Data 670 Data Analytics

Daanish Ahmed

Professor: Dr. Steve Knode

Assignment 6: Final Report

August 12, 2018

Executive Summary

This project involves analyzing storm data collected across the U.S. by NOAA and the NWS during the year 2017 ("Storm Events Database," n.d.). The analysis revolves around studying the characteristics of storms that resulted in loss of life. The goal is to predict the likelihood that any given storm will produce casualties, as well as identify the five primary features that may cause a storm to be deadly. The analysis involves creating five different predictive models and tuning those models until at least one of them has a classification accuracy of 80% or higher as well as a sensitivity of at least 70%. By achieving an accurate and reliable model, our organization will be able to identify characteristics of deadly storms at a much faster rate. This can help to increase the warning times issued for any storms that occur in the future. Thus, a successful project has the potential to help increase the storm survival rate across the country.

The analysis includes several visualizations. Firstly, a geospatial map of the U.S. is used to determine which states in the country have the highest risk of being hit by a dangerous storm. Secondly, a time series graph is used to evaluate the time(s) of the year when life-threatening storms are most common. These two methods are useful for finding the locations that are most vulnerable to deadly storms, as well as the months during which these storms are most likely to appear. The final component consists of a text mining study (word cloud) on the storm event descriptions to determine which words are frequently used to describe dangerous storms. This will allow us to identify additional warning signs for deadly storms in the future.

Through my analysis, I found that all five predictive models attained accuracy rates above 80% and sensitivity rates above 70%. Furthermore, some of these models provided information about the variables that are important for classifying dangerous storms—which will be useful for predicting casualties in future storms. My geospatial analysis revealed that certain states have a higher risk of storm casualties, but not every state that is frequently hit by tornadoes or hurricanes will necessarily have a high casualty rate. My time series graph indicated that most storm deaths do indeed occur during the Summer as expected. However, other months of the year may experience rare outbreaks of severe weather—which highlights the importance of storm preparation during the entire year. Finally, my word cloud revealed that a victim's location during a storm can contribute to their likelihood of survival. Many casualties occurred in permanent homes, outside, or in the water. Likewise, fewer casualties occurred within mobile homes. These results will ideally raise awareness about storm safety measures across the country.

Table of Contents

Project Scope

Problem Description

In this project, I will analyze three datasets containing U.S. storm information from the National Oceanic and Atmospheric Administration (NOAA) and National Centers for Environmental Information (NCEI). The data was obtained from NOAA's storm events database, which consists of storm data collected by the National Weather Service (NWS) from 1950 to the present ("Storm Events Database," n.d.). The data was used by the NOAA to create monthly storm data publications, which document severe weather events and highlight significant storms using photos and text narratives ("Storm Data Publication," n.d.). My datasets consist of all recorded storms in the U.S. during the year 2017. The primary dataset contains the storm details, and it includes information such as the storm date, location(s), type of storm, number of injuries, number of fatalities, and damage (in U.S. dollars). The second dataset includes more details about storm fatalities, such as the date and time of death, age, gender, and death location. The third dataset provides more specific location information, such as the storm's range and direction.

My goal is to create five predictive models to determine the likelihood that a given storm will result in deaths. I will include all storm types included in the dataset, rather than focusing only on tornadoes or hurricanes. Ideally, I want at least one of my models to have a classification accuracy of at least 80%. To achieve this, I will identify the 5 factors that contribute the most towards storm casualties, and I will implement several types of classification models to see which method best fits the data. Next, I will tune the model parameters to ensure that each model has optimal accuracy. I hope that my research will contribute towards promoting faster storm warning times, which should help to gradually reduce storm casualties in the future.

Additionally, I will incorporate a geospatial map of the U.S. to evaluate storm fatalities in different states, and I will also create a time series graph to analyze storm casualties during certain times of the year. These methods will allow me to determine which states have the highest risk of dangerous storms, as well as which months contain the highest storm casualties. Furthermore, I will use text mining to study the word usage in the storm event descriptions for all storms that had casualties. This will allow me to find the terms that are most frequently used to describe dangerous storms. I believe that mining the unstructured data in these text descriptions can allow my organization to identify possible warning signs that would otherwise be neglected. Such a technique may also contribute towards lowering storm casualties in the future.

Business Understanding

Severe weather is a major concern for public safety in the U.S. and around the world. In 2017 alone, 103 Americans lost their lives to hurricanes Harvey and Irma (Johnson, 2017). That same year, tornadoes were responsible for 35 deaths across the U.S. ("Annual U.S. Killer Tornado Statistics," 2018). As such, predicting natural disasters is crucial to saving lives in the future. By understanding the properties of different storms and what makes them lethal, researchers and meteorologists will be able to identify warning signs in future storms much earlier than before. This will help to increase warning times before storms—which will give people more time to prepare for the storm, relocate, or find adequate shelter.

One of the challenges with severe weather predictions is the unpredictability of these events. Some years might experience few major storms, while the very next year may contain numerous catastrophic storms. For example, tornadoes caused 45 deaths in the U.S. during 2010—but in 2011, there were over 550 tornado casualties (Brooks, 2009). Because of this, it may seem difficult to identify trends in a severe weather dataset. However, there are certain characteristics that can cause some storms to be deadly while others are not. This project will not necessarily focus on the number of casualties in each storm. Instead, it will evaluate each storm based on whether any deaths occurred. This will allow us to identify all life-threatening storms rather than only the most destructive ones.

Organization

The National Oceanic and Atmospheric Administration (NOAA) is a government agency within the Department of Commerce that focuses on environmental sciences such as weather forecasting and marine studies (NOAA, n.d.). According to this source, NOAA's goals include predicting weather and climate changes as well as preserving marine ecosystems and coastal areas. It was formally established in 1970, but its roots go back to 1807—when President Thomas Jefferson founded the U.S. Coast and Geodetic Survey to promote safer maritime travel in American waters (NOAA, 2006). NOAA and its predecessors have been conducting severe weather forecasts since the early 19th century, and its techniques have improved due to technologies such as computers, satellites, and Doppler radar ("Severe Weather Watches," 2006).

NOAA merged its three data centers into the National Centers for Environmental Information (NCEI) due to an increasing demand for high-quality environmental data ("About Us," n.d.). This source states that NCEI's responsibility is to store and manage NOAA's archive of environmental data, ensuring its quality and availability to the public.

The National Weather Service (NWS) is an organization within NOAA that is directly responsible for collecting weather and climate data and making forecasts ("The National Weather Service," n.d.). It was founded in 1870 by President Ulysses S. Grant to perform weather observations at military bases throughout the U.S., as well as along the coasts and the Great Lakes area ("History of the National Weather Service," 2015). According to this source, the agency was known as the U.S. Weather Bureau until 1970—when it was renamed as the National Weather Service (NWS). It was incorporated into the Environmental Science Services Administration (ESSA) in 1965, which became NOAA in 1970 ("History of the National Weather Service," 2015). Although the NWS measures its success based on forecast accuracy, a tornado outbreak in 2011 showed that accurate predictions do not necessarily prevent loss of life ("Your National Weather Service," 2017). After the storms claimed over 300 lives, the NWS established the "Weather-Ready Nation" program in which they worked with communities to promote storm preparation across the country ("Your National Weather Service," 2017).

Stakeholders

There are several important stakeholders connected to this problem. One set of stakeholders includes the meteorologists and researchers associated with the NWS, NOAA, or other weather organizations. My research is designed primarily to predict the likelihood of casualties across many types of storms. If successful, my project will help researchers to identify potentially dangerous storms more easily. In the future, it may also allow weather stations to issue storm warnings earlier than before. Another set of stakeholders includes the U.S. National Guard, Coast Guard, and regional organizations devoted to public safety. If my project succeeds, it might help these organizations to know when to expect dangerous storms. This should improve their ability to respond to natural disasters and protect more people.

The largest set of stakeholders consists of the American people, especially those who live in areas that are more prone to natural disasters. Storms can appear in any part of the U.S., but different parts of the country have higher risks of experiencing certain storms. For instance, the

NWS found that Texas experienced an average of 137 tornadoes every year, while Florida and Oklahoma witnessed about 52 and 47 tornadoes each year respectively (NOAA, 2013). While tornadoes are most common in the south and central parts of the country, hurricanes and tropical storms are more likely to affect states along the Gulf of Mexico and Atlantic coast. Likewise, blizzards and Winter storms are more common in northern states. Part of my project will contain a geospatial analysis, which will address the severe weather risk in states across the country. Although my project cannot directly reduce casualties, I hope that my findings will promote awareness of storm severity and encourage Americans to better prepare for future disasters.

Define Business Area

Meteorology is a major science that focuses on predicting weather events. The history of weather forecasting goes back thousands of years, with the ancient Babylonians being among the first to make short-term weather predictions around 650 BC ("Weather Forecasting," n.d.). Major breakthroughs occurred in the 16th and 17th centuries, during which Galileo Galilei developed one of the first thermometers in 1592 and Evangelista Torricelli invented the barometer in 1643 to measure atmospheric pressure ("Weather Forecasting," n.d.). According to this article, the invention of the telegraph in the 19th century revolutionized the industry since it allowed for weather observations to be quickly transmitted across great distances. Another breakthrough was the invention of radar, which was first used during World War II to track aircraft movement (Moran, 2016). According to this article, radar was used for weather forecasting after the war, during which it successfully helped an airplane land during a thunderstorm. Radar usage was expanded significantly in the 1950s to predict hurricanes across the U.S. (Moran, 2016).

One of the most important meteorological developments was the implementation of the Doppler effect into weather radars. The Doppler effect evaluates the change in pitch of a moving object to determine how fast the object is moving ("What is a Weather Radar," n.d.). When used during a rainstorm, a Doppler radar can measure the pitch from falling raindrops to determine the speed of the winds that carry those raindrops ("What is a Weather Radar," n.d.). As a result, this technology makes it easier to determine if a storm will produce destructive winds. Doppler weather radars were developed during a joint effort by NOAA and the U.S. Air Force in 1978, and they became widespread in the U.S. after the project was completed in 1990 (Moran, 2016).

Business Objectives

Although this project focuses on predicting storm casualties, I expect that my findings will help my organization to make progress towards improving storm safety. My first business objective is to contribute towards reducing casualties in any meaningful way. Although I am unable to measure casualty reduction within the scope of this project, I hope that my findings will promote practices that will save lives in the future. To work towards this objective, I will create a predictive model with a classification accuracy of at least 80%. Additionally, I will identify the 5 factors that contribute the most towards fatal storms. I will also incorporate a geospatial map to evaluate storm casualties by state, as well as a time series graph to analyze storms during each month of the year. This will help to raise awareness about storm risks in certain states or during different times of the year.

My second business objective is to help increase the average storm warning time across the U.S. Storm warning times can differ drastically depending on the type of storm. Hurricanes and tropical storms can receive over a week of notifications, while tornadoes usually involve only several minutes of warning time. However, I believe that increasing the storm warning time can help to reduce casualties to a certain extent. While some people may fail or refuse to evacuate or find shelter, many others would take the warnings seriously. To achieve this objective in the future, my model(s) will have to be very effective at classifying storm casualties.

My third business objective is to work towards reducing the number of false alarms. One of the problems with storm warnings is the high frequency of false alarms. According to Erdman (2018), roughly 70% of all tornado warnings in America are false alarms. This can cause many people to disregard tornado warnings, which puts them at a significant risk if the tornado warning is accurate. Solving this issue will be difficult, but we can work towards it by obtaining a model with a high accuracy and sensitivity. The predictive model(s) will have to be effective at classifying both lethal and non-lethal storms to ensure a lower false alarm rate.

Business Success Criteria

To evaluate this project, I will mainly consider the performance of my predictive models. I will label the project a success if I meet at least one of the criteria described in this section. Firstly, the project will be successful if I am able to obtain at least one model with a classification

accuracy of at least 80%. By predicting the likelihood that a storm will be lethal, we will be able to easily classify life-threatening storms in the future. There are many potential implications for achieving an accurate model. For instance, it may help researchers and meteorologists to identify dangerous storms more quickly, which can contribute towards increasing the warning time before a storm arrives. This will give people more time to prepare for storms, which will hopefully result in fewer casualties. And though I cannot directly measure casualty reductions, I would be very pleased if I learned that my research helped to save lives.

Additionally, I will consider this project successful if I can identify the five most important features and how each of them impacts the number of storm casualties. One of my goals is to understand what causes some storms to be so deadly. If I can narrow down the most important variables and understand why they are critical, then my organization will be able to easily identify such characteristics in future storms. Furthermore, the project will be successful if I am able to contribute towards increasing storm warning time by at least 5%. This will be possible if I can accurately identify lethal storms, as well as the characteristics that cause storms to be dangerous. Finally, I will view the project as successful if I can help to reduce the storm false alarm rate by at least 5%. This can be achieved if my model has a high sensitivity rate and can accurately identify dangerous storms.

Background

The data in the storm events database is used primarily by the NOAA and NWS to document severe weather phenomenon each month. This involves creating storm data publications that focus on deadly storms as well as other unusual or newsworthy weather events ("Storm Data Publication," n.d.). In addition, these agencies use the data to improve the quality of their forecasting predictions. However, having accurate forecasts does not guarantee a low casualty rate. In 2011, the NWS accurately predicted a tornado outbreak in Mississippi and Alabama and issued tornado warnings 20 minutes in advance—which is much higher than the average warning time ("Your National Weather Service," 2017). Nevertheless, this article states that over 300 lives were lost during the outbreak. The article implies that reducing storm casualties cannot be achieved by accurate predictions alone, but also by preparing communities for potential storms and ensuring that they take the threat seriously.

Research

Similar studies have been conducted using the data within this database, as well as other related problems. One such problem is the issue of tornado warning false alarms. About 70% of all tornado warnings in the U.S. are false alarms (Erdman, 2018). As a result, many people may not take tornado warnings seriously—which can lead to higher casualties if a tornado does appear. To address this issue, the NWS created the Tornado Warning Improvement Process (TWIP), which aims to reduce the rate of false alarms and increase the accuracy of tornado detections (Erdman, 2018). According to the article, the TWIP considers factors such as wind shear and cloud base height to determine the likelihood of a tornado forming. One of their efforts helped to reduce the false alarm rate by 45% in southern Wisconsin (Erdman, 2018).

Another study involved using the NWS storm data, but covering different years. In 2014, Timothy Coleman and P. Grady Dixon sought to predict the likelihood of a tornado hitting any area in the U.S. (Erdman, 2016). The article states that Coleman and Dixon improved upon previous studies by considering a tornado's path length rather than only its starting location. They found that Southern states such as Arkansas, Louisiana, and Tennessee had the highest tornado risk—which challenges the conventional belief that Great Plains states such as Kansas and Nebraska had the highest risk (Erdman, 2016).

Gaps in this Problem Resolution

The gap that I will fill is to predict whether any given storm will be potentially dangerous. Previous studies have shown that predicting the likelihood of severe weather does not guarantee a high survival rate. However, I believe that identifying lethal storms can help to raise awareness of storm severity. I hope that my findings can encourage readers to take storm warnings more seriously in the future. In addition, many previous studies have focused on specific types of storms such as tornadoes or hurricanes. However, this can cause certain weather events (such as floods and droughts) to be neglected when compared to other storm types. To fill this gap, I will consider all types of storms included in the dataset. Ideally, my research should improve my organization's ability to predict dangerous storms ahead of time.

Proposed Project

This project was selected because of my interest in natural sciences. I am deeply fascinated by weather and natural disasters, and I have wondered if anything can be done to improve public awareness of storm severity. I do not work in a weather-related industry, but it is one of several fields that I may consider working in. This project is very important because of the frequency and severity of natural disasters in the U.S. From 2010 to 2012, there were 19 hurricanes or tropical storms recorded every year ("Top 10 Most Active," n.d.). Several of these storms resulted in a significant loss of life—Hurricane Irene caused 49 deaths (of which 41 were American), while Hurricane Sandy killed 72 Americans and 147 people altogether ("Hurricane Statistics," 2018). More recent storms such as Harvey, Irma, and Maria have caused 68, 44, and 64 U.S. deaths respectively ("Hurricane Statistics," 2018).

When looking at tornadoes, a study using the NWS storm data and other historical tornado statistics found that the average number of tornado deaths has decreased noticeably since 1875 (Brooks, 2009). However, this article also reveals that certain individual years can have significantly more casualties than others—making it difficult to evaluate storm casualties on a yearly basis. For instance, the year 2010 saw only 45 tornado-related deaths in the U.S., but 2011 had 553 casualties—the highest tornado death rate since 1925 (Brooks, 2009). When evaluating the regions with the highest tornado casualties, the NWS and NOAA found that Mississippi, Texas, and Indiana had the highest average death rate with 10, 8, and 7 each (NOAA, 2013). However, other types of storms have resulted in extensive death tolls as well. For instance, the "Snowmageddon" blizzard of 2010 resulted in 41 deaths, while Winter Storm Jonas killed 55 people in 2016 ("10 Worst US Storms," 2016).

One of the issues with predicting severe weather is the ability to issue timely warnings about potentially dangerous storms. According to the NWS, the average tornado warning time is only about 13 minutes (Heberton, 2014). This can make it difficult for people to prepare for the storms and find adequate shelter. But if we can accurately predict a storm's severity, it can allow us to identify warning signs from life-threatening storms much earlier than before. This project will primarily benefit researchers studying different storms and the factors which influence their severity. If successful, it may have the potential to benefit millions of Americans across the country who would otherwise be vulnerable to various natural disasters.

Key Performance Indicators

The first KPI is to select a predictive model that can accurately classify storm casualties at least 80% of the time. My project will involve building five models and tuning their parameters to obtain optimal accuracy. Ideally, at least one of my five models should achieve a classification accuracy of at least 80%. This is important because having a high model accuracy can allow us to quickly and easily classify dangerous storms in the future. Preferably, I would like a model that can accurately identify deadly storms at least 90-95% of the time. But due to the scale of the project and the datasets, it is reasonable to strive for an 80% accuracy rate and further improve the accuracy in a future analysis.

The second KPI is to obtain a model with a sensitivity of at least 70%. Although overall accuracy is important, we want a model that can identify lethal storms with at least some degree of reliability. This is a major concern because of the skewness of my target variable. Since there are very few death and injury cases, it is easy for the model to disregard these cases and focus on classifying non-lethal storms. The result may have a very high accuracy, but it would fail to identify most of the storm casualties. To achieve this sensitivity rate, I will change the cutoff threshold to ensure that the model will accurately classify death cases.

The third KPI is to identify the 5 factors that contribute the most towards fatal storms. I not only seek to predict the likelihood of storm casualties, but I also wish to analyze why certain storms are deadlier than others. Understanding the causes behind storm severity will make it easier to determine if a given storm will be deadly. This can be achieved during the predictive model implementation, during which I will look at the variables with the highest significance rating for each model. But my goal is not just to identify the important factors, but also to understand how they contribute to a storm's severity. To do this, I will explore the important variables and study their relationship with the target. For variables related to location or time, I will create geospatial and time series graphs to analyze their impact on storm casualties.

Project Insights of your Data Analysis

I expect that my analysis will produce numerous insights towards addressing the issue of storm casualties. With regards to my predictive model, I expect that it will be feasible to obtain at least one model with an accuracy of 80% or higher. Since there are many variables in the data,

I will likely have to remove many of them to ensure that only the most relevant features are included. However, I think that achieving a sensitivity of 70% will be more difficult due to the skewness of my target variable. Since there are very few cases of deaths and injuries compared to the number of storm events, I will change the cutoff threshold so that the model will identify cases with casualties. This will reduce the overall accuracy of the model, but it will help me to achieve the desired sensitivity rate. I will try to select an appropriate threshold that greatly improves the model sensitivity without causing the overall accuracy to decrease too much.

When looking at the features which contribute the most towards storm casualties, I expect that some of these variables may include the type of storm, location, storm path, time of year, number of injuries, and amount of property damage. With regards to the storm type, I suspect that tornadoes and hurricanes will result in casualties more often than other types of storms. However, I would not be surprised if floods or blizzards appeared on this list as well. For storm location, I expect that most storms will appear in states within the South, Midwest, Great Plains, and Gulf of Mexico areas. This is primarily due to the high frequency of tornadoes or hurricanes in these areas. I will find out if these assumptions are true when I create my geospatial visualization. For storm path, I believe that storms with longer paths will often result in more deaths and injuries. However, this is assuming that the storms are of the same type—since a thunderstorm with a long path may not be as deadly as a major tornado with a short path.

With regards to time of year, I expect that casualties will peak during the late Spring to early Autumn—since that is when tornadoes and hurricanes are most likely to occur. However, I would not be surprised if frequent casualties occurred during the Winter due to blizzards and Winter storms. I will evaluate this assumption once I create my time series graph on storm casualties throughout the year. When considering injuries, I expect that storms with higher numbers of injuries are more likely to result in deaths. And when looking at property damage, I believe that storms with casualties will often produce higher amounts of damage. This is because storms that result in deaths or injuries are often more powerful and destructive. The final component of my analysis involves performing a text mining analysis on words that are frequently used to describe deadly storms. In this section, I expect that the most frequent words will relate to hurricanes, tornadoes, tropical storms, flooding, and high winds. I will exclude stop words from this list, as well as other generic terms such as “storm” or “weather.”

One of my concerns is that there are relatively few cases of deaths and injuries compared to the total number of storms. Although my data is very robust, the small sample size of casualties will require changing the cutoff threshold to improve the predictive model sensitivity. This will cause my model's accuracy to decrease, and it might prevent me from achieving my goal of having an accuracy rate of 80% or higher. One solution for future analysis is to incorporate storm data from multiple years instead of only using 2017 data. This will allow me to obtain a higher number of storm events with casualties. Another solution would be to focus on a specific type of storm (such as tornadoes or hurricanes) rather than studying all of them together. Although this differs from my initial project goal, it will allow me to have a more specific subject and a more balanced proportion of deaths and injuries.

Project Milestones

Here, I will describe the milestones which I have completed in this project. Firstly, I defined my project's initial scope and submitted the Assignment 1 report on June 3. Next, I made major revisions to my project scope and began examining my data in depth. I completed the Assignment 2 report on my dataset selection on June 17. Afterwards, I presented my project to my group on June 26 and submitted the recorded Presentation 1 on June 28. Their feedback allowed me to make additional changes to my scope—such as changing my target variable to focus only on deaths rather than on both deaths and injuries. I then completed the data preparation steps in R and SAS Enterprise Miner, and I finished the Assignment 3 write-up on July 8. I also updated my preprocessing steps over the course of the project as needed. For instance, I removed correlated variables from the fatalities dataset during the creation of my predictive models for Assignment 5 in Week 10.

My approach was to complete the PowerPoint presentations by the Friday on the week it is due. I finished the second presentation and presented it to the class on July 13, and I finished the third presentation on August 3. This allowed me to present my presentation to my classmates and receive feedback in a timely manner. It also gave me additional time to make changes to my project based on their feedback. With regards to the assignments, I began each step shortly after completing the preceding step. My goal was to have the software-related tasks (SQL, R, SAS EM, Tableau, etc.) completed about 4 days prior to when the respective assignment is due. This is because it often takes me longer to write the reports than to perform the analysis. This approach

gave me the time needed to work on the reports and the presentations, as well as to make necessary changes to my code or models.

For Assignment 4, however, I finished the visualizations by July 13 and completed the written report by July 15. Since Assignment 3 was due on July 8, there was only a week before Assignment 4 is due. Thus, I had to begin work on the visualizations before finishing the Assignment 3 report. Additionally, Presentation 2 was also due that week, so I had to balance working on all three tasks. For Assignment 5, I returned to my approach of completing the analysis 4 days before the report is due. I finished creating the predictive models by July 25, and I submitted the Assignment 5 report on July 29. For Assignment 6, I began working on the report immediately after completing Presentation 3, and I completed the final report on August 12. Over the course of this project, I worked hard and made a few schedule adjustments to ensure that I did not miss any deadlines for these milestones.

Completion History

Week 1	I completed the weekly readings and my initial discussion post for Week 1.
Week 2	I completed the weekly readings, selected my datasets, finished Assignment 1, made my discussion post for Week 2, and replied to posts from other students over the last two weeks (2 responses for each week).
Week 3	I completed the readings for Weeks 3-4 and began working on Assignment 2. I also made my discussion post and 2 responses for Week 3.
Week 4	I completed Assignment 2 and made my discussion post and 2 responses for Week 4.
Week 5	I completed the readings for Week 5 and began working on Presentation 1.
Week 6	I completed Presentation 1 and presented to the group on June 26. I also submitted the recorded presentation on June 28 and made my initial discussion post for Week 6.
Week 7	I completed the weekly readings, finished data preparation and Assignment 3, made my discussion post for Week 7, and replied to posts from other students.
Week 8	I completed Presentation 2 and presented to the class on July 13. Next, I finished the analysis and visualizations, and submitted the Assignment 4 report.
Week 9	I completed the weekly readings, designed the predictive models for Assignment 5, began working on the Assignment 5 report, and made my discussion posts and responses.
Week 10	I completed Assignment 5 and worked on the weekly readings.

Week 11	I completed the weekly discussion posts and responses, and I began working on the Assignment 6 report.
Week 12	I completed the Assignment 6 final report.

Lessons Learned

Week 1	I learned about the importance of using an analytical approach to solve complex problems, as opposed to using a purely instinctive approach. I also learned about the three problem solving approaches (repair, improve, and engineer), the differences between them, and when to use each approach.
Week 2	I re-familiarized myself with the CRISP-DM process and its components. After selecting the storm events datasets, I began studying problems related to storms, such as short warning times, the frequency of false alarms, and the likelihood of certain areas being hit by a tornado. I also learned about some of the approaches that other students used to select their projects and datasets.
Week 3	After receiving feedback from Assignment 1, I learned that my initial project proposal is too ambitious and broad in scope. I need to set goals that are achievable and measurable with the tools given in the course.
Week 4	I explored my project's subject material in greater depth. I learned more about the companies (NOAA, NCEI, NWS), and I developed a deeper understanding about the importance and challenges of severe weather forecasting. I also became more familiar with my three datasets and their properties.
Week 5	Based on the week's readings, I gained a better understanding of how to make an effective presentation and target the right audience.
Week 6	After receiving feedback from Assignment 2 and Presentation 1, I learned how to further improve the scope of my project. For instance, I changed my target variable to focus on deaths instead of both deaths and injuries. This helped my project to be more focused, and it made the "fatalities" dataset more useful for the analysis. I also learned that my presentation needs to focus more on selling the project's importance rather than on technical details (such as variables).
Week 7	I dug deeper into the data quality issues (such as skewness, outliers, etc.). During the data preparation stages, I learned more about each variable, as well as which variables may be more important to the analysis.
Week 8	After creating my visualizations, I gained some insights from the results. For instance, the geospatial map revealed that Nevada had the second highest number of storm casualties behind Texas. After performing some research and creating a visualization to explore the cause of death, I found that nearly all of these deaths were related to heat. This highlights the importance of considering other types of severe weather, rather than focusing only on hurricanes or tornadoes.
Week 9	I refamiliarized myself with the predictive models that I would be using (logistic regression, neural networks, SVM, random forest, and ensemble models). I

	gained initial insights about my models, and I found that most of them can achieve very high accuracy and sensitivity regardless of the skewness of the target variable.
Week 10	I noticed that my models had very high accuracy and sensitivity rates due to correlated variables from the fatalities dataset. After removing these features from the models, their sensitivity rates were much lower and more realistic. Thus, I had to run my models again and re-evaluate the results. To handle the target skewness, I had to select an appropriate cutoff threshold for each model.
Week 11	Based on my lessons learned in Week 10, I realized that the variables from the fatalities dataset were not useful for the predictive models and were only used in the text mining component. I learned that an alternative would have been to use storm data from multiple years, since it would provide a more complete picture of storm casualties over time.
Week 12	One of the lessons I learned from the overall project was the importance of handling correlated variables in the data. I found that all variables in the fatalities dataset were correlated with my target "casualties." This caused the predictive models to have unrealistically high accuracy and sensitivity rates. Thus, I had to remove these variables from the analysis. Additionally, there were several insights from the data that changed the way I view natural disasters. Firstly, I found that the massive death toll of 122 casualties was attributed to heat. This shows the importance of considering different types of severe weather, since heatwaves can be just as deadly as tornadoes or hurricanes (and sometimes even deadlier). Unfortunately, many of these weather phenomena are given less coverage in the media. Secondly, I found a relatively high death toll in January that was largely caused by a large tornado outbreak. Such a finding was surprising due to the rarity of tornadoes during Winter. This finding shows that major storms may not always happen during the times of year when we expect them to, which raises the importance of storm preparation throughout the year. Another lesson from this project was the importance of defining an appropriate scope for your project. The project cannot be too ambitious or broad in scope, and the goals and KPIs must be measurable based on the methods and technologies used in the project. This helped me to outline the problem more effectively, which greatly contributed to the project's overall success.

Data Set Description

Data Set Description

This project uses three datasets that were obtained from NOAA's NWS storm events database ("Storm Events Database," n.d.). The datasets focus on storms that occurred in the U.S. during 2017. The first dataset contains the storm details, and it provides most of the important information for categorizing each storm. It will be the primary dataset for my analysis. It has 51 variables and 56,921 cases, which gives it an appropriate level of complexity for this project. Some of the important variables include the storm's beginning and ending dates, the state in which it occurred, the type of storm (such as hurricane, tornado, or flood), the numbers of direct and indirect deaths and injuries, the amount of property and crop damage, the beginning and ending locations, and text descriptions of the storm event.

My target variable will be a newly-created variable called "casualties." It will utilize both direct and indirect deaths, and its purpose is to indicate whether a storm resulted in any deaths. The reason why I am creating a new target instead of using an existing target (such as the number of direct deaths) is because I want to classify storms based on whether they will produce any casualties at all. I am not trying to predict the number of deaths that a storm will cause. My target will be a binary categorical variable with the values "yes" and "no," where "yes" indicates that at least one death occurred, while "no" indicates that there were no deaths. This variable will be highly imbalanced by default, since there are very few deaths compared to the number of storm events. I will address the imbalance by changing the cutoff threshold to improve the model's ability to identify true positives in the data.

I examined the storm details dataset using R, and I found that numerous variables contain a massive percentage of missing values. For instance, "magnitude" contains over 23,000 missing values, while "category" contains over 56,000 missing values. The reason why such variables have so many missing values is because they are often used to describe a single type of storm (such as hurricanes). To address this issue, I will remove all variables that are missing at least 40% of their records. Additionally, I noticed that the numbers of deaths and injuries are highly skewed. There are only 346 direct injuries and 278 direct deaths in a dataset containing over 56,000 storm events. This skewness also results in these variables having numerous outliers. For instance, there are storms that resulted in 36 deaths and up to 500 injuries, while the average numbers of deaths and injuries are very close to 0. I will address my approach to handling skewness in the

data preparation section of this paper. Furthermore, variables such as "year" are not helpful—since all storms occurred in the same year. Thus, I can remove features such as this. When analyzing the data, I found that the most common types of storms are thunderstorms and hail—each with over 10,000 cases. Also, most storms occurred between March and July, which is not surprising due to the frequency of storms during Spring and Summer.

My second dataset consists of information about storm fatalities in 2017. This dataset contains more detail about the individuals who died in storms last year, making it useful for evaluating the specific factors contributing to storm-related deaths. It is connected to the first dataset through the event ID foreign key. Important variables include the day and time of death, the victim's age, their gender, and the location of death (whether it was in a permanent home, mobile home, outside, or in a vehicle). It is considerably smaller than the first dataset, as it only contains 775 cases and 11 variables. But its small size is most likely because there were relatively few deaths when compared to the number of storms. By examining the data, I found that 490 out of 775 victims are male, and the average age is 49.5. Interestingly, there are only 31 deaths occurring in mobile homes—which is surprising because mobile homes are frequently destroyed by tornadoes and other dangerous storms. Instead, most deaths occur either outside (172 deaths), in vehicles (168 deaths), in the water (165 deaths), or in permanent homes (116 deaths). Age has 104 missing values, while fatality sex has 68 missing values. I will address these missing values in the “data cleansing” section of the paper.

My third dataset includes specific information about the locations and individual episodes of each storm. It is useful because it may show the impact of a storm's range and direction on its severity. It is linked to the other two datasets through the event ID variable, and it shares the episode ID with the storm details data. It has 43,579 cases and 11 variables. Some noteworthy variables include the storm's range, its direction, and the town in which it occurred. It does not contain any missing values. The average storm range is about 2.5 miles, though the maximum range is 138 miles (which is most likely a hurricane). This means that the “range” variable has outliers. Additionally, a significant majority of storms were traveling north compared to any other direction. I am not sure if this has any real importance, but I will find out during my analysis.

These three datasets can easily be joined together since they share the event ID variable as a key. In SQL, the tables can be joined by using a left join statement. To combine the datasets

permanently using R, I can use the "merge" function to combine two data frames at a time. I would use the dataset names and the event ID as input parameters for this function.

High-Level Data Diagram

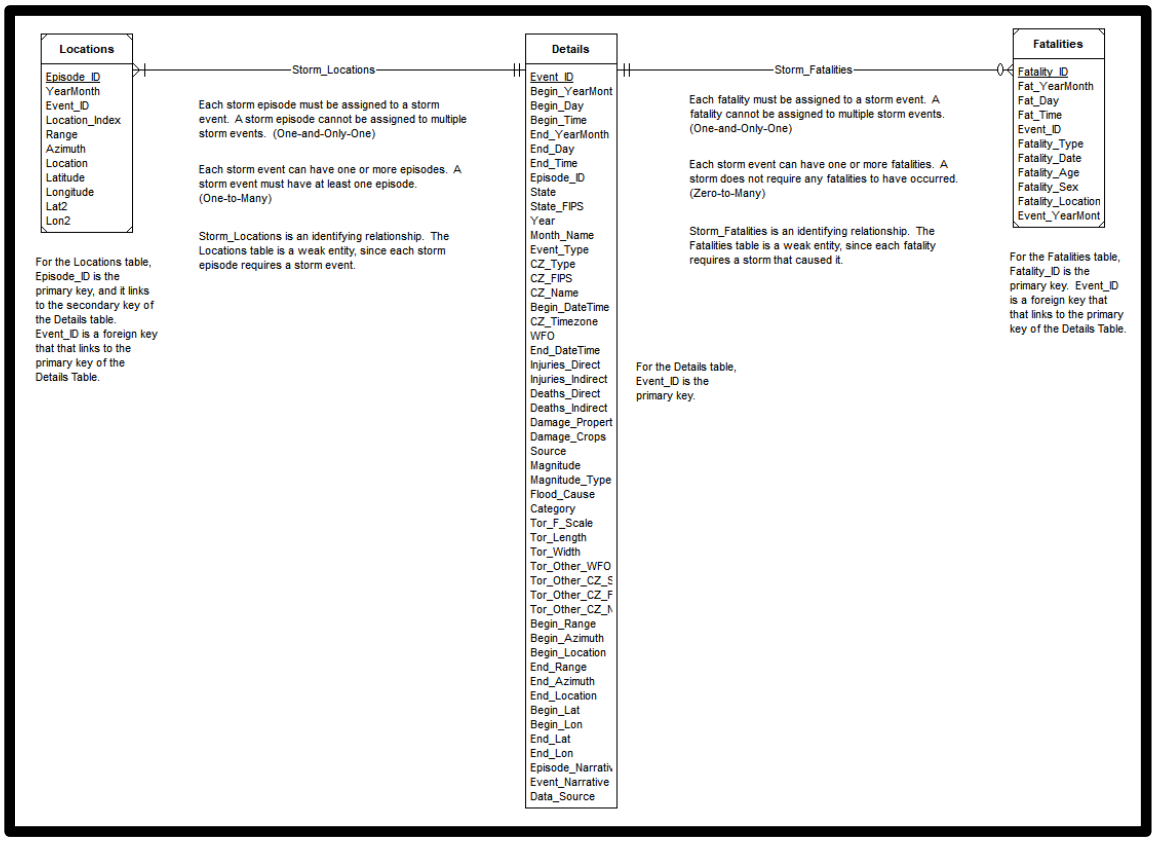

Figure 1. Entity-Relationship Diagram of 2017 Storm Events Data Tables.

This Entity-Relationship Diagram (ERD) shows the relationships between the three datasets in my project (see Figure 1). Based on this figure, we see that the event ID variable is a common key that links all three datasets together. The storm events "details" table contains 51 variables, and its primary key is the event ID. The "fatalities" table contains 11 variables, and its primary key is the fatality ID. It uses the event ID as a foreign key that connects to the "details" table. The "fatalities" table consists of deaths caused by storms, which means that every fatality must be attributed to one storm event. Furthermore, a death cannot be caused by multiple storms. However, any given storm may produce one or more casualties. Likewise, it is possible for a storm to result in no fatalities. Altogether, the "fatalities" table is a weak entity because each fatality requires an associated storm.

The “locations” table also contains 11 variables, and its primary key is the episode ID. The episode ID is also a foreign key within the “details” table, while the event ID is a foreign key in the “locations” table. The locations table is made up of “storm episodes,” which are individual components of a “storm event” (which refers to an entire storm). As such, each storm episode must be part of a storm event, and an episode cannot be assigned to more than one event. However, a storm event may contain one or more episodes—but it must contain at least one episode to exist. Each storm episode within the “locations” table is dependent upon the existence of a storm event, which means that the “locations” table is a weak entity.

Data Definition/Data Profile

I examined all variables within the three datasets using the data description provided by the NWS (“Storm Data Export,” n.d.). This information was used to create three tables providing descriptions of each variable, their data type, and any known data quality issues (see Figures 2, 3, and 4). Additionally, I examined the descriptive statistics of all numeric variables in the datasets, and I created a table showing their minimum, maximum, mean, standard deviation, and skewness (see Figure 5). For the storm events “details” dataset, there are several variables that I would remove. Firstly, the year and data source variables are unnecessary since all of their values are the same. All storms occurred in 2017, and all data comes from a CSV data source. Furthermore, there are several variables that are redundant with each other. The variables “begin year/month” and “month name” both refer to the same month. “Begin year/month” contains the year information, but this is unnecessary since all storms are from 2017. Likewise, “state” and “state FIPS” both refer to the same state, while “CZ name” and “CZ FIPS” refer to the same county or zone. For each pair of redundant variables, I will remove one and keep the other (which is specified in the table below).

As mentioned earlier, the numbers of deaths and injuries are highly skewed due to the small proportion of casualties compared to the number of storm events. This will be addressed by changing the cutoff threshold of the predictive models. I am creating a new target variable “casualties” to determine whether a given storm will produce any deaths. However, I will not remove the numbers of deaths and injuries because they will be useful for visualizations such as the geospatial and time series analysis. Still, I may omit them from some of my predictive models

if they are correlated with my new target variable. Afterwards, I will handle outliers by replacing them with a number that is 3 standard deviations from the mean.

Furthermore, there are numerous variables that are missing much (or most) of their data. Some variables—such as magnitude type and tornado length—are missing between 30,000 and 56,000 values. However, other variables such as beginning range and ending location are only missing about 17,000 values. In fact, those 17,000 rows are missing all records on their beginning and ending locations, range, direction, latitude, and longitude. I will describe my approach to handling missing values in the data cleansing section of this paper. Another data quality issue is that the "category" variable does not have any known definition according to the NWS ("Storm Data Export," n.d.). Since it is missing almost all of its values, I will remove it.

The "fatalities" and "locations" datasets contain fewer data quality issues. For the "fatalities" dataset, the "fatality time" variable can be removed since all recorded values are 0. Also, "fatality age" and "fatality sex" contain 104 and 68 missing values respectively—which is relatively small in proportion to the dataset size. Thus, I can simply compute a replacement value for these missing values. The "locations" dataset does not have any missing values, but the "range" variable contains outliers. It contains ranges as high as 138 miles, which is many times higher than the average of 2.5 miles. As mentioned, I will handle outliers by replacing them with values that are 3 standard deviations from the mean.

Variables in Storm Events Details ("Storm Data Export," n.d.):

Name	Definition	Data Type	Quality Issues
Begin Year/Month	The beginning year and month of the storm event.	Integer	Not in proper date format, redundant with "Month Name." Will remove this variable.
Begin Day	The day (number) of the month that the storm event began.	Integer	None
Begin Time	The starting time of the storm.	Integer	Not in proper time format.
End Year/Month	The ending year and month of the storm event.	Integer	Not in proper date format.

End Day	The day of the month that the storm ended.	Integer	None
End Time	The ending time of the storm.	Integer	Not in proper time format.
Episode ID	A key that refers to a certain episode (component) of a storm event. Links the "details" and "locations" tables together.	Integer	None
Event ID	The primary key of the dataset, the ID assigned to each storm event. Links all three datasets together.	Integer	None
State	The state where the storm occurred.	Character	Redundant with "State FIPS." Will keep this variable.
State FIPS	A unique number assigned to the state.	Integer	Redundant with "State." Will remove this variable.
Year	The year that the storm occurred.	Integer	All storms are from the same year (2017). Will remove this variable.
Month Name	The name of the month in which the storm occurred.	Character	Redundant with "Begin Year/Month." Will keep this variable.
Event Type	The type of storm or weather event.	Character	None
CZ Type	Indicates whether storm occurred in a county, zone, or marine area.	Character	None
CZ FIPS	A unique number assigned to the county, zone, or marine area.	Integer	Redundant with "CZ Name." Will keep this variable.
CZ Name	The name of the county, zone, or marine area.	Character	Redundant with "CZ FIPS." Will remove this variable.
WFO	The NWS forecast office that is responsible for this area.	Character	None

Begin Date/Time	The starting date and time of the storm event.	Date	Redundant with "Begin Year/Month," "Begin Day," and "Begin Time." Might remove this variable.
CZ Timezone	The time zone of that location.	Character	None
End Date/Time	The ending date and time of the storm event.	Date	None
Injuries (Direct)	The number of injuries directly caused by the storm.	Integer	Highly skewed, very small proportion of injuries. Contains outliers (56 injuries).
Injuries (Indirect)	The number of injuries indirectly caused by the storm.	Integer	Highly skewed, very small proportion of injuries. Contains outliers (500 injuries).
Deaths (Direct)	The number of deaths directly caused by the storm.	Integer	Highly skewed, very small proportion of deaths. Contains outliers (36 deaths).
Deaths (Indirect)	The number of deaths indirectly caused by the storm.	Integer	Highly skewed, very small proportion of deaths. Contains outliers (20 deaths).
Damage (Property)	The estimated amount of property damage in U.S. dollars.	Character	Not in numeric format (ex: 650.00K, 1.00M, etc.). Contains 10,589 missing values.
Damage (Crops)	The estimated amount of crop damage in U.S. dollars.	Character	Not in numeric format (ex: 650.00K, 1.00M, etc.). Contains 10,720 missing values.
Source	The type of source that reported the weather event.	Character	None
Magnitude	The magnitude of storms based on their wind speeds and hail size.	Numeric	Contains 23,228 missing values.

Magnitude Type	The type of wind measured.	Character	Contains 33,666 missing values.
Flood Cause	The actual or estimated cause of the flood.	Character	Contains 50,604 missing values (only used for floods).
Category	Unknown.	Integer	Contains 56,876 missing values. Will remove this variable.
Tornado F-Scale	The Enhanced Fujita (EF) scale used to describe a tornado's strength.	Character	Contains 55,277 missing values (only used for tornadoes).
Tornado Length	The length of the tornado while it is on the ground.	Numeric	Contains 55,277 missing values (only used for tornadoes).
Tornado Width	The width of the tornado (in feet) while it is on the ground.	Integer	Contains 55,277 missing values (only used for tornadoes).
Tornado Other WFO	The NWS forecast office that is responsible when a tornado moves from one forecast office region to another.	Character	Contains 56,702 missing values (only used for tornadoes).
Tornado Other CZ State	The state abbreviation for the state that the tornado crosses into.	Character	Contains 56,702 missing values (only used for tornadoes).
Tornado Other CZ FIPS	The county number for the county or zone that the tornado crosses into.	Integer	Contains 56,702 missing values (only used for tornadoes).
Tornado Other CZ Name	The name of the county or zone that the tornado crosses into.	Character	Contains 56,702 missing values (only used for tornadoes).
Begin Range	The storm's starting range, to the nearest 1/10 mile.	Integer	Contains 17,566 missing values.
Begin Azimuth	The storm's initial direction (N, NW, SE, etc.).	Character	Contains 17,566 missing values.
Begin Location	The storm's starting location.	Character	Contains 17,566 missing values.

End Range	The storm's ending range, to the nearest 1/10 mile.	Integer	Contains 17,566 missing values.
End Azimuth	The storm's final direction (N, NW, SE, etc.).	Character	Contains 17,566 missing values.
End Location	The storm's ending location.	Character	Contains 17,566 missing values.
Begin Latitude	The storm's starting latitude.	Numeric	Contains 17,566 missing values.
Begin Longitude	The storm's starting longitude.	Numeric	Contains 17,566 missing values.
End Latitude	The storm's ending latitude.	Numeric	Contains 17,566 missing values.
End Longitude	The storm's ending longitude.	Numeric	Contains 17,566 missing values.
Episode Narrative	A text description of the storm episode (part of the event).	Character	None
Event Narrative	A text description of the entire storm event.	Character	Contains 12,555 missing values.
Data Source	The type of data file.	Character	All cases are from CSV file. Will remove this variable.

Figure 2. Description of Variables in Storm Details Dataset.

Variables in Storm Events Fatalities ("Storm Data Export," n.d.):

Name	**Definition**	**Data Type**	**Quality Issues**
Fatality Year/Month	The year and month of the fatality.	Integer	Not in proper date format.
Fatality Day	The day (number) of the month that the fatality occurred.	Integer	None
Fatality Time	The time that the fatality occurred.	Integer	All values are 0. Will remove this variable.

Fatality ID	The primary key of this dataset, an ID assigned to each casualty.	Integer	None
Event ID	The ID assigned to each storm event. Links all three datasets together.	Integer	None
Fatality Type	Whether the death was caused directly or indirectly by the storm.	Character	None
Fatality Date	The date that the fatality occurred.	Date	None
Fatality Age	The victim's age in years.	Integer	Contains 104 missing values.
Fatality Sex	The victim's gender.	Character	Contains 68 missing values.
Fatality Location	The type of location that the fatality occurred (such as permanent home, mobile home, vehicle, outside, etc.).	Character	None
Event Year/Month	The year and month of the storm event.	Integer	Not in proper date format.

Figure 3. Description of Variables in Fatalities Dataset.

Variables in Storm Events Locations ("Storm Data Export," n.d.):

Name	**Definition**	**Data Type**	**Quality Issues**
Year/Month	The year and month of the storm event.	Integer	Not in proper date format.
Episode ID	The primary key of this dataset, refers to a certain episode (component) of a storm event. Links the "details" and "locations" tables together.	Integer	None
Event ID	The ID assigned to each storm event. Links all three datasets together.	Integer	None

Location Index	A number used by the NWS to certain locations within the same storm event.	Integer	None
Range	The storm's range at the time of recording, to the nearest 1/10 mile.	Numeric	Contains outliers (range of 138 miles).
Azimuth	The storm's direction at the time of recording (N, NW, SE, etc.).	Character	None
Location	The storm's location at the time of recording.	Character	None
Latitude	The storm's latitude coordinate.	Numeric	None
Longitude	The storm's longitude coordinate.	Numeric	None
Latitude 2	Unknown.	Integer	None
Longitude 2	Unknown.	Integer	None

Figure 4. Description of Variables in Locations Dataset.

Descriptive Statistics of Important Numeric Variables:

Name	Minimum	Maximum	Mean	St. Dev	Skewness
Begin Range	0	136	2.347	4.371264	10.21334
Deaths (Direct)	0	36	0.00919	0.2691112	93.97529
Deaths (Indirect)	0	20	0.004093	0.1332104	82.26315
End Range	0	136	2.382	4.461389	10.28824
Fatality Age	0	95	49.56	23.03584	-0.2923497
Injuries (Direct)	0	56	0.0212	0.5834658	57.43582
Injuries (Indirect)	0	500	0.0133	2.112789	233.1148
Magnitude	0.25	173	36.08	24.47783	-0.5245527

Range	0	138.070	2.534	4.705244	9.500566
Tornado Length	0.01	41.88	3.17	4.088775	2.865184
Tornado Width	1	2464	214.2	303.7691	3.159766

Figure 5. Descriptive Statistics of Numerical Variables from all Datasets.

Data Preparation/Cleansing/Transformation

Data Preparation

I will use R to perform most of the data preparation and cleansing steps. This is because R provides a strong framework for performing tasks such as feature creation and handling missing values. However, some data preparation steps will be done using SAS Enterprise Miner. This is because SAS Enterprise Miner provides an easy-to-use interface that allows for fast computations on multiple models. This makes it effective for handling outliers, transforming variables, and partitioning the data. The first step is to import the three datasets into R and initialize the packages used in the project. These packages include "tm," "SnowballC," and "wordcloud," which will be used in the text mining component of my analysis. After initializing the data, I will perform data cleansing steps such as handling missing values, outliers, and correlated variables. These steps will be described in greater detail in the data cleansing section. I will also create new features for this analysis, which will be described in the data transformation section.

After cleaning the data, I will combine the three datasets in R. This will be done using the "merge" command, which merges two data frames into a single data frame. As mentioned, the event ID variable links all three datasets together. Thus, it will be used as a common key to combine the datasets. But in addition, I will use the condition "all.x = TRUE," which functions as a left outer join statement (De Vries & Meys, n.d.). This is useful because it allows me to keep all rows from the storm details dataset (the left data frame), while only keeping the matching rows from the other datasets (the right data frame).

After the datasets have been merged, I will use SAS Enterprise Miner to import the merged dataset and prepare the data for model implementation. Since the data is in .csv format, it needs to be converted into a .sas7bdat file to be compatible with SAS Enterprise Miner. Thus, I imported the file using the "file import" node and converted it using the "save data" node. Next, I used the "replacement" node to handle outliers in the data. I will describe my technique for handling outliers in the data cleansing section of this paper. I will also use SAS Enterprise Miner to explore the important variables and identify possible correlations between the inputs and the target variable. This will be described further in the variable exploration section. After this step, I will partition the data into training, validation, and test sets. I will allocate 50% of data to the training set, 30% to the validation set, and 20% to the test set. I have separate data partition nodes for each model to prevent the models from using identical data samples. The only model

that does not involve data partitioning is the random forest model. This is because random forests and other ensemble models only need to be evaluated on the training set (Knode, 2016a).

Next, I will use SAS Enterprise Miner to handle variable skewness before building the predictive models. As mentioned, I intend to change the cutoff threshold to handle the skewness of my target variable "casualties"—which is imbalanced due to the small proportion of deaths in the dataset. Although the storm details dataset contains almost 57,000 rows, there are only 387 rows with either direct or indirect deaths. Based on these numbers, the ideal cutoff threshold to maximize sensitivity is 0.0068. However, this might drastically reduce the overall model accuracy since it may classify too many storms as having casualties. Thus, I select an optimal cutoff threshold that maximizes both the model's sensitivity and its accuracy.

The final component of my analysis will consist of a bag-of-tokens text mining analysis on the words that are frequently used to describe deadly storms. It will involve analyzing words in the "event narrative" variable, which contains text descriptions of each storm event. The text mining analysis and preprocessing will be done using R. This is because R provides a strong framework for cleaning texts, analyzing word frequencies, and creating visualizations such as word clouds. First, I will create a corpus that only contains the event narratives for storms with casualties. I will then clean the texts by removing numbers, punctuation, special characters, and extra whitespace. I will set all letters to lowercase, remove stop words and other trivial terms (such as "weather"), and perform stemming to reduce terms to their root word. Lastly, I will build the document-term matrix (DTM) and remove sparse terms from the DTM.

Data Cleansing

As mentioned earlier, R provides an effective foundation for cleaning the data. As such, I will use R to perform most of the data cleansing tasks. The first step is to handle missing values in the data. The storm details dataset—which has almost 57,000 cases—contains numerous variables that are missing vast amounts of data. For instance, property damage and crop damage are missing about 10,000 cases each. Furthermore, the beginning and ending ranges and locations are missing over 17,000 cases, while the tornado length and F-scale are missing over 55,000 cases. The reason why some of these variables are missing so many cases is because they only refer to a specific type of storm. For instance, tornado length is only used to describe tornadoes, while flood cause is only applicable to floods.

To handle missing values in the storm details dataset, I will use the following approach. If the variable is missing more than 20,000 cases, then I will remove it from the analysis. If it is missing fewer than 20,000 cases, then I will handle them depending on whether the variable is numeric or categorical. If it is categorical, I will remove the rows with missing values. This is because the categorical variables with missing values are the beginning and ending location and azimuth (direction). And by examining the dataset, I found that these variables have missing values within the same 17,566 rows. Thus, it is reasonable to remove these rows from the data. If the variable with missing values is numeric, I will simply replace the missing values with the mean. The exceptions to this approach are the beginning and ending range, latitude, and longitude, which are numeric. The 17,000 rows that have missing values for location and azimuth are also missing data on range, latitude, and longitude. Thus, it makes more sense to remove these rows than to compute replacement values for these variables.

In the fatalities dataset, there are two variables with missing values. The victim's age is missing 104 values, while the victim's sex is missing 68 values. The fatalities dataset contains 775 cases, which means that these variables are missing a relatively small proportion of data. Thus, I will handle these missing values using the following approach. Since the victim's age is numeric, I will replace its missing values with the mean. And since the victim's sex is categorical, I will remove the rows with missing values. One alternative is to replace the missing gender value with the mode, which is male. But this will result in too many male victims, which can lead to misinterpretation of the relationship between storm casualties and gender. Lastly, the locations dataset does not contain any variables with missing values.

The next step is to handle outliers in the data. The storm details dataset contains several variables with outliers, such as direct and indirect deaths and injuries, property damage, and crop damage. For instance, one storm caused 500 indirect injuries—even though the vast majority of storms produced 0 injuries. The fatalities dataset does not have any variables with outliers. But in the locations dataset, storm range has outliers. The average storm has a range of about 2.5, but there are storms with ranges as high as 138. To handle outliers, I will compute a replacement value that is 3 standard deviations from the mean. This will be done using SAS Enterprise Miner, since it allows for fast computations and can be used on multiple variables at a time.

Additionally, there are several pairs of variables that are redundant or correlated. For instance, the state name and state FIPS (i.e. state number) are both referring to the same state.

Likewise, the CZ name and CZ FIPS are referencing the same county, zone, or region. To address this issue, I will remove one of the correlated variables and keep the other. Furthermore, there are a few variables that have the same value for all cases. These include year and data source, since all storms occurred in 2017 and all cases are recorded in a .CSV file. Also, "fatality time" has the value 0 for all of its cases. I will use R to remove unnecessary variables from the analysis. Finally, the event ID and episode ID variables are unique identifiers that add little to the analysis. However, the event ID is required to join all three datasets together. Therefore, I will remove these variables only after merging my three datasets in R.

Data Transformation

There are at least two features that I will create in my analysis. First, I will create my target variable "casualties," which states whether a storm produced any direct or indirect deaths. As mentioned, my goal is not to predict the number of casualties from a given storm. Instead, I want to determine whether a storm will result in any casualties at all. Therefore, this feature is extremely important because it allows us to easily identify all lethal storms in the data—regardless of whether the death was direct or indirect. This variable is a binary categorical variable with the values "yes" and "no," and it will be created using R. It will be assigned "yes" if either the number of direct deaths or indirect deaths are greater than 0. Otherwise it will be assigned "no," which indicates that the storm resulted in 0 deaths. As mentioned earlier, this variable will be highly skewed due to the small proportion of deaths in the data. This skewness will be addressed by changing the cutoff threshold to improve the model's sensitivity without causing the accuracy to decrease too significantly.

I will also use R to create another feature called "injuries." This variable is very similar to my target variable, but the difference is that it describes whether a storm resulted in any injuries (direct or indirect). It is important for the project because it allows us to easily identify whether a storm will produce any injuries at all. This makes it useful for comparison with the target variable casualties. Although I am expecting that there is a strong relationship between deaths and injuries, I would not be surprised if certain storms resulted in many injuries but no deaths. This feature is also a binary categorical variable with the values "yes" and "no." It will be assigned "yes" if any direct or indirect injuries occurred. Otherwise, it will be assigned "no." This feature

will also be highly skewed at first, which is due to the small number of injuries in the dataset. Its skewness will be addressed by changing the cutoff threshold in SAS Enterprise Miner.

After creating these new features, I will have to address the skewness of other variables in the data. Based on the descriptive statistics for numeric variables (see Figure 5), we see that the most highly skewed variables are the numbers of direct and indirect deaths and injuries. This will be resolved once I change the cutoff threshold for the predictive models. However, we also see that the beginning and ending ranges are skewed as well—with skewness values of around 10.2. As such, it might be necessary to apply transformations before building the predictive models. The variable transformations can be performed using SAS Enterprise Miner. This is because it features a "transform variables" node for implementing several types of transformations on the data. Some of the most useful transformations include the "best" and "maximum normal" methods. The "best" transform selects the transformation type with the strongest impact on the target variable, while the "max normal" method uses the transformation that maximizes the normality of each input. However, the issue of applying variable transformations is that it causes the newly-created features to be very complex and difficult to explain. Thus, I will avoid using transformations unless it is necessary for the model.

Variable Exploration

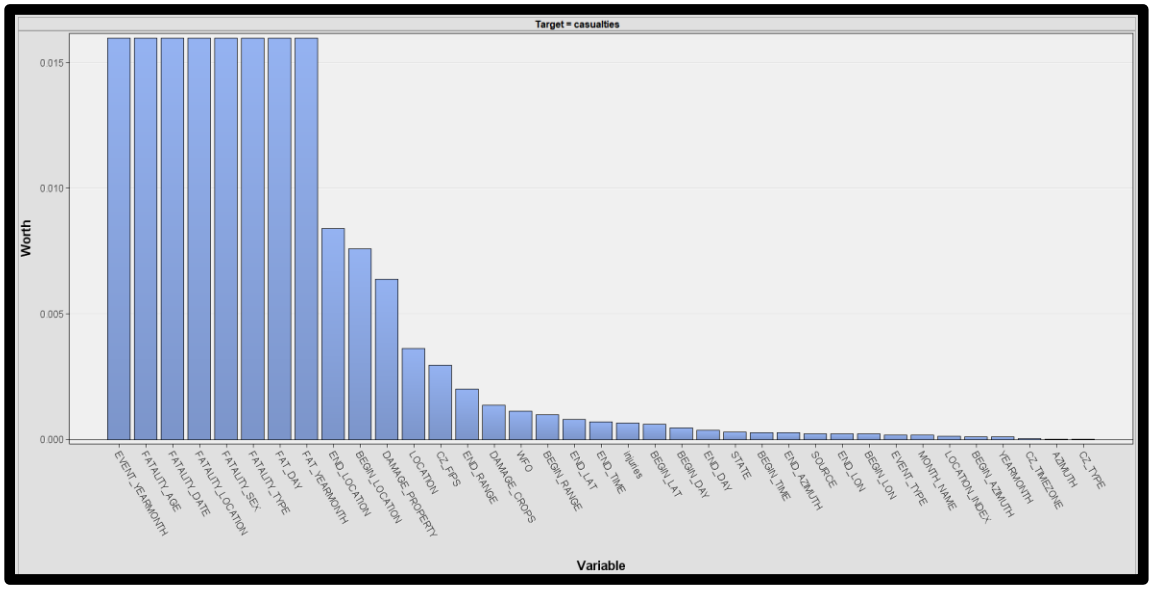

Figure 6. Plot of Input Variable Worth.

Before implementing my predictive models in SAS Enterprise Miner, I will explore the variables using the "StatExplore" node to evaluate the most important features (see Figure 6). This is because one of my goals is to identify the five features which contribute the most towards storm casualties. Based on this figure, we see that the variables with the highest worth include the year and month of the event, the date and location of the death, and the age and gender of the victim. One observation is that the most important features all come from the fatalities dataset, which consists of confirmed death cases. Since all of these variables refer to deaths, it is clear that they are correlated to the target variable "casualties." This would result in the models having abnormally high accuracy and sensitivity rates, since any row with a value in these variables is guaranteed to have casualties. I tested this theory on my models using SAS Enterprise Miner, and I found that nearly every model had an accuracy and sensitivity of around 100% (even without changing the cutoff threshold). This confirms that these variables are correlated to my target, so I removed them from the models.

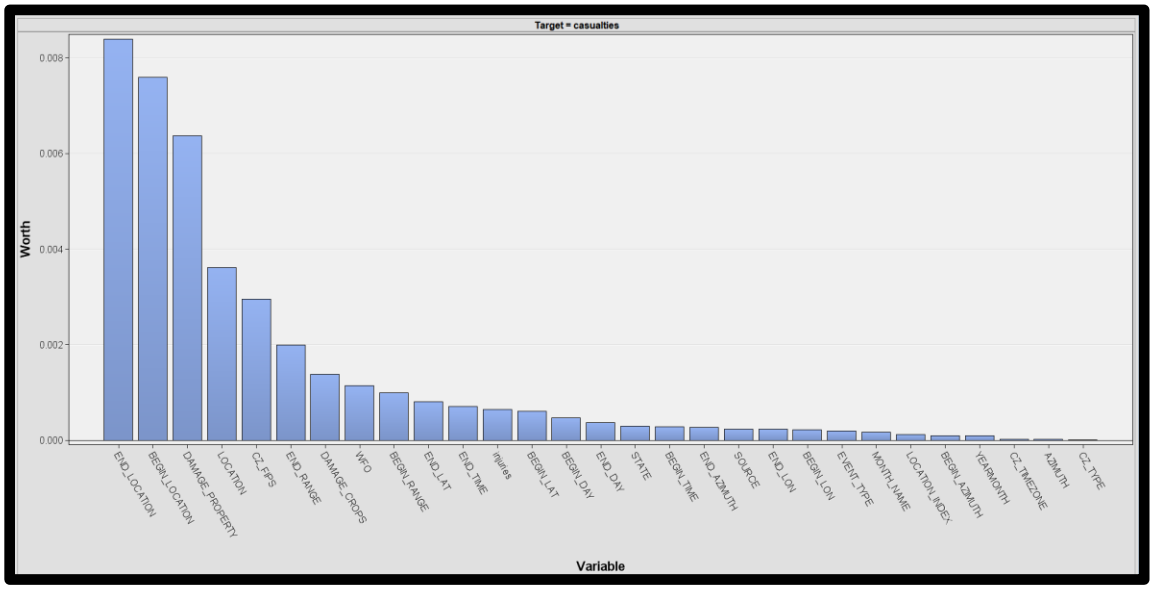

Figure 7. Plot of Input Variable Worth After Removing Correlated Inputs.

After removing the correlated variables, I once again examined the plot of variable importance (see Figure 7). Based on this figure, we see that several of the most important variables are referring to location. This suggests that geographical location plays a significant role in the formation of deadly storms. Additionally, we see that property damage is the third most significant variable for identifying storm casualties. This makes sense, since it implies that the most destructive storms are more likely to result in deaths. Furthermore, the storm's

range has a relatively high ranking in this list. This could mean that storms with larger ranges will impact more locations, which increases the chance of casualties. Some of the least important variables include the storm's direction (azimuth) and the month that it occurs. The low importance of direction is understandable, but I am somewhat surprised that the month is less important. I would assume that certain times of the year (such as late Spring and Summer) would have a higher risk of dangerous storms. However, this may suggest that deadly storms can occur during any time of the year.

Data Analysis

With the data preparation stages complete, I will now describe my plans for the analysis portion of the project. With regards to visualizations, my geospatial and time series graphs will be created using Tableau. This is because Tableau provides a very strong framework for generating many types of insightful visualizations. My geospatial map will show the number of storm casualties that occurred in each state in the U.S. during 2017. It is helpful for determining which parts of the country are at the highest risk of having life-threatening storms. Tableau is useful for this task because it can identify location-specific information from the data, and it can create maps that accurately match the data entries to the appropriate state, country, or city. My time series graph will show all deadly storms that occurred in 2017, along with the date that each storm occurred. It is important for determining the months of the year that have the highest frequency of dangerous storms. Tableau provides powerful tools for a time series analysis, including a forecasting feature which allows us to make future predictions about the data within a given range of error. Due to its capabilities, Tableau is an ideal tool for generating these graphs.

Tableau will not be the only tool used for visualizations, as I will also use R to create graphs that are relevant to the text mining component. As mentioned earlier, my text mining analysis will focus on analyzing keywords that are used in the storm event narratives for all storms that resulted in casualties. R is an ideal tool for this section because it includes many powerful text mining packages such as "tm," "SnowballC," and "wordcloud." My text mining visualization will be a word cloud that shows the most frequent terms which appear in the storm descriptions. By finding words that are commonly used to describe dangerous storms, we can identify early warning signs for future storms.

After creating these visualizations, I will focus on building the predictive models. My goal is to create at least 5 models and tune their parameters to obtain at least one model with an accuracy of at least 80%. Additionally, I want one of my models to attain a sensitivity of 70%. I will implement these predictive models using SAS Enterprise Miner because its interface is easy to use and allows users to quickly build multiple models. It also allows for easy adjustment of model parameters, and it allows model results to be compared using the "model comparison" node. Since I am building multiple predictive models, SAS Enterprise Miner will make the analysis easier and less time-consuming to implement.

I will implement a variety of classification models to see which type of model best fits the data. These models will include a logistic regression model, neural network, support vector machine (SVM), and a decision tree model (such as bagging or random forest). Furthermore, I will also include an ensemble model that combines the results of the other four models. These models can easily be implemented in SAS Enterprise Miner using nodes such as "regression," "SVM," "neural network," "HP forest," and "ensemble." From here, the parameters of each model can be tuned to obtain the highest possible accuracy and sensitivity. Building these models would be very time-consuming if using a program such as R or Python. This is because the models will have to be created individually through typing the appropriate code. But due to SAS Enterprise Miner's streamlined interface and minimal coding, it is ideal for implementing this task.

Data Visualization

Data Visualization 1: Geospatial Analysis

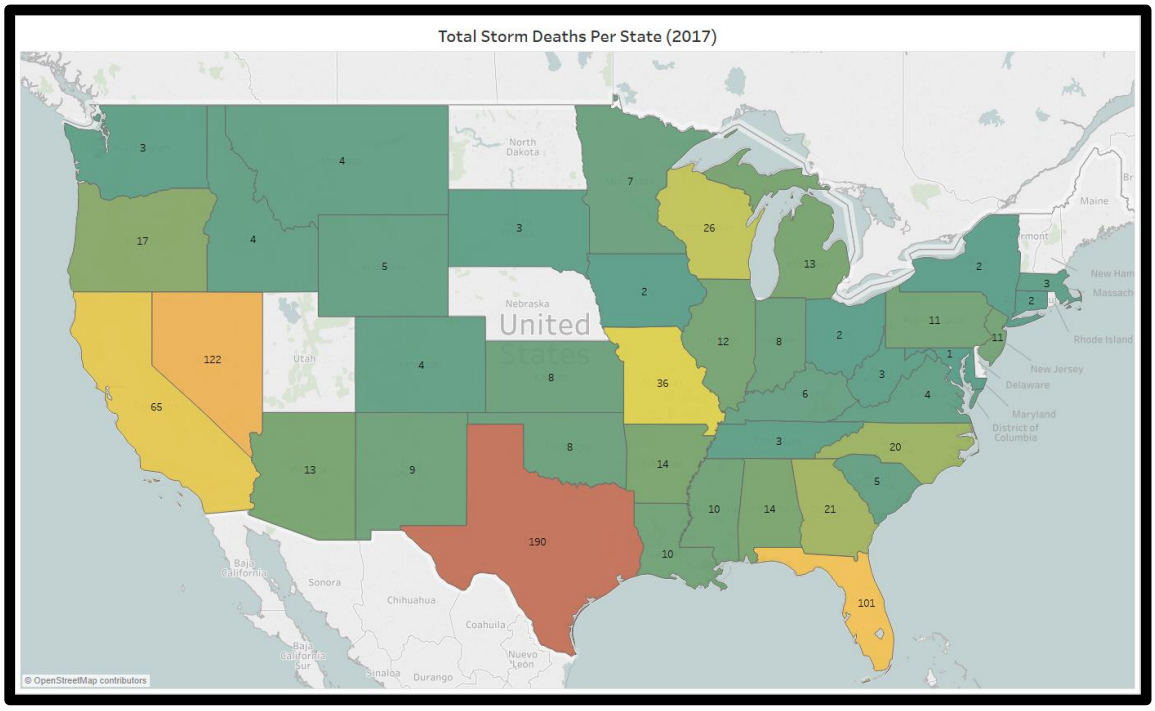

Figure 8. Map of the Total U.S. Storm Casualties in Each State During 2017.

My analysis on storm casualties involves three main visualizations. The first visualization is a geospatial map of the U.S. which shows the total number of storm-related deaths in each state during 2017 (see Figure 8). It combines the number of direct and indirect deaths into the total number of deaths for each state. This visualization is important because location plays a major role in the occurrence of severe weather. Certain storms are more likely to occur in specific parts of the country. For instance, hurricanes and tropical storms are most likely to threaten states along the Gulf of Mexico and Atlantic coast. Studying the storm casualties per state can provide insights on the parts of the country with the highest risk of having dangerous storms. Although storm locations have been studied thoroughly in the past, I expect that my analysis can yield new insights or shed light on information that may have been overlooked.

This visualization shows a map of the continental U.S. with states colored according to their death count (see Figure 8). I used a color scheme such that states with few casualties are in green while states with the most casualties are in red. State colors will transition from green to yellow to orange as the number of casualties increases. This color scheme was chosen because it

draws attention to the states with higher numbers of deaths. It is very easy for the reader to identify the states that suffered the most from deadly storms during 2017. In addition to the colors, the states are labeled by their total numbers of storm casualties. This allows the reader to see the exact number of people who died in each state. However, states with no casualties were excluded from the labeling and coloring scheme. This is done to provide a distinction between states with no casualties and states with at least one casualty.

This visualization was built using Tableau. It was created by importing the storm details and locations datasets and combining them using a left join. The fatalities dataset was not included because it is not needed during this part of the analysis. I used the original versions of the datasets prior to preprocessing. This is because many of the records were removed during the data cleansing process, and the cleaned versions of the data are mainly needed for the predictive models and text mining analysis. From here, I created a new variable "total deaths" which adds the number of direct and indirect deaths for each record. The reason why I use this variable is because it provides a more complete picture of the storm-related deaths and eliminates the need to make separate maps for direct and indirect deaths. After making the "total deaths" variable, I created the map using the state names. Next, I added a color scheme based on the sum of total deaths, and I customized the color scheme to obtain the map shown above (see Figure 8). I then added the sum of total deaths as a label to show the number of deaths in each state. Finally, I filtered out states with no casualties at all during 2017.

Based on the visualization above, there are several important observations that I made. Firstly, we see that Texas has by far the highest number of casualties with about 190 direct and indirect deaths combined. This is not surprising due to Hurricane Harvey, which claimed at least 68 lives ("Hurricane Statistics," 2018). Likewise, Florida experienced 101 storm-related casualties last year—which makes sense due to Hurricane Irma which resulted in at least 44 deaths ("Hurricane Statistics," 2018). What is surprising is the relatively low numbers of storm casualties in other states in the Great Plains, which is commonly known as "Tornado Alley" (Erdman, 2016). Aside from Texas and Missouri, there are relatively few states with a high number of storm casualties. For instance, Kansas and Oklahoma experienced only 8 storm casualties during 2017, while South Dakota had only 3 deaths and Iowa had only 2 deaths. By looking at other tornado-prone states in the Midwest and South (aside from Missouri and Florida), we find that only Wisconsin, Georgia, and North Carolina had more than 20 storm casualties. However, there is a

simple explanation for the relatively low numbers of casualties in these states. Although these areas are more likely to experience tornadoes, it does not mean that there will be large death tolls in these states every year. It is likely that some of these states will have more storm casualties next year, while other states in this region will have fewer casualties.

One of the most surprising findings from this map is the massive number of storm casualties from Nevada. 122 people were killed in Nevada during 2017, making it the state with the second highest casualty rate behind Texas. To understand why this is the case, I created a simple pie chart in Tableau which shows the number of deaths per storm type for the three states with over 100 casualties—Florida, Nevada, and Texas (see Figure 9). This visualization is useful because it allows us to compare the deadly storms in Nevada to the other states which suffered from massive numbers of storm casualties. Based on this graph, we see that Nevada differs drastically from the other two states since nearly all of its casualties were caused by severe heat. Of the 122 casualties from Nevada, 63 people were killed by "heat" and 52 were killed by "excessive heat." Only 7 deaths were caused by weather events unrelated to heat. This finding reveals a major phenomenon that I did not consider initially. According to KTNV and the NWS, an average of 131 people are killed by heat in the U.S. each year—a number that is twice as high as hurricane casualties and higher than tornado and lightning deaths combined ("Nevada has highest rate," 2017). And according to this source, Nevada has experienced many more heat-related deaths than other states in the country. For instance, Nevada saw 50 heat-related deaths in 2016, while neighboring Nevada only had 9 deaths caused by heat ("Nevada has highest rate," 2017).

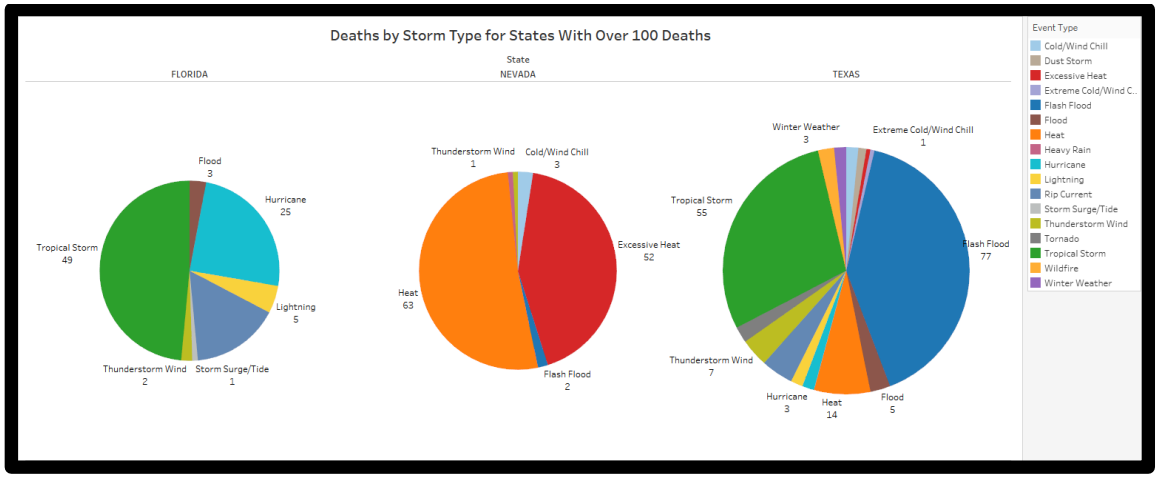

Figure 9. Breakdown of Deaths by Storm Type for States with Over 100 Deaths.

The pie chart in Figure 9 is also useful for exploring other components of my geospatial map. This is because it helps to provide more information about the types of storms that caused so many deaths in Florida and Texas. When breaking down the 101 casualties from Florida, we see that 49 people were killed by a tropical storm, 25 were killed by a hurricane, and 16 were killed by rip currents (see Figure 9). The higher number of tropical storm deaths compared to hurricane deaths may imply that more people were killed by Hurricane Irma after it weakened into a tropical storm—which happened while it was moving north from Florida into Georgia (NOAA, 2017). It could also mean that some of these deaths may have been caused by other tropical storms aside from Irma. Also, some of the deaths from rip currents might be attributed to Hurricane Irma as well, though certainly not all of them. Rip currents are powerful currents that move away from the shore, and they have caused many swimmers to drown by pulling them away from the coast ("Rip Currents," n.d.). Deadly rip currents have been reported in Florida during Hurricane Irma (Gillis, 2017), so it is likely that they may have resulted in some deaths.

When analyzing the deadly types of storms in Texas (see Figure 9), we see that the vast majority of casualties result from either flash floods (77 deaths) or tropical storms (55 deaths). The catastrophic flash flooding deaths were most likely caused by Hurricane Harvey, which produced historic flooding in the state. It is estimated that around 81 percent of the storm's victims in Texas drowned due to the extensive flooding (Wright, 2018). And much like Florida, there are far fewer deaths in Texas from hurricanes than from tropical storms. In fact, only 3 deaths are attributed to hurricane conditions, while 55 were caused by tropical storm weather. As was the case with Irma, Hurricane Harvey weakened into a tropical storm once it hit southern Texas ("Historic Hurricane," 2017). It was as a tropical storm—not a hurricane—that Harvey produced most of its flooding and caused most of the casualties and destruction ("Historic Hurricane," 2017). But one surprising finding from the pie chart is that only 4 people in Texas were killed by tornadoes during 2017. This is surprising because Texas experiences more tornadoes than any other state—with roughly 132 tornadoes occurring in the state every year ("Texas Tornado Facts," n.d.). However, just because the state is prone to tornadoes does not mean that tornado casualties will be high each year.

Based on these results, there are several findings that alter the scope of my analysis. For instance, I found that storm casualties were relatively low in most states in the Great Plains, Midwest, and South. This is surprising because these regions are frequently hit by tornadoes each

year. One of the only states in this region with a massive death toll was Texas—but most of these casualties are due to flash flooding and tropical storm conditions from Hurricane Harvey. And although Texas experiences more tornadoes than any other state, there were only 4 people killed by tornadoes in Texas last year. Thus, while such states may be more vulnerable to tornadoes, it does not mean that casualties will be consistently high each year. This highlights the unpredictability of severe weather, which makes it extremely challenging to accurately predict storm casualties on a yearly basis. Since the number of deaths can fluctuate drastically over the years, I would recommend expanding the timeline of the data in a future analysis. Instead of focusing only on a single year, I would suggest analyzing storm data at least over a 10-year period. This will provide more consistency to the overall results—making it easier to identify the regions of the country with the highest risk over an extended time frame.

Another important finding is the massive death toll in Nevada during 2017. 122 people were killed by severe weather in Nevada last year, which is the second highest death toll of any state. This is surprising because Nevada is far outside of the geographic range where I expected to see the highest numbers of storm casualties. I predicted that most storm deaths would occur in the Great Plains, Midwest, South, and Gulf of Mexico areas. This is because these regions are more likely to experience destructive storms such as tornadoes or hurricanes. But by breaking down the Nevada casualties using pie charts, I found that 63 people died from "heat" and 52 died from "excessive heat." This finding affects the scope of my analysis because it indicates that tornadoes and hurricanes are not always the deadliest forms of severe weather. According to KTNV and the NWS, there are far more casualties caused by heat each year compared to the number of deaths from tornadoes or hurricanes ("Nevada has highest rate," 2017). As such, this is a major insight that is overlooked not only by my analysis, but by the public as well. Most media outlets focus on covering destructive storms such as hurricanes, while giving less attention to heat-related deaths ("Nevada has highest rate," 2017). This shows the importance of considering other types of severe weather in the analysis, rather than only hurricanes and tornadoes.

Data Visualization 2: Time Series Graph

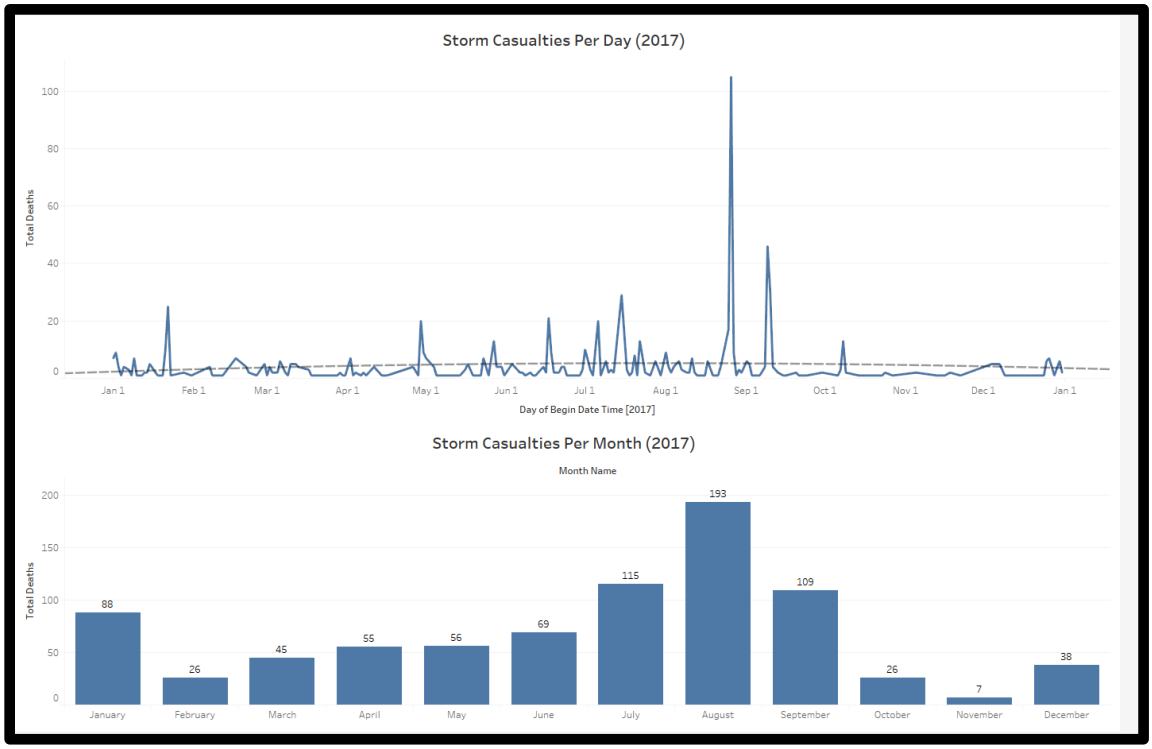

Figure 10. Daily and Monthly Storm Casualties During 2017.

My second visualization is a time series graph that shows the storm casualties throughout 2017 on a daily and monthly basis (see Figure 10). The visualization is divided into two separate graphs—the top graph shows the number of deaths during each day of the year, while the bottom graph is a bar graph showing the number of storm deaths that occurred each month. Like my geospatial graph, this visualization also combines the number of direct and indirect deaths into the total number of deaths. The time series approach is important to this project because it highlights the impact of the time of year regarding severe weather. Many types of storms are more likely to form during certain times of the year. For instance, hurricanes typically form in the Summer due to factors such as reduced wind shear and increased water temperatures (Shepherd, 2017). Therefore, studying storm casualties throughout the year can reveal which months have the highest risk of fatal storms.

This visualization is split into two components—with each component having a specific purpose. The top graph shows the number of storm casualties each day for all storms with at least one death. It also includes a trend line which computes the expected number of deaths throughout the year. This part of the graph is useful because it shows a complete breakdown

of the storms that occurred during various times of the year. It makes it easy for the viewer to identify all major storms, which are represented as the larger spikes in the graph. Some spikes are much larger than others—these represent the deadliest storms such as Harvey and Irma. The bottom graph is a bar graph that shows the number of storm casualties during each month of the year. It also labels the exact number of deaths per month, which makes it easier to understand. It is helpful for the analysis because it summarizes and simplifies the information from the top graph into individual months. This allows us to determine which months are most likely to experience life-threatening storms.

This visualization was also created using Tableau. It uses the storm details and locations datasets, which were combined by using a left join. It also uses the original versions of the datasets to preserve the cases that may have otherwise been removed during data preparation. Of course, this means that outliers exist in variables such as the number of deaths. For instance, the catastrophic death toll from Hurricane Harvey is an outlier. The reason I did not remove outliers from the visualization is because doing so would remove major storm events from the graph. This would produce an inaccurate depiction of the storms which occurred during 2017. Instead, outlier removal will be done for the predictive models. The top graph was created by plotting the beginning date and time with the total deaths (a variable I created for the geospatial analysis). Next, I scaled the beginning date and time to show the deaths during each day of the year. I then created a trend line on the graph using a polynomial of degree 2 to provide the best possible fit. Afterwards, the bottom graph was created by plotting the month name with the sum of total deaths and setting the graph to a bar graph. I then added the total deaths as a label to display the exact number of casualties for each month. Finally, I added both graphs into a single dashboard.

There are several important observations from this graph. Firstly, we see that August has the highest number of casualties compared to any other month, and most of these deaths occurred around August 26, 2017. This is the date when Hurricane Harvey downgraded into a tropical storm, which is when most of the rainfall took place ("What Happened on August 26," n.d.). According to my previous findings, most deaths from Hurricane Harvey (at least in Texas) were caused by severe flash flooding and tropical storm conditions which the storm produced. Thus, it makes sense that most storm deaths occurred in August. However, one flaw of this graph is that the dates used in these visualizations are the beginning dates and times of the

storm. For storm systems that lasted longer than a single day, the deaths represented in the graph may not have occurred exactly on the specified date. For hurricanes, it is most likely that the deaths would have been spread out over a period of several days or weeks, rather than occurring all on the same day (as the graph would suggest).

Another noteworthy observation is the moderate number of casualties that occurred from March to June. These months average between 45 and 70 deaths—which makes them considerably less deadly than the period from July to September. This is significant because the period known as "tornado season" lasts considerably longer than "hurricane season." The peak of Atlantic hurricane season generally occurs around September due to high water temperatures and low wind shear (Shepherd, 2017). However, tornado season usually spans different months depending on the region of the country—with Southern states having the most tornadoes from March to May, while Northern Great Plains and Midwest states peak around June or July (Berkowitz, 2011). This means that tornado casualties should be expected between the months of March and July. By looking at the graph, we see that the casualties during these months (while still relatively high) are small compared to those from other months. This raises the question of whether tornado casualties were low during 2017.

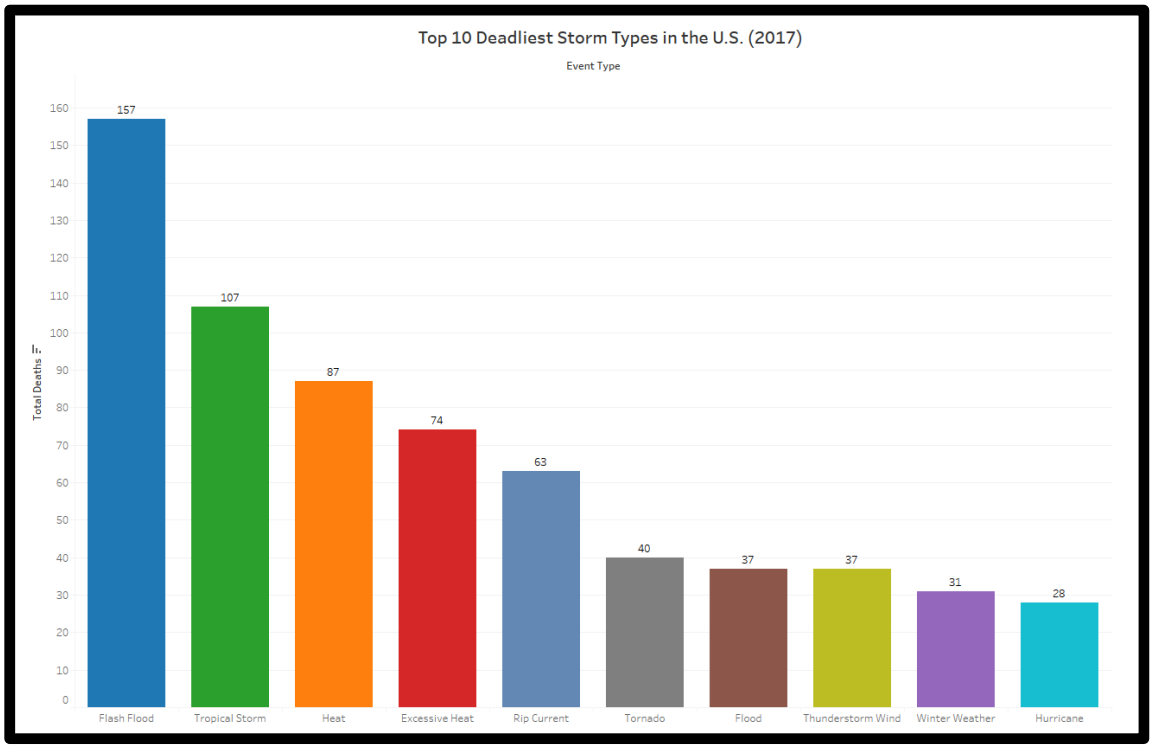

Figure 11. Death Tolls of the 10 Deadliest Storm Types During 2017.

To answer this question, I created a supplementary bar graph that shows the number of deaths caused by the top 10 deadliest types of storms during 2017 (see Figure 11). The types of storms are sorted from highest to lowest casualties, with the total number of deaths labeled for each storm type. This graph is useful because it quantifies the exact number of people killed by each storm type throughout the entire country. According to this graph, there were 40 people killed by tornadoes last year—which is considerably smaller than the deaths from flash flooding, tropical storms, heat, and rip currents. While 40 casualties is still significant, it pales in comparison to the 553 people killed by tornadoes in 2011 (Brooks, 2009). Of course, it is understandable that tornado casualties are inconsistent on a yearly basis—since tornadoes do not always hit major population centers. But once again, this highlights the difficulty of accurately predicting casualties from severe weather.

One of the most interesting findings from my time series analysis is the surprising number of deaths during January. 88 casualties occurred during the month of January, with the highest number of deaths occurring from January 21 to 22 (see Figure 10). It is likely that some of these deaths can be attributed to Winter weather conditions, since Winter weather claimed 31 lives during 2017 (see Figure 11). But in addition, there was a massive tornado outbreak in Georgia and neighboring states from January 21 to 23, during which at least 20 lives were lost ("January 21-23, 2017 Tornado Outbreak," 2017). This indicates that a large percentage of deaths during January can be attributed to tornadoes. Although it is rare for tornadoes to form during the Winter, it is possible for them to appear during any month of the year (Dolce, 2017). In fact, January tornadoes are most likely to form in Southern states due to the air moisture from the Gulf of Mexico (Dolce, 2017). Although deadly storms are more likely to form during the Summer, this finding shows that storm casualties can occur during any time of the year.

Altogether, several of these findings have major implications on the scope of this project. Of course, there were many findings that line up with my initial expectations. For example, I expected that most storm casualties would occur during Summer due to the high frequency of tornadoes and hurricanes during this time of year. My time series analysis confirmed this assumption, since the highest numbers of storm casualties occurred during July, August, and September. The trend line on the time series graph also peaks between June and late August, which reinforces my assumption. I also expected fewer storm deaths during Fall

and Winter. This assumption was mostly true, since only January had more than 50 casualties. However, there were lower storm casualties from March to June than I originally predicted. The reason why I expected high casualties during these months is due to the higher frequency of tornadoes in this time of the year. But ultimately, the low number of tornado casualties during 2017 can once again be explained by the unpredictability of severe weather. Tornadoes may occur frequently during the year, but the number of casualties will depend on whether the storm hits a populated area and whether the residents are adequately prepared. To obtain a more accurate depiction of storm casualties, I would again recommend using data that spans multiple years instead of relying on a single year of storm data.

Another major discovery comes from the high number of storm casualties that occurred during January. My initial expectation was that Winter months would have fewer storm deaths compared to months during the Spring and Summer. Furthermore, I expected that any deadly storms produced during Winter would be blizzards, Winter storms, or ice storms. Thus, I was surprised to learn that a major tornado outbreak occurred during January and caused at least 20 deaths. This finding affects the scope of my analysis for two reasons. Firstly, it challenges my expectation that fewer deadly storms would occur during Winter. Secondly, it reveals that life-threatening tornadoes can occur during any time of the year instead of only forming during Spring or Summer. As such, this finding draws attention to the relationship between storm casualties and the time of year. Although deadly storms are more likely to form during Summer months, this does not eliminate the risk of severe weather during other parts of the year.

Data Visualization 3: Word Cloud

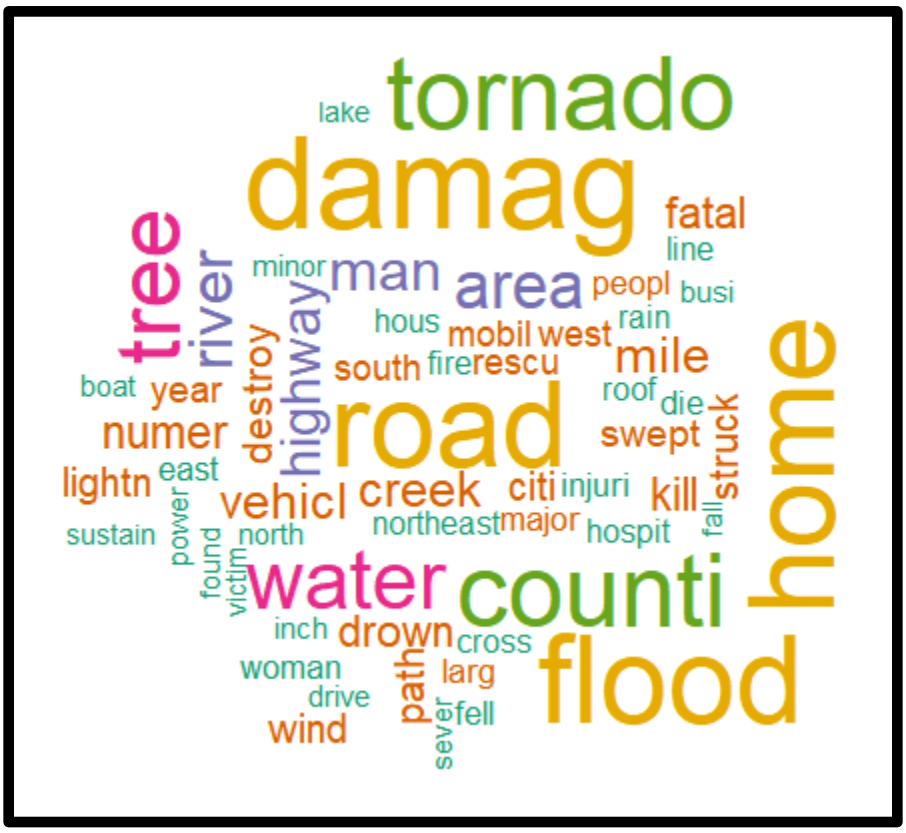

Figure 12. Word Cloud of the 60 Most Frequent Words in Storm Event Narratives.

My third visualization is a word cloud that shows the 60 most frequent terms in the storm event narratives for all storms with at least one casualty (see Figure 12). It is the main component of my text mining analysis, which focuses on studying terms that are frequently used to describe dangerous storms. It uses the bag-of-tokens approach, which involves breaking a document down into a set of words and evaluating the frequencies of these words (Bramer, 2016). This approach is useful for gaining insights about a text document, since it is much faster than manually examining a text to understand its meaning. By identifying the most frequent terms, we can make inferences about which terms are important to the text. Since this problem involves studying the terms that describe deadly storms, my results can help us to determine which factors are most frequently attributed to storm casualties. The shortcoming of this approach is that it provides very limited context due to the focus on keywords which are separated from their sentences. Therefore, we can only make inferences regarding a word's context.

The word cloud is useful for showing word frequency in a visually-appealing way. It scales each word based on its frequency, such that the most frequent terms appear the largest. One of the limitations of this approach is that it can sometimes be difficult to quantify the exact frequencies of certain terms. Unlike other visualizations such as bar graphs, word clouds do not label the exact size or frequency of each term. If two words are of similar sizes, it may not be apparent which of these words has a higher frequency. To help alleviate this issue, I used color labels that color terms based on their frequency. My word cloud assigns six colors: yellow, lime green, magenta, blue, orange, and blue-green. These colors are from highest frequency to lowest—where yellow indicates the most frequent terms and blue-green indicates the least frequent of the included words. The word cloud will be limited to the 60 most common words, which ensures that the visualization is adequately detailed while being easy to read.

This visualization—along with all of my text mining work—was created using R. As mentioned earlier, this required several preprocessing steps to clean the texts and prepare them for analysis. The storm event narratives were compiled into a corpus, which was stripped of extra whitespace, numbers, punctuation, and special characters. Stop words were removed, and words were stemmed to their root words. But in addition, I also removed other words that are unnecessary to the analysis. For instance, the word "storm" is not helpful since the entire dataset refers to storms. Afterwards, I created a document-term matrix and removed sparse terms from the DTM. The word cloud itself requires installing and activating the "wordcloud" package. I created this visualization by making a list of all unique words and their frequency counts. I then set the color scheme with 6 colors to label words based on their frequencies. Next, I used a random seed to reproduce the results. Finally, I generated the word cloud using the word frequency list as an input, and I set the maximum number of words to 60.

With the word cloud created, we can now examine the results. The four most common words (labeled in yellow) are "damage," "road," "flood," and "home." The high frequency of the word "flood" implies that floods have been amongst the deadliest natural disasters in the U.S. during 2017. Additionally, the word "home" could imply two possibilities. Firstly, it may suggest that many (or most) storm-related deaths occurred in people's homes. This would differ from my findings in the data exploration stage, during which I found that most deaths occurred either outside or in vehicles. The other possibility is that many homes were damaged or destroyed during the storm, but the deaths may have occurred elsewhere. This is supported by the high

frequency of the term “damage,” which has a similar frequency as the word “home.” A highly destructive storm is likely to produce casualties and destroy many homes, but this does not necessarily mean that the deaths occurred within homes.

Another frequent term used to describe deadly storms is “tornado” (in green). However, the word “tornado” has a lower frequency than the term “flood” (in yellow). This implies that floods occurred in a higher number of deadly storms than tornadoes, which differs from my initial assumption that tornadoes and hurricanes produced the most casualties. The words “water,” “drown,” and “swept” also appear in this word cloud, which further indicates that many people drowned or were swept away in flood waters. Another key finding is that the term “tree” is mentioned with a relatively high frequency compared to other words. This could mean that fallen trees contributed to some deaths in the dataset. But it could also mean that the same storm that resulted in casualties also caused trees to fall or be uprooted. This highlights the difficulty of interpreting context when performing a keyword-based text analysis. Additionally, the term “vehicle” has a much lower frequency than expected—especially since 168 deaths occurred inside vehicles during 2017. But the fact that it appears on this word cloud indicates that it is mentioned at least somewhat consistently in the storm event narratives.

By further examining the word cloud, we notice that the term “mobile” has a relatively low frequency compared to other terms. The word “mobile” most likely refers to mobile homes, which are frequently destroyed during tornadoes and hurricanes due to their lack of a solid foundation. The term’s low frequency may seem surprising at first, but it is consistent with my earlier findings that only 31 deaths occurred in mobile homes. Instead, most deaths occurred either outside (172 deaths), in a vehicle (168 deaths), in the water (165 deaths), or inside a permanent home (116 deaths). Just because mobile homes are often destroyed during storms does not mean that people are killed inside those mobile homes. But in addition, modern mobile homes are being built to withstand higher wind speeds. Regulations require new mobile homes in the U.S. to be able to survive wind speeds up to 70 miles per hour, while those built in hurricane-prone areas are built to withstand speeds of about 100-110 miles per hour (“5 Mobile Home Myths,” 2013). This does not protect them from the most severe storms, but it may explain why mobile homes are mentioned less frequently.

Overall, some of these findings will have implications regarding the scope of my project. Of course, many of the results make sense logically. For instance, the frequency of the term

"damage" suggests that highly destructive storms are much more likely to produce casualties. But other findings differ drastically from my initial expectations—such as the fact that tornadoes are mentioned less frequently than floods. I originally believed that tornadoes and hurricanes would be the deadliest types of storms by far, particularly due to their highly-destructive nature. But it is no surprise that floods caused a massive percentage of deaths, since my initial data exploration revealed that 165 casualties occurred in the water. Interestingly, the term "hurricane" does not appear at all in this word cloud. This is surprising due to the number of deaths caused by Hurricanes Harvey, Irma, and Maria in 2017. However, I believe that much of the flooding can be attributed to hurricanes—especially from Hurricane Harvey. Of the roughly 70 people killed in Texas during the storm, about 81 percent of the victims drowned due to the large-scale flooding (Wright, 2018). This highlights the difficulty of identifying the most dangerous types of storms. Certain years may involve more tornado-related deaths, while other years will have more casualties caused by hurricanes or flooding. One suggestion for future analysis would be to analyze storm casualties over a longer period of time, since focusing on a single year can lead to different results from previous years.

Another key finding is the location of the death (whether it was outside, in a vehicle, etc.). Location is important to evaluating storm casualties because it can significantly affect a person's ability to survive. Being outside during a storm provides little to no protection against the elements, but being in a vehicle can make you vulnerable to flooding or high winds. My initial data exploration revealed that most deaths occurred outside (172), followed by deaths in vehicles (168), in the water (165), and in permanent homes (116). Mobile homes saw a much smaller number of deaths, with only 31 casualties in 2017. But although there were more deaths in vehicles than in permanent or mobile homes, the word "home" has a much higher frequency than the term "vehicle." One possible explanation is that homes were frequently damaged or destroyed during these storms, but there were fewer deaths that actually occurred at home. This suggests that the bag-of-tokens approach may not be accurate for determining a storm victim's cause of death. This is because it provides too little context to determine whether a keyword such as "home" is directly related to storm casualties. This approach has been more effective at identifying the types of storms which frequently resulted in casualties.

Proposed Visualizations

The visualizations that I created so far have explored the relationships between storm casualties and other important variables. My geospatial map showcased the impact of location on the number of deaths, while my time series analysis showed the relationship between time of year and severe weather. My word cloud revealed the importance of the victim's location of death (whether they were outside, in a vehicle, etc.). And all three of my visualizations explored which types of storms were more likely to produce storm casualties. For further analysis, I would recommend building additional visualizations to explore the relationships between storm deaths and other interesting variables. One variable that I would recommend exploring is the amount of property damage. Storms that result in high casualties tend to produce significant amounts of damage as well. Hurricanes Harvey, Irma, and Maria not only resulted in over 100 confirmed deaths, but they also collectively caused at least $202 billion in damage—which makes 2017 the costliest hurricane season in U.S. history (Drye, 2017). But while a given storm might be highly destructive, it does not guarantee that the storm will result in massive numbers of casualties (or vice versa). Thus, it would be interesting to further explore the relationship between these variables.

One useful approach would be to create a geospatial visualization that shows the amount of property damage caused by storms in each state. This map can be added to a dashboard next to my original geospatial map that shows the storm casualties per state. This technique would allow for a side-by-side comparison of the amount of property damage and the number of deaths in each state. This would make it easier to see if there is a correlation between the amount of damage and the number of casualties—as well as how strong that correlation is. The visualization would be created using a similar approach as my first geospatial map. It would label each state by its total amount of property damage, and it would use a color scheme that colors states from the least damage to the most damage. However, one of the obstacles to creating this visualization is the format of the property damage variable. This feature is not in numeric format, since entries are written in formats such as "200.00K" or "1.00M." Transforming this feature into a numeric variable will be difficult, since I would have to change the "K" and "M" symbols into their corresponding numbers (thousands, millions) while also handling decimal points properly. Thus, I will have to explore methods to accurately convert this variable into its correct numeric form.

Another interesting variable to explore is the storm range. I expect that dangerous storms with longer ranges will often produce higher casualties since they are more likely to hit multiple populated areas. This would depend on the type of storm, since a weak thunderstorm with a large range will likely cause fewer deaths than a powerful tornado with a short range. Thus, I would recommend creating a visualization that compares the storm range to the number of casualties for certain storm types. For instance, I would suggest building a series of scatterplots that show each storm's range plotted over the number of casualties. There would be one scatterplot each for hurricanes, tropical storms, flash floods, tornadoes, heatwaves, thunderstorms, and winter storms. Such a graph would ensure that storm ranges are only compared for storms of the same type. Furthermore, it would help us to determine whether a storm's range is a major contributor to its severity for each type of storm.

I would recommend creating this visualization using Tableau, since it provides a powerful interface that allows for easy creation and modification of visualizations. This visualization would be created by plotting the total number of deaths over the storm range. From there, the storm types can be added as separate rows, and we can use a filter to include only the types of storms we are interested in. This will result in graphs for each storm type included, which will make it easier to compare the results for each type of storm. Next, trend lines can be added to show whether casualties will increase or decrease as the storm's range grows. The completed visualization should provide us with a deeper understanding of the relationship between storm type, range, and the number of deaths.

The visualizations I have proposed so far are useful for understanding the importance of certain variables in the data. However, I would recommend one additional visualization to enhance my findings from my text mining analysis. My recommendation would be to build a correlation plot to identify correlations between important terms in the storm event narratives. Correlation plots involve drawing lines to connect pairs of correlated words such that the strength of their correlation is greater than the minimum correlation threshold ("Text Mining Analysis," 2018). This visualization is useful because it can help to provide a stronger sense of context to the words in this analysis. One of the weaknesses of the bag-of-tokens approach is the difficulty of interpreting context, since keywords are isolated from their sentences. This means that we lose all information about the word's usage within the context of its original sentence, and we are only able to make inferences about the word's possible

meaning. Correlation plots help to address this issue since they allow us to find words that are commonly used together. If two words are more likely to appear within the same body of text, it is easier to infer what those words might possibly mean.

This visualization would be created using R, since that is the tool that I am using to perform my text mining work. R provides a powerful package "Rgraphviz," which allows us to implement correlation plots ("Text Mining Analysis," 2018). Using a higher correlation threshold will result in fewer words in the graph, but it will also result in pairs of terms that are more strongly correlated with each other. Additionally, correlation plots can be weighted so that the width of each line indicates the strength of the correlation ("Text Mining Analysis," 2018). This is useful because it can help us identify not only the related pairs of terms, but also how strongly those terms are related. Overall, this visualization might provide a stronger sense of context regarding the words that are most important for identifying dangerous storms.

Predictive Models

Predictive Model 1: Logistic Regression

In the following sections of this paper, I will create five predictive models using SAS Enterprise Miner to determine the likelihood that a given storm will be lethal. My first model is a logistic regression model, which is a form of regression that is used for categorical target variables. It uses a logarithmic function with values restricted between 0 and 1 to calculate the probability that the target will have a certain class (Knode, 2016b). Logistic regression works well on binary targets, and its performance can be enhanced by using variable selection methods such as forward selection, backward elimination, and stepwise selection. Forward selection starts with an empty model and adds inputs one at a time according to their significance, while backward elimination begins with all variables and removes the least significant variable in each iteration (Knode, 2016b). Stepwise selection begins similarly to forward selection, but it allows for variables to be added or removed during each iteration (Knode, 2016b). Since my goal is to maximize the model accuracy and sensitivity, I will choose the variable selection method that has the lowest misclassification rate and the smallest number of false negatives.

The logistic regression model was created using the regression node in SAS Enterprise Miner. As mentioned, I previously partitioned the data into training, validation, and test sets using the data partition node. I used three regression nodes—one for each of the three variable selection methods. For each regression node, I set the regression type to "logistic regression" and set the selection criterion to "validation misclassification," which helps to minimize the number of misclassified cases in the validation set. I then set the selection models to "forward," "backward," and "stepwise" for the respective model type. These three regression models will be compared using a model comparison node—for which I will mainly focus on their validation misclassification rates and the number of true positives, true negatives, false positives, and false negatives. If any models have identical results on the validation data, I will consider their training or test results as well. I will then select the strongest model as my primary regression model for the analysis. After selecting this model, I change its cutoff threshold to improve its classification performance. I will then evaluate its overall accuracy and sensitivity, as well as explore some of the most important variables that were included in the model.

Model	Data	False Negatives	True Negatives	False Positives	True Positives	Misclassification Rate
Forward	Training	119	27,932	6	112	0.004438
	Validation	76	16,759	4	62	0.004733
Backward	Training	69	27,924	14	162	0.002947
	Validation	61	16,733	30	77	0.005384
Stepwise	Training	131	27,936	2	100	0.004722
	Validation	81	16,761	2	57	0.004911

Figure 13. Comparison of Logistic Regression Variable Selection Methods.

The table above shows the model results for all three variable selection methods in the training and validation sets (see Figure 13). It reveals that all three models are extremely accurate, with miniscule misclassification rates in both the training and validation sets. I am not surprised that the models were able to achieve high overall accuracy, since the target variable is extremely skewed—making it easy for the model to overclassify the majority class. Based on the numbers of true negatives compared to false positives, the models misclassified only a small percentage of storms which had no casualties. But when comparing false negatives and true positives, the models clearly struggled to identify fatal storms. The backward regression model has 61 false negatives in the validation set, which is the lowest among these models. There are 138 death cases in the validation data, which means that the backward model has a sensitivity of 55.8%. This is significantly lower than my goal of 70%, but it is the highest of the three regression models. The forward model has a sensitivity of 44.9%, while the stepwise model's sensitivity is only 41.3%. Since the overall misclassification rate is negligible, I will focus primarily on the validation sensitivity. Therefore, I will select the backward regression model as my primary logistic regression model.

Data	False Negatives	True Negatives	False Positives	True Positives	Accuracy (%)	Sensitivity (%)
Training	10	26,921	1017	221	96.35	95.67
Validation	28	16,072	691	110	95.75	79.71

Test	21	10,750	424	72	96.05	77.42

Figure 14. Backward Logistic Regression Results with 0.01 Cutoff Threshold.

To enhance the model's classification performance, I will change its cutoff threshold using the cutoff node. I previously mentioned that an ideal cutoff value is 0.0068. But since 0.01 is the minimum cutoff threshold, I set this model's cutoff to 0.01. Using the cutoff results, I created a table for evaluating the logistic regression model's performance (see Figure 14). Based on this table, we see that the model has become much more effective at classifying storm casualties. It only failed to classify 28 false negatives in the validation set and 21 in the test set, which results in a validation sensitivity of 79.71% and a test sensitivity of 77.42%. The model now fulfills my minimum sensitivity level of 70%. Furthermore, it achieves this without causing the overall accuracy to decrease too much. Its accuracies in the training, validation, and test sets are around 96%, which is only slightly lower than the original model accuracy of over 99%. The only concern from this model is that the validation and test sensitivities are clearly lower than the training sensitivity. Since the training and validation accuracies are very close, it does not appear that overfitting took place. But the lower sensitivity rates may be due to the small percentage of death cases in the data. The change in the cutoff threshold clearly improved the model's ability to identify the minority class, but it does not necessarily eliminate the difficulty of predicting a skewed target variable.

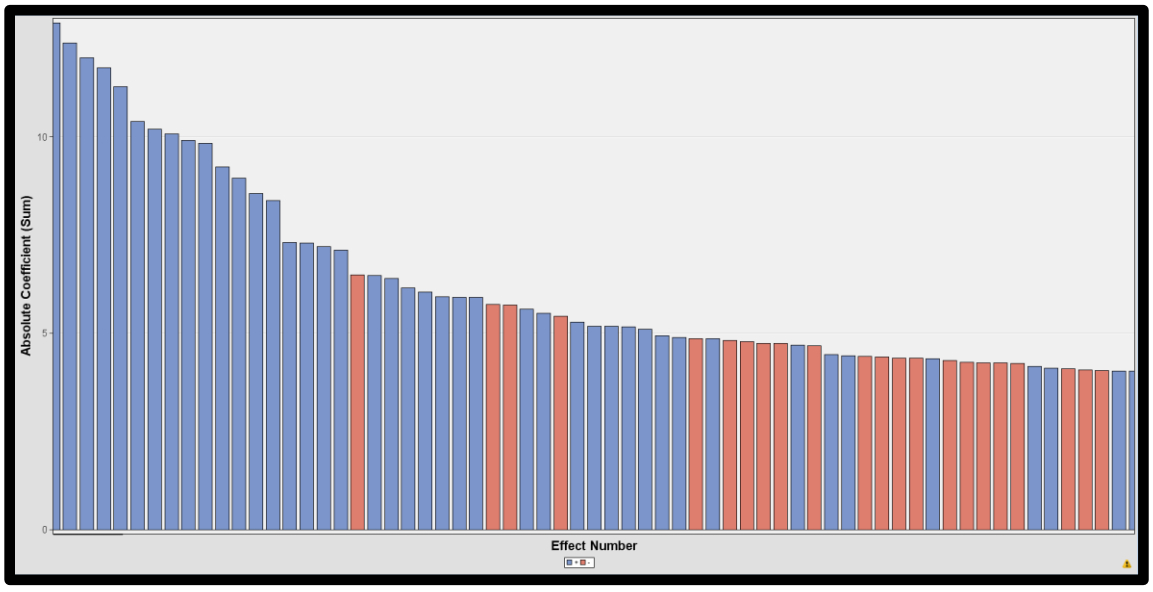

Figure 15. Graph of Regression Coefficients for Backward Regression Model.

One of my goals in this analysis is to identify the variables that most strongly contribute to whether a storm will be deadly. To do this, I will examine the graph of the regression coefficients, which shows the variables that were used to build the regression model (see Figure 15). Regression coefficients are not only helpful for identifying important variables, but they also show how strongly that variable is related to the target and whether that relationship is positive or negative (Frost, 2013). By examining the graph in Figure 15, I found that most of the largest coefficients refer to the amount of property damage. One of the coefficients refers to property damage of around $10 billion, and it has a positive coefficient strength of 12.39. Another refers to property damage of $1.5 billion, and it has a positive coefficient strength of 11.76. The strong positive coefficient implies that highly destructive storms contribute strongly towards casualties. This makes sense logically and is supported by the data—since most U.S. storm deaths from 2017 can be attributed to flash floods and hurricane conditions caused by Harvey and Irma.

There are several findings which alter the scope of the project expectations. One of my findings was the high accuracy and low initial sensitivity of my logistic regression model. The model has an accuracy rate of over 99% and a sensitivity rate of 55.8% in the validation data. Of course, it is not surprising that the skewness of my target variable would cause the model to perform poorly when identifying true positives. However, I believe that the model's initial sensitivity is higher than I would expect, given the extent of my target's skewness. My target, "casualties," has only 387 death cases in a dataset containing 56,921 rows. This means that deaths take up only 0.68% of the dataset. Thus, it is impressive that the model accurately identified over half of these cases. This finding is important because it shows that the model was somewhat able to overcome the target's skewness on its own. Furthermore, after changing the cutoff threshold to 0.01, the model was able to obtain a sensitivity of 79.71% while still having an accuracy of 95.75%. This means that the model fulfilled two of my KPIs: to obtain a model with an accuracy of at least 80%, and to have a model with a sensitivity above 70%. This implies that logistic regression using backward selection is a very effective approach for categorizing storm casualties in this dataset.

When looking at the graph of regression coefficients, I found additional important insights as well. Many of the features with the highest positive coefficients are related to the amount of property damage. This suggests that storms with higher property damage are more likely to result

in casualties. But one of the most interesting findings is the large number of features included in the backward selection model. By zooming out of the graph in Figure 15, I found that hundreds of features were included in the model. Typically, having large numbers of variables in a model can weaken its overall performance. However, this model is still able to achieve high accuracy regardless of the variable count. It is possible that the large selection of features caused my model to have a lower sensitivity. But even with this in consideration, the backward regression model had a higher sensitivity and similar accuracy rates to the forward and stepwise models. Thus, I still recommend using the backward logistic regression model for this problem.

Predictive Model 2: Neural Network

My second predictive model is a neural network, which is an algorithm designed to imitate the behavior of a human brain. It is a collection of input and output nodes with weighted connections that link them together (Han, Kamber, & Pei, 2011). According to this source, neural networks can predict the class of the target variable by using a process called backpropagation—which involves adjusting the weights of its connections until the weights converge. The advantages of neural networks are that they are easily able to handle noisy data, and they can identify patterns in the data without being previously trained (Han et al., 2011). However, one of its weaknesses is that it is a black box algorithm—which means that it is difficult to interpret and troubleshoot (Bati, 2015). Due to its nature, it provides little information about the variables that are important to the model.

In my analysis, I will create one neural network using SAS Enterprise Miner. It uses a data partition node to allocate 50% of data to the training set, 30% to the validation set, and 20% to the test set. I created the model by using the "neural network" node and preserving the model's default parameters. The model architecture is a multilayer perceptron, which involves using one or more hidden layers between the input and output layers of the neural network (Bati, 2015). My model will use 3 hidden layers, and its model selection criterion was set to "misclassification" to minimize the misclassification rate in the validation set. After creating the model, I will evaluate its initial performance prior to changing its cutoff threshold.

There are several tools that I will use to evaluate the model results. I will first examine the iteration plot, which shows the number of steps required to train the model, as well as the misclassification rates for the training and validation sets during each iteration (see Figure 16).

This figure is useful for checking if overfitting took place, as the model's validation misclassification rate would be much higher than its training misclassification rate. Next, I will analyze the model's initial classification results on the training and validation data (see Figure 17). This table shows the number of correct and incorrect predictions, as well as the overall misclassification rate. It is used to determine not only how accurate the model is, but how well it performed when classifying fatal storms. Finally, I will evaluate the results after changing the model's cutoff threshold (see Figure 18). This was done by using the cutoff node, for which I found that the ideal cutoff value is 0.01. This table will be useful for assessing the model's viability when compared to my other predictive models.

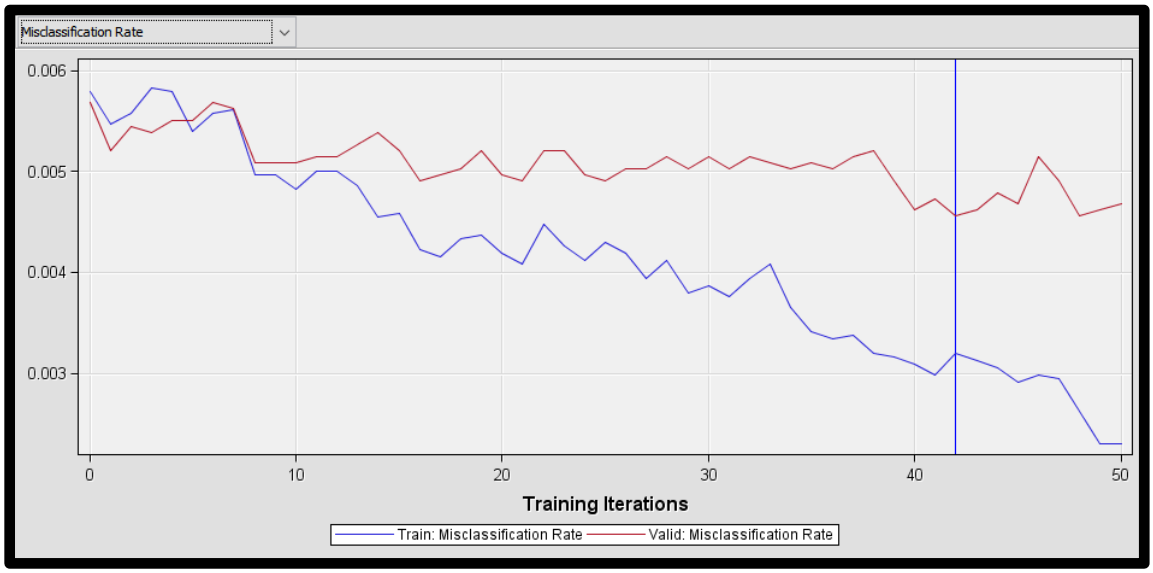

Figure 16. Iteration Plot for Neural Network Model.

First, I will evaluate the model's iteration plot of the training and validation data (see Figure 16). This graph reveals that the model required 42 iterations to train, and that the misclassification rate is higher in the validation set. By moving the cursor onto the graph in SAS Enterprise Miner, I found that the misclassification rates for the training and validation sets were 0.0032 and 0.0046 respectively. This difference does not seem large enough to confirm that overfitting took place. Next, I examined the model's initial classification table to evaluate its performance on the data (see Figure 17). Based on this figure, we see that the model has an extremely low misclassification rate since it identified nearly all true negatives in the data. This is expected due to the target's skewness, and it is consistent with the performance of my previous model. The neural network also misclassified 75 false negatives

in the training data and 66 in the validation data. This results in a training sensitivity of 67.4% and a validation sensitivity of 52.5%. These sensitivity rates are below my minimum rate of 70%, which shows that the model is currently ineffective at classifying fatal storms.

Data	False Negatives	True Negatives	False Positives	True Positives	Misclassification Rate
Training	75	27,923	15	155	0.003195115
Validation	66	16,752	11	73	0.004555674

Figure 17. Initial Neural Network Classification Results.

To improve the model's performance, I will once again change its cutoff threshold using the cutoff node. Like my previous model, the neural network has an ideal cutoff value of 0.01. Using this cutoff threshold, I created a table of the model's updated results (see Figure 18). Based on this figure, we see that the model is much stronger when classifying true positives. It only misclassified 17 fatal storms in the training set, 36 in the validation set, and 23 in the test set. This results in a sensitivity of 92.6% for the training data, 74.1% for the validation data, and 75.2% for the test data. Furthermore, it achieves these sensitivity rates while still maintaining high accuracy rates of over 96%. When compared to my logistic regression model, I found that the neural network's final accuracies were slightly higher in all three data partitions (see Figure 14). However, its sensitivity rates are clearly lower than those of the regression model. For instance, the neural network has a validation sensitivity of 74.1%, while the logistic regression model's validation sensitivity is 79.7%. This suggests that the regression model is more effective at classifying fatal storms in this dataset.

Data	False Negatives	True Negatives	False Positives	True Positives	Accuracy (%)	Sensitivity (%)
Training	17	27,024	914	213	96.69	92.61
Validation	36	16,241	522	103	96.70	74.10
Test	23	10,829	345	70	96.73	75.27

Figure 18. Neural Network Results with 0.01 Cutoff Threshold.

Altogether, there are a few findings that alter the scope of expected results in my project. Firstly, the neural network initially had an accuracy of over 99% and a sensitivity of

only 52.5% in the validation set. This shows that the model was relatively weak at classifying storm casualties prior to changing the cutoff threshold. This is expected due to my target's skewness. But after changing the cutoff threshold to 0.01, the model achieved an accuracy of 96.7% and a sensitivity of 74.1% in the validation data. These figures are higher than the minimum accuracy (80%) and sensitivity (70%) which I selected as KPIs. This means that both models developed so far have exceeded my initial expectations. I originally expected difficulty in fulfilling these KPIs, especially for having a model with a sensitivity of at least 70%. But since every model developed so far has surpassed these expectations, I would not be surprised if my remaining models are also able to fulfill my KPIs. Of course, this was only possible by changing the cutoff threshold—since neither model originally had a sensitivity above 70%.

Another important finding occurs when comparing my neural network to my logistic regression model. I noticed that the neural network has a slightly higher accuracy but a lower sensitivity than the regression model. Sensitivity is more important for evaluating a model's classification performance, since I am more interested in a model that can identify lethal storms. Although the neural network fulfills my KPIs, it is nevertheless a weaker model than my logistic regression model. Furthermore, the neural network is a black box algorithm, which makes it difficult to interpret. This means that it provides little to no information about the variables that are important to the model. Since the logistic regression model shows important inputs through its regression coefficients, this makes the regression model more useful to the analysis. Of the two models created so far, I would recommend the logistic regression model.

Predictive Model 3: Support Vector Machine

I will now describe the third predictive model that I created, which is a support vector machine (SVM). Support vector machines are classification models that involve separating classes using a hyperplane (such as a line or plane) to ensure that records of the same class are on the same side of the hyperplane (Knode, 2016d). Additionally, this hyperplane should maximize the distance between the closest data points of each class—which is known as the "margin" (Knode, 2016d). If the dataset is not linearly separable, then a kernel function can be used to transform the data into a higher dimensional space which would allow the hyperplane to separate classes more accurately (Kane, 2015). Some of the most useful kernel functions include linear, polynomial, radial basis function, and sigmoid kernels. In this analysis,

I will experiment with these kernels and select the SVM with the highest accuracy and sensitivity. One of the strengths of SVMs is that they perform well on datasets with many input variables (Knode, 2016d). Since my combined datasets contain over 70 features, I expect that this method might perform well. However, SVMs are often less accurate on larger datasets and can be computationally demanding (Knode, 2016d). Since my data has over 57,000 cases, it will be interesting to see how the SVM compares to my other models.

The SVMs were created using the "HP SVM" node in SAS Enterprise Miner. Like my other models, they use a data partition node that allocates 50% of data to the training set, 30% to the validation set, and 20% to the test set. I initially used four HP SVM nodes, one for each of the four kernel types. My linear model used the "interior point" optimization method, and its kernel type was set to "linear." The other three models used "active set" as their optimization methods, with the kernels set to "polynomial," "radial basis function," and "sigmoid" respectively. I initially left each model with its default parameters under the "active set options." The polynomial kernel had its degree set to 2, the radial basis function had its RBF parameter as 1.0, and the sigmoid kernel had its sigmoid parameters set to 1.0 and -1.0. When running the models, I found that only the linear kernel executed correctly—since the other three models yielded errors. Though I am unsure why this is the case, it is possible that the massive data size makes it difficult for some of these SVM models to function properly. Since the linear SVM was the only working model, it will be my focus in the analysis.

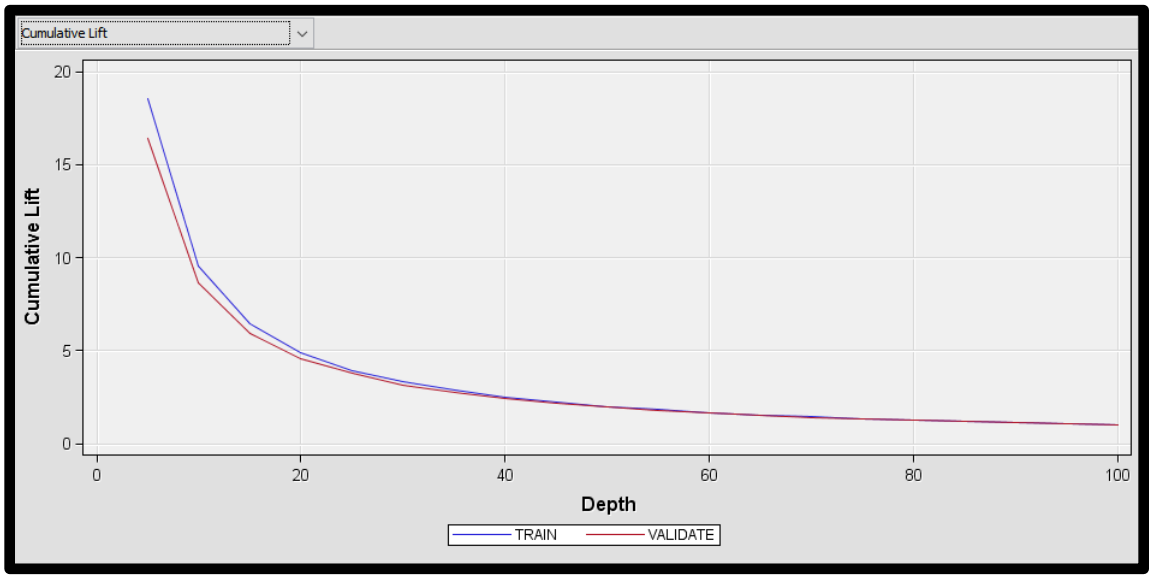

Figure 19. Cumulative Lift Curve for Linear Kernel SVM.

There are several tools that I will use to evaluate this model. Firstly, I will examine the lift curve, which shows the cumulative lift in both the training and validation data (see Figure 19). A strong model will have training and validation curves that match very closely, since this would suggest that no overfitting took place. Another useful tool is the classification chart, which shows a bar graph of the correct and incorrect predictions in the training and validation sets (see Figure 20). This graph is useful for visualizing the model's ability to classify storm fatalities. Furthermore, I will evaluate the classification matrix and fit statistics, which show a breakdown of the predictions compared to the actual values (see Figure 21). An optimal model will have both of its predicted classes match the actual number of records in each class. This figure is useful not only for determining the model's overall accuracy, but its sensitivity as well. Like neural networks, SVMs are considered black box algorithms, which makes it difficult to understand them (Kane, 2015). This means that the model provides no information about which of the input variables are the most important.

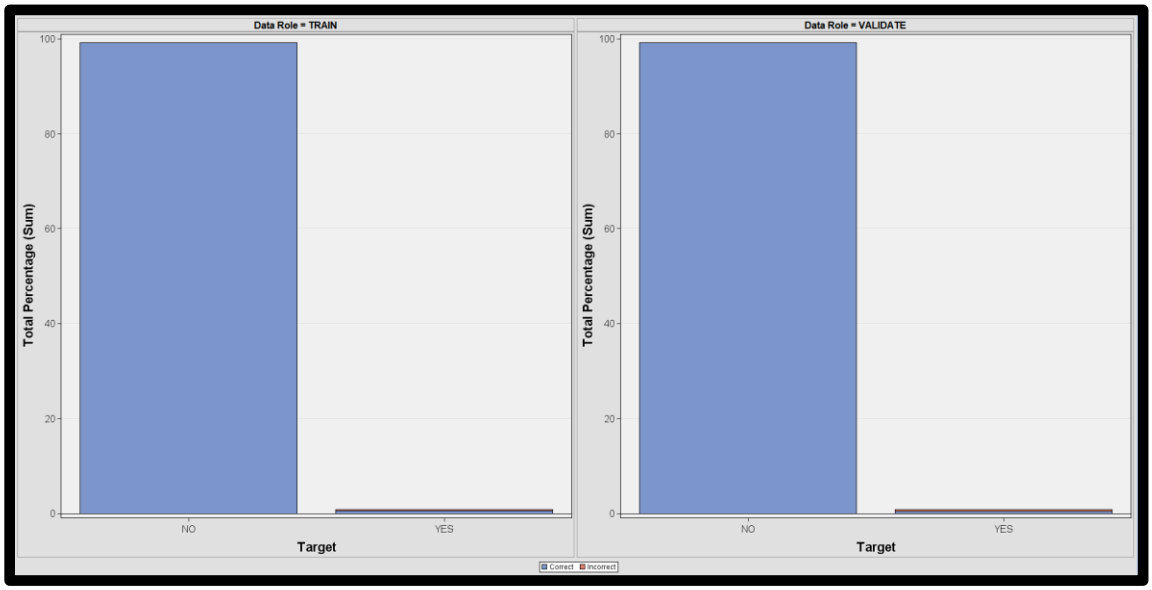

Figure 20. Classification Chart for Casualties in Linear Kernel SVM.

First, I will look at the model's cumulative lift curve (see Figure 19). According to the graph, the validation lift curve follows a similar trajectory as the training lift curve, but the two curves do not align perfectly. This suggests that the training and validation results are different, but the curves are not distant enough to confirm that overfitting took place. Next, I will examine the model's classification charts for the training and validation sets (see Figure 20). This figure shows the correct classifications in blue and incorrect classifications in red.

We notice that nearly all cases refer to storm events with no casualties, and there are almost no misclassifications visible for this class. It makes sense that the model would accurately classify "no" cases, since the vast majority of records are "no." However, the cases listed as "yes" (which indicate deaths) clearly have misclassified entries—especially in the validation set. This shows that the model does not perform as well when identifying the minority class, which makes sense due to the target's skewness.

Classification Matrix

	Training Prediction			Validation Prediction		
Observed	yes	no	Total	yes	no	Total
yes	149	81	230	82	57	139
no	2	27936	27938	9	16753	16762
Total	151	28017	28168	91	16810	16901

Fit Statistics

Statistic	Training	Validation
Accuracy	0.9971	0.9961
Error	0.0029	0.0039
Sensitivity	0.6478	0.5899
Specificity	0.9999	0.9995

Figure 21. Classification Matrix and Fit Statistics for Linear Kernel SVM.

To gain further insight about these findings, I examined the model's classification matrix and fit statistics (see Figure 21). This figure shows that the model has an extremely high accuracy of 99.7% in the training set and 99.6% in the validation set. This high accuracy is due to the model successfully classifying nearly all "no" cases. It only misclassified 2 "no" cases in the training data and 9 cases in the validation data. However, the model currently performs very poorly when identifying "yes" cases—with a sensitivity of 64.8% in the training set and 59.0% in the validation set. Due to its low sensitivity, it is apparent that the dataset is not linearly separable. And since the other kernel types did not work correctly, it seems that SVMs are not the ideal classification model to use on this dataset. However, this does not make the model useless—as its performance can be improved by changing the cutoff threshold.

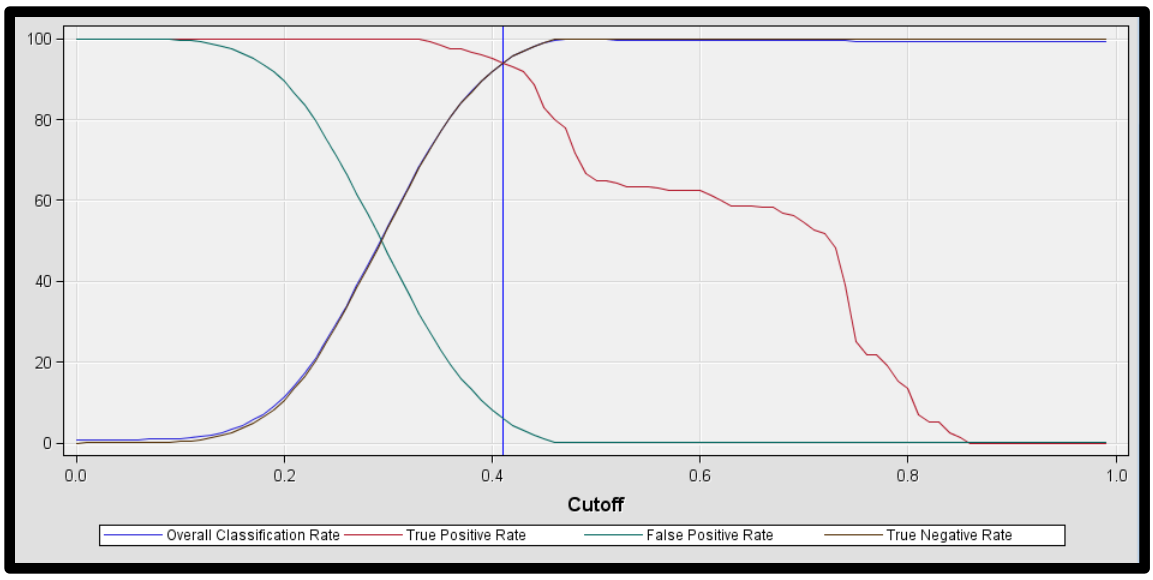

Figure 22. Classification Rates for Linear SVM with 0.41 Cutoff Threshold.

I changed the linear SVM's cutoff threshold by using the cutoff node. Although I had used the cutoff threshold of 0.01 for every model so far, I found that this cutoff value results in a model with a very poor accuracy (see Figure 22). This is because at low cutoff values, the model is able to predict nearly every "yes" case, but it loses its ability to predict the "no" cases. Since over 99% of records consist of "no" cases, it is better to choose a different cutoff threshold to maximize both the accuracy and sensitivity. Based on Figure 22, an ideal cutoff value is 0.41 since it has a true positive rate of 93.91% and an overall classification rate of 93.90% in the training set. Thus, I selected 0.41 as my cutoff threshold for this model.

Data	False Negatives	True Negatives	False Positives	True Positives	Accuracy (%)	Sensitivity (%)
Training	14	26,233	1705	216	93.90	93.91
Validation	22	15,714	1048	117	93.67	84.17
Test	18	10,497	678	75	93.82	80.65

Figure 23. Linear SVM Results with 0.41 Cutoff Threshold.

After changing the cutoff threshold, I examined the model's classification performance (see Figure 23). This table indicates that the model has become much stronger at classifying storm casualties. It only has 14 false negatives in the training set, 22 in the validation set, and 18 in the test set. Its validation sensitivity is 84.17%, which is considerably higher than its

original sensitivity of 59.0%. Additionally, the choice of cutoff threshold ensures that the model still has a very high overall accuracy. Its training accuracy is 93.90%, while its validation accuracy is 93.67%. This means that the SVM fulfills my KPIs of achieving a predictive model with minimum accuracy and sensitivity rates of 80% and 70% respectively. Furthermore, the model's validation sensitivity (84.17%) is higher than that of the logistic regression model (79.71%) and neural network (74.10%). This means that the linear SVM is the strongest model so far when considering validation sensitivity alone.

There are several implications regarding the scope of expected results in this project. The first implication is that that only the linear kernel executed successfully without errors. As mentioned earlier, the polynomial, radial basis function, and sigmoid kernels all yielded errors. Changing the model parameters did not solve this issue. This is the only time in my analysis where I encountered this issue, which seems to suggest that these kernel types are not compatible with my dataset. One possible explanation is that these models have difficulty with handling the large size of my dataset—which contains almost 57,000 cases. One of the weaknesses of SVMs is that they perform poorly on large datasets. One way to test this theory is to create smaller data samples and implement these kernels on the sample data. If the models still fail to run, then the data size is most likely the cause.

Furthermore, the SVM's original performance implies that the storm events dataset is not linearly separable when evaluating storm casualties. The model's low initial sensitivity (59.0%) means that the linear kernel cannot separate the target variable's classes perfectly. This is not too surprising, since most datasets are not linearly separable. But since the polynomial, radial basis function, and sigmoid kernels all failed to run, this implies that the SVM algorithm is not the most effective model for this dataset. However, the linear SVM's classification performance increased significantly after changing the cutoff threshold. In fact, the model's validation sensitivity increased to 84.2%, which is higher than that of my logistic regression and neural network models. Although the model performed poorly initially, it was able to outperform my other models with regards to sensitivity. And since the sensitivity is one of the most important measures, the SVM appears to be one of the strongest predictive models that I developed so far.

Predictive Model 4: Random Forest

My next predictive model is a random forest, which is a form of ensemble model that uses decision trees. Ensemble models involve combining two or more models, and they tend to be more accurate than the original models. The random forest algorithm uses multiple decision trees with random data samples—where each split in the tree will use a random subset of input variables instead of using all inputs (Knode, 2016a). One of the advantages of random forests is that they are capable of handling large datasets with many variables without compromising their performance ("Ensemble models," n.d.). Additionally, this source states that they can easily handle outliers and missing values, and they can maintain high accuracy even when the data changes. Furthermore, random forests can overcome the high variance from individual decision trees by generating random trees that are not correlated with each other—leading to a lower variance between each tree (James, Witten, Hastie, & Tibshirani, 2013). However, one of its weaknesses is that it can be difficult to understand and deploy ("Ensemble models, n.d.).

In this analysis, I will implement one random forest model using the "HP Forest" node in SAS Enterprise Miner. I used the node's default parameters, which sets the maximum number of trees to 100 and uses a maximum tree depth of 50. Unlike my other models, the random forest does not partition the data into training, validation, or test sets. This is because ensemble models such as random forests only need to be evaluated on the training data (Knode, 2016a). After creating the model, I examined the output and recorded the important results into a table (see Figure 24). Next, I used a cutoff node to change the model's cutoff threshold. To balance this model's accuracy and sensitivity, I chose a cutoff threshold of 0.02. Afterwards, I recorded the new results into the same table in Figure 24. This allows for comparison between the model's initial and final performance.

I will use two approaches to evaluate this model. Firstly, I will compare the model output before and after changing the cutoff threshold (see Figure 24). This table is useful for seeing how well the model performed initially, as well as how much it improved by changing the cutoff threshold. One of the main differences between this model and my previous models is that the random forest does not involve validation or test sets. This means that the data is evaluated only on the training set, which makes it somewhat harder to compare with my other models. However, it is still possible to gain an understanding of this model's performance on

the data. Afterwards, I will evaluate the variables that were used to build the model (see Figure 25). One of the benefits of the random forest algorithm is its ability to provide information about variable importance ("Ensemble models," n.d.). Therefore, this figure will be useful to identify some of the features that contribute the most towards storm casualties.

Cutoff Threshold	False Negatives	True Negatives	False Positives	True Positives	Accuracy (%)	Sensitivity (%)
0.5	208	55,875	0	254	99.63	54.98
0.02	20	54,636	1239	442	97.77	95.67

Figure 24. Random Forest Results Comparison Based on Cutoff Threshold.

First, I will evaluate the random forest's performance prior to changing the cutoff threshold (see Figure 24). The original model uses a default cutoff threshold of 0.5. According to this table, the original model once again has a near perfect accuracy since it accurately identified all true negatives in the data. However, it also failed to identify 208 casualty cases—resulting in a sensitivity of 54.98%. This sensitivity is comparable to those of my previous models prior to changing their cutoff thresholds. It shows that the target's skewness makes it difficult to obtain a model with a sensitivity above 70%. But after changing the model's cutoff threshold to 0.02, its classification performance increases significantly. The model only has 20 false negatives, which results in a sensitivity of 95.67%. And it achieves this with a minimal decrease in accuracy, as its final accuracy rate is still 97.77%. And although the model does not use a validation or test set, its training accuracy is higher than that of any previous model. Likewise, its training sensitivity is higher than most of my other models. This suggests that the random forest is one of the strongest algorithms for this dataset.

Variable Name	Train: Gini Reduction ▼
REP_DAMAGE_PROPERTY	0.002329
REP_WFO	0.001328
REP_CZ_FIPS	0.000685
REP_DAMAGE_CROPS	0.000516
REP_SOURCE	0.000507
REP_END_DAY	0.000484
REP_END_RANGE	0.000369
REP_END_TIME	0.000356
REP_injuries	0.000334
REP_BEGIN_LON	0.000317

Figure 25. Top 10 Input Variables Based on Gini Reduction.

Next, I will examine the table of significant inputs that were used to create this model (see Figure 25). This figure shows the 10 most important variables used in this model, which are sorted according to their Gini reduction score. Gini impurity is a measure used to evaluate decision trees, and it measures the likelihood that a random variable from the dataset will be incorrectly classified if it were given a random classification from the existing classes (Ambielli, 2017). Furthermore, the reduction in Gini can be used to rank a variable's importance, since having a higher decrease in Gini suggests that the variable is more important (Thoplan, 2014). Based on this table, we see that the most important variable is the amount of property damage, which has a Gini reduction score of 0.0023. This strongly suggests that the deadliest storms tend to be the most destructive, and it is similar to my finding in the logistic regression model in which many regression coefficients referred to high property damage.

Additional important variables include the WFO and CZ FIPS, which are the second and third most important variables respectively. Interestingly, both variables refer to location. WFO indicates the NWS forecasting office that is responsible for covering that region, while CZ FIPS is a number assigned to a specific county or zone (see Figure 2). This highlights the importance of location towards storm casualties, since certain geographic areas of the country experienced higher numbers of deadly storms. The fourth most important variable is the amount of crop damage. This relates to my finding about property damage—since costlier storms are more likely to produce casualties. However, its Gini reduction score is only 0.0005, which is several times smaller than that of property damage. This suggests that the amount of property damage is a stronger indicator of storm casualties, as high property damage usually

indicates a more powerful storm than high crop damage. Other important features in this table include the storm's range and whether it produced injuries. The importance of storm range suggests that storms with greater ranges are more likely to result in deaths. Likewise, the presence of injuries within a given storm implies that the storm is dangerous—which increases the likelihood of casualties.

Altogether, there are several findings that alter the scope of expectations in this project. Firstly, I found that the random forest model is one of the strongest models that I developed so far. It initially had a low sensitivity, which is consistent with every model created in this project. But after I changed the model's cutoff threshold, I found that its accuracy and sensitivity were higher than those of my other models. Of course, this model does not involve a validation or test set, which makes it somewhat more difficult to compare to the other models. But when looking at the training results alone, the model's accuracy is 97.77% and its sensitivity is 95.67%. These figures strongly indicate that the random forest model is among the most effective algorithm for classifying casualties in this dataset. Yet its high performance is reasonable, since ensemble models tend to be among the most accurate types of models.

Another major finding involved the variables which are the most important to the model. I found that some of the most significant variables referred to the property damage, crop damage, location, range, and injuries. Many of these findings are consistent with earlier observations in this analysis. For instance, I expected that location would have a major impact on whether a given storm will be deadly. Likewise, the importance of property damage makes sense because more destructive storms generally result in higher casualties. However, the use of the Gini reduction score helped to quantify each variable's importance. By studying this measure, I found that property damage far outweighed any other input variable. This feature was also the most important input in my logistic regression model as well. These findings seem to suggest that property damage might be the most significant input in the dataset.

Predictive Model 5: Ensemble Model

The final model in my analysis is a heterogeneous ensemble model, which is a type of model that combines the outputs of different models to obtain a single result. They are typically more accurate than the individual models, but they can also be more difficult to interpret and deploy (Knode, 2016a). Heterogeneous models differ from decision tree-based

ensemble models such as random forests since they allow combining different types of models, such as logistic regression and neural networks. There are four methods that are used to combine models: average, maximum, voting average, and voting proportion. The average method computes the average of all models' posterior probabilities, while the maximum method involves choosing the maximum probability of the models included (Knode, 2016c). The voting average method involves calculating the mean probability from the most common class (Maldonado, Dean, Czika, & Haller, 2014). Lastly, the voting proportion method calculates the ratio of models in the majority group compared to the total number of included models (Knode, 2016c). I will experiment with all four methods and select the model with the highest accuracy and sensitivity.

My ensemble models will involve combining the backward linear regression model, the neural network, and the support vector machine. They do not use the random forest because that model did not partition the data into training, validation, and test sets. Including it in the heterogeneous models would yield errors, since every other model used a data partition node. To combine the models, I will use the "ensemble" node in SAS Enterprise Miner. For the average ensemble model, I set the interval target predicted values and the class target posterior probabilities to "average." For the maximum model, I set both of these properties to "maximum." For the voting average and voting proportion methods, I set the interval target predicted values to "average" and the class target posterior probabilities to "voting." Next, I set the voting posterior probabilities to "average" and "proportion" for each respective model. I will compare these models using a model comparison node. After selecting the strongest model, I will change its cutoff threshold to optimize its classification performance.

I will now describe the tools that I used to evaluate the models. First, I compared their ROC curves in the training, validation, and test sets (see Figure 26). The ROC curves show the models' sensitivity plotted over their specificity, and the optimal model will have the highest sensitivity rates. Next, I created a table that compares the initial classification performance of each model (see Figure 27). This table is useful for determining which model combination method is the most effective at classifying storm casualties in my dataset. After selecting the strongest ensemble model, I will choose a cutoff threshold that optimizes the model's accuracy and sensitivity. I will then create a new table to evaluate the updated model's classification performance (see Figure 28).

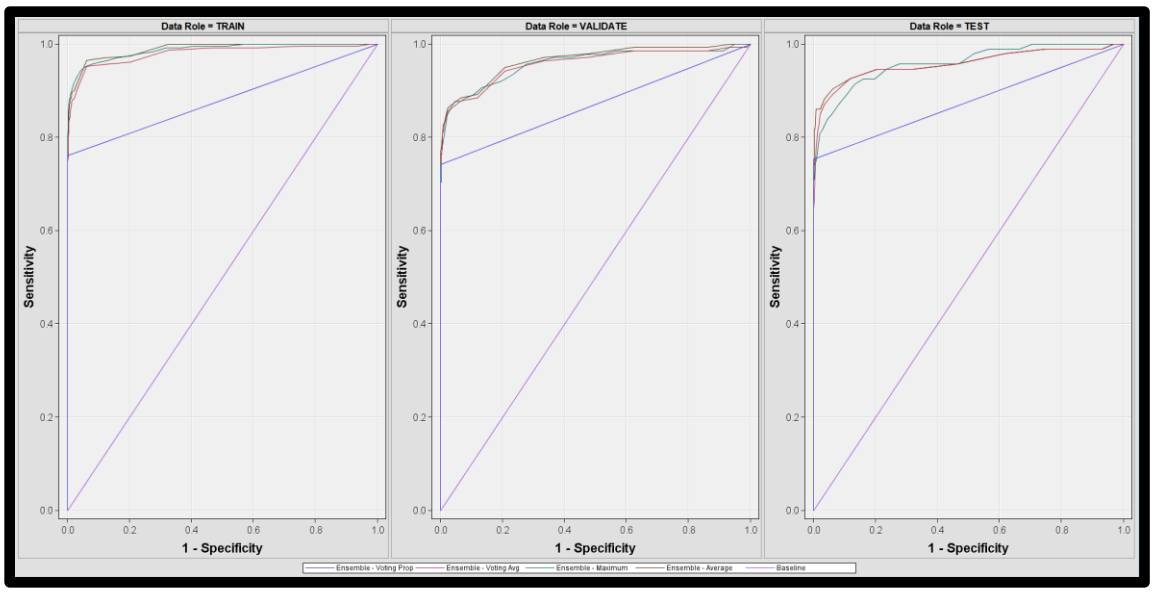

Figure 26. ROC Curves for Heterogeneous Ensemble Models.

First, I will compare the ROC curves for each heterogeneous model (see Figure 26). This figure reveals that the average ensemble model has the highest sensitivity per specificity in the training and validation sets, and it is followed closely by the maximum and voting average models. In the test data, the maximum method clearly has the highest overall sensitivity over specificity. In all three data partitions, the voting proportion method is the weakest by far since it has much a lower sensitivity than any other model. Next, I will evaluate the model comparison table for all four ensemble models (see Figure 27). According to this table, every model has a very low misclassification rate while also having a high number of false negatives. This is once again expected due to the model skewness. However, this figure reveals that the maximum model has the lowest numbers of false negatives in both the training and validation sets. It misclassified 46 storm casualties in the training set and 41 in the validation set—resulting in a training sensitivity of 80.1% and a validation sensitivity of 70.3%. This may seem to conflict with the ROC curves—which indicated that the average model had the highest training and validation sensitivities. However, this is because the models are plotted over their specificity rates (the true negative rate). Since the maximum model has a lower number of true negatives compared to the average method, this explains why the maximum model has a lower ROC curve. But since sensitivity is more important for evaluating model performance, we see that the maximum model is the optimal choice. Thus, I will use the maximum ensemble model as my primary heterogeneous model.

Model	Data	False Negatives	True Negatives	False Positives	True Positives	Misclassification Rate
Average	Training	80	27,928	10	150	0.003195
	Validation	46	16,756	6	93	0.003077
Maximum	Training	46	27,901	37	185	0.002947
	Validation	41	16,725	38	97	0.004674
Voting Average	Training	85	27,932	6	145	0.003231
	Validation	53	16,757	5	86	0.003432
Voting Proportion	Training	85	27,932	6	145	0.003231
	Validation	53	16,757	5	86	0.003432

Figure 27. Comparison of Heterogeneous Model Combination Methods.

One interesting finding is that the maximum model already has a validation sensitivity of 70.3%—meaning that it fulfills my KPI of having a minimum sensitivity of 70%. This would mean that the model does not require changing the cutoff threshold. However, I will still change the cutoff threshold to allow for comparison with my other four models. By using the cutoff node in SAS Enterprise Miner, I found that the ideal cutoff value is 0.41 (similar to the SVM). Thus, I evaluated the model using a cutoff threshold of 0.41, and I recorded the model's results in a table (see Figure 28). Based on this table, we see that the model gained a much higher sensitivity without causing the accuracy to decrease too significantly. It now only has 11 false negatives in the training set and 17 in the validation set—resulting in training and validation sensitivities of 95.2% and 87.7% respectively. This means that the model has a higher validation sensitivity when compared to logistic regression (79.7%), neural network (74.1%), and SVM (84.2%). These three models were used to create my heterogeneous model. Thus, it is not surprising that its sensitivity is higher—since ensemble models are almost always more accurate than the individual models used to create them (Knode, 2016a).

Data	False Negatives	True Negatives	False Positives	True Positives	Accuracy (%)	Sensitivity (%)
Training	11	26,234	1704	220	93.91	95.24

Validation	17	15,695	1068	121	93.58	87.68
Test	14	10,501	673	79	93.90	84.95

Figure 28. Maximum Ensemble Model Results with 0.41 Cutoff Threshold.

There were a few findings that have implications on my project expectations. Prior to changing the cutoff threshold, the maximum ensemble model had a very strong accuracy and a somewhat high number of false negatives. However, I found that its initial validation sensitivity was 70.3%, which is higher than my minimum sensitivity rate of 70%. This is important because it indicates that the heterogeneous model was the only model to fulfill this KPI before changing its cutoff threshold. This means that the model does not require changing the cutoff threshold, and the only reason I have done so is to allow for fair comparison with my other models. And after I changed the model's cutoff threshold, I noticed that its validation sensitivity was higher than that of my logistic regression, neural network, and SVM models. This finding makes sense, since these three models were used to create my heterogeneous model, and ensemble models are typically more accurate than their individual models.

Another implication involves comparison between my two ensemble models: the random forest and the heterogeneous model. Since the random forest does not involve a validation and test set, the only way to compare the models is to use the training results. When evaluating the results on the training data, I noticed that the heterogeneous model has a lower sensitivity (95.24%) than the random forest (95.67%). Of course, the heterogeneous model has less data in its training set due to its data partitions. But since its validation and test sensitivities are lower than its training sensitivity, it is likely that the heterogeneous model's overall sensitivity would be lower if it only used a training set. Thus, the random forest is a more effective model when considering the sensitivity rates. Furthermore, the random forest might be easier to create as well—since it does not require building three separate predictive models and combining their results. However, the main advantage of the heterogeneous model is that it is the only model with a sensitivity above 70% prior to changing the cutoff threshold. The random forest had a validation sensitivity of only 54.98% before changing its cutoff threshold to 0.02. This means that the heterogeneous model has the highest sensitivity rate by default, and it is the only model that does not require changing the cutoff threshold. In the following section, I will offer a detailed comparison of all five predictive models to justify which model is the best choice for classifying storm casualties in this dataset.

Predictive Model Review

Now that I have created all five predictive models, I will evaluate their performances to determine which model is most effective for classifying fatal storms in this dataset. The first tool I will use is the ROC curve, which shows the models' sensitivity over specificity prior to changing their cutoff thresholds (see Figure 29). This figure is useful for evaluating the models' initial performance. We see that the neural network is the weakest model in all three data partitions, followed by the backward regression model. In the training set, the SVM has the highest sensitivity per specificity, followed by the maximum heterogeneous model. In the validation set, we see that the heterogeneous model and the SVM have the highest ROC curves. And in the test set, we see that the heterogeneous model by far has the highest sensitivity over specificity. Based on these results, the strongest initial model appears to be either the SVM or the heterogeneous ensemble model. Since the validation and test results are more significant than the training results, I would recommend using the heterogeneous model instead of the SVM. However, we need to consider the models' results after changing their cutoff thresholds to confirm which model is the strongest.

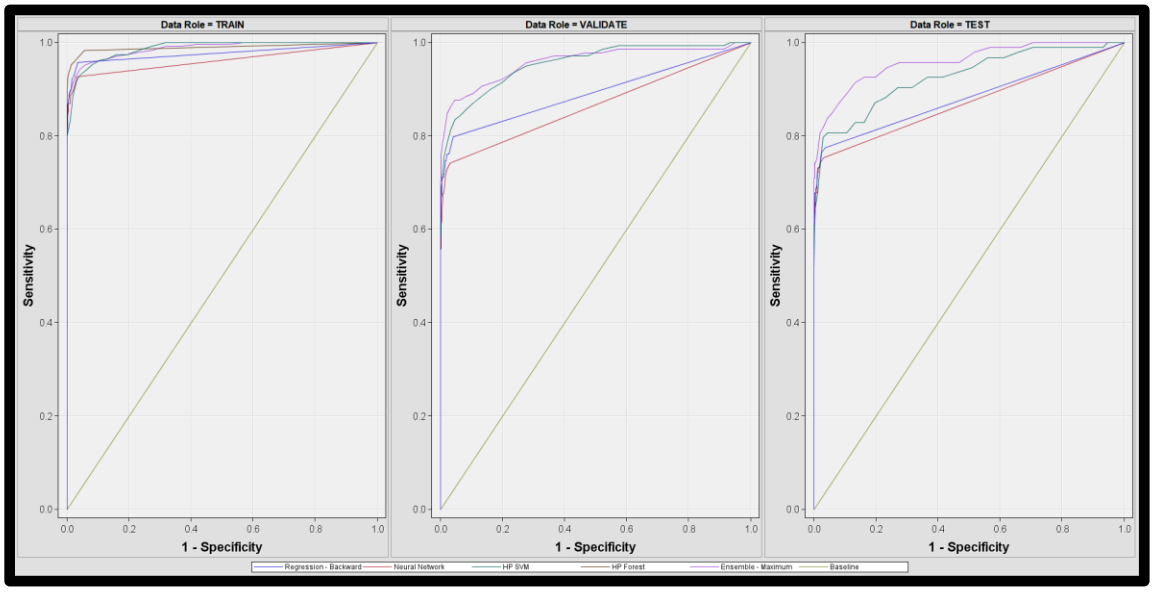

Figure 29. ROC Curves for 5 Completed Predictive Models.

The next tool that I will use to evaluate model performance is the model comparison table (see Figure 30). This table shows each model's accuracy and sensitivity rates in the training,

validation, and test sets using the selected cutoff values. Since all models have excellent accuracy rates, I will place more emphasis on their sensitivity rates. I will primarily focus on the models' validation and test sensitivities—though I will consider training results when evaluating the random forest. Firstly, we see that the neural network has the lowest sensitivity of any model in the training, validation, and test sets. This confirms that it is the weakest model in the analysis, although it still exceeds my minimum accuracy and sensitivity KPIs of 80% and 70% respectively. The logistic regression model's validation and test sensitivities are lower than those of any other model except for the neural network. This suggests that it is the second weakest model in my analysis. Furthermore, we see that the SVM's validation and test sensitivities are lower than those of the maximum ensemble model. This indicates that the heterogeneous model is the stronger model after changing the cutoff thresholds.

Model	**Cutoff Threshold**	**Data**	**Accuracy (%)**	**Sensitivity (%)**
Logistic Regression (Backward)	0.01	Training	96.35	95.67
		Validation	95.75	79.71
		Test	96.05	77.42
Neural Network	0.01	Training	96.69	92.61
		Validation	96.70	74.10
		Test	96.73	75.27
SVM (Linear Kernel)	0.41	Training	93.90	93.91
		Validation	93.67	84.17
		Test	93.82	80.65
Random Forest	0.02	Training	97.77	95.67
Ensemble (Maximum)	0.41	Training	93.91	95.24
		Validation	93.58	87.68
		Test	93.90	84.95

Figure 30. Final Model Comparison Table.

When evaluating the training sensitivities, I found that the random forest performs slightly better than the maximum ensemble model (see Figure 30). In fact, the random forest's training sensitivity is approximately equal to that of the logistic regression model. However, this highlights the issue of evaluating models using their training results. The regression model has the second weakest sensitivity rate in the validation and test sets, which suggests that using the training results can produce unreliable findings. But ensemble models such as random forests do not require validation and test sets, since they are evaluated on the training data. Thus, I will compare the random forest and heterogeneous model by evaluating their training results. The random forest has a training sensitivity of 95.67%, while the maximum ensemble model's training sensitivity is 95.24%. Additionally, the random forest has a training accuracy of 97.77%, which is higher than the ensemble model's training accuracy of 93.91%. Based on these numbers alone, the random forest is the most effective model in the analysis. However, the heterogeneous model's validation and test results cannot be ignored. Thus, the champion model will be selected from one of these two models.

Overall, the random forest has a few advantages over the heterogeneous model. Firstly, it has a higher accuracy and sensitivity in the training set after changing the cutoff threshold. Additionally, it may be somewhat easier to create since it does not require building three separate predictive models and combining their results. My heterogeneous model requires implementing a logistic regression model, neural network, and a support vector machine. This approach is not only expensive computationally, but it also results in a difficulty interpreting individual model results—as well as making the model more difficult to deploy (Knode, 2016c). However, this does not mean that the random forest does not suffer from these issues as well. Random forests are also black box algorithms, which makes it hard to interpret their results and deploy them ("Ensemble models," n.d.). Yet, one of the benefits of the random forest approach is that it provides information about the important variables used in the model. For instance, my random forest indicated that property damage, location, and storm range are some of the most significant features for classifying deadly storms. This information helps to make the random forest more useful for the analysis.

The heterogeneous model also has several important advantages. One advantage is the high sensitivity in the validation and test data. This model may be more time consuming to build, since it requires implementing three separate models. But the benefit of this approach is that the

final ensemble model performs better than any of the individual models. The heterogeneous model attained the highest validation and test sensitivity rates of any model used in this analysis. But more importantly, the heterogeneous model was able to obtain a validation sensitivity of 70.3% before even changing the cutoff threshold. This model fulfilled my KPIs of having an accuracy of at least 80% and a sensitivity of at least 70% while using the default cutoff value of 0.5. This is the only model to achieve this feat, which makes it the only model that does not require changing the cutoff threshold. This is important because it shows that the model can fulfill my project expectations without taking the additional step of using another cutoff value. And although the model does not provide information about variable importance, the logistic regression algorithm used to create this model does provide such information. Altogether, both the random forest and the heterogeneous models are useful for classifying deadly storms. But since the heterogeneous model is the only method to achieve a high sensitivity by default, I will select the heterogeneous model as the champion model in this analysis.

Final Results

Analysis Justification

In this analysis, I created five predictive models to analyze the 2017 storm events data and determine the likelihood that a storm will produce casualties. I compared these models based on their classification performance to identify the strongest model for classifying fatal storms. My goal was to obtain at least one model with an accuracy of at least 80%, as well as a model with a sensitivity above 70%. This is because having a high accuracy and sensitivity ensures that the model is effective at classifying dangerous storms. Furthermore, I explored the causes of deadly storms by identifying the five features that contribute the most towards whether a given storm will be deadly. To do this, I examined the features of the models that provide information about their input variables—such as the regression coefficients for logistic regression and the Gini reduction scores for the random forest. I also created a geospatial map, time series graph, and word cloud to explore the relationship between the target variable (casualties) and other important inputs (such as location or time of year). These visualizations were intended to show how strongly these variables contribute to storm casualties.

The purpose of this project was to contribute towards addressing three major problems regarding storm safety. My three business objectives were to help reduce the overall number of storm casualties, increase the average storm warning time across the U.S., and lower the number of storm false alarms. Although these issues are difficult to solve, I believe that my project will allow us to make progress in these areas if it is implemented successfully. Reduction of casualties is possible if we have a predictive model that is very effective at identifying dangerous storms. Additionally, having knowledge of the variables that contribute towards fatal storms will help us to identify characteristics that would make classifying storms easier in the future. These achievements can also contribute towards increasing the average storm warning time, since having effective predictions can allow us to identify deadly storms much faster than before. Finally, we can help lower the number of false alarms across the country by having a model that can successfully identify both lethal and non-lethal storms.

Over the course of my analysis, I found that the data used for this project was appropriate for the project scope. The storm details dataset was the most widely used dataset since it contains most of the observations and many important variables. I used this dataset to create my target variable “casualties,” which combines direct and indirect deaths to assess whether a given storm

will produce any fatalities. My target variable was appropriate for the analysis because it focuses on predicting whether a storm will cause any loss of life, rather than determining the number of casualties. All of my predictive models performed well when classifying my target, which means that the algorithms are a good fit for my dataset. The locations dataset was also helpful in this analysis, as some of its variables (such as storm range) were among the most important inputs for my predictive models. The fatalities dataset was useful as well, especially during my word cloud analysis. This is because many of the frequent terms refer to the victim's location of death, which is found in the fatalities dataset. However, the shortcoming of including the fatalities dataset is that its variables are correlated with my target—since all of its records involve only death cases. As a result, I omitted this dataset from my predictive models.

The visualizations used in this project were very helpful for addressing my project goals. The geospatial analysis yielded useful information about the states with the highest risk of having dangerous storms. It enhanced my analysis by showing not only the regions with the most storm deaths, but also the types of storms that affect different parts of the country. This information can help us to promote storm safety practices that are appropriate for each state. The time series graph was useful for showing the times of year with the highest numbers of storm casualties. Additionally, it brought attention to the likelihood of severe weather events that occur during unexpected months—such as the January tornado outbreak. This makes it useful for raising awareness about deadly storms during any time of the year, rather than only during Spring or Summer. Lastly, the word cloud was helpful for showcasing the terms most commonly used to describe dangerous storms. It emphasized that the victim's location (outside, in a vehicle, etc.) has a major impact on their likelihood of survival. Information such as this could have been easily overlooked if we lacked the ability to access the unstructured text data.

The predictive models themselves were critical to my analysis. Having high accuracy and sensitivity rates allow for the models to effectively categorize lethal and non-lethal storms. This makes it feasible to predict storm casualties in the future, which will ideally promote safety measures that can save lives across the country. I used five predictive models: logistic regression, neural network, support vector machine (SVM), random forest, and heterogeneous ensemble model. Through my analysis, I found that all of my models achieved accuracy rates much greater than 80% and sensitivity rates above 70% after changing their cutoff thresholds. This means that all five methods used in my analysis are effective for classifying dangerous storms in this dataset.

It also shows the importance of changing the cutoff threshold to optimize the model performance. Since my target variable is highly skewed, most of my models could not achieve the minimum sensitivity of 70% without changing their cutoff threshold. The only model to do so was the maximum heterogeneous model, which is my recommendation for the champion model.

One of the most important components of the predictive models is the information about the inputs that are significant to each model. Identifying the variables that most strongly impact storm casualties can make it easier to determine whether future storms will be deadly—since it allows us to find characteristics that are associated with deadly storms in the dataset. One of my goals in this project was to identify the five features that contribute the most to whether a storm will be deadly. Although many of my models did not provide information about the most important inputs, the logistic regression model and random forest have methods for evaluating variable importance. The logistic regression method features a graph of regression coefficients that indicates the variables' relationships with the target. The random forest allows inputs to be evaluated using their Gini reduction scores, in which higher values indicate a more important variable. By using these methods, I found that the most important variables include the property damage, storm location, crop damage, range, and whether injuries occurred. Altogether, my findings indicate that these models are effective for fulfilling the goals of my project.

Findings

I will now provide a brief overview of my findings throughout this project. First, I will summarize the results from my three visualizations, beginning with my geospatial map. This visualization showcased the number of storm-related casualties in each state of the U.S. during 2017 (see Figure 8). The image revealed that the states with the highest casualties were Texas, Nevada, and Florida. By examining the types of storms responsible for each death, I found that most deaths in Texas and Florida were attributed to hurricane and tropical storm conditions (including flash flooding) produced by Hurricanes Harvey and Irma. However, Nevada's high death toll appears to be an anomaly, since it is rarely targeted by tornadoes or hurricanes. Interestingly, most of the 122 casualties in Nevada were caused by heat—which seldom receives the same attention that hurricanes or tornadoes receive. This highlights the importance of considering other types of natural disasters rather than focusing only on tornadoes or hurricanes. Additionally, I found that most states in the Great Plains, Midwest, and South experienced

relatively few storm-related deaths. Although these states frequently experience tornadoes, the low death toll shows an unpredictability regarding severe weather. Just because certain storms commonly occur in the region does not mean that casualties will be high each year. Altogether, this visualization suggests that location does play a role in storm casualties (as evidenced by the high death tolls in Texas and Florida). However, it may not be a deciding factor—since there is a strong possibility of storm-related deaths in other parts of the country.

Next, I will discuss the results from my second visualization: the time series graph. This visualization showed the number of storm casualties during 2017 on a daily and monthly basis (see Figure 10). I found that the number of deaths peaked during August, which was around the time when Hurricane Harvey was ravaging Texas. Overall, storm casualties were highest during Summer, which is the peak of hurricane season. Since 2017 saw three devastating hurricanes in the U.S. (Harvey, Irma, and Maria), it is no surprise that most casualties occurred during Summer. One of the surprises was the relatively low death toll from March to June, which is commonly known as tornado season. The low casualties during this period reinforces my earlier statement, for which I claimed that a high frequency of storms does not guarantee that the death toll will be high as well. The biggest surprise from this visualization was the high number of casualties that occurred during January. Interestingly, many of these deaths were caused by a tornado outbreak in Georgia and the surrounding states. This once again highlights the unpredictability of severe weather, since major storms can occur during any time of the year. Overall, the time of year does play a role in storm casualties—since most deaths generally occur during the Summer. However, deadly storms can still appear during the months when we least expect them to.

I will now review my findings from my third visualization—a word cloud that shows the 60 most frequent terms in the storm event narratives (see Figure 12). As mentioned, this image only includes words used to describe storms with at least one casualty. One of my findings was that the term "flood" has a higher frequency than the term "tornado." This implies that floods were responsible for more casualties than tornadoes in 2017. This finding makes sense, as many deaths in 2017 were caused by flash flooding attributed to Hurricane Harvey. Another important finding was the importance of the location of death (whether it was outside, at home, in a vehicle, etc.). Many common terms seem to reference a victim's location of death—including words such as "home," "water," "road," or "vehicle." The high frequency of the term "home" may imply that many people were killed inside their homes. Likewise, it may mean that the same storm which

claimed lives was also responsible for damaging or destroying homes. This highlights one of the method's shortcomings—a lack of context for each term in the word cloud. This means that we can only make educated guesses regarding each word's possible meaning.

Model	Data	Accuracy (%)	Sensitivity (%)	Ease of Understanding	Ease of Implementation
Regression (Backward)	Validation	95.75	79.71	Easy	Easy
Neural Network	Validation	96.70	74.10	Difficult	Moderate
SVM (Linear Kernel)	Validation	93.67	84.17	Difficult	Easy
Random Forest	Training	97.77	95.67	Moderate	Difficult
Ensemble (Maximum)	Validation	93.58	87.68	Difficult	Difficult

Figure 31. Comparison of Model Performance and Ease of Use.

From here, I will provide an overview of the results from my five predictive models. All five models were optimized by changing their cutoff thresholds, which helps them achieve higher sensitivity rates when using a skewed target variable. I evaluated the models primarily using their validation sensitivities. The only exception was the random forest, which was evaluated on its training set since it does not use a validation or test set. Based on the model comparison table (see Figure 31), I found that the maximum ensemble model had the highest validation sensitivity of any model. Furthermore, it was the only model to have a sensitivity above 70% before changing its cutoff threshold—as its original validation sensitivity was 70.3%. The neural network clearly has the lowest sensitivity rate, followed by the logistic regression model. Regarding the random forest, I found that it had one of the highest training sensitivity rates of any model (95.67%). But since it only uses a training set, it is difficult to compare its classification performance to the other models. Thus, I had to consider other factors when choosing the champion model.

One important factor to consider is the difficulty of implementing each model. The logistic regression model is one of the simplest models to implement, and it is easy to understand since it provides information about the important input variables. However, it is the second

weakest model with regards to classification performance. The neural network is both difficult to understand and has the lowest sensitivity rate. The SVM is also difficult to understand since it is a black box technology (Kane, 2015). However, it is easy to implement since it only uses a linear kernel. But since the model sensitivity is one of the most important factors, I recommend using either the random forest or the heterogeneous model. The random forest is somewhat easier to understand than the heterogeneous model since it provides information about the important input variables. And although it is challenging to implement, it may be easier to deploy than the heterogeneous model since it does not require building three separate predictive models. But although the heterogeneous model does not provide information about its important variables, one of its individual models (logistic regression) does provide this information. Furthermore, since the heterogeneous model achieved a sensitivity above 70% prior to changing its cutoff threshold, it is the only model that does not require changing its cutoff value at all. This gives it a major advantage in model implementation when compared to the random forest. Due to these factors, I recommend the heterogeneous model as the champion model in this analysis.

Lastly, I will discuss my findings on the models which provide information about the important variables for classifying dangerous storms. Two of my models describe their important inputs: the logistic regression model and the random forest. The logistic regression model included this information in its graph of regression coefficients (see Figure 15). By examining this graph, I found that most of the strong positive coefficients referred to property damage. Cases with high property damage in the billions or hundreds of millions of dollars had the strongest positive coefficients—meaning that the most destructive storms were typically the deadliest. The random forest provides information about its important inputs according to their Gini reduction scores (see Figure 25). In this figure, the most important variables will have the highest Gini reduction scores. I found that the most important variable was property damage, which is consistent with my findings in the logistic regression model. Other important inputs include the storm's location, its range, the amount of crop damage, and whether it caused injuries. To select the five most important variables for classifying storm casualties, I considered these results along with the plot of variable worth from the StatExplore node (see Figure 7). Since many of the same variables are included in both figures, I conclude that the five most important inputs are the property damage, location, crop damage, storm range, and the presence of injuries.

Review of Success

Overall, this project was successful in many ways—since it fulfilled all three of my key performance indicators and many of my business success criteria. With regards to my KPIs, all five of my models easily surpassed my first two KPIs of having a minimum accuracy of 80% and a minimum sensitivity of 70%. Having an accuracy rate of at least 80% is important because a highly accurate model is more effective at classifying dangerous storms. I noticed that all five models surpassed this minimum accuracy rate prior to changing their cutoff thresholds—which means that the models were already very accurate. In fact, every model initially had an accuracy rate of around 99%. It is not surprising that these models were able to fulfill this KPI so easily, since my target variable "casualties" is highly skewed. There are only 387 rows out of 57,000 cases which involve storm deaths. This means that the models can easily attain a high accuracy rate by identifying all cases that do not involve deaths. Of course, having a high accuracy alone does not result in a strong model, since it needs to be effective at identifying death cases in the data. Thus, achieving a sensitivity above 70% was the primary focus of my predictive models.

To fulfill this KPI, I changed each model's cutoff threshold to improve its ability to identify deaths in the data. I selected a cutoff threshold that maximizes the model's sensitivity without causing its accuracy to decrease too significantly. By doing so, I was able to get all five models to have sensitivity rates above 70% in their validation and test sets. Initially, I expected some difficulty in obtaining at least one model to fulfill this KPI due to the extreme skewness of my target variable. But since all five models surpassed my minimum sensitivity rate, this means that every model is effective on my dataset to some degree. However, only one of my models fulfilled this KPI prior to changing its cutoff threshold. Since the maximum heterogeneous model had an initial validation sensitivity of 70.3%, it is the only model that can fulfill my first two KPIs without having to change its cutoff threshold. This is one of the main reasons why I selected it as the champion model. Not only does it have one of the highest accuracy and sensitivity rates of any model, but it can also achieve this feat without the one extra step that the other models require. Nevertheless, all five of my models fulfill my accuracy and sensitivity KPIs—which means that all of them are useful for predicting storm casualties in this dataset.

This project was able to fulfill my third KPI as well, which involves identifying the five features that contribute the most towards deadly storms. Having knowledge of the variables that contribute to storm severity makes it easier to identify characteristics of life-threatening storms

that occur in the future. One of the methods I used to fulfill this KPI was to evaluate the variables that were the most significant to each predictive model. The logistic regression and random forest models were the only models to provide information about their important inputs. Through my analysis, I found that property damage was the most important indicator of deadly storms—as storms with higher amounts of property damage were much more likely to cause casualties. The other important inputs include the storm's location, its range, the amount of crop damage, and whether it produced injuries. However, my goal was not only to identify the most important features, but also to study how some of the variables contribute to storm severity. I explored these relationships using my three visualizations. My geospatial analysis indicated that geographic location does contribute to storm casualties, but it is not the single deciding factor. My word cloud indicated that a victim's location during the storm (at home, outside, in a vehicle, etc.) strongly contributed to their likelihood of survival. My time series analysis indicated that the time of year has an impact on the likelihood of deadly storms. However, it was not a deciding factor and was not included in the top 5 input variables according to the predictive model results. But altogether, this analysis provided many powerful insights about the features that influence the likelihood of deadly storms.

Altogether, I believe that this project will contribute greatly towards addressing my three business objectives: reducing the storm casualty rate, increasing the average storm warning time, and lower the number of false alarms across the U.S. While these results may not be measurable within the scope of my project, I believe that my project's success will allow researchers to make progress in each of these areas. With regards to the overall casualty rate, my champion model's high accuracy and sensitivity rate allows it to easily predict whether a storm might be deadly. This will make it useful for predicting storm fatalities in the future. Additionally, understanding the five most important input variables will make it easier to identify fatal storms. Furthermore, my geospatial analysis will be useful for promoting safety in the states with higher risks of having deadly storms. It also emphasizes the risk of other types of severe weather, such as heatwaves. This should help to raise awareness about severe weather in states that rarely experience tornadoes or hurricanes. Additionally, my time series graph is useful for not only showcasing the times of year with the deadliest storms, but also for showing the risk of severe weather during other times of the year. All of these components have the potential to promote storm safety and preparation across the country, which should help to save lives in the future.

Likewise, this project's success can strongly contribute towards increasing the average storm warning time across the country. This can be achieved by having a method to accurately predict dangerous storms prior to their formation or landfall. By developing a model with a high accuracy and sensitivity rate, I believe that my methods can be applied to future storm data to determine whether a storm will be deadly. Furthermore, I believe that having knowledge of the important variables will also help us to identify dangerous storms more quickly. Of course, some of these variables (such as the amount of damage and injuries) are only relevant after the storm occurs. However, other variables (such as location, range, and time of year) can be studied prior to the storm causing any harm. By analyzing such features in historic storms, we can identify similar characteristics of future storms to determine whether they will be deadly. This will allow us to issue warnings more quickly. Additionally, the project can contribute towards reducing the number of false alarms across the country. False alarms occur because of the difficulty of accurately predicting whether a storm will appear. I believe that my model can address this issue since it attained both a high accuracy and sensitivity. This means that it can easily determine whether a dangerous storm will form, but not at the expense of its overall accuracy.

Finally, I mentioned in the beginning of this paper that there are several criteria that I will use to evaluate the success of my project, and the project would be considered successful if it fulfills at least one of these criteria. Interestingly, the project fulfilled multiple criteria—which makes it very successful. My first criterion was to obtain at least one model with a classification accuracy of at least 80%. This was easily fulfilled, as all five models achieved accuracies much higher than 80%. Secondly, I sought to identify the five variables that contribute the most towards storm casualties and explore their relationship with the target variable. Through my analysis, I found that the most important inputs are property damage, location, crop damage, range, and injuries. By using the graph of logistic regression coefficients, I found that higher amounts of property damage corresponded with higher casualties. And by using my geospatial visualization, I found that certain locations were at higher risk of having dangerous storms. Some of these variables (such as storm range) were not thoroughly explored due to time constraints. However, such variables can be explored further in a future analysis. My final two success criteria are to contribute towards increasing the storm warning time by at least 5%, and to help lower the false alarm rate by 5% or more. These cannot be measured within the scope of this project, but they can be fulfilled if the project is implemented effectively.

Recommendations for Future Analysis

Although I consider my project to be successful overall, there are some areas that could be improved or expanded upon in a future analysis. One of my recommendations is to expand the data to span multiple years rather than only a single year. This is because having only one year of storm events may not paint the full picture of storm behavior in the U.S. Severe weather can often be unpredictable, and certain types of storms may not always occur in the same regions or during the same times of the year. I noticed this when implementing my geospatial and time series visualizations. In my geospatial visualization, I found that storm-related deaths in 2017 were defined by the major disasters of that year. For instance, the catastrophic death tolls in Texas and Florida can be attributed to Hurricanes Harvey and Irma respectively. Additionally, certain regions of the country that frequently suffer from tornadoes had a relatively low death toll—which may be attributed to a quieter tornado season. In my time series graph, I found that certain months of the year that frequently experience dangerous storms had fewer casualties than expected. Likewise, months that rarely receive tornadoes (such as January) suffered from higher casualties than expected. Due to the unpredictability of severe weather, I believe that expanding the timeline of the data will help to produce more consistent results overall.

I would recommend using at least 10 years of data spanning from around 2008 to 2018. I believe that this will help to provide a more accurate picture of storm casualties with regards to both location and time. It will help my geospatial analysis since it will showcase the trends of storm casualties over a longer period of time. I expect that the average number of casualties will be higher across the Midwest, Great Plains, South, and Atlantic coast. This is because the data will have more instances of major tornadoes and hurricanes across different states. Likewise, this approach will help my time series analysis to have more realistic results on average. I expect that there will be a higher number of storm casualties during Spring and Summer, and there may be fewer casualties during January and other Winter months. But although expanding the timeline of the data might produce more consistent results, the tradeoff is that the predictive models will take much longer to train. This is because the amount of data will be many times larger, and the predictive models already run slowly in SAS Enterprise Miner using the current dataset. However, the benefit of having more accurate data may outweigh the cost of having a higher training time. Thus, I would consider exploring this possibility in a future analysis.

Another recommendation for future analysis would be to expand the text mining section with additional useful components. My text mining analysis involved generating a word cloud to study the most common terms used in the storm event narratives for all storms with casualties. The reason I did not incorporate other text mining techniques is because they may have produced a project scope that is too broad, which could have taken away attention from other components such as the predictive models or visualizations. However, one of the shortcomings of my text mining approach is that it provides little information about the context of each word. Since it only shows words and their frequencies, we are limited to making inferences about the significance or meaning of each important term. Therefore, I would suggest using additional components such as word associations and correlation plots. Word associations are used to show how often two terms appear together, which makes it easier to determine the context of these terms. For instance, the words "home" and "flood" can have many possible meanings when examined alone. But if these terms are frequently found together, we can infer that homes were damaged or destroyed in a flood. Correlation plots have a similar role as word associations, but they show the relationships between words in a visual format—which makes the results easier to interpret. As such, I believe that these methods can provide more context about each word's importance.

But in addition, I would recommend exploring other natural language processing (NLP) techniques in the future, rather than relying only on the bag-of-tokens approach. Word associations and correlation plots make it easier to determine the context of each word, but they are still limited to individual terms and lack the ability to understand the meaning of an entire sentence or document. Even if the terms "home" and "flood" are used together, the actual sentence may state the following: "very few homes were damaged by the flooding." Thus, the assumption that many homes were destroyed in the flood is false. One alternative is to use cognitive computing techniques, which involve algorithms that can process human languages and imitate human problem solving in ways that traditional computing cannot (Gortcheva, 2018). One example of cognitive computing technology is IBM Watson, which is a computer system designed by IBM to understand the structure of the human language. Watson has a massive collection of texts in its database, and it is able to understand context by extracting a text's features, examining its stored texts to determine possible responses, comparing the responses using algorithms, and selecting the strongest response (High, 2012). This makes it much more effective for text mining operations when compared to traditional approaches, and I recommend exploring the possibility of implementing Watson techniques in the future.

Furthermore, I would recommend building additional visualizations to explore the relationships between my target variable and other important inputs. The visualizations I created in this analysis focus on inputs such as location and time of year. But another interesting variable to explore would be the storm's range. It would be easy to assume that storms with higher ranges would be deadlier than those with lower ranges. This is assuming that the storms are of the same type, since a powerful tornado with a short range will likely be deadlier than a weak thunderstorm with a large range. One suggestion would be to build a series of scatterplots comparing the storm range to the number of casualties for certain types of storms (such as hurricanes, tornadoes, floods, and thunderstorms). There would be one scatterplot for each type of storm, and the results would indicate how the storm's range affects the casualty rate for each storm type. Additionally, I would recommend exploring the property damage in greater depth. Property damage is the most important variable for determining whether a storm will be deadly, and thus exploring it in greater detail can provide valuable insights. One approach would be to create another geospatial visualization similar to the one in this analysis, but it would instead showcase the amount of property damage in each state. It can be compared to my original geospatial map using a dashboard to determine how strongly property damage and casualties are related.

Finally, one of the issues in my analysis revolved around the usage of the fatalities dataset. Although this dataset was useful for my text mining analysis, it caused problems when used in my predictive models. This is because of the strong correlation between my target "casualties" and the variables in the fatalities dataset. This caused my models to initially have accuracy and sensitivity rates close to 100% prior to changing their cutoff thresholds. The reason for the correlation is that every observation in the fatalities dataset refers to a confirmed death case. This means that any record which has a non-empty value for these variables will automatically be classified as a death case. In essence, it causes the model to "cheat" since it only has to look at the variables in the fatalities dataset and can disregard all other inputs. Due to this finding, I had to remove these variables from the predictive model development in SAS Enterprise Miner. This means that one of my datasets was not used in its full capacity, since it was only used for the text mining visualization. Therefore, I would recommend using an additional dataset to fulfill the project scope. For instance, I recommended using datasets that contain storm data for multiple years. This would ensure that all of the data is used for the models, which will improve its overall performance. Altogether, I believe that my suggestions can improve the quality of this analysis, which will make it even stronger for predicting dangerous storms in the future.

References

10 Worst US Storms in the 100 Years of Winter Weather History. (2016, December 15). Retrieved June 14, 2018, from https://www.earthnetworks.com/blog/worst-blizzards-winter-weather-history/

5 Mobile Home Myths Busted. (2013, December 08). Retrieved July 12, 2018, from https://mobilehomeliving.org/5-mobile-home-myths-busted/

About Us. (n.d.). Retrieved June 12, 2018, from https://www.ncdc.noaa.gov/about

Ambielli, B. (2017, October 29). Gini Impurity (With Examples). Retrieved July 25, 2018, from https://bambielli.com/til/2017-10-29-gini-impurity/

Annual U.S. Killer Tornado Statistics. (2018, January 16). Retrieved June 12, 2018, from http://www.spc.noaa.gov/climo/torn/fatalmap.php

Bati, F. (2015, Fall). Classification using Artificial Neural Networks (ANN). Lecture presented at UMUC. Retrieved July 9, 2017.

Berkowitz, S. F. (2011, October 7). When is tornado season? Retrieved July 14, 2018, from https://www.mnn.com/family/protection-safety/stories/when-is-tornado-season

Bramer, M. (2016). Principles of Data Mining. Retrieved February 7, 2018.

Brooks, H. (2009, March 1). US Annual Tornado Death Tolls, 1875-Present. Retrieved June 12, 2018, from https://blog.nssl.noaa.gov/nsslnews/2009/03/us-annual-tornado-death-tolls-1875-present/

De Vries, A., & Meys, J. (n.d.). How to Use the merge() Function with Data Sets in R. Retrieved July 5, 2018, from https://www.dummies.com/programming/r/how-to-use-the-merge-function-with-data-sets-in-r/

Dolce, C. (2017, January 18). Where January Tornadoes Are Most Likely in the United States. Retrieved July 14, 2018, from https://www.wunderground.com/storms/tornado/news/january-tornado-threat-area

Drye, W. (2017, November 30). 2017 Hurricane Season Was the Most Expensive in U.S. History. Retrieved July 15, 2018, from https://news.nationalgeographic.com/2017/11/2017-hurricane-season-most-expensive-us-history-spd/

Ensemble models and portioning algorithms in SAS Enterprise Miner. (n.d.). Retrieved November 6, 2017, from http://www.sas.com/apps/webnet/video-sharing.html?bcid=4363855671001

Erdman, J. (2016, March 18). Your Odds of Being Hit By a Tornado. Retrieved June 1, 2018, from https://weather.com/storms/tornado/news/tornado-odds-of-being-hit

Erdman, J. (2018, May 30). How the National Weather Service Is Working to Reduce Tornado Warning False Alarms. Retrieved June 1, 2018, from https://weather.com/storms/tornado/news/2018-05-30-tornado-warning-false-alarms-research-nws

Frost, J. (2013, December 12). Regression Analysis Tutorial and Examples. Retrieved October 4, 2017, from http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-tutorial-and-examples

Gillis, C. (2017, September 9). Hurricane Irma: Tornadoes possible, rip current 'life-threatening'. Retrieved July 13, 2018, from https://www.news-press.com/story/news/2017/09/09/hurricane-irma-tornadoes-possible-rip-current-life-threatening/649520001/

Gortcheva, E. (2018, February 13). *DATA 650 NLP & Cognitive*. Retrieved February 14, 2018, from https://www.youtube.com/watch?time_continue=474&v=xvBg4ZCJIS8

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Retrieved June 27, 2017.

Heberton, B. (2014, April 24). 13 Minutes: The Average Warning-Time Before a Tornado Hits. Retrieved June 1, 2018, from https://www.theatlantic.com/technology/archive/2014/04/13-minutes-thats-the-average-warning-time-before-a-tornado-strikes/361195/

High, R. (2012, December 12). The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works. Retrieved February 19, 2018, from http://www.redbooks.ibm.com/redpapers/pdfs/redp4955.pdf

Historic Hurricane Harvey's Recap. (2017, September 2). Retrieved July 13, 2018, from https://weather.com/storms/hurricane/news/tropical-storm-harvey-forecast-texas-louisiana-arkansas

History of the National Weather Service. (2015, February 20). Retrieved June 12, 2018, from https://www.weather.gov/timeline

Hurricane Statistics Fast Facts. (2018, May 31). Retrieved June 12, 2018, from https://www.cnn.com/2013/05/31/world/americas/hurricane-statistics-fast-facts/index.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Retrieved June 28, 2017.

January 21-23, 2017 Tornado Outbreak One of Largest Winter Outbreaks on Record in U.S. (2017, January 26). Retrieved July 14, 2018, from https://www.wunderground.com/news/severe-weather-forecast-south-high-risk-tornadoes-january-2017

Johnson, D. (2017, September 24). Is 2017 the Worst Hurricane Season Ever? Retrieved June 12, 2018, from http://time.com/4952628/hurricane-season-harvey-irma-jose-maria/

Kane, D. (Performer) (2015, January 26). *Data science part IX: Support Vector Machine* [Web]. Retrieved October 24, 2017, from https://www.youtube.com/watch?v=fMWjhQ2UcNs

Knode, S. (2016a, November 1). *Ensemble Models (bagging, boosting, random forest)*. Lecture presented at UMUC. Retrieved November 6, 2017.

Knode, S. (2016b, August 19). *Regression Models*. Lecture presented at UMUC. Retrieved October 3, 2017.

Knode, S. (2016c, November 8). *SAS Enterprise Miner Ensemble Models (combining models)*. Lecture presented at UMUC. Retrieved November 11, 2017.

Knode, S. (2016d, August 25). *Support Vector Machine Models*. Lecture presented at UMUC. Retrieved October 3, 2017.

Maldonado, M., Dean, J., Czika, W., & Haller, S. (2014). *Leveraging ensemble models in SAS Enterprise Miner*. Retrieved from https://support.sas.com/resources/papers/proceedings14/SAS133-2014.pdf

Moran, D. (2016, April 14). A Brief History of Weather Radar. Retrieved June 15, 2018, from https://blog.wdtinc.com/a-brief-history-of-weather-radar

Nevada has highest rate of heat related deaths. (2017, June 23). Retrieved July 12, 2018, from https://www.ktnv.com/news/contact-13/nevada-has-highest-rate-of-heat-related-deaths

NOAA. (2006, December 12). About the Celebration. Retrieved from https://celebrating200years.noaa.gov/about.html

NOAA. (2013, August 03). Tornadoes and Averages Deaths per Year. Retrieved June 12, 2018, from https://www.weather.gov/cae/lgaverages.html

NOAA. (2017, September 11). Tropical Storm IRMA. Retrieved July 13, 2018, from https://www.nhc.noaa.gov/archive/2017/al11/al112017.discus.050.shtml

NOAA. (n.d.). About Our Agency. Retrieved June 12, 2018, from http://www.noaa.gov/about-our-agency

Rip Currents. (n.d.). Retrieved July 13, 2018, from https://www.usla.org/page/ripcurrents

Severe Weather Watches and Warnings. (2006, December 12). Retrieved June 12, 2018, from https://celebrating200years.noaa.gov/foundations/severe_weather/welcome.html

Shepherd, M. (2017, November 06). Why Atlantic Hurricanes Don't Form In Winter And Spring (Or Do They?). Retrieved July 14, 2018, from https://www.forbes.com/sites/marshallshepherd/2017/11/05/why-atlantic-hurricanes-dont-form-in-winter-and-spring-or-do-they/#6ec4d5a128ba

Storm Data Export Format, Field names. (n.d.). Retrieved May 31, 2018, from https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Export-Format.docx

Storm Data Publication. (n.d.). Retrieved June 12, 2018, from https://www.ncdc.noaa.gov/IPS/sd/sd.html

Storm Events Database. (n.d.). Retrieved May 31, 2018, from https://www.ncdc.noaa.gov/stormevents/details.jsp

Texas Tornado Facts. (n.d.). Retrieved July 13, 2018, from https://www.tornadoalleyarmor.com/locations/dallas/texas-tornado-facts

Text Mining Analysis Using R: Analyzing Course Descriptions (2018). Retrieved February 5, 2018.

The National Weather Service (NWS). (n.d.). Retrieved June 12, 2018, from https://www.weather.gov/about/

Thoplan, R. (2014). Random Forests for Poverty Classification. *International Journal of Sciences: Basic and Applied Research*. Retrieved July 24, 2018, from https://www.researchgate.net/publication/264785074_Random_Forests_for_Poverty_Classification.

Top 10 Most Active Hurricane Seasons. (n.d.). Retrieved June 12, 2018, from https://www.wunderground.com/hurricane/top10.asp

Weather Forecasting Through the Ages. (n.d.). Retrieved June 15, 2018, from https://earthobservatory.nasa.gov/Features/WxForecasting/wx2.php

What Happened on August 26, 2017. (n.d.). Retrieved July 14, 2018, from https://www.onthisday.com/date/2017/august/26

What is a Weather Radar? (n.d.). Retrieved June 15, 2018, from http://www.hko.gov.hk/m/article_e.htm?title=ele_00189

Wright, P. (2018, April 20). Most Hurricane Harvey Drowning Deaths Occurred Outside Flood Zones, Study Finds. Retrieved July 12, 2018, from https://weather.com/news/news/2018-04-20-hurricane-harvey-deaths-outside-flood-zones

Your National Weather Service: Evolving to Build a Weather-Ready Nation. (2017, August 11). Retrieved June 12, 2018, from https://www.weather.gov/about/wrn