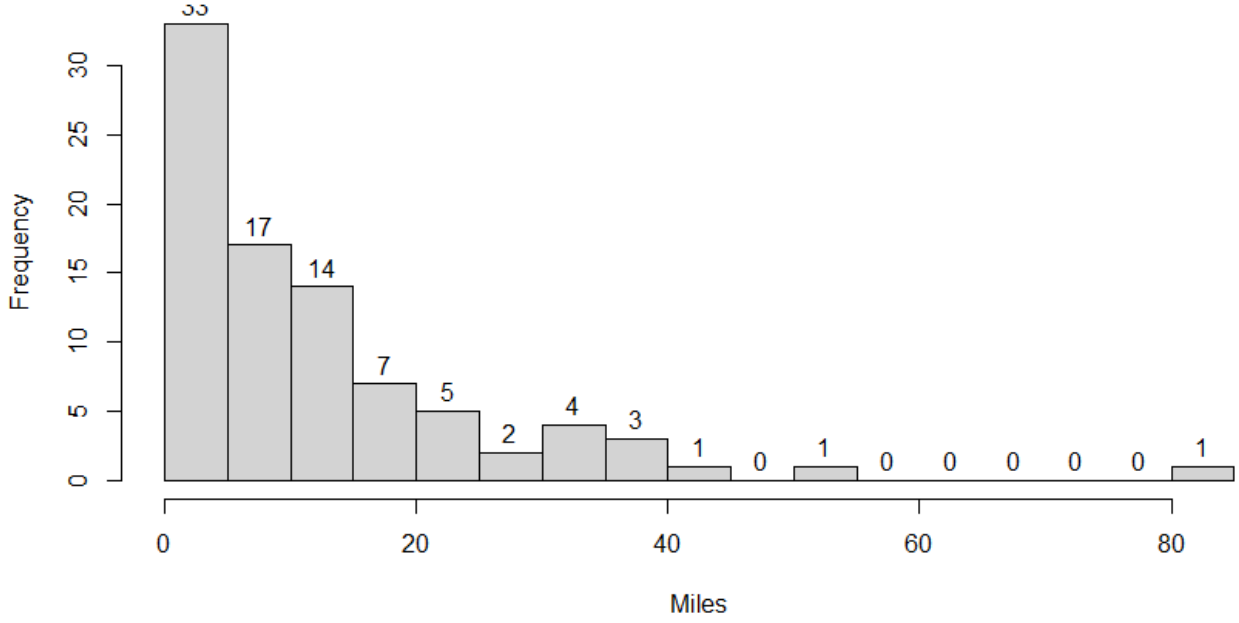
Statistic	Value
Unit	Miles
Mean	12.050
Std Dev	13.816
Minimum	0.250
1st Quartile	2.095
Median	7.240
3rd Quartile	17.565
Maximum	82.530
Range	82.280
Interquartile Range	15.355
Lower Fence	-20.908
Lower Outliers	0.000
Upper Fence	40.508
Upper Outliers	3.000

Table 3. Frequency table of the length of the path of harmful tornadoes.

Classes	Frequency	Relative Frequency
[0,5)	33	0.375
[5,10)	17	0.193
[10,15)	14	0.159
[15,20)	7	0.080
[20,25)	5	0.057
[25,30)	2	0.023
[30,35)	4	0.045
[35,40)	3	0.034
[40,45)	1	0.011
[45,50)	0	0.000
[50,55)	1	0.011
[55,60)	0	0.000
[60,65)	0	0.000
[65,70)	0	0.000
[70,75)	0	0.000
[75,80)	0	0.000
[80,85)	1	0.011

This skewed data is best exhibited by the resultant histogram (Figure 3) that shows the tallest bars to the far left (lower values) and shorter and shorter bars as the number of miles increases. This distribution that tails off to the right is termed “right skewed” and therefore the best method of central tendency is the median and not the mean.

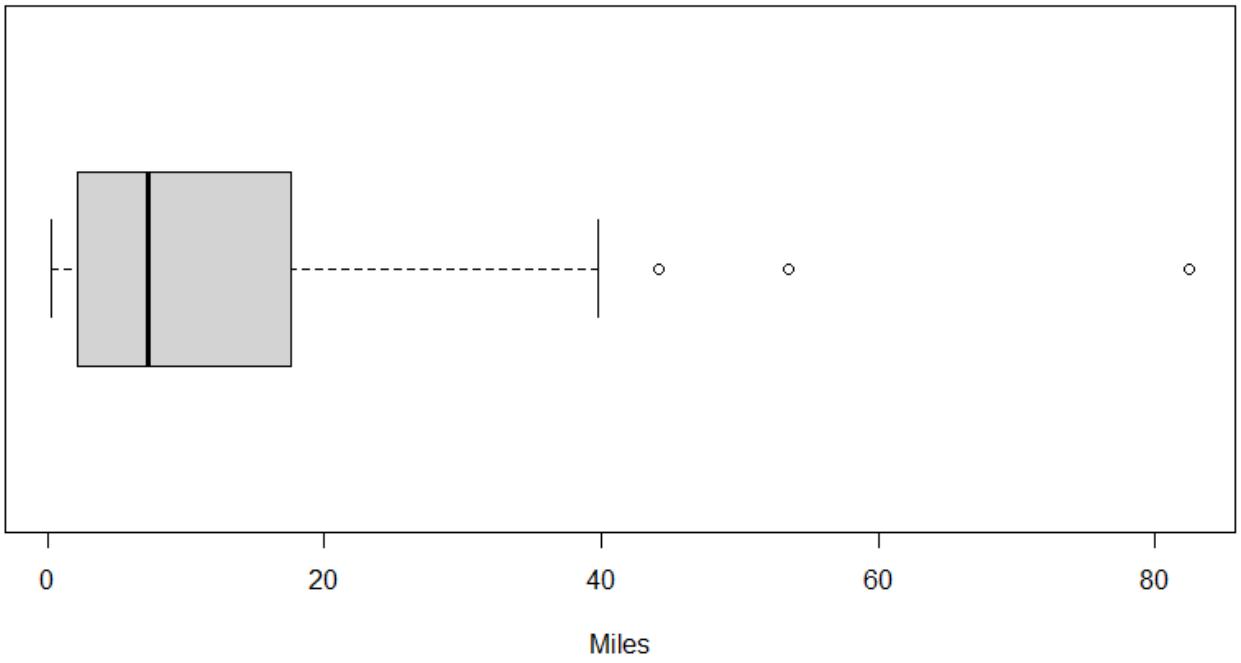
Figure 3. Histogram of the length of the path of harmful tornadoes.



As the distribution of the Length data is not normal, we cannot apply the Empirical rule to describe its variance. Instead, we can employ Chebyshev's Inequality to calculate that at least 75% (66 of 88) of the tornadoes' lengths should be between 0 (mean of 12.05 – 2 * standard deviation of 13.816) and 39.682 (mean of 12.05 + 2 * standard deviation of 13.816) miles. This is confirmed as the data reveals that 84 of the 88 (95.5%) of the observations are within this range.

The quartile statistics (Table 2) show that 50% of harmful tornadoes are between 2.095 and 17.565 miles in length, and that 25% are less than 2.095 miles in length, and the remaining 25% are more than 17.565 miles in length. This once again demonstrates the right-skewedness of the data that is also exhibited in its box plot (Figure 4) with a right whisker that is much longer than the left whisker. (Also noted is the median line that is left of the center of the box).

Figure 4. Boxplot of the length of the path of harmful tornadoes.



With heavily skewed data and a relatively large range of observations, the presence of outliers is expected. The calculated lower (0: Quartile 1 of 2.095 – 1.5 * Interquartile Range of 15.355) and upper (40.508: Quartile 3 of 17.565 + 1.5 * IQR of 15.35) outlier fences reveal three observations that are considered outliers. These coincide with tornadoes that were 44.13, 53.47, and 82.53 miles in length, indicated by the circles of the box plot (Figure 4).

In summary, the length of harmful tornadoes ranges from 0.25 to 82.53 in length and are best described as having a median length of 7.24 miles. This means that harmful tornadoes are relatively shorter in length, but longer ones do exist, but are much rarer.

Injuries and/or Fatalities, in number of people. The combined number of people injured or killed by each harmful tornado is a quantitative variable because it is a numeric measurement that can be subjected to arithmetic operations to provide meaningful results (Sullivan, page 7). Further, it is a discrete variable because there are a countable number of values (Sullivan, page 8) as people are whole units unto themselves.

The summary statistics of the Injuries and Fatalities data (Table 4) of the 88 harmful tornadoes reveal that the least harmful resulted in one injury or fatality, and the most harmful resulted in 61 injuries and/or fatalities. Within that range of 60 people, the mean is 5.466 people, while the median is just 1 person. A median that is less than the mean indicates that most of the harmful tornadoes injure and/or kill fewer people than the average, but that there are several that are very harmful and draw the average of the data up. This is exhibited by the frequency table (Table 5) that shows that 48 of the 88

Table 4. Summary Statistics of the injuries and fatalities per harmful tornado.

Statistic	Injuries & Fatalities per Tornado
Unit	People
Mean	5.466
Std Dev	10.608
Minimum	1.000
1st Quartile	1
Median	1
3rd Quartile	3.5
Maximum	61
Range	60
Interquartile Range	2.5
Lower Fence	-2.750
Lower Outliers	0.000
Upper Fence	7.250
Upper Outliers	15.000

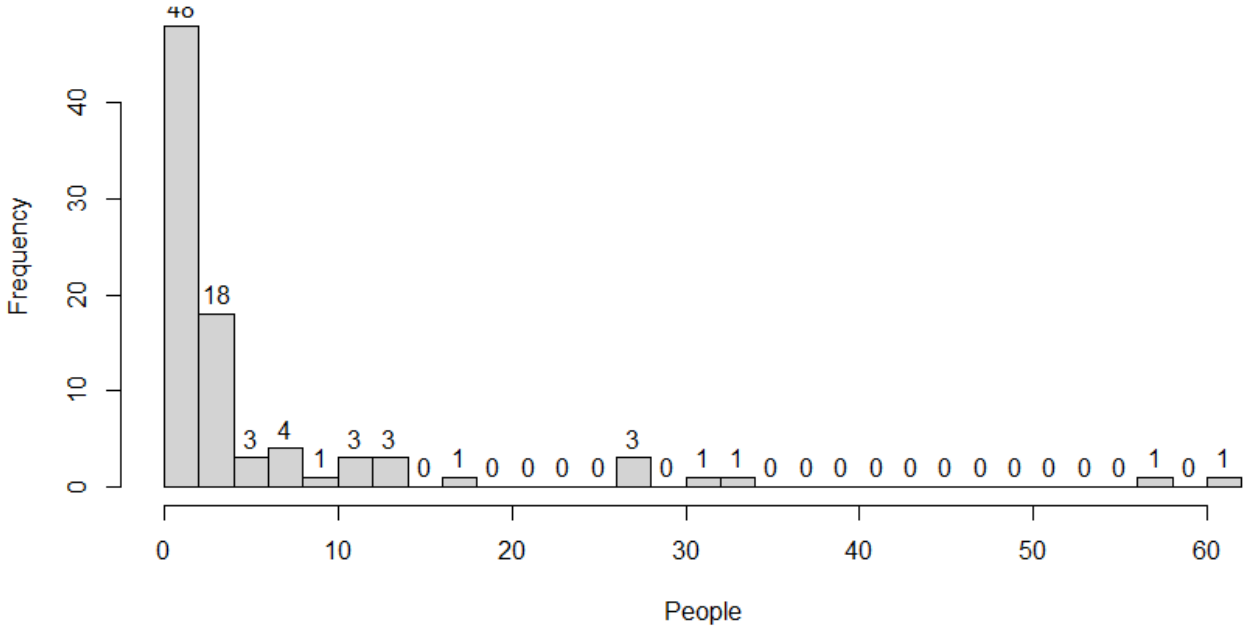
Table 5. Frequency table of the injuries and fatalities per harmful tornado.

Classes	Frequency	Relative Frequency
[0,2)	48	0.545
[2,4)	18	0.205
[4,6)	3	0.034
[6,8)	4	0.045
[8,10)	1	0.011
[10,12)	3	0.034
[12,14)	3	0.034
[14,16)	0	0
[16,18)	1	0.011
[18,20)	0	0
[20,22)	0	0
[22,24)	0	0
[24,26)	0	0
[26,28)	3	0.034
[28,30)	0	0
[30,32)	1	0.011
[32,34)	1	0.011
[34,36)	0	0
[36,38)	0	0
[38,40)	0	0
[40,42)	0	0
[42,44)	0	0
[44,46)	0	0
[46,48)	0	0
[48,50)	0	0
[50,52)	0	0
[52,54)	0	0
[54,56)	0	0
[56,58)	1	0.011
[58,60)	0	0
[60,62)	1	0.011

tornadoes (55%) injured or killed 1 person, 66 of 88 (75%) injured or killed between 1 and 3 people, inclusive, and the remaining 22 (25%) observations are spread out over a range of 4 to 61 people.

This skewed data is best exhibited by the resultant histogram (Figure 5) that shows the tallest bars to the far left (lower values) and shorter and shorter bars as the number of people injured or killed increases. This distribution that tails off to the right is termed "right skewed" and therefore the best method of central tendency is the median and not the mean.

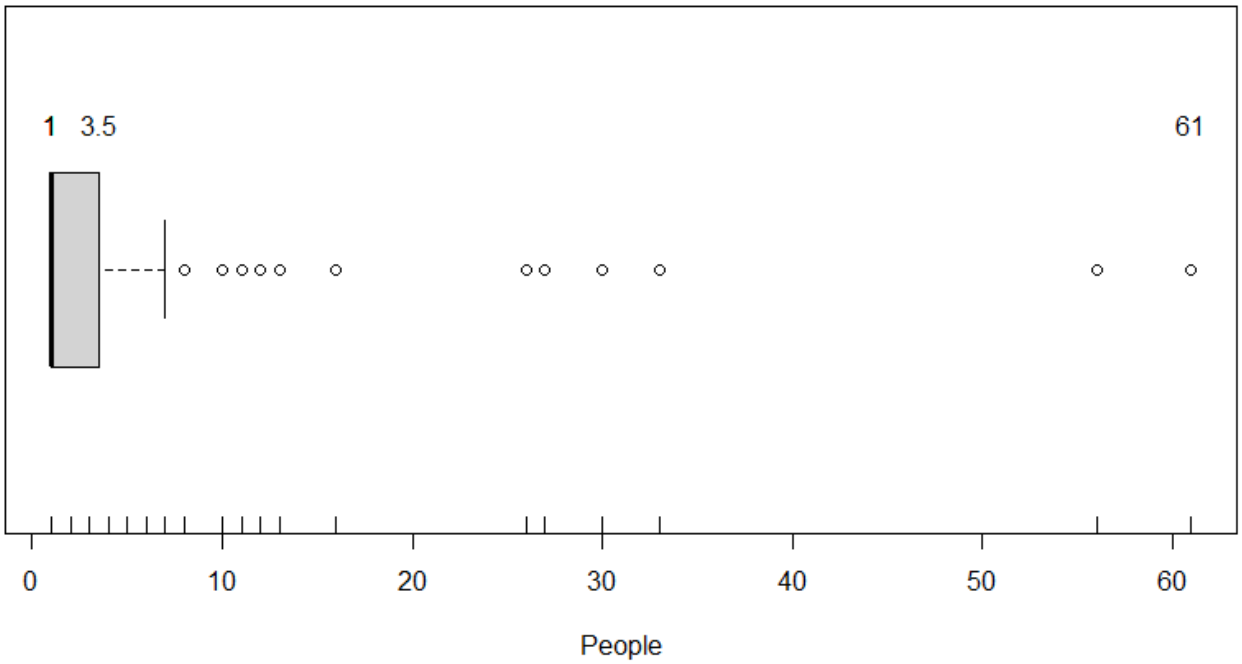
Figure 5. Histogram of the number of Injuries & Fatalities per harmful tornado.



As the distribution of the Injuries and Fatalities data is not normal, we cannot apply the Empirical rule to describe its variance. Instead, we can employ Chebyshev's Inequality to calculate that at least 75% (66 of 88) of the tornadoes should account for injuring or killing between 0 (mean of 5.466 – 2 * standard deviation of 10.608) and 26.682 (mean of 5.466 + 2 * standard deviation of 10.608) people. This is confirmed as the data reveals that 83 of the 88 (94.3%) of the observations are within this range.

The quartile statistics (Table 4) show that 75% of harmful tornadoes injure or kill between 1 and 3 (calculated as 3.5) people, and that 25% injure or kill 4 (calculated as 3.5) or more people. This once again demonstrates the right-skewedness of the data that is also exhibited in its box plot (Figure 6) with a right whisker that is technically much longer than the left whisker that does not exist. (Also noted is the median line that is left of the center of the box).

Figure 6. Boxplot of the number of Injuries & Fatalities per harmful tornado.



With heavily skewed data and a relatively large range of observations, the presence of outliers is expected. The calculated lower (0: Quartile 1 of 1 – 1.5 * Interquartile Range of 2.5) and upper (7.25: Quartile 3 of 3.5 + 1.5 * IQR of 2.5) outlier fences reveal 15 observations that are considered outliers. These are note by the circles of the box plot (Figure 6) and listed with details in Table 6.

Table 6. Detail of the most harmful (outlier) tornadoes.

Month	Day	State	Length	Injuries	Fatalities	Injuries & Fatalities
1	21	MS	31.060	57	4	61
1	22	GA	24.660	45	11	56
2	7	LA	10.090	33	0	33
8	6	OK	6.900	30	0	30
4	29	TX	21.420	25	2	27
4	29	TX	39.710	24	2	26
5	16	WI	82.530	25	1	26
2	28	IL	11.500	14	2	16
2	28	MO	17.390	12	1	13
2	28	MO	53.470	12	1	13
3	6	MO	17.740	12	0	12
5	16	OK	18.000	10	1	11
3	1	IN	2.150	10	0	10
4	29	TX	11.680	10	0	10
11	5	OH	5.410	8	0	8

It is noted that the two most tragic tornadoes occurred on two consecutive days. These both occurred during the tornado outbreak of January 21–23, 2017, which "was a prolific and deadly winter tornado outbreak that occurred across the Southeast United States. Lasting just under two days, the outbreak produced a total of 81 tornadoes, cementing its status as the second-largest January tornado outbreak and the third-largest winter tornado outbreak since 1950" (Tornado Outbreak, 2021).

In summary, harmful tornadoes injury or kill between 1 and 61 people and are best described as having a median injury and/or fatality rate of 1 person. This means that harmful tornadoes effect a relatively small number of people, but more tragic ones occur, but are much rarer.

Quantitative Bivariate Analysis

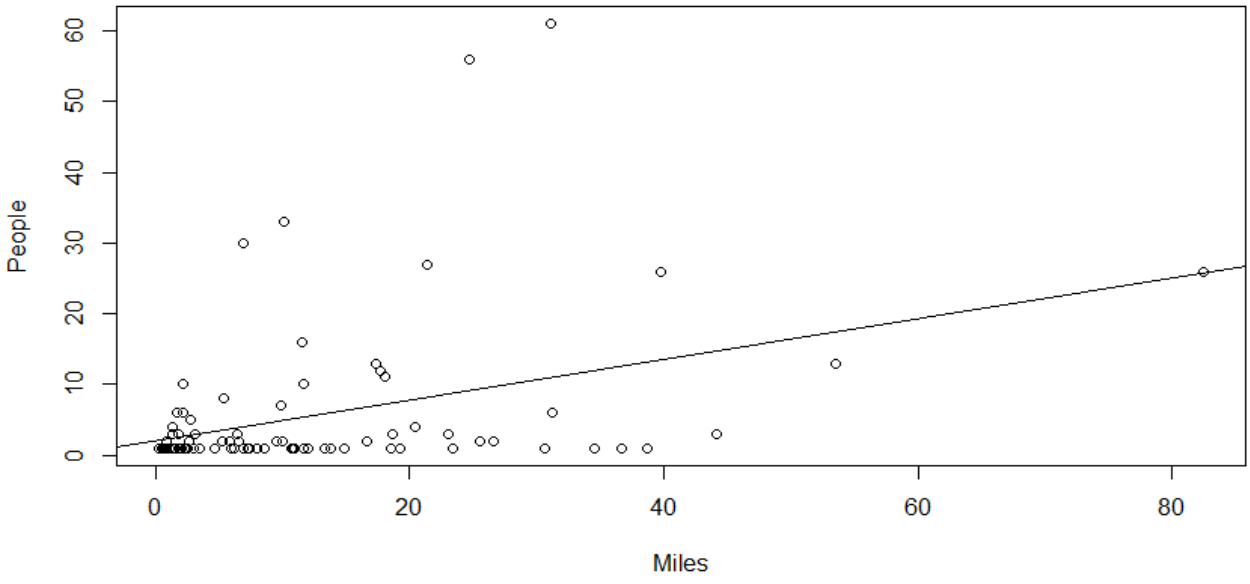
To determine if the number of injuries or fatalities are related to the length of a tornado's path, or if there is a linear correlation between these two variables, a Pearson Correlation model was calculated in R as

$$y = mx + b + \epsilon \text{ , or}$$

Injuries & Fatalities (People) = 0.2893 * Length of Tornado Path (miles) + 1.9801 + Error

The Pearson Correlation Coefficient (r) was calculated as 0.363 (r-squared = 0.132) which indicates that the relationship between the two variables is significant. However, this relationship is relatively weak when viewed graphically in Figure 7 due to the presence of the outliers of the Injuries & Fatalities variable. The number of injuries or fatalities is not likely to be directly related to or caused by the length of a tornado's path alone. There are many other factors that can make a tornado more or less harmful that have not been discussed here, such as the width, wind speed, severity (F scale), population density, warning systems, time of day and more.

Figure 7. Plot of the number of Injuries & Fatalities as a function of the length of a harmful tornado



Endnotes

[1]The data analyzed in this project was selected from a complete data set of the 1,472 tornadoes that occurred in the United States in 2017. Figures 8 and 9 display the Injury and Fatality data of the entire data set. The data is extremely skewed right due to the number of tornadoes that do not cause injuries nor fatalities. Removing those data points made the data more interesting in terms of descriptive analysis and basic plotting in R.

Figure 8. Histogram of the number of Injuries & Fatalities per tornado for all tornadoes.

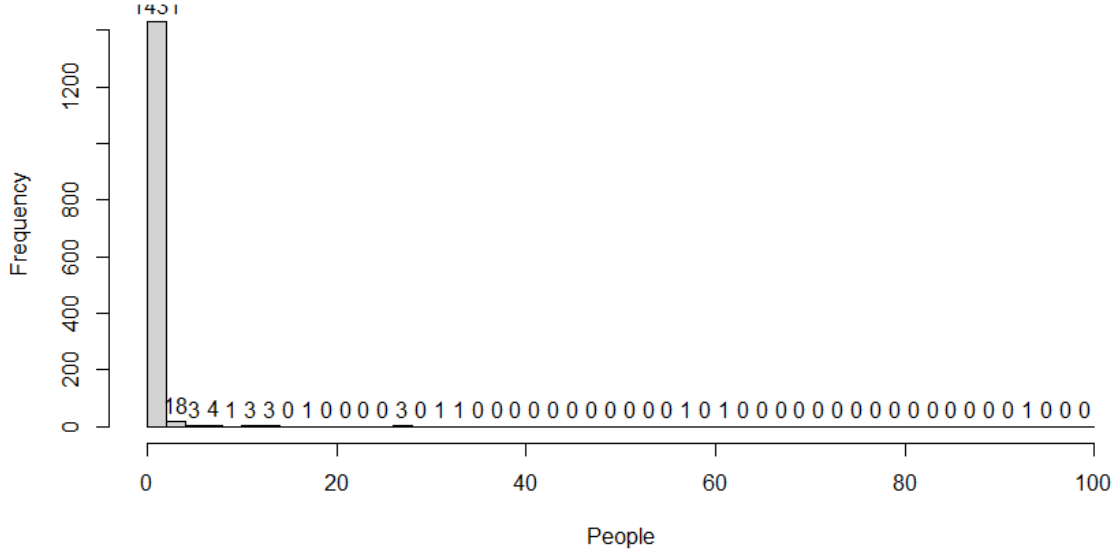
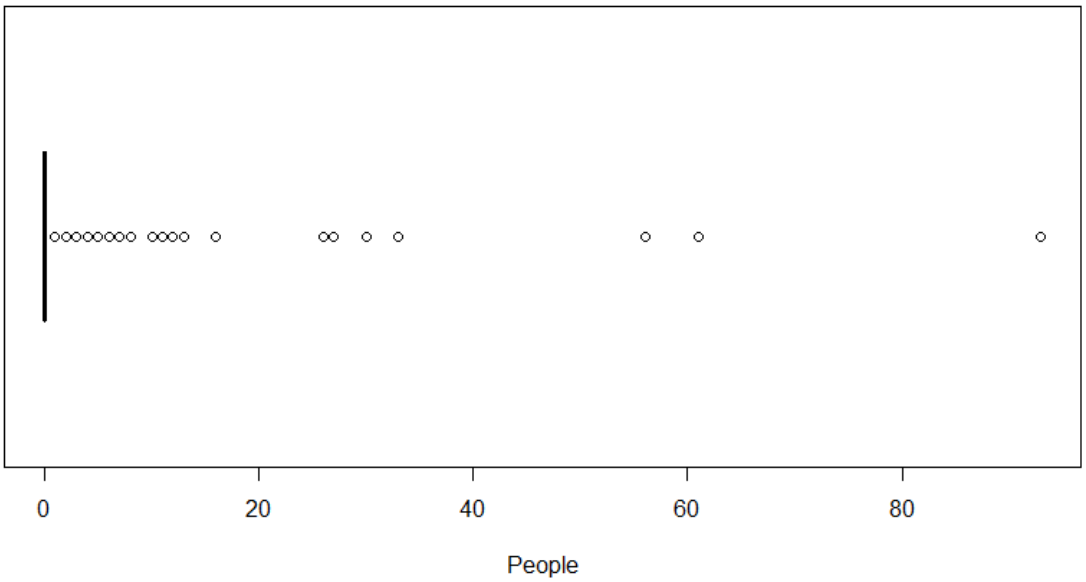


Figure 9. Boxplot of the number of Injuries & Fatalities per tornado for all tornadoes.



References

Sullivan III, Michael. Statistics: Informed Decisions Using Data; 6th Edition, ISBN-13: 9780135780121.

Tornado Outbreak of January 21–23, 2017. (2021, March 7). In Wikipedia. https://en.wikipedia.org/wiki/Tornado_outbreak_of_January_21%E2%80%9323,_2017