DAT-204 - R for Data Analytics: Final Project

Zach Trozenski

12/8/2020

An analysis of the previous 5 years worth of storm events

Introduction For my final project I wanted to revisit a data set I had originally queried for DAT-202 from the National Climatic Data Center (NCDC) Storm Events Database. The data set was composed of self-selected severe weather events for a given time range in a given area. When I ran the query to create this data set it was with the purpose of attempting to forecast severe weather based on historical data, however my plans changed and I chose to forecast a different data set still within the weather domain (as it turns out forecasting rain is best left to the professionals). Revisiting the storm events data presented a good opportunity however to conduct analysis in R.

Here are some of the questions I was hoping to answer:

- Which storm events were the most common in our area over the past five years?
- How much damage (as measured in USD) do these storms cause on average?
- When did the most destructive storms occur?
    - Does the data indicate they are increasing?
- In the data set, is there a correlation between storm severity and damage in USD?
- Who reports and documents the storms?
    - Are there any observable changes over time?

Please note the code is not being echoed in the markdown document for brevity and cleanliness. The raw markdown file is posted on my Github account here. For reference here is the data dictionary,

Most common storms in our area

Using a combination of a `for` loop and the `unique()` function I created a dataframe of the total counts of all storm events in descending order:

```
##             Weather.Event Total.Count
## 2       Thunderstorm Wind         182
## 4             Flash Flood         167
## 5                   Flood          48
## 6              Heavy Rain          18
## 1 Extreme Cold/Wind Chill           5
## 3             Strong Wind           4
## 9               High Wind           3
## 7              Heavy Snow           2
## 8            Winter Storm           2
```

The most common storm events in our area over the previous 5 years have been Thunderstorm Winds and Flash Floods. The two events seem to go hand in hand, as it suggests that thunderstorms can cause both the high damaging winds as well as flash floods. (Please note the index at the far left of the data frame is the original order in which the events were listed and doesn't factor into the count.)

Now that we know the most common storm events let's shift our focus onto the damage they can cause.

Storm destruction measured in dollars

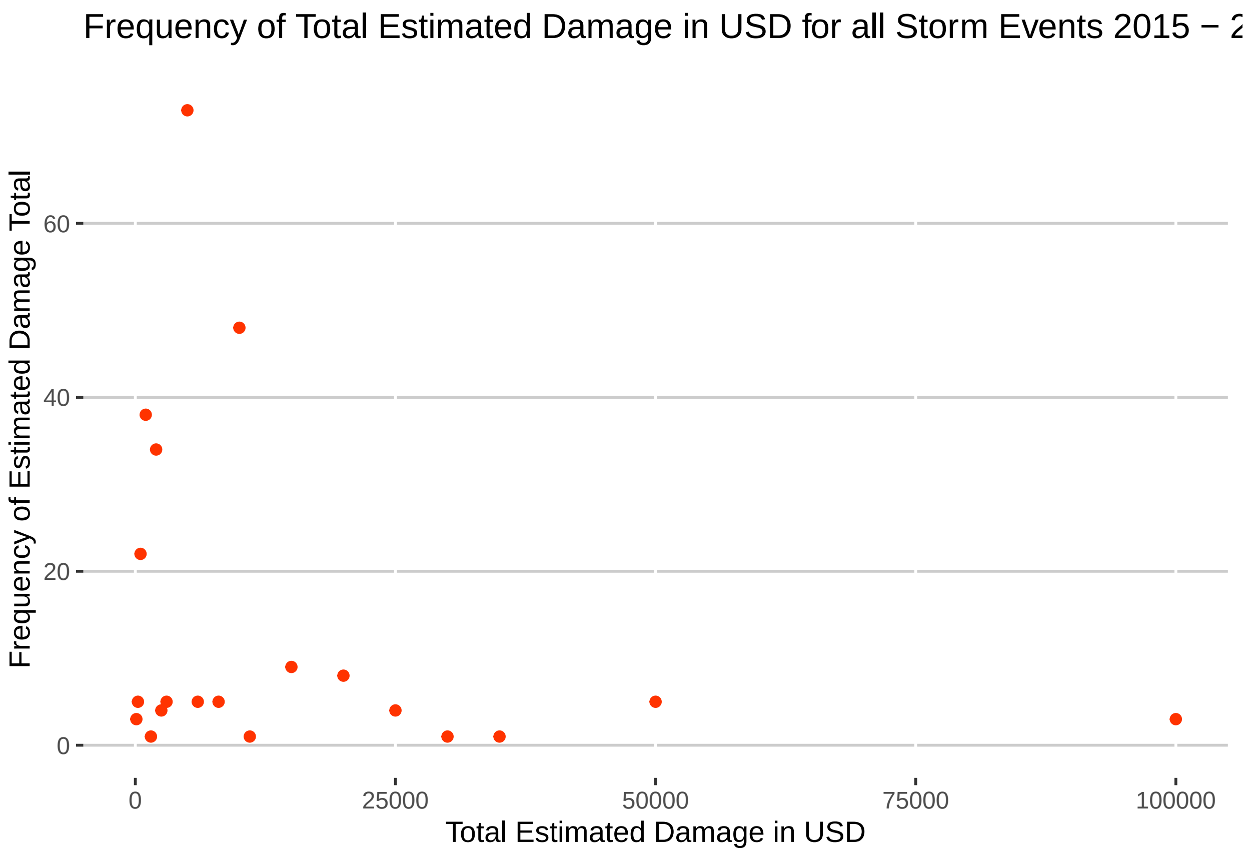
The data set contains estimations for the total amount of crop and property damage a storm event caused. Additionally the data set also accounts for injuries and deaths directly caused by the storm. Luckily the number of injuries and deaths relative to the total number of storm events logged is low. Instead let's focus on the damage estimates.

To get a total count of each storm event's estimated damage, I created a new series in the data frame which summed the crop and property damage estimates and added it to my original data frame. From there I ran the `summary()` function to get descriptive statistics on the the total estimates of storm event damage.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0        0     1000    33262     5000 10100000
```

```
## [1] "Standard deviation: " "491486.882977695"
```

The descriptive statistics indicate that the most of the estimates are quite low coming in at $1000 USD. What is interesting is that the Maximum value, which taken into consideration with the median and mean values seem to indicate it is an extreme outlier which would also explain the mean value being much higher than the 3rd quartile and the median. That being said, we must confron the absurdly large standard deviation. While we are getting a low median value the standard deviation is massive which tells us our data is distributed on a very large range. I think it would be helpful to visualize the distribution of the total storm damage to augment our understanding of the descriptive statistics.

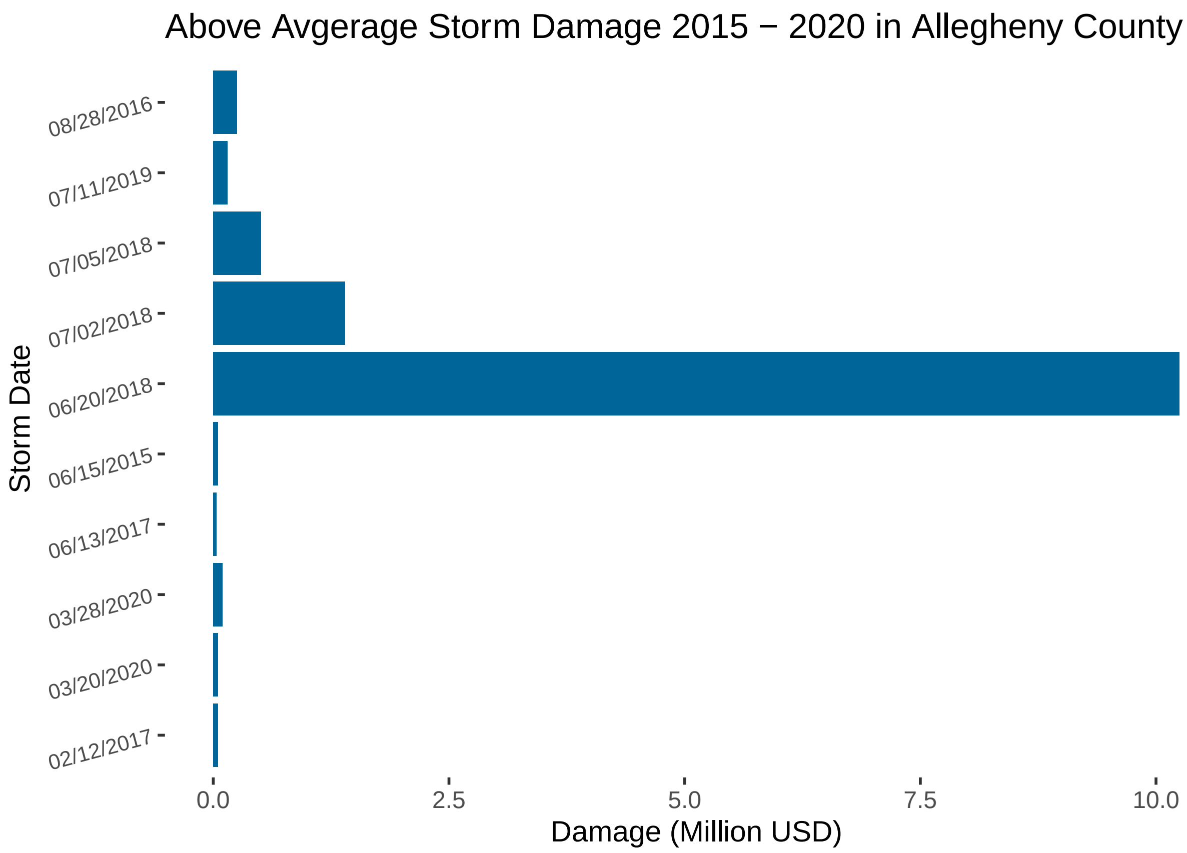


(Note: I've cut off the outliers at the top and bottom of the cost range (`$0 = 152`, `$10.1MM = 1`) for clarity.)

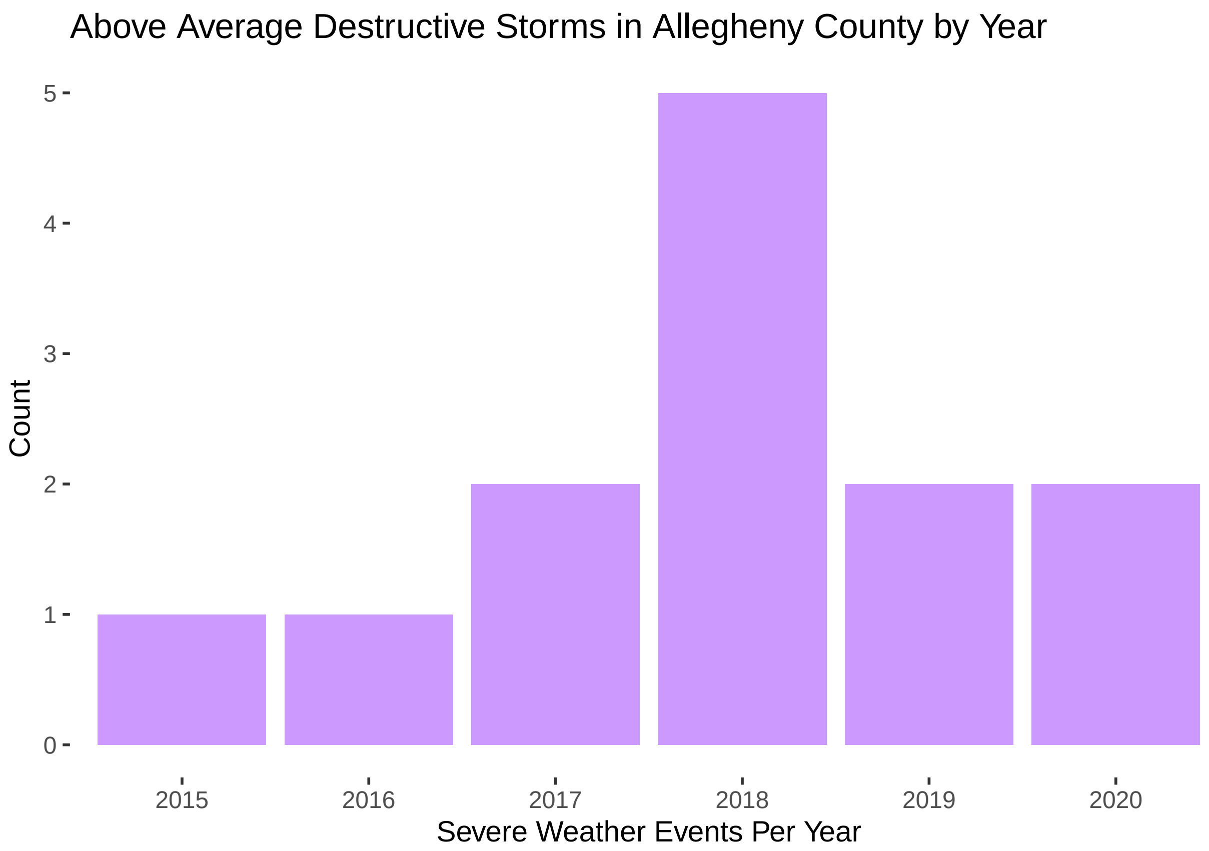
Based on this plot we can see that our assumptions about the distribution of data, that it's concentrated heavily at the lower range but with a large distribution. Given the descriptive statistics together with the plot we can assume that a given severe storm event won't cause an extreme amount of damage. While that may be the case, we only know the frequency of a storms damage. We don't know if the damage is increasing over time or when the most severe storms occurred.

Most destructive storms in the data set

Let's create a subset of the data to only look at the most destructive storms and see if we can detect any patterns. First, we'll slice our data set by pulling all storms that have total storm damage higher than the mean of $33,262. I noticed when I was plotting this subset that the maximum value was creating some difficulty in making the graph legible, so I scaled the the total damage so that it would reflect values in millions of dollars rather than the absolute value.



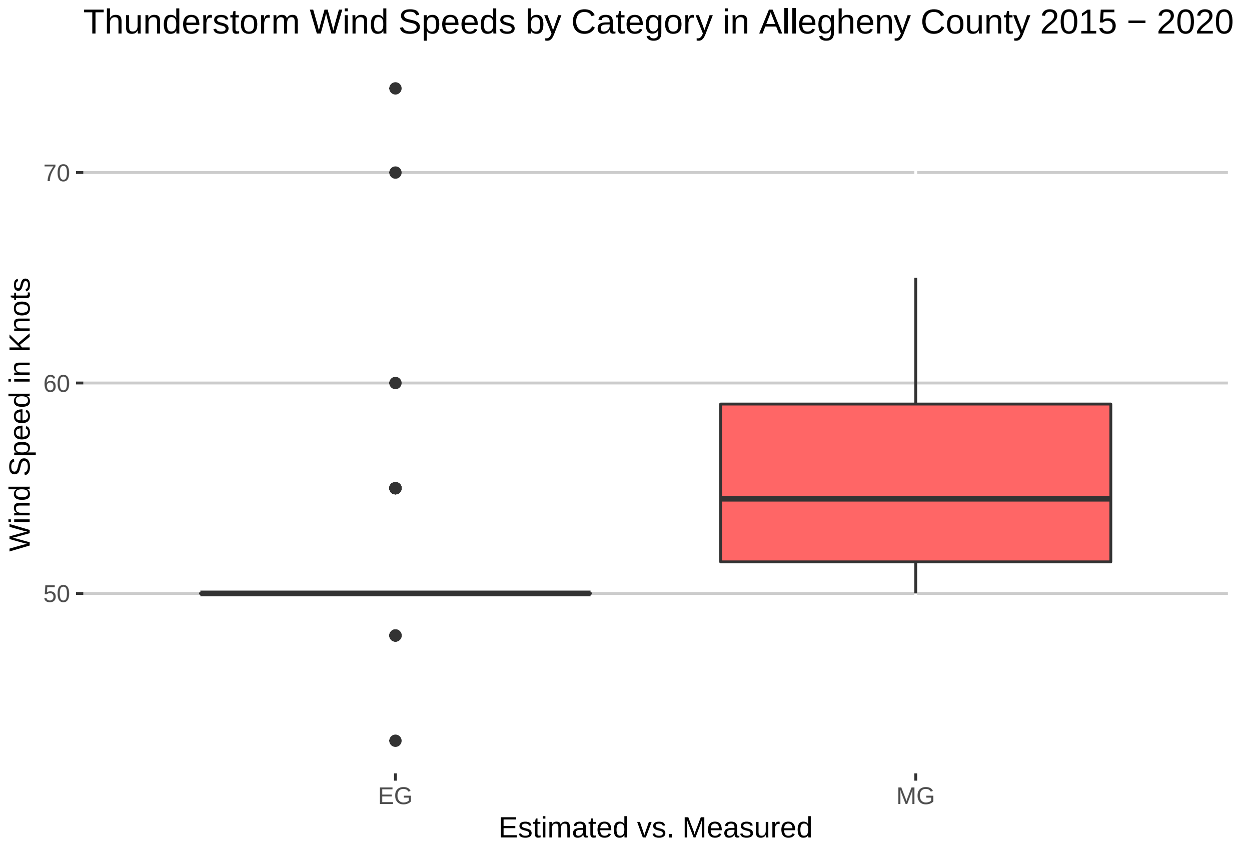
We can see from this graph there are 10 storms above the mean of $32,262 in total estimated damage. These storms are all the storms that are causing the mean to skew higher, significantly higher than the median, as well as stretch the standard deviation. It certainly may appear like storms are getting more destructive as time passes but let's plot these storms by year to see if there is a trend.



There isn't much evidence or data to back the claim up, but it does appear that there is a trend of storms increasing in destruction, as measured in USD, over time. There would need to be a much heavier analysis into a multitude of data sets to determine why the storms are increasing in severity and furthermore why 2018 of the six available years was the most destructive in terms of dollar value.

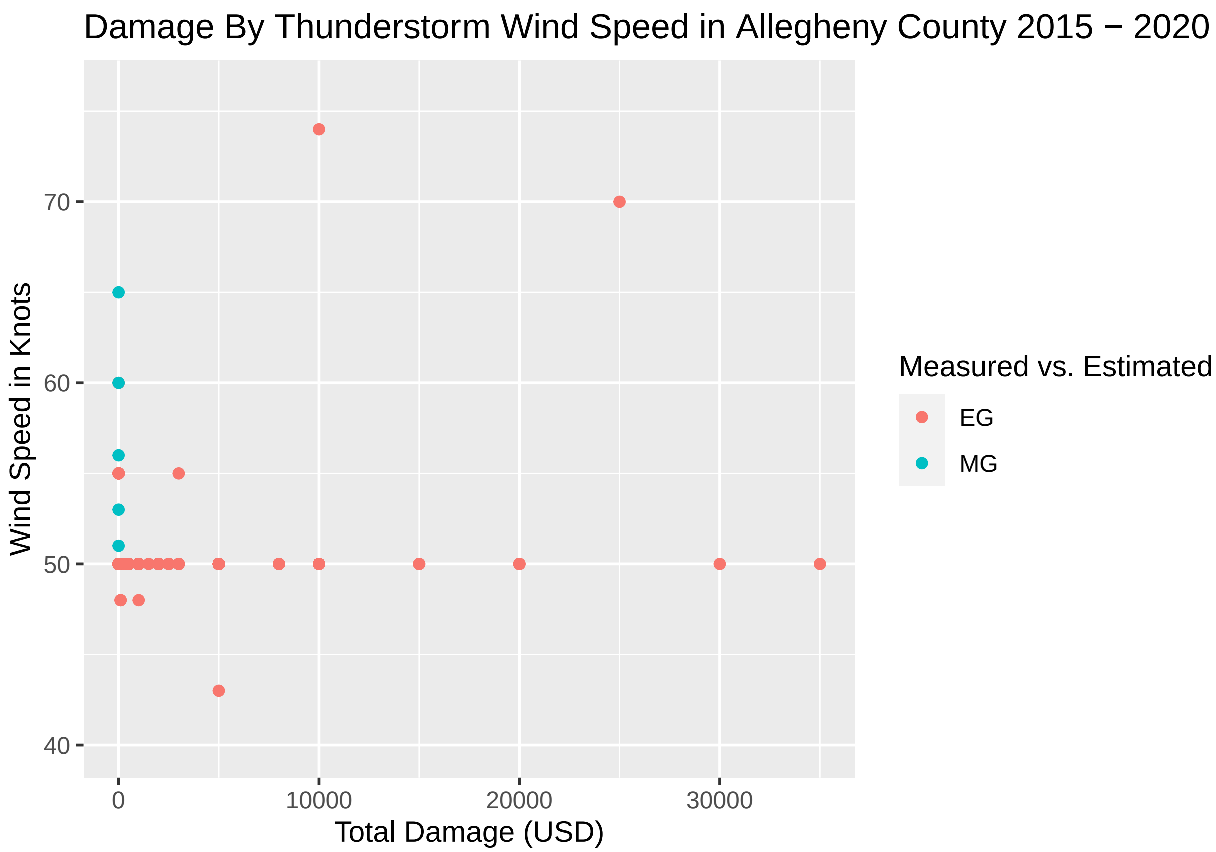
Thunderstorm winds

Seeing as thunderstorm winds are the most prevalent severe storm events, it would be a poor analysis of the data to ignore them completely as a facet. The thunderstorm wind has a measurement of magnitude as captured in two different categories: estimated magnitude and measured magnitude. Magnitude here refers to wind speed in knots. In order to get an understanding of the magnitude of the thunderstorm winds let's plot the two categories as box plot.



There doesn't appear to be much consistency in the estimated data, which stands to reason since there by all accounts there was no instrument used to verify the wind speed. There is however much more consistency in the thunderstorm wind speed as it was measured, the majority of the winds reached between 50 and 60 knots (roughly 57 - 69 miles per hour). While these aren't hurricane force winds, they are however classified as damaging winds.

Now let's try to get an understanding for the damage these winds cause in terms of dollar values by measurement category.



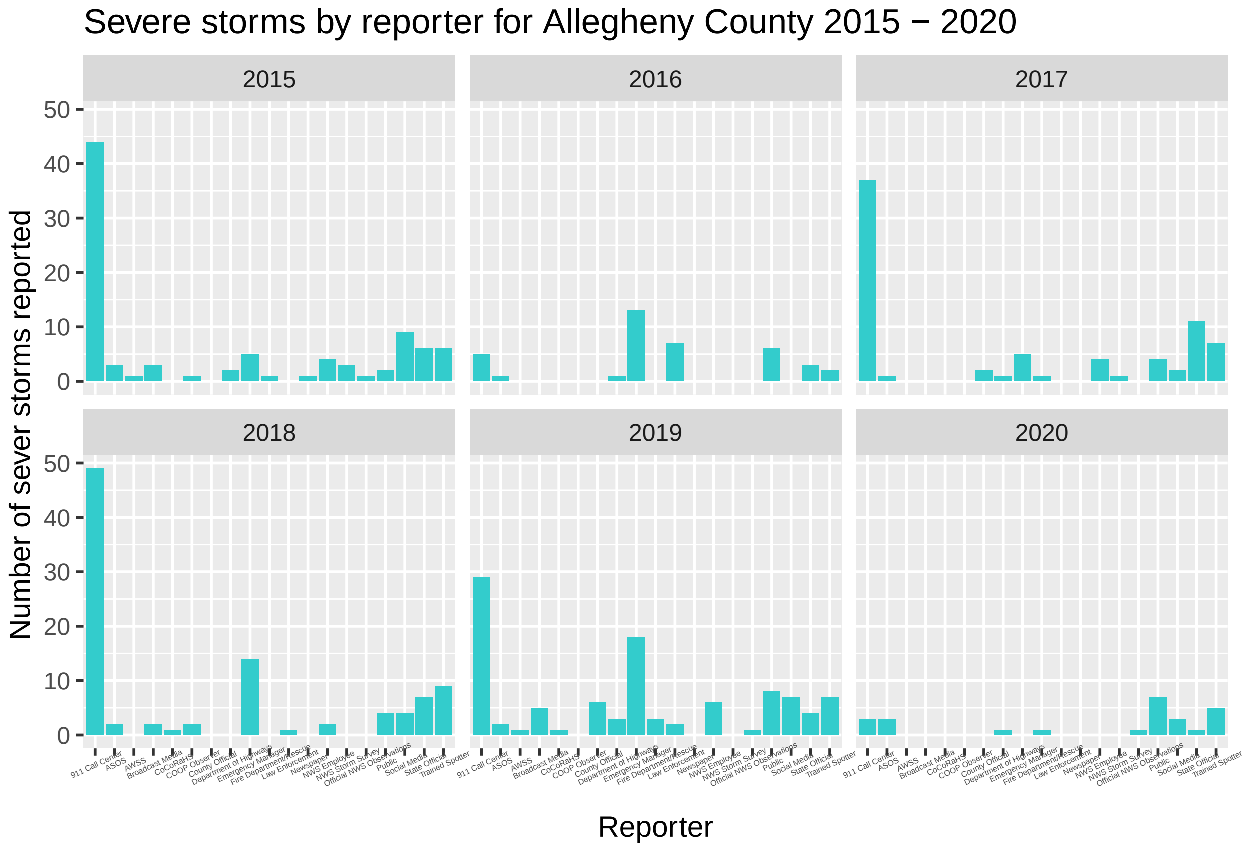
Here we see the measured data reporting no damage but we are seeing a fair amount of distribution across our dollar range. We can calculate correlation, in particular Pearson's correlation coefficient, to produce a number that can tell us whether in our data higher wind speed is correlated to higher damages.

```
## [1] "Correlation coefficient: " "0.0935446393402838"
```

Unfortunately the data does not indicate a correlation between wind speed and estimated damage as the value is much closer to zero, indicated no relationship. As an aside I thought this was interesting since it stands to reason that there would be higher damages with higher wind speeds.

Reporting of severe storms

One point of interest that I had was how storm events are reported and whether or not it has changed from 2015 to 2020. Are there any trends or changes that we can observe? Does social media have a larger presence now than it had as it becomes more integral in daily life?



Unfortunately there doesn't appear to be much in terms of trends that are clearly observable. Reporting does appear to consistently come from 911 call centers and social media does have a fairly established presence but there are no apparent patterns or relationships we can see.

Conclusion

Overall the severe storm data reveals some interesting nuances and insights into the weather we experience in our region. I didn't expect that thunderstorm winds would be the most commonly reported storms or that the median total damage estimate was roughly $33k. I had naively thought there would be a trend of increasing instances of severe storms as climate change is causing the severity of weather to increase across the nation. Additionally, I thought that maybe there would be some observable trends in how the storms are reported, or at least there would be some correlation in the data between wind speed and damage. I think this exercise has been a valuable lesson in the difference between reality and expectation. I went into this looking to validate a particular narrative and was proven they aren't plainly evident in the data. While that doesn't mean my thinking is wrong prima facie, rather it reaffirms that there are complex systems at work in the world that need to be understood with measured nuance and skepticism.