Name: **Muley, Tushar**
Assignment: **DSC 680 - Week 4 Milestone 3 White Paper Final**
Date: **January 9, 2022**

Final White Paper – Milestone 3

What is the Human and Economic toll of Hurricanes?

Topic:

The topic selected is deaths caused by hurricanes that make landfall in the United States. As some time permitted secondary analysis was done on storm damages.

Background:

The topic of natural disaster has come up a few times this year. The US as a whole saw a lot of natural disaster in 2021 from heatwaves, flooding, winter storms, tornados and hurricanes. It seems like we have seen an increase in natural disasters in the last 10 to 15 years. Based on news report you would think things are getting worse. You hear newscasters make headline grabbing announcements like, “storms we have not seen in the last 100 years.” Is there any truth to what newscasters 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 I wanted narrow the scope to just hurricanes.

I selected hurricanes because I experience one a few years back. The other reason is based on my readings and news reports they seem to be increasing in intensity and becoming commonplace. If the number of hurricanes increase and people become desensitize what will happen? Based on the data I have found can the data be used to predict deaths by a hurricane? Could we also use the information to predict the amount of economic damage hurricanes might cause?

Figure 1: Hurricane Isabel from the as seen from International Space Station (Sept 15, 2003)

Images like Figure 1 have become common to see on the news reports. They are pretty cool to look at. Images of a hurricane's aftermath are both amazing, scary and saddening to look at. Humans have been tracking hurricanes since the 1850s. Hurricane forecasting models have been around since the 1950s. The forecasting of hurricanes was accomplished by two major developments. One was the aircrafts (to collect data) and the other was computer technology (to crunch all that data). As computing power advanced, so did statistical models. The first statistical-dynamic tracking model appeared in 1973.[6] Hurricanes by numbers. The hurricane season start June 1st and ends November 30th. The average season has about 5.9 hurricanes. There are 10.1 named storms.[7] In 2020 hurricanes caused 47 deaths and in 2021 they caused 96 deaths. One cannot forget the hurricanes also cause a large amount of damage. In Table 1 is a list of the costliest hurricanes in recent history and the deaths they caused.

Name	Normalized damage (Billions USD)	Deaths	Year	Storm classification at peak	Area Affected
Katrina	$116.90	1,557	2005	Category 5 hurricane	Louisiana
					Mississippi
					The Bahamas
					United States Gulf Coast
					South Florida
					Northeast
					Eastern Canada
Harvey	$62.20	68	2017	Category 4 hurricane	Texas
					Louisiana
					South America
					Central America
					The Caribbean
					Yucatan Peninsula
Maria	N/A	3,057	2017	Category 5 hurricane	Puerto Rico
					Lesser Antilles
					Greater Antilles
					Caribbean Sea
					Eastern United States
Irma	$31.00	134	2017	Category 5 hurricane	Lesser Antilles
					Greater Antilles
					Caribbean Sea
					Eastern United States
Sandy	$73.50	157	2012	Category 3 hurricane	The Caribbean
					United States East Coast
					Eastern Canada

Table 1: Most destructive storm to date

The thing about hurricanes is they are more than the initial storm. Hurricane bring rain, flooding, winds and often spawn tornados. So, a hurricane is just more than a single natural

disaster, but a group of disasters occurring at once or one after the other. Using the data, I plan to perform Exploratory Data Analysis (EDA) to determine the impact of hurricanes disasters as it relates to human lives taken and damage done. 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 to see if we can determine futures deaths and economic damages hurricanes might cause.

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 only. This data has 168 rows and 8 columns. This needed to be mapped to other file which had only 99 rows.

Once loaded I ran the usually checks on the data for nulls and odd field like integers as string I found w few issues that needed to be corrected. 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.

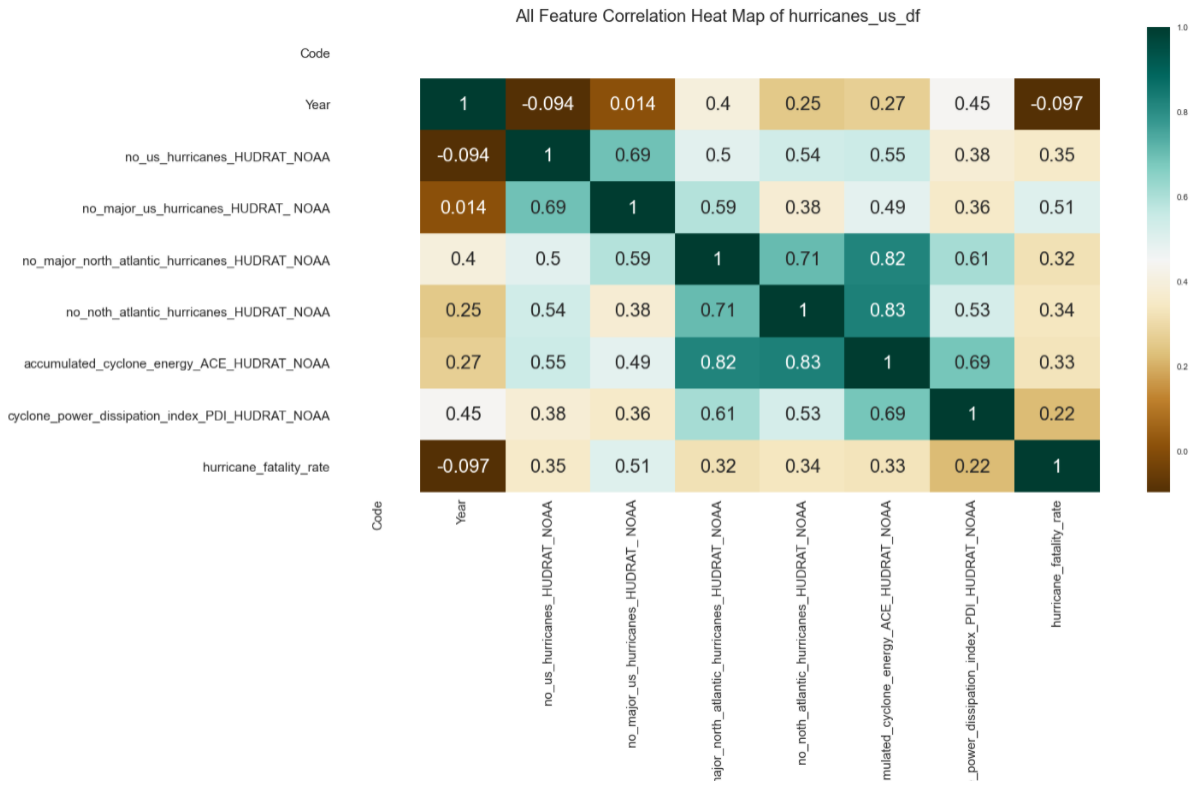

Figure 2: Correlation of hurricane_us_df file

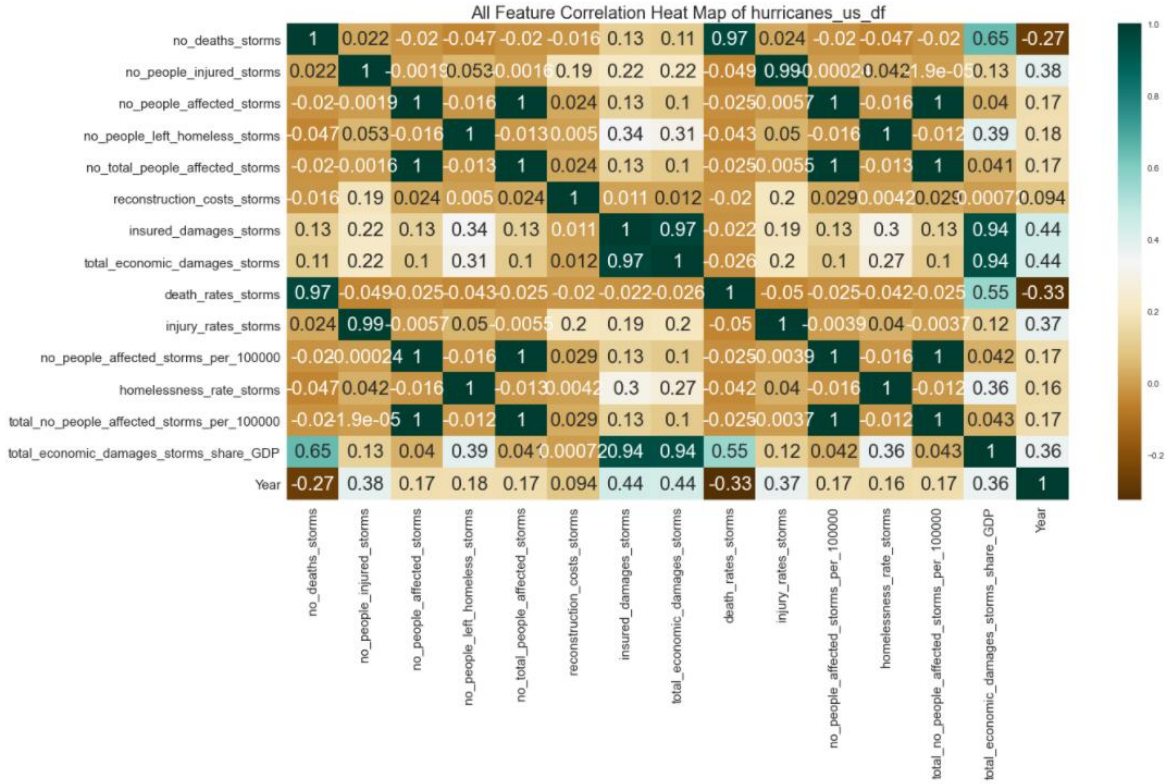

Figure 3: Correlation of disaster_us_storm_df file

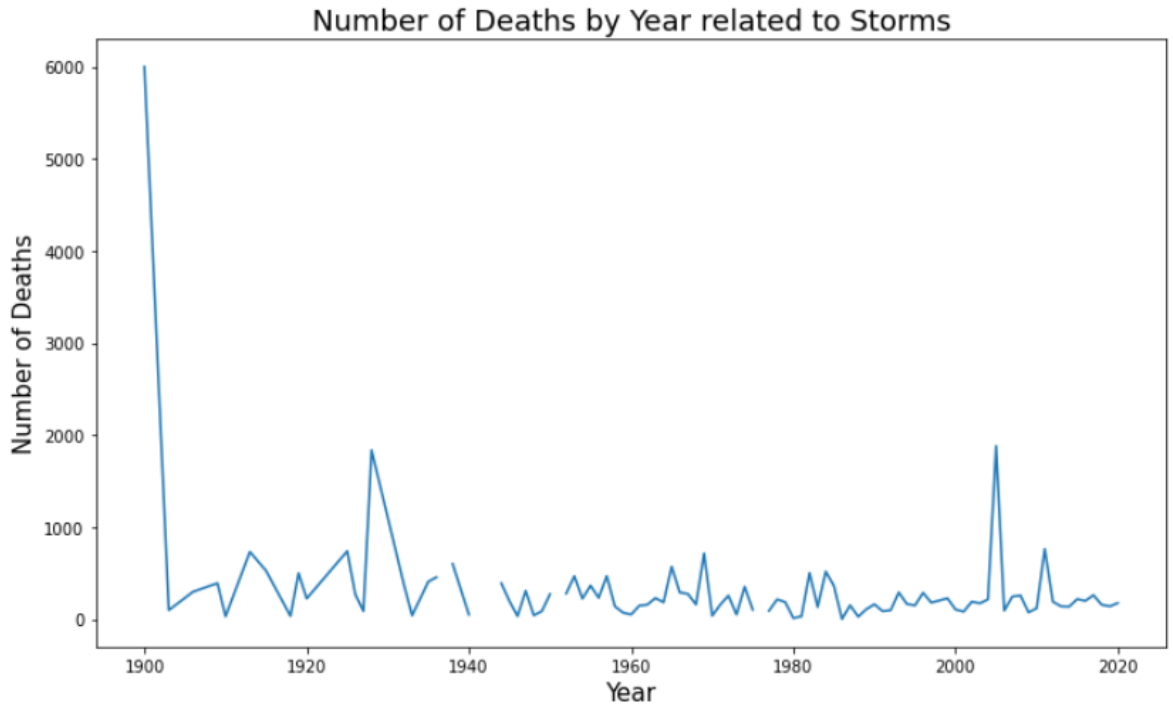

Figure 4: Number of deaths from hurricanes by year

The above graph (Figure 4)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. The data is more continues from 1980 forward.

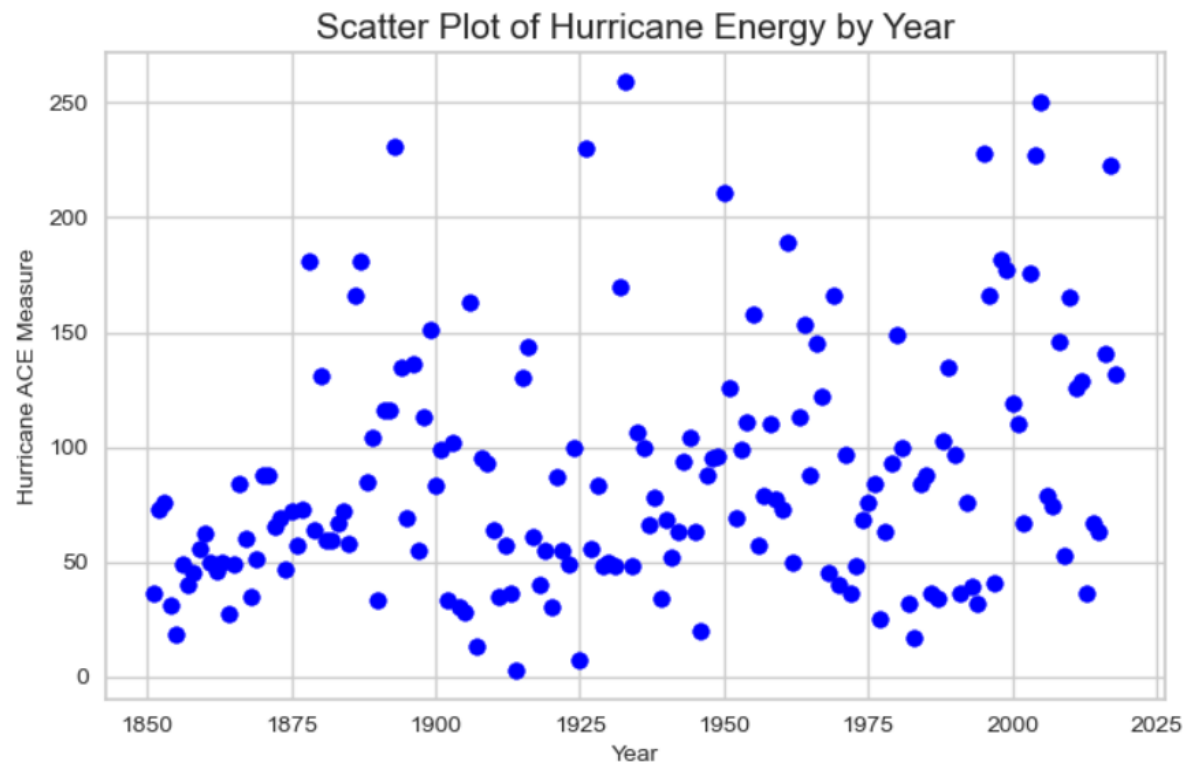

Figure 5: Scatter plot of hurricane energy by year

Another interesting fact found is level of Accumulated Cyclone Energy (ACE) by year. In Figure 5 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 number of years. This might indicate increasing amounts of dangers storms.

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 a 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.

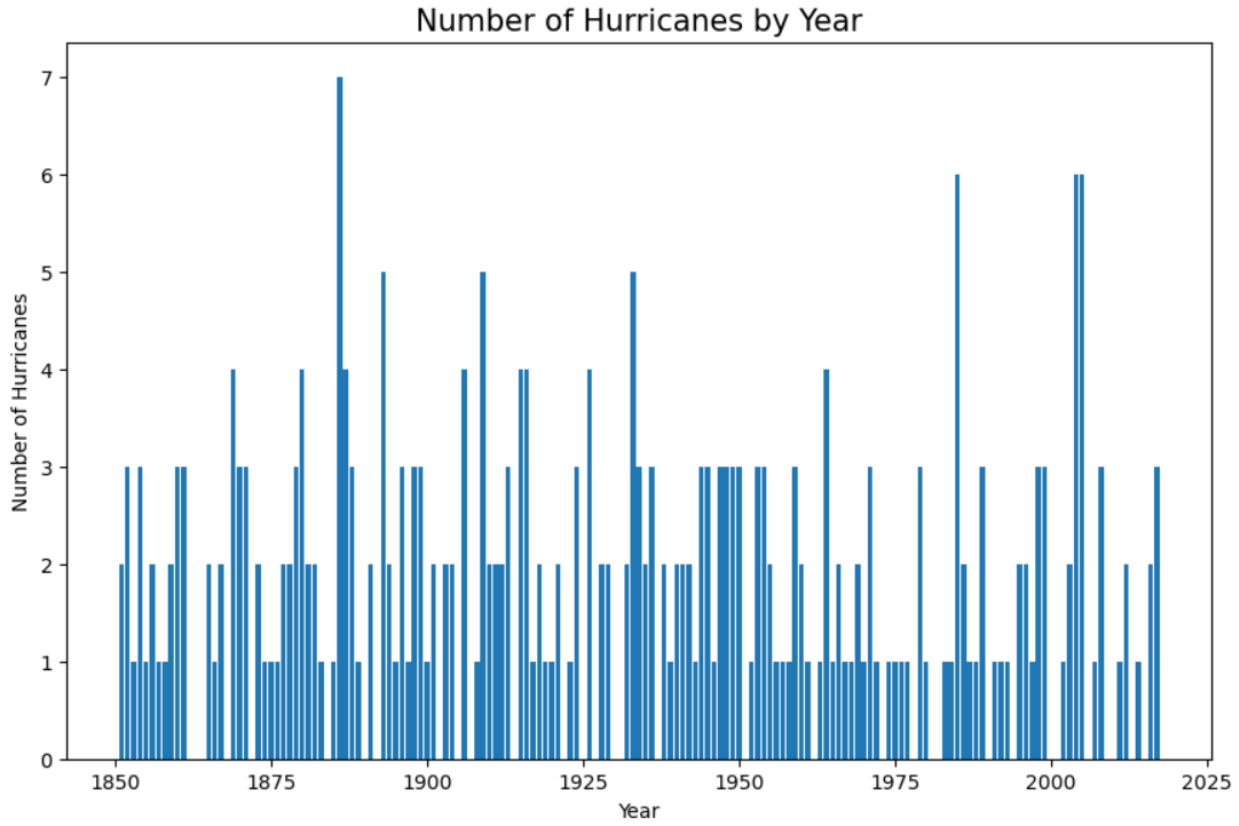

Figure 6: Bar plot of the number of hurricanes by year

Figure 6 shows the number of hurricanes that have been named in the Atlantic Ocean. Based on the bar graph 3 is the most common number of hurricanes in a given year. Years with 4 and 5 hurricane is less common.

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 of the coastal areas.

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 three models that gave me increasing accuracy. I ran various versions of Linear Regression, Decision Tree Regressor model and Ordinary Least Square. I also, ran binning of death to see if that reduction would help. The accuracy of my model when I first ran the model I got a baseline which was a R-square value of -2.48 all the way up to 0.224 for the OLS model. After grouping death by bin, the Linear Regression model increased to -1.079, the other models also saw an increase. The OLS model saw a R-Square of 0.095. I believe a little more time would yield a better result with the grouping. From Table 2 you can see the ungrouped data using OLS returned 0.224 R-square value.

Model Type	R-Squared
Linear Regression	-2.41
Decision Tree	0.071
OLS	0.224
with Binning	
Linear Regression	-1.079
Decision Tree	0.00004102
OLS	0.095

Table 2: Comparison of Model with Binning and without Binning

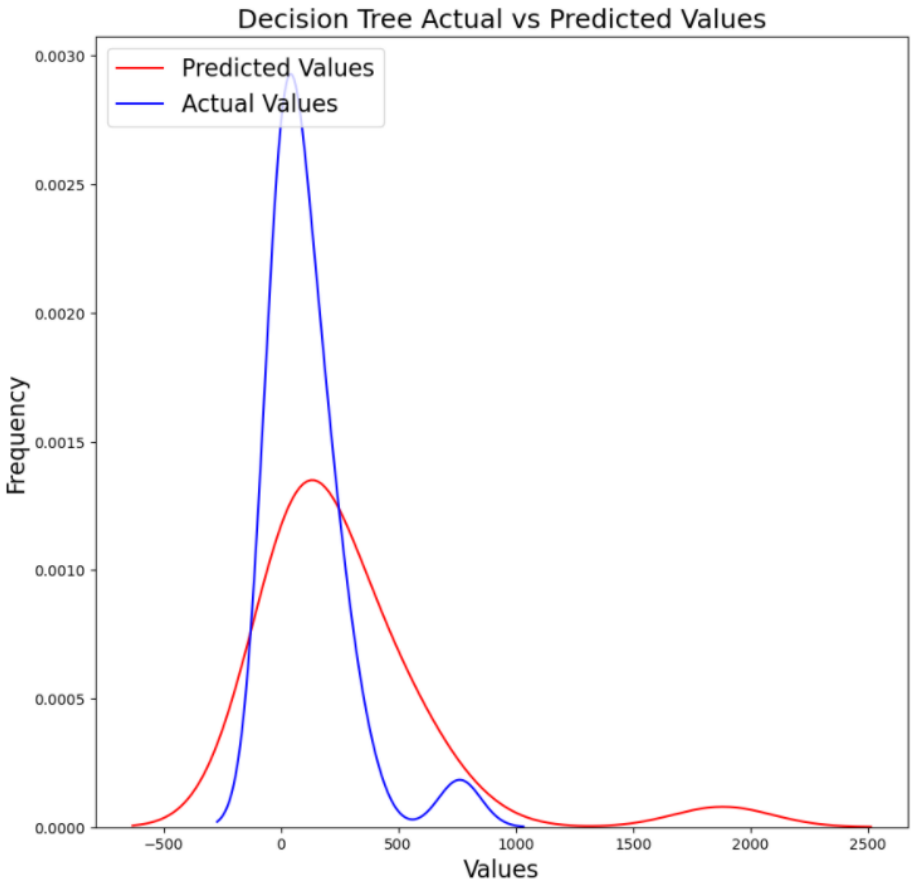

Figure 7: Decision Tree Model Analysis

Figure 7 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. The R-squared of -2.41 as a baseline.

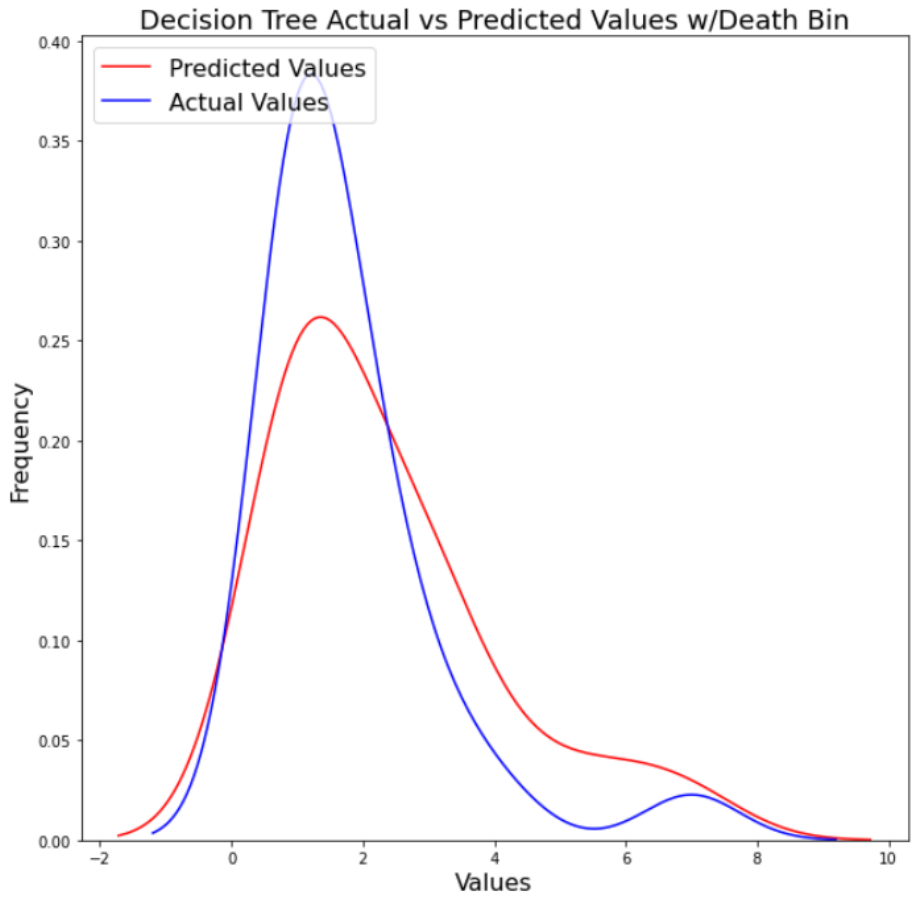

Figure 8: Decision Tree Model Analysis using death_bin

Figure 8 shows the results of prediction compared to actuals with binned data which combined a range of deaths into a set bin. The bin table is located in the Appendix section. The R-squared value increase to -1.07. Better but not good enough.

One additional model was run using Ordinary Least Squares. The R-square for this model was 0.224. This was done without the binning of deaths. I do believe additional features and more detail will get me better results. Figure 9 shows statics for the model. Figure 10 shows a line graph with actuals and prediction values. The redline show predicted values while the blueline show actual values. They are not very close.

OLS Regression Results

Dep. Variable:	deaths_hurricanes_us	R-squared:	0.224
Model:	OLS	Adj. R-squared:	0.163
Method:	Least Squares	F-statistic:	3.695
Date:	Tue, 28 Dec 2021	Prob (F-statistic):	0.00532
Time:	17:41:53	Log-Likelihood:	-477.21
No. Observations:	70	AIC:	966.4
Df Residuals:	64	BIC:	979.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	54.4223	75.425	0.722	0.473	-96.256	205.101
no_us_hurricanes_HUDRAT_NOAA	-6.8552	25.447	-0.269	0.788	-57.691	43.980
no_major_us_hurricanes_HUDRAT_ NOAA	107.4614	49.510	2.171	0.034	8.554	206.369
no_major_north_atlantic_hurricanes_HUDRAT_NOAA	25.7825	33.102	0.779	0.439	-40.346	91.911
no_noth_atlantic_hurricanes_HUDRAT_NOAA	23.6061	22.647	1.042	0.301	-21.636	68.848
accumulated_cyclone_energy_ACE_HUDRAT_NOAA	-0.8826	1.293	-0.683	0.497	-3.465	1.700

Omnibus:	52.265	Durbin-Watson:	2.239
Prob(Omnibus):	0.000	Jarque-Bera (JB):	242.618
Skew:	2.190	Prob(JB):	2.07e-53
Kurtosis:	11.000	Cond. No.	330.

Figure 9: Results from OLS model without binning

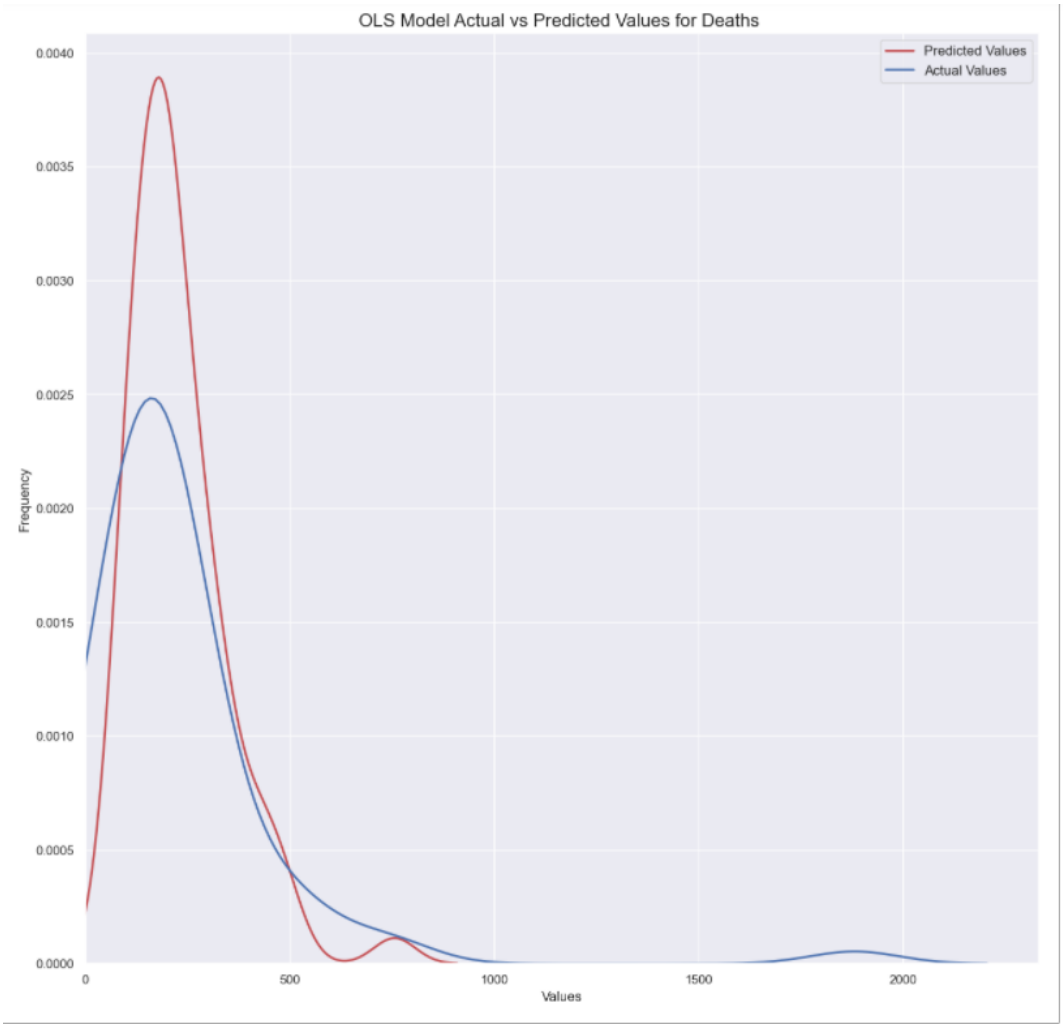

Figure 10: OLS with without death_bin

As a side analysis (later made part of the original question) I used the data to run a similar set of models to determine economic damage. This seemed a better use of the data. Below is the analysis using same data to predict economic damage. Running the data through various models the R-square value went from 0.025 using a Decision Tree model all the way to 0.410 with a OLS model. Below is the results (Figure 11) and line graph (Figure 12) of the OLS model.

OLS Regression Results

Dep. Variable:	total_economic_damages_storms	R-squared:	0.410
Model:	OLS	Adj. R-squared:	0.354
Method:	Least Squares	F-statistic:	7.302
Date:	Tue, 28 Dec 2021	Prob (F-statistic):	6.00e-06
Time:	16:40:54	Log-Likelihood:	-1285.6
No. Observations:	70	AIC:	2585.
Df Residuals:	63	BIC:	2601.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.727e+07	7.91e+06	-2.182	0.033	-3.31e+07	-1.46e+06
no_us_hurricanes_HUDRAT_NOAA	6.438e+05	2.66e+06	0.242	0.810	-4.67e+06	5.96e+06
no_major_us_hurricanes_HUDRAT_ NOAA	5.942e+06	5.36e+06	1.108	0.272	-4.77e+06	1.67e+07
no_major_north_atlantic_hurricanes_HUDRAT_NOAA	2.477e+05	3.48e+06	0.071	0.943	-6.7e+06	7.19e+06
no_noth_atlantic_hurricanes_HUDRAT_NOAA	1.432e+05	2.39e+06	0.060	0.952	-4.63e+06	4.91e+06
accumulated_cyclone_energy_ACE_HUDRAT_NOAA	1.368e+05	1.36e+05	1.009	0.317	-1.34e+05	4.08e+05
deaths_hurricanes_us	4.017e+04	1.31e+04	3.075	0.003	1.41e+04	6.63e+04

Omnibus:	69.945	Durbin-Watson:	1.411
Prob(Omnibus):	0.000	Jarque-Bera (JB):	640.020
Skew:	2.817	Prob(JB):	1.05e-139
Kurtosis:	16.700	Cond. No.	988.

Figure 11: OLS Results for Economic Damage

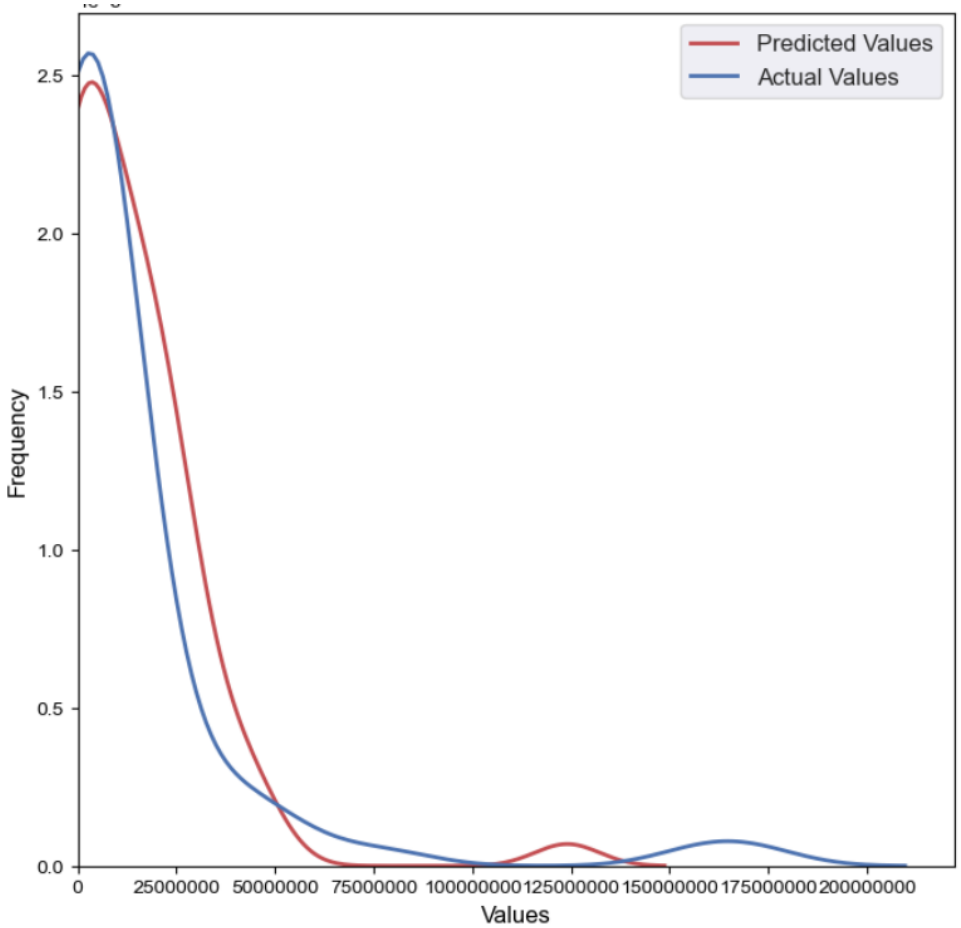

Figure 12: Line graph of OLS Actuals compared to Predicted values for Economic Damage

It appears using the same data to predict economic damage would have been use of the data from the start compared to predicting deaths. As you can see from the redline the predicted values trend close to the actual values from the data as show by the blueline. I still believe more data is needed to enhance this model to make it more accurate. Below Figure 13 shows an alternative graph of actuals compared to predicted values. As you can see there is a room for improvement. The blue bars are the actuals values and the gold or orange bars are the predicted values.

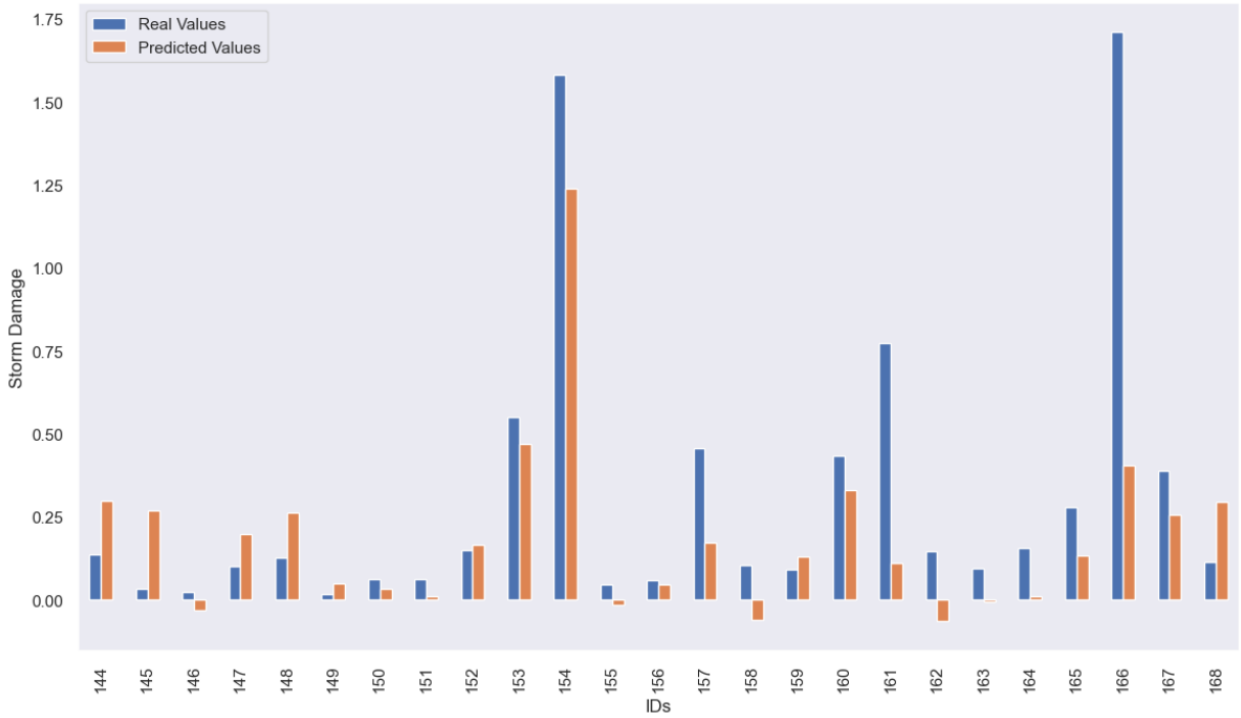

Figure 13: Bar graph as an alternative to line graph showing Predicted value for Economic Damage

Some varied close while other are way off leading to the 0.410 R-squared value. I believe additional data like storm tracks, population density, inflation, and region-specific information I could build a better model. This model would be region based. For example, Florida area of Orlando and we could gauge the expected damage in that area. For now, this was best that could be accomplished with the information available.

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
10. Erdman, J. (July 2018). “The 15 Most Iconic Hurricane Images of All Time”. From The Weather Channel.com. https://weather.com/storms/hurricane/news/hurricane-images-photos-most-iconic
11. Rappaport, E. (March 2014). “Fatalities in the United States from Atlantic Tropical Cyclones: New Data and Interpretation”. From American Meteorological Society.com https://journals.ametsoc.org/view/journals/bams/95/3/bams-d-12-00074.1.xml
12. “Fast Fact/Hurricane Costs”. (January 2022). From Office for Coastal Management. https://coast.noaa.gov/states/fast-facts/hurricane-costs.html

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

Death_bin grouping

Number of Deaths	Bin Number
0-100	1
101-200	2
201-300	3
301-400	4
401-500	5
501-600	6
601-700	7
701-800	8
801-900	9
901=>	10

Questions:

1. Where was the data obtained from?
   The data was obtained from 'Our World in Data' the 'Natural Disaster' subsite. The Our World in Data' site did do some transformation and combining of data to get their data. It is all documented in the reference section of the White Paper.
2. Why where two different datasets used to perform the analysis?
   The reason for the two different dataset was the information on the Natural Disaster subsite needed to be combined with additional hurricane data like ACE information. Everything was split out and I need a more cohesive data set.
3. Why did the models have such low R-Squares?
   I believe this can be attributed to two factors. The data was robust. But more detail information is needed. As I started going into this further I learned having storm strength data, population and common storm tracks might have help determine the number deaths a little better. I tried binning the deaths so we could put a range on the expectation. This help the R^2 value went from -2.41 to -1.07. An improvement but still falling short of anything meaning full. The binning of data would be too broad.

   On the economic impact I saw a good starting value of 0.410. That was after using OLS model. I think I can improve this by selecting a better model and identifying common storm tracks. That information was a little hard to find.
4. What other different predictive models could you have run?
   I ran linear regression, decision tree and OLS. OLS seems to produce the best results for the limited amount of time I have to work on this. I would try using larger amounts of

data and running classification model for deaths and maybe for economic damage as well. If you knew which areas these storms might head in you would be able to determine the amount of damage they might cause. Deaths would be hard because human are becoming desensitized to warnings and requests to leave when big storms like this approach. Having sentiment analysis along with running ensemble model might help.

5. What different data points could you have used to get better correlation?
   The correlation is not bad. It could be better. I believe having some additional data like population, common storm tracks and maybe even some type of storm sentiment. Storm sentiment would be how people are perceiving the threat alerts. How they feel about the first storm of the season compared to the second, third all the way to average. This might get better death rates. Or even areas where deaths have occurred.

6. What understanding is needed to perform this analysis?
   When I started the analysis I thought it was straight forward. There is a lot to understand about hurricanes, their tracks, what cause changes and other details. This was something I did not expect. The volume of data needed is large and I believe a review of how various data effect the correlation would need to be done. This was not simple and unexpected. With the amount of data that would be need an the different types of analysis that would need to be done before coming to a predictive model is a lot. This would have been better served as an EDA project.
7. Can additional detail data at a lower-level help make the analysis better?
   Yes, more detail is always better but too much detail will probably slow the model down and cause other issues. For example, when you watch Metrologies show different storm tracks they are taking into account so many different variables like water temp, pressure and water currents in real time it is difficult to gauge where a storm will go. Detail would help but deaths might be harder to predict. You could come up with a range or even category but it would be wide and saying we can expect 0 to 100 deaths and be off by more than 50%.
8. Would you be able to obtain better data from NOAA?
   NOAA has a lot of the data. But they don't just make it available as it is very large. They use different source to gather the data like airplanes, boats, and sensors in the water. It can be done but it will be a large amount of data and my PC would not be able to handle it.
9. Can hurricanes deaths be predicted?
   Yes. Even if categorized into say small, medium or large can be done. It will be very inaccurate on the level that it would help the public make the connection that a storm is bad enough to warrant leaving the area. That every storm is like finger print. They are all a little different and different enough that you can't get a large population to understand the deadliness of it.

10. Can anything else be predicted from the data gathered?

Yes, I had some success but limited. I was able to gauge economic damage from the storm. But it is low at 41% accuracy. I believe that can be increased if I knew what additional data I could get and changing the model to classification or running a ensemble of models.