Name: **Muley, Tushar**
Assignment: **DSC 680 - Week 3 Milestone 2 White Paper**
Date: **December 5, 2021**

Draft White Paper – Milestone 2

What is the Impact of Natural Disasters?

Topic:

The topic selected is deaths caused by hurricanes that make landfall in the United States.

Background:

The topic of natural disaster has come up a few times this year. It seems like we have seen an increase if natural disasters in the last 10 to 15 years. We have had earthquakes, volcanic eruptions and violent storms activity increasing based on news reports. You hear the broadcaster make headline grabbing announcements like, "storms we not seen in the last 100 years." Is there any truth to what news broadcaster are saying? When looking for data concerning natural disasters it led me to a site called Our World in Data. This site contains a large amount of information on natural disasters and their effects on humans. For this research project based on the data I found I want to propose a basic question.

I wanted to use the data to predict one thing. Can we use the data to predict the future deaths caused by hurricane?

Hurricane tracking started back in the 1950s. To be more accurate forecasting models were developed in the 1950s. Hurricane tracking has been going on for many years before it. The forecasting was accomplished by two major developments. One was the aircrafts and the other was computer technology. As computing power advanced so did statistical models. The first statistical-dynamic tracking model appeared in 1973.[6] Hurricanes by numbers, on average each hurricane season (start date of June 1st and end date of November 30th) has about 5.9 hurricanes. There are 10.1 named storms.[7] In 2020 hurricanes caused 47 deaths. Hurricanes also cause a lot of damage and expense.

Have to keep in mind hurricanes are more than the initial storm. Hurricane bring rain, flooding, wind and tornados. So, a hurricane is just more than a single natural disaster, but a group of disasters occurring at once. Using this data to perform Exploratory Data Analysis (EDA) to determine the impact of natural disasters, more specifically impact storms have on United States citizens. The analysis will look at number of deaths, reconstruction cost and total economic damage storms have. Using additional data, we can perform additional analysis on

number of hurricanes that hit the United States. Once the analysis has been complete I will use the data gathered to build a predictive model.

Data Explanation:

The data was obtained from Our World in Data[1] site and the subsite called Natural Disasters[2]. When starting out with the data it contained 5,569 rows and 169 columns. Not all the columns were used. The data was also narrowed to the United States region. This reduced the data to 99 rows and 18 columns. The columns are described in the Appendix section of this paper. A second file which was used in the modeling included 168 rows and 8 columns. The columns are described in the Appendix section of this paper. The second file contained hurricane specific data for the United States.

Method/Analysis:

I started by reviewing the data from Our World in Data subsite 'Natural Disasters'[2]. The data appeared to be pretty robust and usable for what I had in mind. I cleaned up the data removing 5,470 rows of data from the raw file. It contained disaster counts from all around the world. I reduced it to just U.S. related data. After that I reviewed the columns. I removed everything that was not relate to storm/hurricanes. This included wind, heat, volcanic, drought and some others disasters not part of this analysis. This reduced the column count to 18. I updated human readable column names to model ready names. Once that was completed I uploaded a second file which was specific to hurricanes. This data included U.S. Virgin Islands data as well. Once again reduce the data down to U.S. mainland data. This data has 168 rows and 8 columns.

Once loaded I ran usually checks on the data for nulls and odd field like integers as string I did not find any issues. I updated missing values to zero so there would not be any NaNs. I ran auto visualization package to get a better understand the data. I ran a correlation of the full data from both files. Those can be viewed below as Figure 1 and Figure 2.

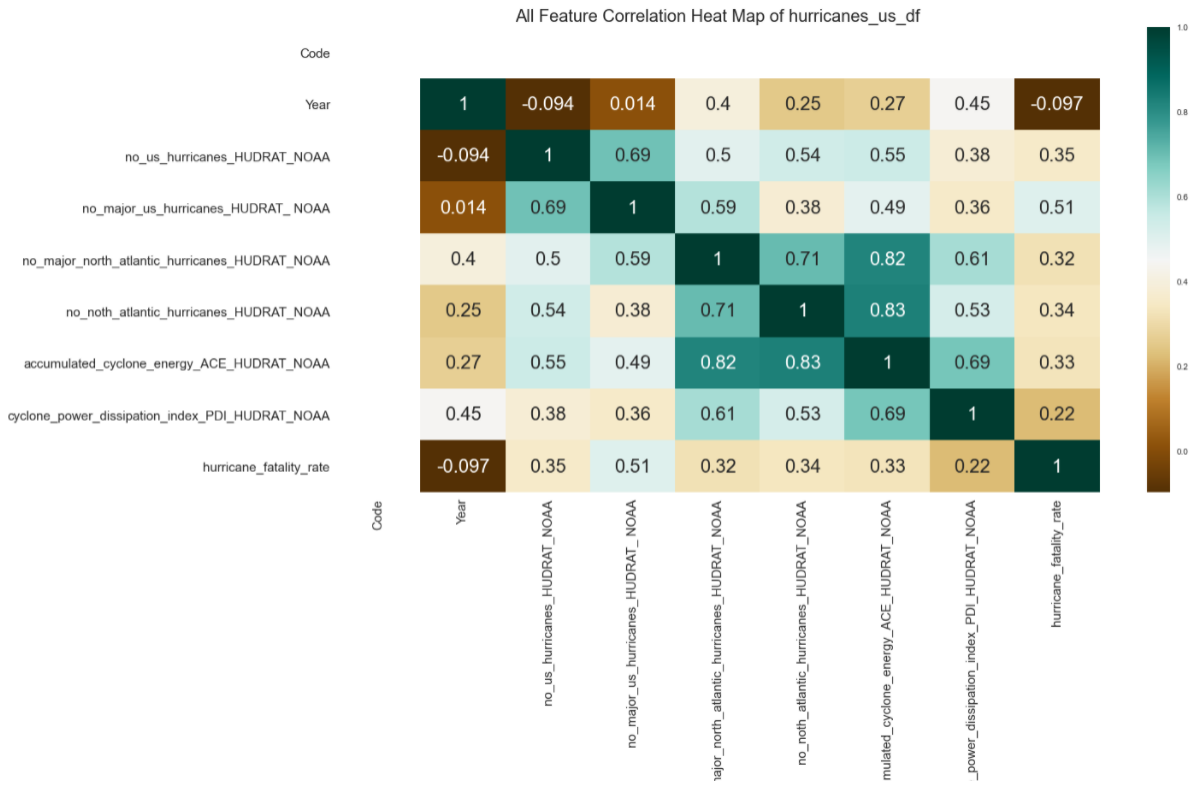

Figure 1: Correlation of hurricane_us_df file

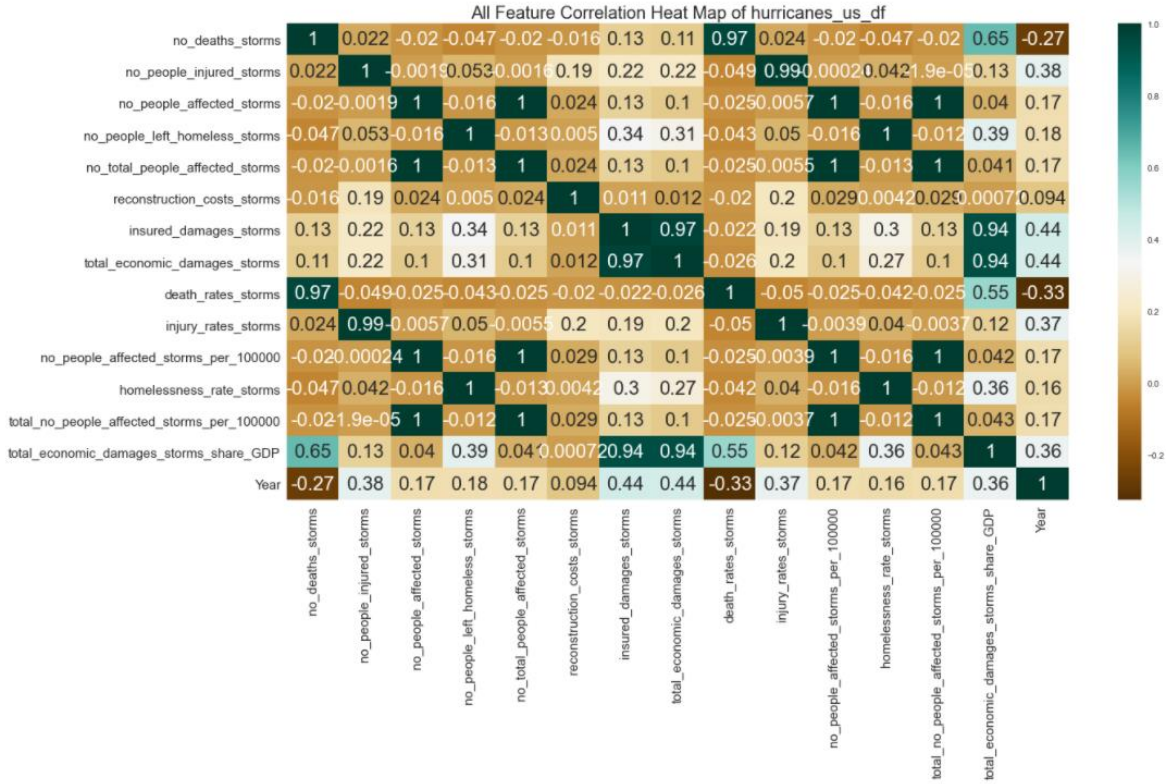

Figure 2: Correlation of disaster_us_storm_df file

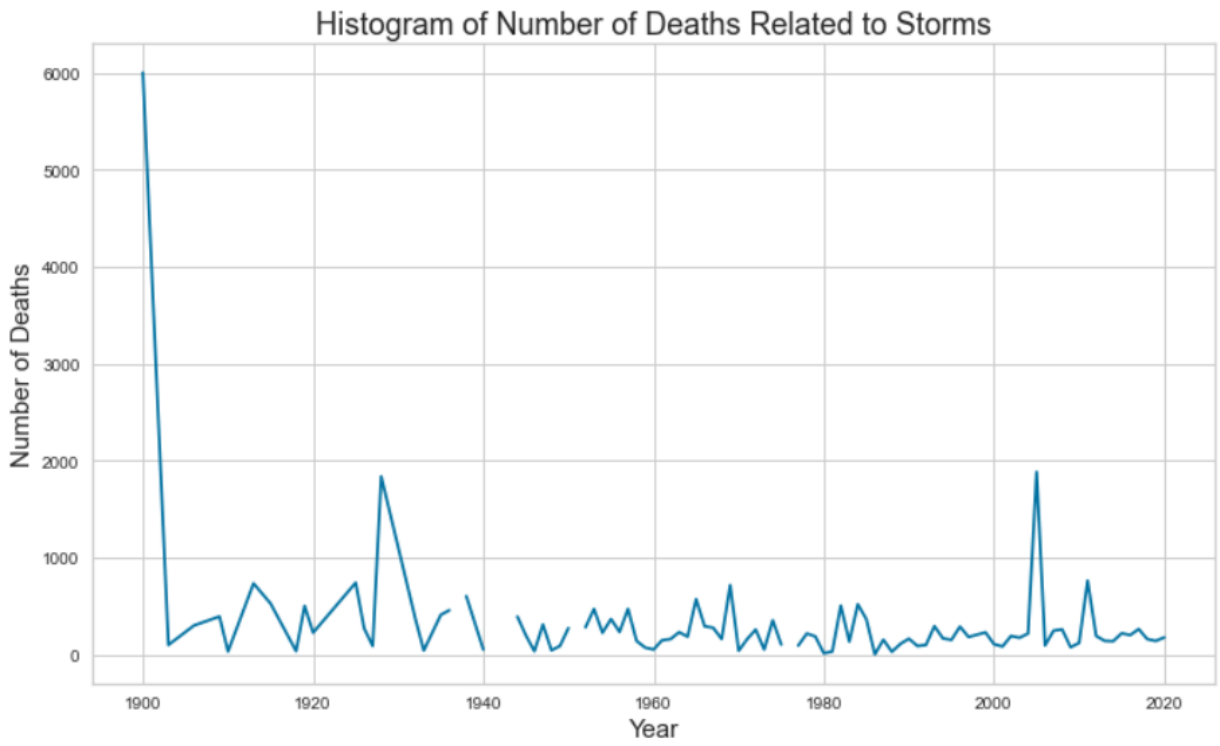

Figure 3: Number of deaths from hurricanes by year

The above graph (Figure 3)shows number of deaths caused by hurricanes by year. As you can see there are missing segments in the data. Those years either did not have deaths or the data is missing. That is in the earlier years compared to more modern times from 1980 forward.

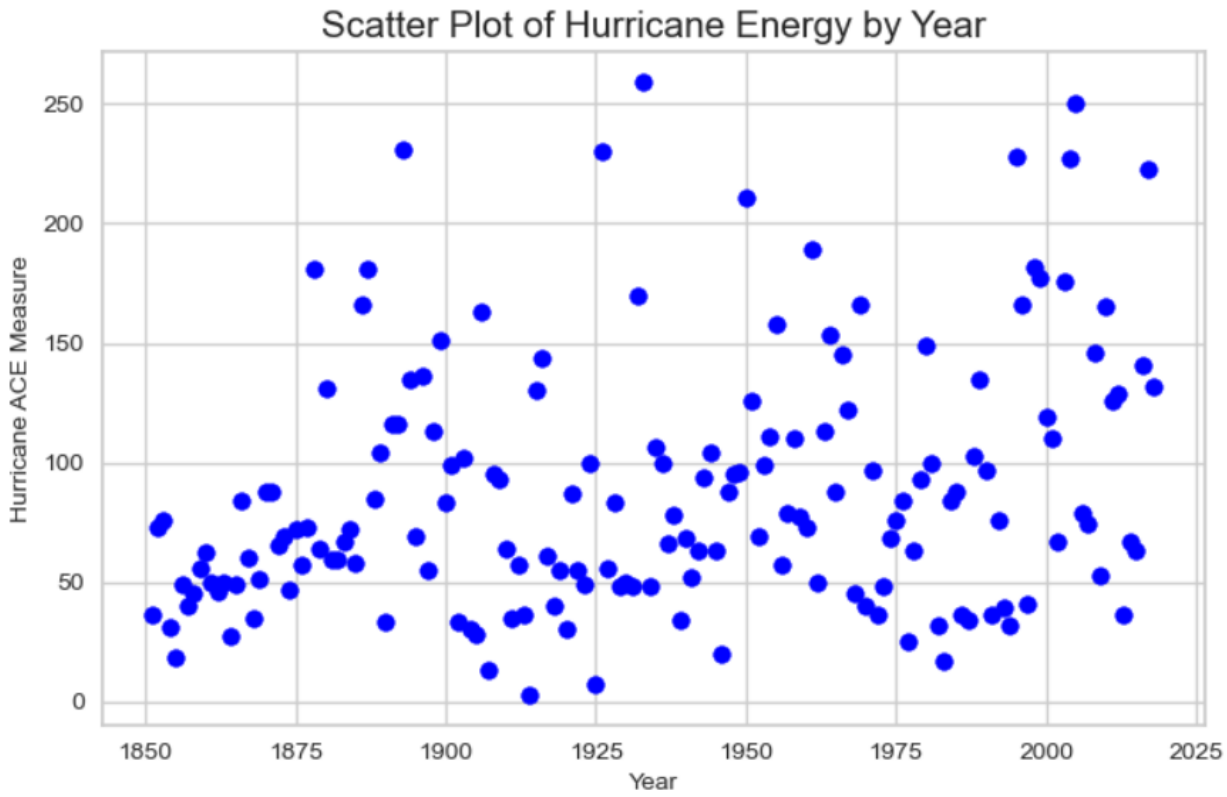

Figure 4: Scatter plot of hurricane energy by year

Another interesting fact found is level Accumulated Cyclone Energy (ACE) by year. In Figure 4 you can see number of high ACE storms is increasing based on scatter plot. Before it appears the US had one major ACE storm (above 200) every 10 to 50 years between 1850 and 1950. But between 2000 and 2020 we have had same number but in a smaller amount of year. This might indicate increasing amounts of dangers storms.

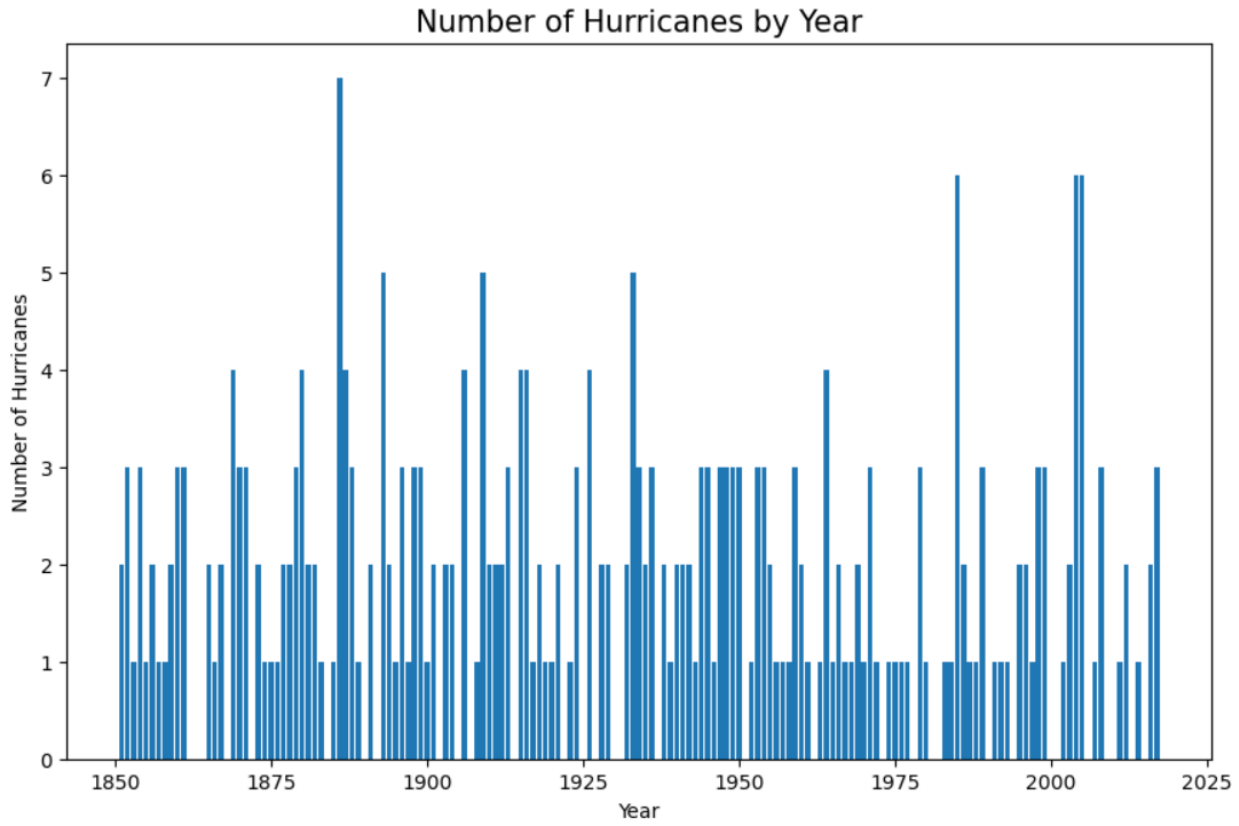

Figure 5: Bar plot of the number of hurricanes by year

I concentrated on data that had number of hurricanes and Accumulated Cycle Energy numbers. The reason for this is it goes back to our questions of can we predict the number of deaths caused by hurricanes?

Accumulated Cycle Energy (ACE) is calculated metric used by various agencies to express the energy used by tropical storms which turn into hurricanes.[9]. This metric was created by William Gray of the Colorado State University as the Hurricane Destruction Potential Index.

Assumption/Limitation/Challenges:

I had pointed out one issue with the data. Even though I got the data from a single site after digging into the information further I learned the data was gathered from other sites and blended together. This was not a surprise but a challenge. Earlier I thought I could overcome this by going to the various sites, but I was not able to get full detail of how the blending of the data was performed. This is a big issue when questions of my analysis are presented.

Issues that were encounter with the data is the lack of detail. I noticed in my early analysis that the data contains rates for global data. The initial data has been combined with information from other agency's data to derive it.

The challenge with this data is the need for detail. Hurricane are more than just ACE numbers or number of hurricanes. There are measurements need like air temps, water temps and ocean currents to name a few. In order to predict the number of deaths I need information concerns common hurricane tracks along with population density.

Ethical Considerations:

This data does not appear to have a lot of ethical challenges. I have scanned the data and there is no personally identifiable information. Everything is based at country level and arrogated numbers. The data appears to have been collected from government agencies like National Oceanic and Atmospheric Administration (NOAA), Global Health Data Exchange (GHDx) and World Bank.

On a side thought my analysis could be used to increasing insurance premiums. This might affect individual living closer to the prover line or finically challenged. If my model works to show an increase in the rate related to hurricane it could be used differently than initial intended. Companies and municipalities could use the data to drive people out of certain areas of the city or country. This might cause migration issues not just across boarders but human migration leading to more condensed living areas. That in turn might cause other human made disasters.

Conclusion:

Having run two models that gave me an increase accuracy. I ran various versions of Linear Regression and Decision Tree Regressor models planning to run one more(revised for final version). The best of those outcomes is displayed in this paper. The accuracy of my model when I first ran the model to get a baseline was a R-square value of -2.48. I used a linear regression model on non-linear data just to rule it out and have a starting point. This showed the data is not linear which does not come as a surprise. Running the second model a Decision Tree Regressor model. This returned a R-square value of 0.0709 which is better than Linear Regression model. I believe a better R-square can be obtained with different data and additional time. Below are the plots of the Decision Tree Regression model. I am missing information I need which will make this analysis more complex than originally thought.

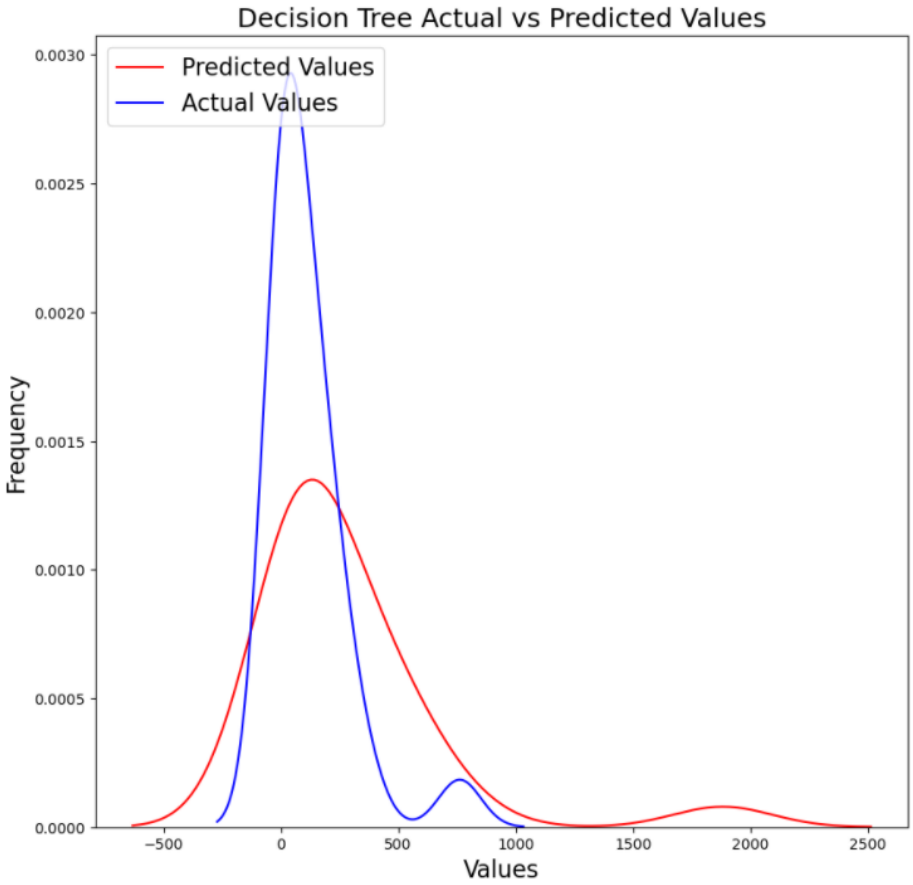

Figure 6: Decision Tree Model Analysis

Figure 5 show the result of prediction compared to actuals via line graph. I do believe some change to the data and a different model selection will yield some better results.

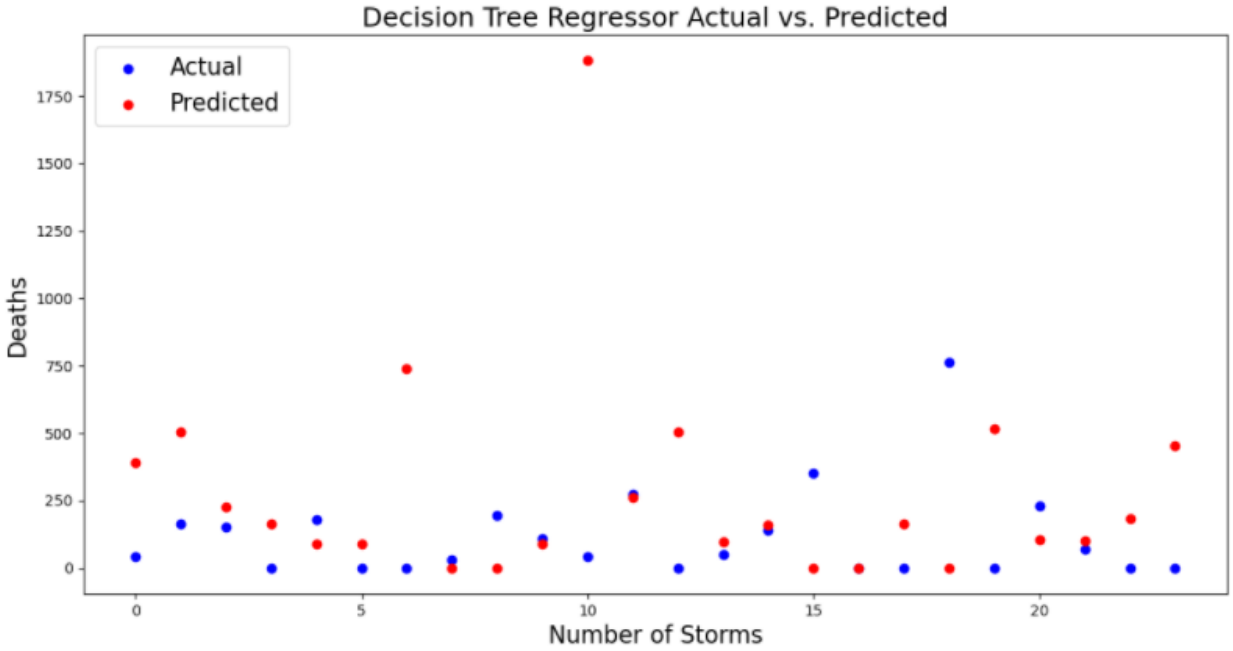

Figure 6: Scatter plot of Decision Tree Model Actual compared to Predicted

Figure 6 above shows a scatter plot of actuals compared to predicted values. While some of the points are close majority of they are very far apart. For my final I am planning to run a different model with some different data.

Reference:

1. "Our World in Data" (November 2014). https://ourworldindata.org/
2. Ritchie H. & Roser M. (November 2014). "Natural Disasters". From OurWorldInData.org. Retrieved from: https://ourworldindata.org/natural-disasters
3. Global Health Data Exchange (November 2021). From University of Washington. http://ghdx.healthdata.org/gbd-results-tool
4. Pielke R. (October 2018). "Tracking progress on the economic costs of disasters under the indicators of the sustainable development goals" From Taylor & Francis Online. https://www.tandfonline.com/doi/abs/10.1080/17477891.2018.1540343?journalCode=tenh20
5. Ache', M. (December 2021). "Natural Disaster Data Explorer". From Kaggle.com. https://www.kaggle.com/mathurinache/natural-disasters-data-explorer
6. University of Rhode Island. (2021). "Hurricane: Science and Society". From U of RI http://www.hurricanescience.org/
7. "Atlantic hurricane season". (December 2021). From Wikipedia.com. https://en.wikipedia.org/wiki/2021_Atlantic_hurricane_season
8. "Hurricanes". (2020). From Statista.com https://www.statista.com/statistics/203729/fatalities-caused-by-tropical-cyclones-in-the-us/
9. "Accumulated cyclone energy". (December 2021). From Wikipedia.com. https://en.wikipedia.org/wiki/Accumulated_cyclone_energy

Appendix:

File 1: disaster_us_storm.csv/ disaster_us_storm_df

Column Name	Column Description
no_deaths_storms	Number of deaths attributed to storms in the year indicated
no_people_injured_storms	Number of injured people attributed to storms in the year indicated
no_people_affected_storms	Number of people affected by storms in the year indicated
no_people_left_homeless_storms	Number of people left homeless by storms in the year indicated
no_total_people_affected_storms	Total number of people affected by storms in the year indicated
reconstruction_costs_storms	Total reconstruction costs for the year indicated
insured_damages_storms	Amount of insurance damage reported
total_economic_damages_storms	Total economic damage reported
death_rates_storms	The death rate from the storms in the year indicated
injury_rates_storms	Total injury rate for the year indicated
no_people_affected_storms_per_100000	Number of people affected per 100,000
homelessness_rate_storms	Total homeless rate for the year indicated
total_no_people_affected_storms_per_100000	Total number of affected by the storms per 100,000
total_economic_damages_storms_share_GDP	Total economic damage as share of the US GDP
Entity	Country Name
Year	Year in which the storm occurred

File 2: north_atlantic_hurricanes_stats.csv/ hurricanes_us_df

Column Name	Column Description
Entity	Country Name
Year	Year in which the storm occurred
no_us_hurricanes_HUDRAT_NOAA	Number of US hurricanes reported by NOAA
no_major_us_hurricanes_HUDRAT_ NOAA	Number of major US Hurricanes reported by NOAA
no_major_north_atlantic_hurricanes_HUDRAT_NOAA	Number of major North Atlantic Hurricane reported by NOAA
no_noth_atlantic_hurricanes_HUDRAT_NOAA	Number of North Atlantic Hurricane reported by NOAA

accumulated_cyclone_energy_ACE_HUDRAT_NOAA	The accumulated cyclone energy or ACE as reported by NOAA
cyclone_power_dissipation_index_PDI_HUDRAT_NOAA	The amount of power or energy the cyclone dissipated as an index reported by NOAA
hurricane_fatality_rate	The fatality rate of hurricanes for the year indicted
deaths_hurricanes_us	Number of deaths related to hurricanes in the US

Questions:

1. Where was the data obtained from?
2. Why where two different datasets used to perform the analysis?
3. Why did the models have such low R-Squares?
4. What other different predictive models could you have run?
5. What different data points could you have used to get better correlation?
6. What understanding is needed to perform this analysis?
7. Can additional detail data at a lower-level help make the analysis better?
8. Would you be able to obtain better data from NOAA?
9. Can hurricanes deaths be predicted?
10. Can anything else be predicted from the data gathered?