Tornadoes in the United States

Jacob McKenney, Luke Sala

June 2022

I. Summary of Questions and Results

Question 1: In which areas of the nation are tornadoes most likely to form? Which states have the highest tornado density?

Result: Tornadoes are significantly more likely to occur in the central and southeast United States than anywhere else in the country. Florida has the highest tornado density of any state in the nation.

Question 2: What time of year is it most common for tornadoes to occur? Is there a time of year in which tornadoes tend to be of a higher magnitude?

Result: Tornadoes tend to to occur most often in the late spring, between the months of May and June.

Question 3: What visualizations are helpful for better understanding our tornado data set?

Result: Geopandas plots are highly effective for tornado plotting. More specifically, a Geopandas plot of the United States can be used as an effective tool to point out trends and hot-spots of tornado activity across the country.

Question 4: How are poverty/income levels affected in tornado-dense areas?

Result: It seems that tornado-dense territories have a lower average income then those that are not. Whether this is causal or just coincidental, we haven't investigated far enough to tell. Most likely it is a factor in the complex equation that determines economic status.

Question 5: Can we estimate injuries, fatalities, dollar loss, and other factors for a given tornado using historical data and machine learning models? We could use this information in conjunction with Questions 1 2 results to figure out when and in what regions it would

be wisest for inhabitants to be on alert given various devastation thresholds.

Result: While we can make some reasonable predictions for devastation metrics like injuries and crop loss it seems to be difficult to use just the features our dataset contained to make sound predictions. Although, models seem to be able to predict particularly devastating tornadoes with some accuracy as there is much more variance in features like magnitude, size, and crop loss for these larger tornadoes.

II. Motivation

By looking at the past we may be able to predict events in the future. We are hoping to gain some insight into long-term patterns of tornadoes and what subsets of people are more greatly impacted by tornadoes. In addition, we are looking to provide more information via visualizations and machine learning to illustrate the trends and effects of tornadoes in the United States.

III. Dataset

1950-2020_all_tornadoes (https://www.spc.noaa.gov/wcm/data):
This is our main data set that documents all tornadoes that have occurred within the United States between 1950 and 2020. It has many columns each describing a feature of tornado which are described in the pdf here:
https://www.spc.noaa.gov/wcm/data/SPC_severe_database_description.pdf. Some of the most important features are starting longitude and latitude (slon, slat), ending longitude and latitude (elon, elat), date, state of occurrence, magnitude, fatalities, injuries, and amount of damage done (characterized 1-10 for dollar amount).

Census Data: Census data can be retrieved via an api (guarded by api key which can easily be attained here: https://api.census.gov/data/key_signup.html) which serves the desired data. The python library 'census'
(https://pypi.org/project/census/) has been developed to make this process smoother. This dataset holds all sorts of demographic information for people in all regions (states, counties, tracts) such as sex, income, age, occupation, etc.

U.S. geography information: The dataset provided in the Hurricane Florence exercise

for geopandas will be used to map a backdrop of the United States. The name of the file is 'gz_2010_us_040_00_5m.json'

IV. Method

Question 1: In order to achieve this goal, a function was written in our analysis file called *plot_tornadoes_by_state* which takes in a dataframe containing tornado information and a dataframe with state geometries. The function groups tornadoes together by state and then joins the counts of each state to the state geometry to be plotted (so that we have the corresponding state geometry for each state tornado count). Then the function does two plots: one plot is the count of tornadoes per state and the next plot is the normalized count (count divided by square mileage of the state).

Question 2: Our analysis showed two different trends for the relationship between time and tornadoes: one for yearly trends (how frequency of tornadoes change over a year), and another for historical change (how frequency of tornadoes has changed on a year by year basis over the past 70 years). We were able to do both of these analyses with one function: *most_likely_time_period*. This function takes in a pandas datetime index (month, day, year, etc.) and then groups over the index, taking the count for each group. The function then plots this data. We simply pass different indexes to get different trends or clarity of trends.

Question 3: In order to figure out which data visualizations were most helpful we had to take into account the specific research question we were asking and how the data analysis lends itself to one visualization or another. For *plot_tornadoes_by_magnitude* and *plot_tornadoes_by_state* a geopandas plot made the most sense as the data was geospatial. For *most_in_year*, *poverty_and_tornadoes*, and *devastation_predictions* the result was text and figures so we wrote this information to a results file. Finally, for *most_likely_time_period* the goal was to show trends in tornadoes so we used seaborn to accurately plot those trends.

Question 4: To look at poverty levels within tornado dense regions we incorporated census data into our tornado dataframe about average median income and ratio of income to poverty. We then were able to create a dataframe consisting of unique counties, which included the pertinent poverty information, and another dataframe just consisting of the counts of tornadoes per county. We then split this data into various portions (top 10% of tornado dense counties and bottom 90, top 20 bottom 80, etc.) and calculated the mean metric for each of these portions by joining the aforementioned unique county and county

count split dataframes together.

Question 5: To see if we could predict devastation metrics like fatalities, injuries, loss, and crop loss we instantiated three different regression models from the sklearn library to compare and contrast: DecisionTreeRegressor, KNeighborsRegressor, and LinearRegression. We split our data into labels and features (the label changed depending on the function call), and then split the labels and features into training and testing sets. From there we trained each model and predicted outcomes for a non-tuned model. Next, for the DecisionTree and KNeighbors models we tuned them via an exhaustive search using the sklearn GridSearchCV class. We then trained our models with the found best parameters and compared the accuracy between a tuned and a non-tuned model for the same dataset. The results were written to the *analysis-results.txt* file.

V. Results

Question 1:

The resulting plot from *plot_tornadoes_by_state* is shown below:

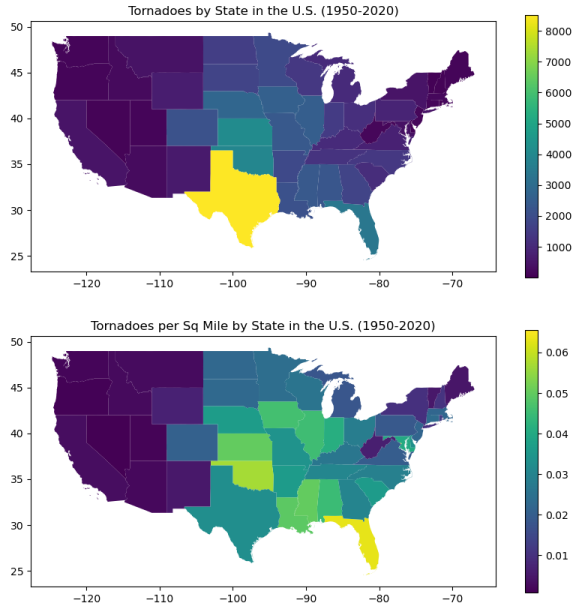


Fig. 1: Total Tornadoes and Tornado Density in each U.S. State

It is evident from the topmost plot that Texas has the greatest total number of tornadoes by far. It is followed by some of the central U.S. states such as Oklahoma and Kansas. This plots usefulness is limited, however, by the fact that a a tornado has a significantly larger chance of forming in a state or larger area like Texas. The bottom plot shows the number of tornadoes in each state normalized by its area; a much more informative visualization when comparing state to state. The overall picture looks somewhat different here as we can that Florida actually is home to the most tornadoes per square mile. It is followed closely by Oklahoma, Louisiana, Georgia, and some other central states. Interestingly, Texas has a relatively low tornado density compared to many of the states around it.

Question 2:

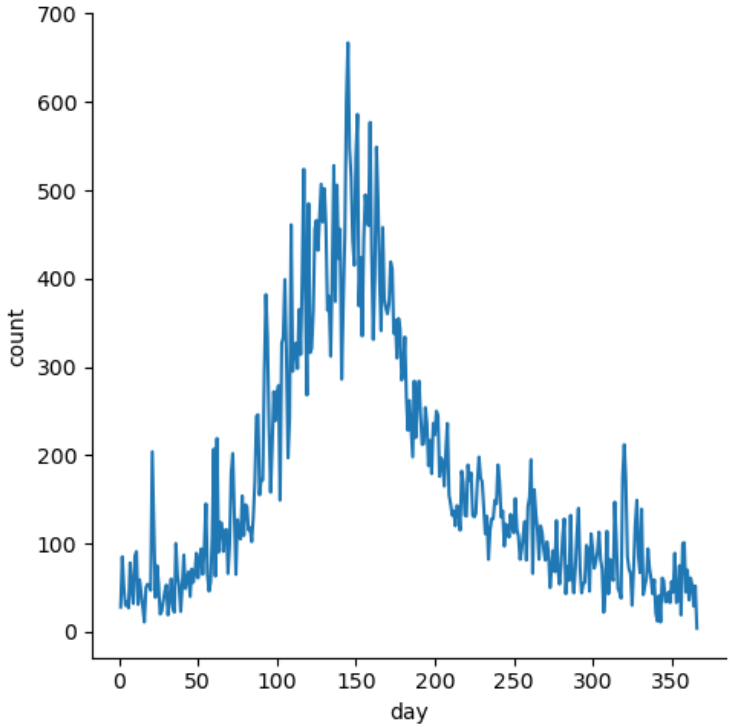
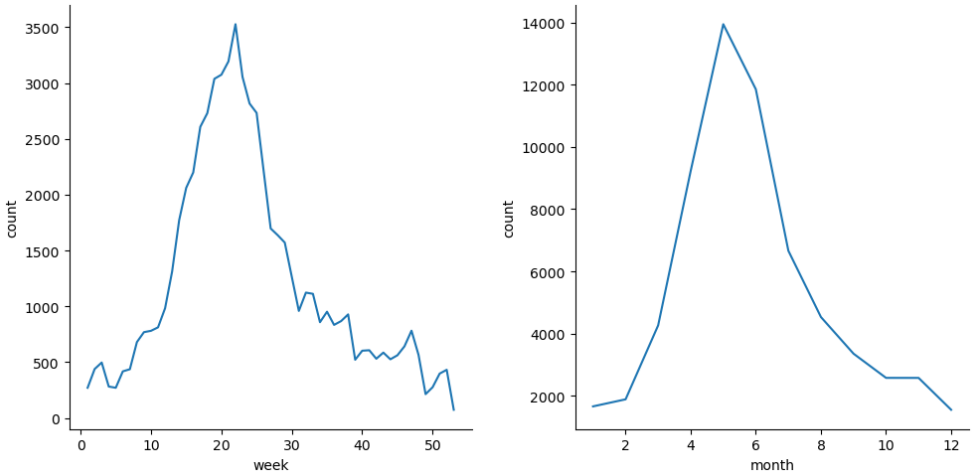


Fig. 2

The plot above of tornado count on each day of the year highlights a clear spike in tornado occurrences toward the middle of the calendar year. Though it highlights a clear trend in the data, this plot is noisy in the sense that there are huge fluctuations in the tornado count on any given day of the year. This makes intuitive sense, as it is entirely possible that any given day may not have any tornado at all. In order to smooth this data and refine the trend, the date index was changed to both weekly and monthly as shown respectively in the graphs below:



(a) **(b)**

Fig. 3

As the time increments are increased the peak of the tornado count can be more precisely pinpointed. From the monthly plot, it can be seen that the maximum tornado count occurs toward the middle of May. This observation aligns with the time period of "tornado season" a period of the year lasting from March to June with the statistically most active day being May 25th.

By once again changing the the time index, we can use our function to gain insight into various other date based trends. One of particular interest is the relationship between tornado count and year. Figure 4 shows this trend. There is a steady rise in tornado count year after year. This is a peculiar observation and further analysis is needed to see if this is a true increase or rather a result of improved tornado measurement technology.

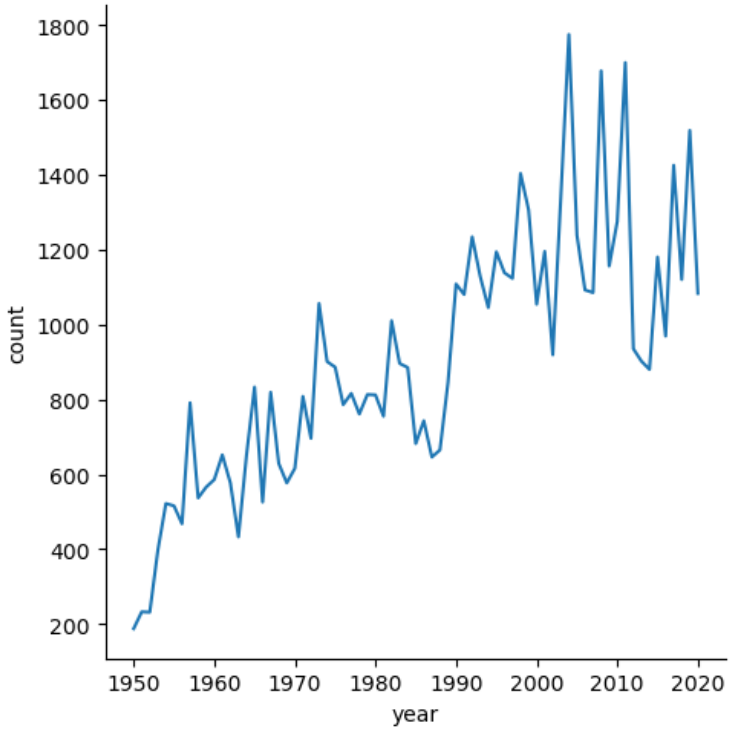


Fig. 4

Question 3:

As tornadoes are relatively localized events geologically, a geopandas plot with point markers is an extremely effective way to mark their locations. Additional information about each tornadoes magnitude can be easily factored in by changing each tornado's point marker's size and color. Figure 5, shown below, provides a clear visualization of the tornadoes that occurred in the U.S. from 2009-2019. The legend that accompanies it gives viewer a clear indication of how each tornado's color represents its relative magnitude.

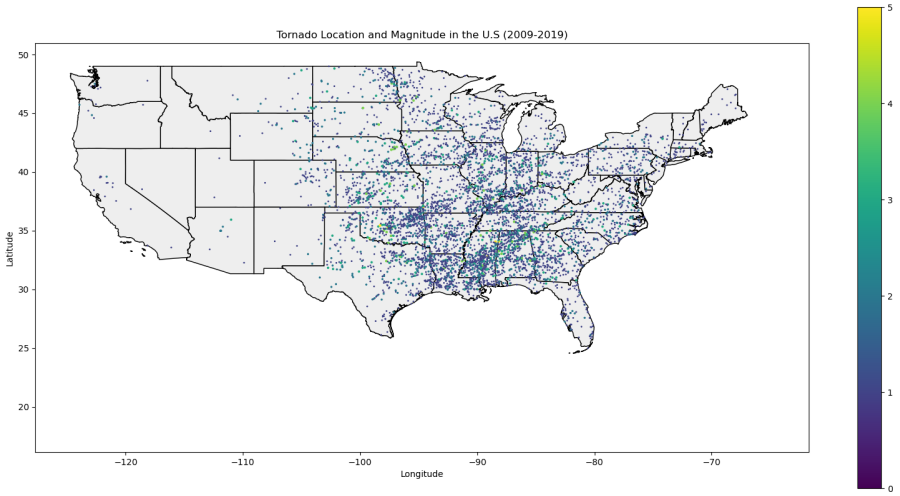


Fig. 5

Question 4:

It can be seen from analysis-results.txt that for each split, the community grouping with less frequent tornado occurrences has a higher average median income. This result heavily implies that areas hit more frequently by tornadoes struggle to a greater degree economically, indicating higher levels of poverty in those locations. While, we can't determine causality due to economic issues being enormously complicated we can say that impact of tornadoes seems to make an impact.

Question 5:

Through a tuning process, we tried to achieve better results from all of the machine learning models. However, the results were insignificant and in many cases the untuned models outperformed their tuned counterparts on the testing datasets for each of the 3 machine learning methods. The models had an especially difficult time predicting property loss figures. In order to improve upon the machine learning model's results, access to more data would be incredibly valuable. We hypothesized that without more data, the models are unlikely to achieve high accuracy predictions due to the large variance in the damage that tornadoes cause. This was a good practice in using machine learning models and taught us the importance of gathering lots of data prior to doing analysis. In addition, we believe it is the case that finding the correct features that correlate to a specific label is more important than brute-force tuning the models to try and maximize accuracy.

VI. Impact and Limitations

Our study of tornadoes, though only scratching the surface, could have huge implications for keeping people safe. Using data science to peer into the past, allow us to pick up on trends such as when tornadoes are most freuqnet, or where they occur most often. Metrics such as these can be invaluable for educating people on where and where not to locate themselves. Even government agencies could make use of this information in situations such as amending laws that limit where towns, or tornado venerable structures can be developed. Though many people could be helped by this information, home development or other companies invested in real estate could be harmed if tornado data is used to limit where structure can be built. Bias may come into play in our analysis in the types of things we chose to study. We chose to study the effects of tornadoes on poverty levels, because we thought there might be a correlation, but the apparent correlation might be due to other factors instead.

All of our tornado data came from the same source. This induces an inherent limitation as the collectors of our dataset have left out certain parameters from the data that didn't interest them. Therefore, our analysis should be compared to data found from other sources to prove its validity.

VII. Challenge Goals

New Libraries - 'census'
Census is a library for dynamically obtaining census information via api calls. Requires a valid api key. Our valid key: d3687e22fcd51ce480482a5caa07e0ae239c77a5
Machine Learning:
We would like to use machine learning to make predictions the quantitative effects of tornadoes based on their qualities
Multiple Datasets:
We are going to join multiple datasets to provide more information about the environment through which the tornados are traveling.
Messy Data:
We are querying the census api to obtain data which will need to be cleaned and joined with other data in order to be useful.

VIII. Work Plan Evaluation

Our work plan ended up working our very closely to how we had planned it. We kept on track throughout the course of the project and were able to achieve all of the goals we set out to hit.

IX. Testing

To test our analysis we took a small subset of the huge tornado dataset and ran many of our functions using this smaller dataset (all of the functions that really could be tested - the devastation predictions function used ml and wasn't really easily testable). The main goal was to ensure valid counts and groupby's were being created and shown on the various plots we made.

X. Collaboration

We did not collaborate with any sources outside of the course.