Storm Data

An Exploratory Data Analysis Project

Amie Davis

Bellevue University

I wanted to tackle whether "storms are getting more intense." In my comparisons between 1999 and 2019 data, I found only slight differences in the intensity of storms, as measured by wind speed and hail size. The magnitude of the size of storms, however, decreased by almost 50%. The value of property damage also decreased. This led me to conclude that storms are not getting more intense when looking at this specific twenty-year comparison.

I was surprised to discover that storms with low winds caused more property damage in 2019 than storms with high winds. Performing a hypothesis test, however, I was able to conclude that the null hypothesis that storms with high winds cause the same damage as other storms to be false. However, this does not determine if high winds cause more damage.

During correlation analysis, I discovered two positive correlations. Property damage and hail size are positively correlated with a medium effect, whereas property damage and wind speed are positively correlated with a lesser effect. Using regression analysis, I was able to create a statistically significant multiple regression model to predict property damage, given wind speed and storm range.

Trend analysis was missing from my study, reviewing each year's dataset. However, I felt I could perform more Exploratory Data Analysis (EDA) using just two datasets. I also omitted performing time series analysis. This could have provided insight into when storms occur. In addition, I was interested in using geodetic data to perform k-means clustering to help predict where storms occur, but that was out of scope and the time alloted for this project.

In addition to assuming that my original hypothesis was true, that storms were getting more intense, there were other assumptions I made that were incorrect. I assumed that storms with high winds caused more damage than storms with low winds, which was true when initially reviewing the mean and mode of the dataset. However, this was skewed by outliers in property damage. Another assumption was that property were valued similarly in different locations. However, a severe storm in a rural area may result in lower property damage value than a small storm in an extremely high-valued area. Property values were not considered in this study.

I had several initial challenges with data formats, particularly with dollar value fields, which I overcame by creating derived numeric fields. I had other challenges with property and crop damage data. It was difficult to view histograms with the wide variety in dollar amounts. Probability Mass Functions (PMFs) were not much better. I need to better understand a clear way to view data with such a vast range.

To conclude, I found that data can overrule assumptions. Common conceptions often depend on perception not data. The data tells its own story. The most important lesson learned in the study, however, was that is it okay to change your initial quest. As data unfolds, more specific questions form to better address the problem at hand.

References

Agarwal, R. (2019, July 21). Apply and Lambda usage in pandas. Retrieved from https://towardsdatascience.com/apply-and-lambda-usage-in-pandas-b13a1ea037f7.

Downey, A. (2014). Think Stats. Sebastopol, CA: OReilly Media.

National Centers for Environmental Information. (n.d.). Storm Events Database. Retrieved December 14, 2019, from https://www.ncdc.noaa.gov/stormevents/details.jsp.

pandas.DataFrame.join. (n.d.). Retrieved January 25, 2020, from https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.join.html

Saffir–Simpson scale. (2020, January 6). Retrieved from https://en.wikipedia.org/wiki/Saffir–Simpson_scale.

Wind. (2019, September 9). Retrieved February 1, 2020, from https://www.weather.gov/safety/wind