Ralph Parlin
Project Report
Prof. Ying Lin
IST 652 Scripting for Data Analysis
Summer 2020

Recreational Boating Accidents:
Causes, Insights, and Ways to Improve Boater Safety

INTRODUCTION

Recreational boating is one of America's favorite pastimes. The annual economic impact of recreational boating averages around $170.3 billion a year. It also supports more than 691,149 workers and 35,277 business. It is estimated that 141.6 million Americans go boating each year. However, being on the water, although fun, is also dangerous. In 2018 there were over 4,415 boating accidents on US waters. These accidents resulted in $45.9 million worth of damages, and more importantly, lead to 633 deaths and 2,511 serious injuries.

What are the leading causes of boating accidents? How do certain maritime factors interact to explain the variation of accident outcome: injury or death. What can we learn from modeling this data that may help the Coast Guard teach prevention methods, or be better prepared when responding to these unfortunate incidents. These are just a few of the questions this analysis sheds light on. There are key factors that impact boating accidents that if properly addressed, would result in less damages, injuries, and deaths.

Through clustering analysis this report discovered patterns in the data that provide insight and helped inform classification efforts. Using regression analysis, the analysis determined which factors are significant when trying to predict the financial cost of boating accidents (Research Question 1). Using decision tree classification, this analysis modeled whether the outcome of an accident results in death (Research Question 2).

LITERATURE REVIEW

A brief review of existing literature on boating accidents is primarily provided via the US Coast Guard. Each year they complete a summary report about the accidents they respond to. Although valuable, the reports offer data at the summary level and as such, is simply descriptive in nature. I was unable to find any attempts at modeling the data in order to make predictions or better determine the factors that lead to accidents when controlling for other variables.

DATA SOURCE AND COLLECTION

The US Coast Guard is the authoritative data source for boating incidents on US waterways. As such, I decided to reach out to their statistics team and ask for access to their data. I established a point of contact and was granted access to their boating accident database. The Coast Guard records very specific details on every boating accident they are called to. Day, time, location, weather conditions, as well as many other features that could be used to model specific outcomes. The data I have been granted access to are Coast Guard accident response data for the years 2017 and 2018 (they are still compiling 2019) for every state in the US except for California as they do not allow public release. All personal identifiable information has been removed from each observation.

The database schema consists of four tables that contain varying information related to accident details, vessels, injuries, and deaths involved in boating accidents. In order to develop a single datafile that included information from each of the four data tables, using MS Access I had to move the tables to second normal form by creating queries to merge the data. After joining the tables I was left with 11,720 observations across 35 columns of data. I exported the table as a MS Excel file for upload to Python for wrangling and analysis.

Below is the data dictionary of the consolidated dataframe made from the database query.

Feature Name	Feature Description
Year	Year in which the accident occurred.
Death	1 for if accident resulted in at least 1 death, 0 otherwise
Injury	1 if accident resulted in at least one injury, 0 otherwise
Numberdeaths	The number of people who died in the accident.
Numberinjured	The number of people who were injured in the accident.
NumberVesselsInvolved	The number of vessels involved in the accident.
TotalDamage	The total amount of damage that resulted from the accident.
Date	The date of the accident.
TimeofDay	Time of day when the accident occurred. If unknown, the time is left blank.
State	State in which the accident occurred. Please note the following codes: AT- Atlantic Ocean GM- Gulf of Mexico PC- Pacific Coast FE- Federal Jurisdiction (used because there was an accident on waters subject to the jurisdiction of the Army). MP- Northern Marianas Islands VI- U.S. Virgin Islands PR- Puerto Rico GU- Guam AS- American Samoa
WaterConditions	Calm 6 in and less; Choppy >6 in to 2 ft; Rough >2 ft to 6 ft; Very Rough > 6 ft; Unknown
Wind	Refers to the wind conditions at the time of the accident None = 0 mph Light >0 to 12 mph Moderate >12 to 25 mph Strong >25 to 55 mph Storm >55 mph
Day	1 if the accident happened during the day, 0 otherwise
Visibility	Good, fair, or poor are words used to describe the level of visibility at the time of the accident.
DayofWeek	Indicates the day of the week.
AccidentCause	The primary cause of the accident
AccidentEvent	The events in an accident.
Operatorgender	The gender of the operator. m = male, f = female, u = unreported
OperatorAge	Age of the operator.
OperatorUsingAlcohol	Notes whether the operator was drinking. This information was first collected in 2009. y = yes, n = no
VesselType	The type of vessel.
Length	The length of the boat, expressed in whole numbers.
Operation	The operation of the vessel, what the vessel was doing at time of accident.
Activity	The activities in which the operator/passenger were involved.
YearBuilt	The year the vessel was built.
NumberPersonsonboard	The number of people onboard the vessel.
DeceasedAge	Age of the victim.
DeceasedGender	The gender of the deceased person. M = male, F = female, U = unreported
CauseofDeath	The cause of death. Common entries include drowning, trauma and carbon monoxide poisoning.
DeceasedPFDworn	Indicates whether the victim was wearing a life jacket at the time of recovery. Y = yes, PFD worn; N = not worn
DeceasedRole	The role of the deceased person. Common entries include operator, occupant, and swimmer.
InjuredGender	The gender of the injured person. This information was first collected in 2008. M = male, F = female, U = unreported
InjuredAge	Age of victim.
InjuryType	The primary injury of the injured victim. Common entries include abrasion, contusion, laceration, amputation, back injury, spinal injury, and broken bone.
InjuredRole	The role of the injured person. Common entries include operator, occupant, and swimmer.

The collection of features extracted from the USCG database provide a valuable dataset with information along four key areas of information with respect to boating accidents: accident details, vessels involved, injuries sustained, and deaths occurred.

METHOD

With the merging of database complete, I had organized a specific dataset to answer each my

research questions. The first step was to inspect, wrangle and clean this primary dataset. I used this dataset to conduct exploratory data analysis and become more familiar with each of the features and their distributions. This analysis helped me determine which features I would use for each modeling effort and what transformations were required. Once familiar with the data, I created derivative datasets to support the analysis and help answer each research question by customizing the features and their formats, as well as splitting the data 70/30 train and test for validation.

ANALYSIS

Data Cleaning: Treatment for Missing Values

The data cleaning phase of this analysis revealed the dataset had a total of 85,183 missing values. There were a total of 11,534 records with missing data and 13 columns with missing data. The columns missing data include: OperatorAge, Length, Yearbuilt, NumberPeopleOnboard, DeceasedAge, DeceasedGender, CauseofDeath, DeceasedPFDWorn, DeceasedRole, InjuredGender, InjuredAge, InjuryType, and InjuredRole. However, this is to be expected after moving the database to second normal form. For example, not all accidents resulted in injury or death so for each of those observations, a great deal of data was missing. I began by addressing the columns that are unique to all accidents: OperatorAge, Length, Yearbuilt, NumberPeopleOnboard.

For OperatorAge there were 2,118 records (approximately 18% of the data) missing data. A distribution plot shows the distribution is roughly normal so I decided to use the mean age of 44 to replace the NA values. For Length there were 588 records (approximately 5%) missing data. A distribution plot of this feature shows a right skewed distribution so I used the median of 20 feet for replacement. For Yearbuilt there were 1,189 records (approximately 10%) missing data. A distribution plot of this feature shows a left skewed distribution so I used the median of 2004 for replacement. Lastly, for NumberPeopleOnboard there were 472 records (approximately 4%) missing data. A distribution plot of this feature shows a right skewed distribution so I used the median of 2 people for replacement.

For categorical values I recorded NAs as “Unknown” rather than interpolate from the mode. This decision is based on the business rules used by the Coast Guard where if the reporting does provide a value for a given field, they typically record it as “Unknown” in many cases. Many of these categorical features also required additional cleaning for consistency. For example, the female gender was recorded as both “f” and “F”, so cleaning was required to make the values consistent.

For the features missing data that are conditional on the circumstances of the accident (DeceasedAge, DeceasedGender, CauseofDeath, DeceasedPFDWorn, DeceasedRole, InjuredGender, InjuredAge, InjuryType, and InjuredRole) I filtered for “Death” and “injured”, check for missing values, and replaced the missing values with mean and median estimates where appropriate. For additional details, please reference the Python code file.

Data Cleaning: Duplicates and Outliers

An inspection for duplicates resulted in the finding of 94 duplicate records. The duplicate records were dropped from the dataframe resulting in a new shape of 11,626 rows by 35

columns. Next I needed to check for and address the numeric columns for potential outliers. The details and the mitigating actions taken are highlighted in the table below:

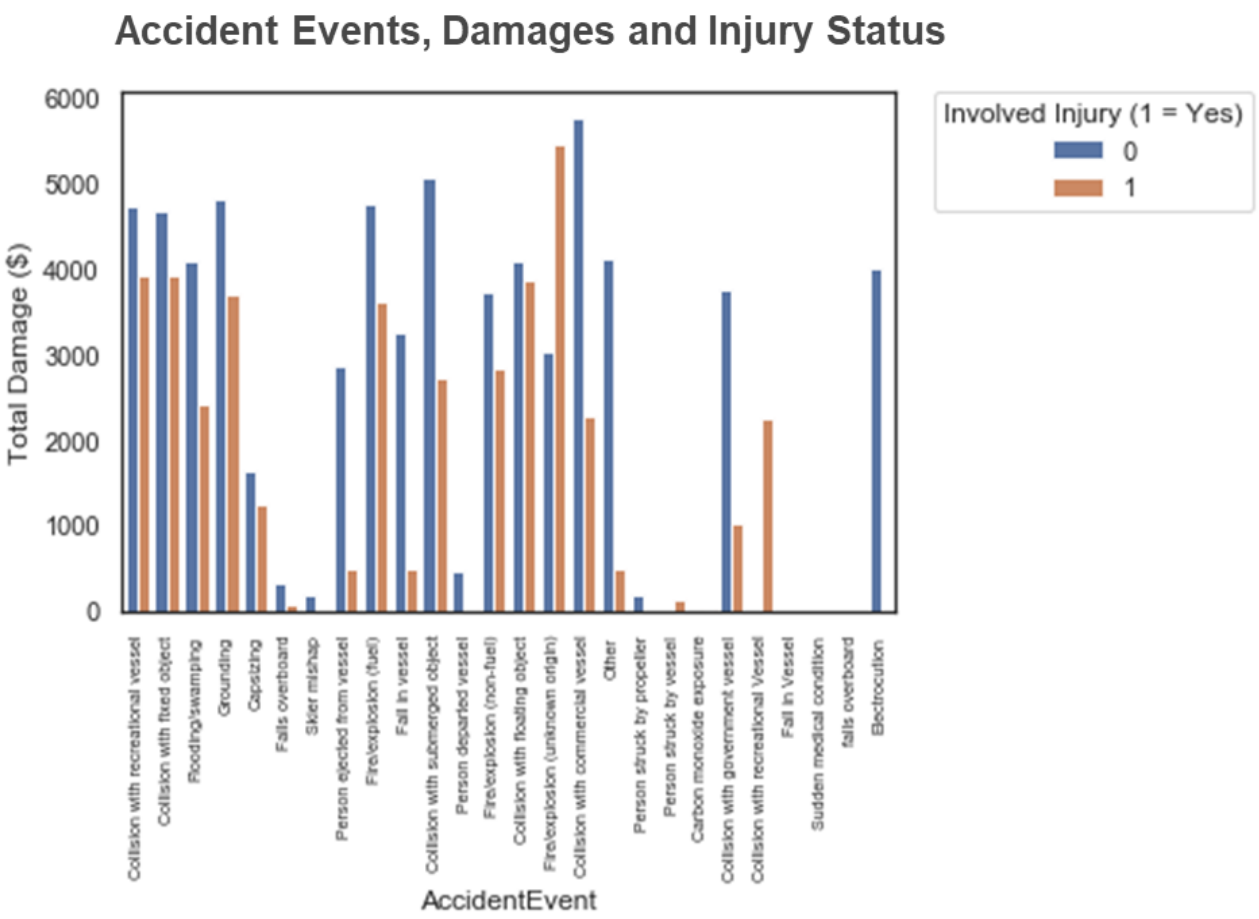
Feature	Outlier	Mitigation
NumberDeaths	Values ranged from 0 to 4, all seem reasonable	None
NumberInjured	Values range from 0 to 17	IQR Fence
NumberVesselsInvolved	Values range from 1 to 23	IQR Fence
TotalDamage	Values range from 0 to 2,303,000	IQR Fence
OperatorAge	Values range from 4 to 91, although low and high values, all seem reasonable considering this data also includes kyacks	None
Length	Values range from 5 to 767	IQR Fence
Yearbuilt	Values range from 1929 to 2019, all seem reasonable	None
NumberPeopleOnboard	Values range from 0 to 69	IQR Fence
DeceasedAge	Values range from 1 to 90, all seem reasonable	None
InjuredAge	Values range from 0 to 91, all seem reasonable	None

Exploratory Data Analysis

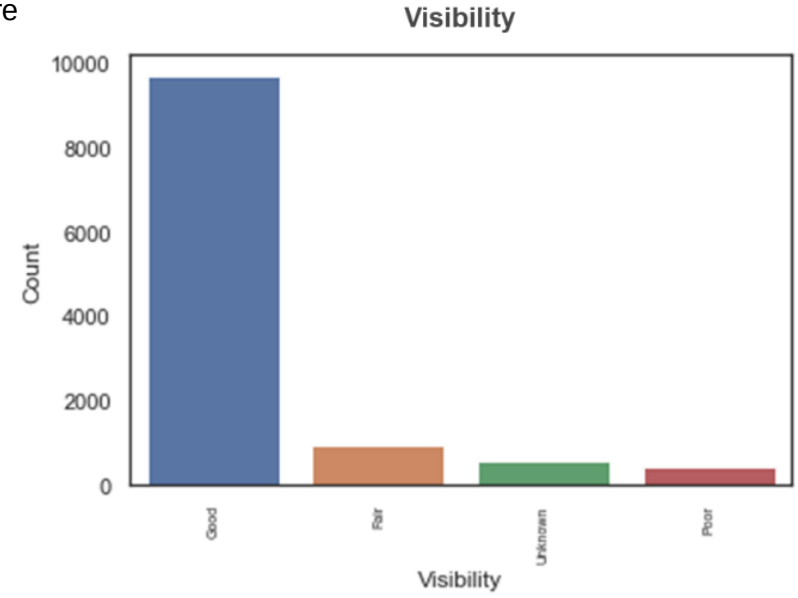
The exploratory data phase consisted of three distinct activities. The first was to create visualizations for both numeric and categorical features, the second was to provide summary statistics for each of the features in the numeric dataset, and the third was to examine the correlations between the numeric features. Each of these actives helped better understand the features, their relationships, and provided intuition during the modeling phase.

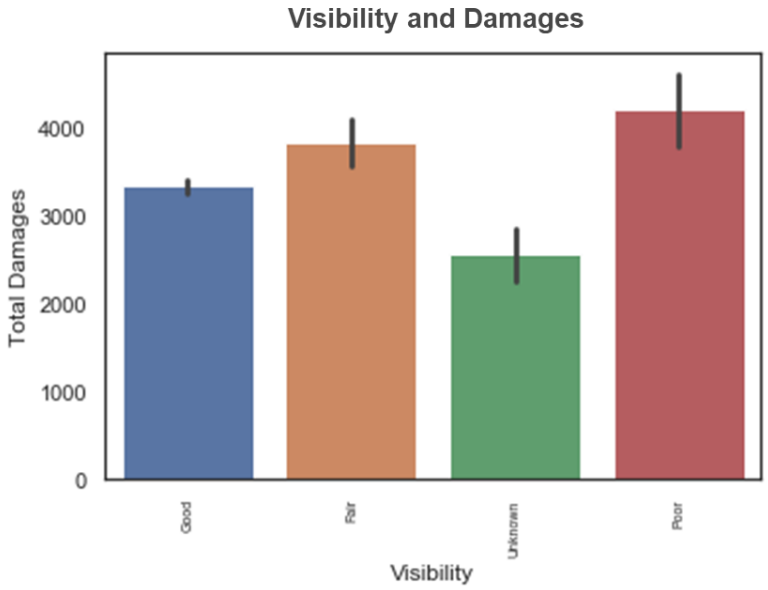
Data Visualizations

I began the exploration by reviewing what sort of accidents are happening, to get an appreciation for the damages they cause, and if these accidents lead to death or injury. The plot below shows the types of accidents, the damages ($) they incur, and are categorized by injury or no injury. Interestingly, it appears that collisions with other vessels is the leading accident event. And, although the plot is not shown, the same can be said for when a death occurs.

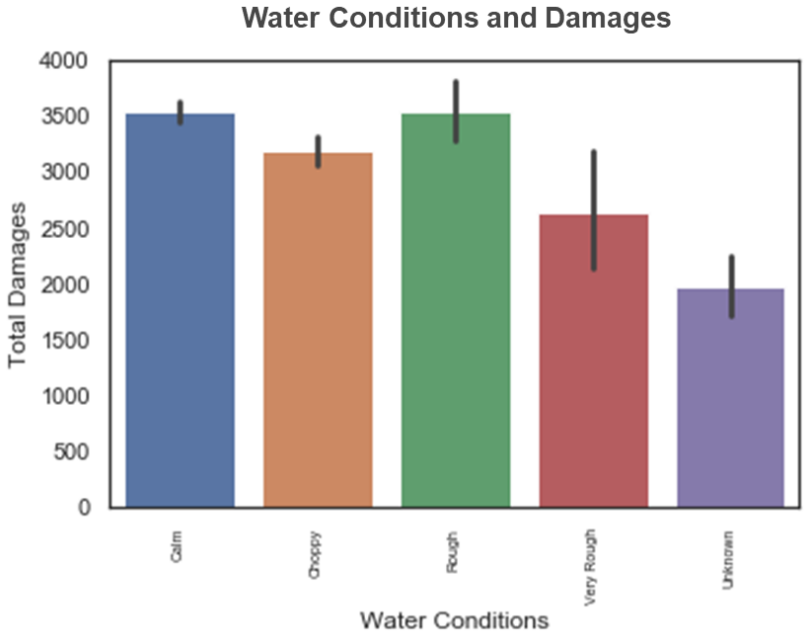


To better understand why boats are colliding I first explored visibility conditions. As we can see from the chart to the right, most often accidents occur when the weather is good. However, if put in terms of damages, we see a different story as depicted in the chart below (next page).

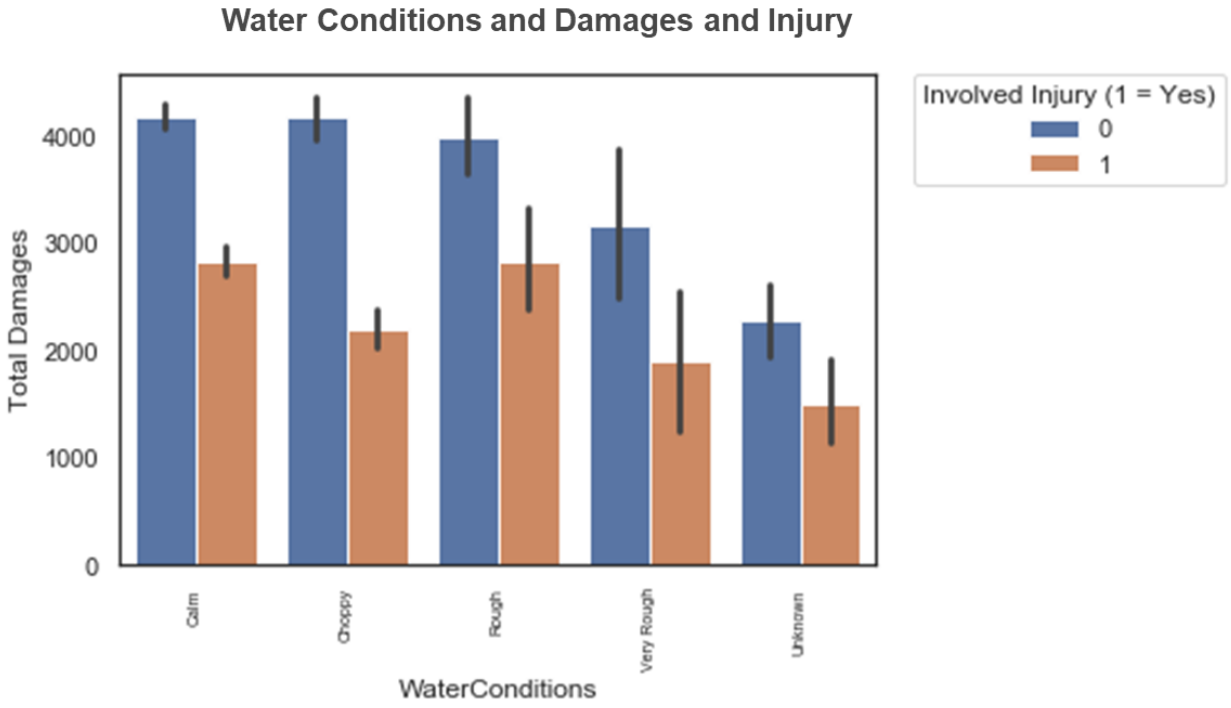




The chart above helps show that with respect to damages, poor visibility plays a big role. With boating, visibility can also be reduced by water conditions. Large waves or spray can severely degrade a captain's visibility. Below is a look at the impact of water conditions.

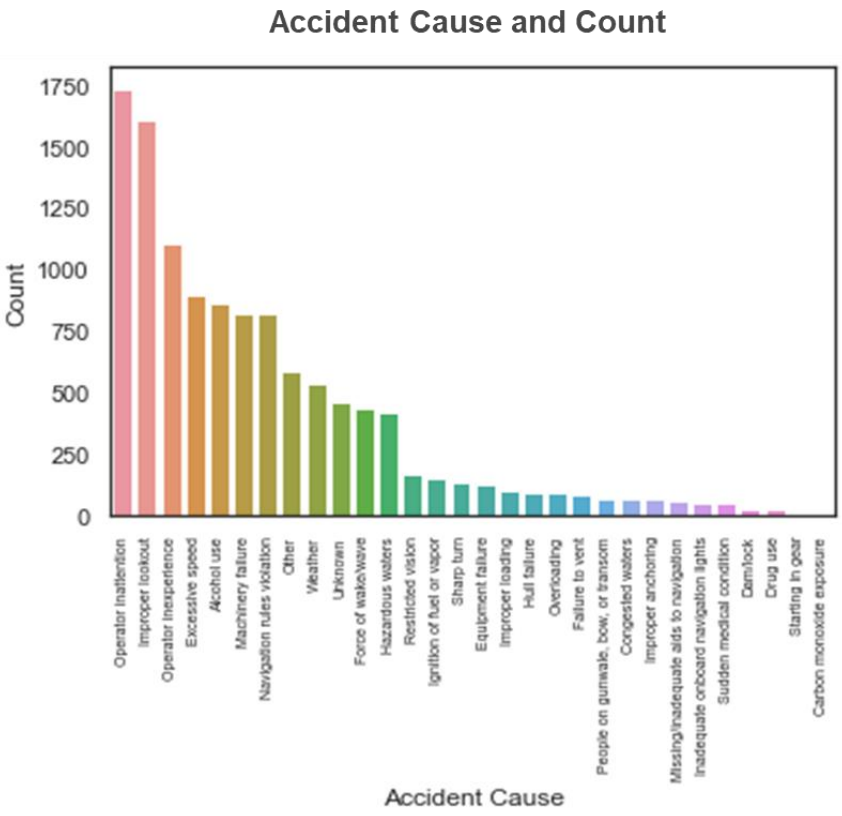


And if we review this with respect to injuries (below), we can see that rough and very rough water conditions combined make up most of the damages, both with or without injury.

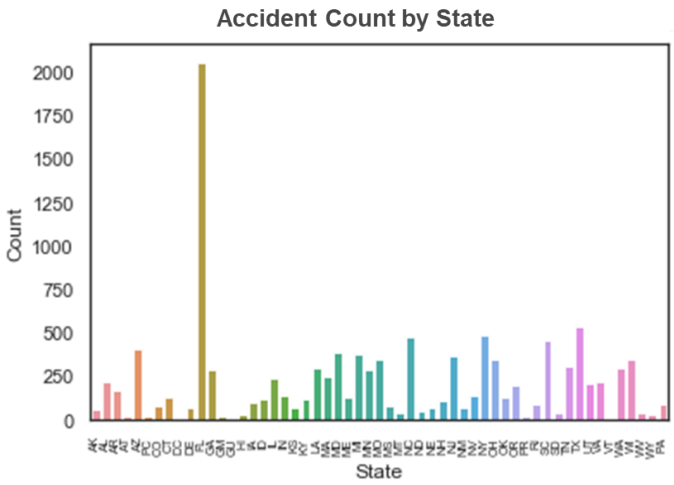


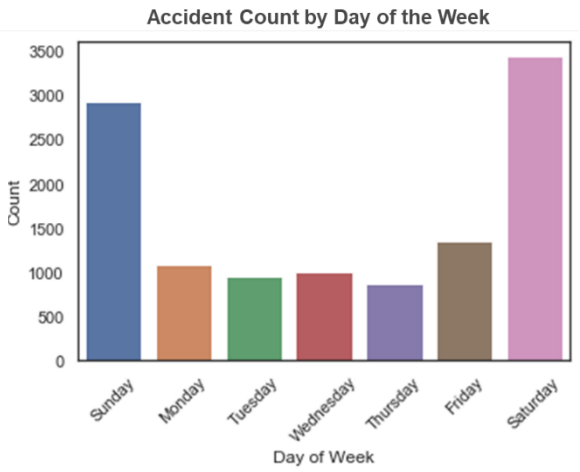
Based on these graphs, it certainly looks like visibility and weather play an important role in boat collisions. Additionally, a review of the reported causes provides additional detail on why these boats may be running into each other.

As we can see in the plot to the right, Operator inattention is the leading factor for accidents. Additionally, the close second is improper lookout. These findings show that most accidents involve boats colliding, the cause of which is poor weather and the Captain failing to pay attention or properly assigning a lookout.

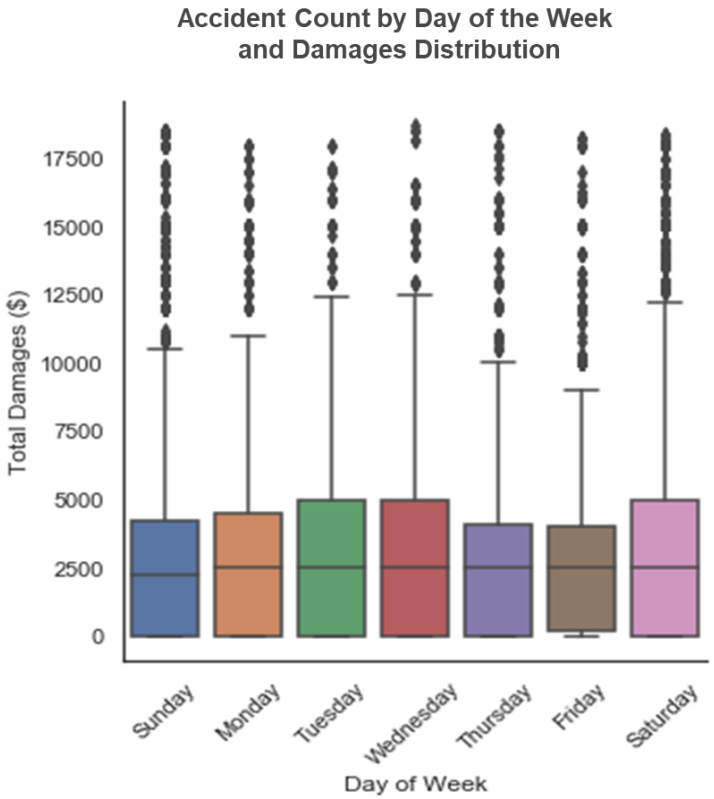


A few more contextual graphs include where and when these accidents are happening. The two charts below shows in which states boating accidents are more common, and on which days of the week they occur. No surprises here, Florida is certainly a very popular boating state and the weekends are the times recreational boaters use their boats most often.





An additional look at the damages by day of week can be seen below. We can see the median damage cost is about the same for all days of the week, with are larger average cost on Tuesdays, Wednesdays and Saturdays.



Summary Statistics

A review of the numeric variables and their descriptive summaries are provided below:

	NumberDeaths	NumberInjured	NumberVesselsInvolved	TotalDamage	NumberPeopleOnboard
count	11626.0	11626.0	11626.0	11626.0	11626.0
mean	0.0	0.0	1.0	3365.0	2.0
std	0.0	1.0	1.0	3887.0	1.0
min	0.0	0.0	1.0	0.0	0.0
25%	0.0	0.0	1.0	0.0	1.0
50%	0.0	0.0	1.0	2500.0	2.0
75%	0.0	1.0	2.0	4700.0	3.0
max	4.0	2.0	3.0	18700.0	6.0

	OperatorAge	Length	Yearbuilt	DeceasedAge	InjuredAge
count	11626.0	11626.0	11626.0	1404.0	5506.0
mean	44.0	19.0	2002.0	46.0	35.0
std	15.0	6.0	12.0	18.0	16.0
min	4.0	5.0	1929.0	1.0	0.0
25%	33.0	15.0	1996.0	32.0	23.0
50%	44.0	20.0	2004.0	46.0	33.0
75%	54.0	22.0	2012.0	60.0	46.0
max	91.0	37.0	2019.0	90.0	91.0

We can see that out of the 11,626 reported incidents in 2017 and 2018, 1,404 (12%) resulted in at least one death, while 5,5506 (47%) resulted in at least one person being injured. A detailed breakdown by type and average age for both deaths and injuries are listed below.

Cause of Death and Count

CauseofDeath	
Carbon monoxide poisoning	11
Cardiac arrest	25
Drowning	854
Hypothermia	19
Other	3
Trauma	242
Unknown	250

Cause of Death and Avg Age of Deceased

CauseofDeath	
Carbon monoxide poisoning	42.0
Cardiac arrest	63.0
Drowning	47.0
Hypothermia	38.0
Other	50.0
Trauma	41.0
Unknown	46.0

Injury Type and Count

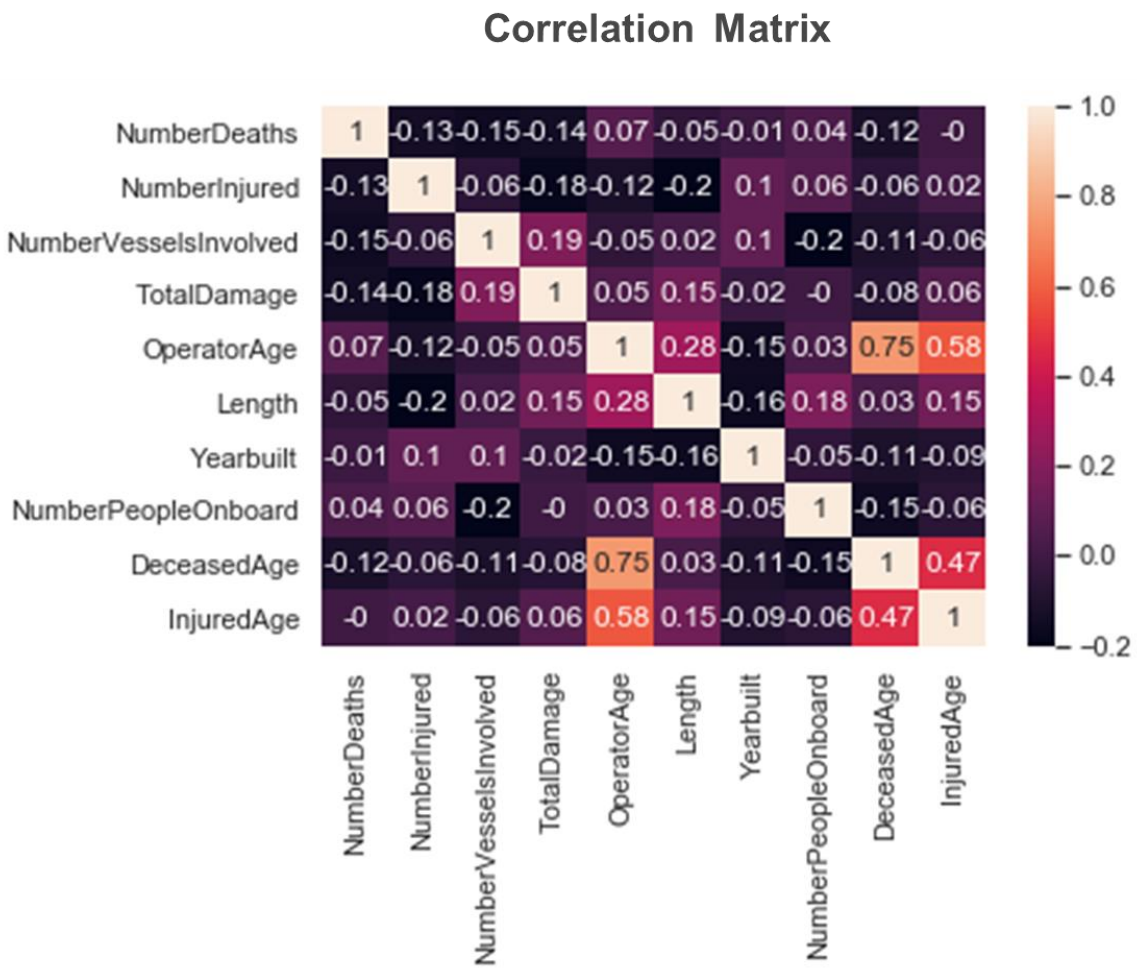
InjuryType	
Amputation	47
Broken bone	861
Burn	153
Carbon monoxide	21
Concussion	443
Cut	1032
Dislocation	91
Hypothermia	326
Internal organ injury	211
Other	12
Scrape/bruise	616
Shock	14
Spinal cord injury	83
Sprain/strain	226
Unknown	1370

Injury Type and Avg Age of Deceased

InjuryType	
Amputation	36.0
Broken bone	37.0
Burn	36.0
Carbon monoxide	26.0
Concussion	30.0
Cut	34.0
Dislocation	34.0
Hypothermia	42.0
Internal organ injury	38.0
Other	39.0
Scrape/bruise	34.0
Shock	43.0
Spinal cord injury	39.0
Sprain/strain	36.0
Unknown	34.0

Correlation Analysis

The correlation matrix plot below shows the correlation between numeric features in the dataset. The color intensity and number indicate the strength of the correlation. As seen below, there do not appear to be any major correlations between the numeric variables. This is good to know prior to regression modeling as highly correlated features often lead to a regression model having high multi-collinearity. The only correlation of note here is that Operator Age and Deceased Age have the greatest positive correlation. This correlation could be explained by the potential situation where boat operators are often in the company of people their own age when someone dies, or the operator themselves is the one that dies.



Modeling

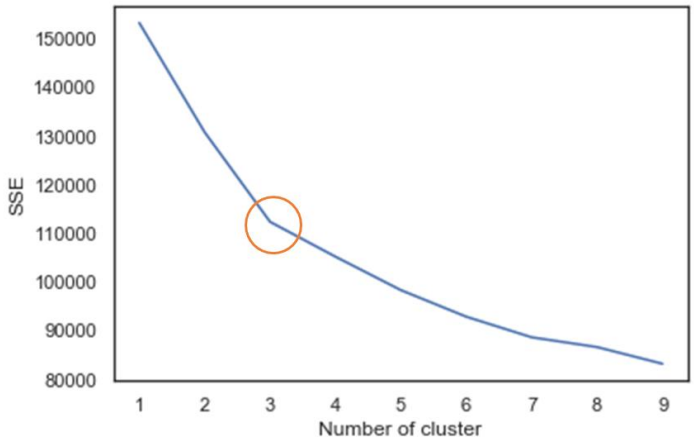
Cluster Analysis

The objective of the cluster analysis was to increase familiarization with the data, and to uncover potential unseen relationships in the data that may not have been recognized during EDA. For this cluster analysis used Kmeans as part of the sklearn.cluster submodule. To support this analysis, I created a copy of the primary dataframe and named it clustdf. Next I dropped several of the features I decided not to include in the analysis. These features include:

Date, ReleasableAccidents_Year, Time, State, Wind, AccidentCause, AccidentEvent, VesselType, Operation, Activity, CauseofDeath, DeceasedPFDWorn, DeceasedRole, InjuredGender, InjuryType, DeceasedAge, DeceasedGender, InjuredRole, InjuredAge, WaterConditions, Visibility, DayofWeek, OperatorGender, TotalDamage

With those features removed, I then normalized the data using standardize Z transformation. This is an important step as clustering analysis is grounded in distance formulas and having a wide range of values between numeric features in the dataset may result in some features having more weight than they should. After normalizing, Dummy variables for WaterConditions, Visibility, DayofWeek, OperatorGender were added to the dataset and it was renamed clustdf_dummies. I added these categorical features to the clustdf dataframe because I was curious on how these features impact the clustering of the data.

With a complete dataframe, the next step was to identify how many clusters to use in the clustering analysis. The technique I used was to initialize the analysis by iterating over several combinations of cluster (1-10) and compare the MSEs in a plot. As seen in the plot below, a good "elbow" in the curve appears at the 3 cluster mark. As such, I decided to run my cluster analysis with K = 3.



After resetting my dataframe, I fit a new cluster analysis model with K=3, and merged the cluster assignment of each observation back to the dataframe and renamed it clustdf_dummiesCluster. As a function of unsupervised learning, the labels for our data and clusters are unknown. However, taking a look at a few of the features of interest with respect to their assigned cluster will tell us a little bit about how the data are grouped together. Several of the feature assignments can be seen below.

Death (1 = Yes)	0	1
Cluster		
0	5040	0
1	5182	0
2	0	1404

Injury(1 = Yes)	0	1
Cluster		
0	0	5040
1	5074	108
2	01046	358

Day (1 = Yes)	0	1
Cluster		
0	784	4256
1	964	4218
2	366	1038

WaterConditions_Rough	0	1
Cluster		
0	4747	293
1	4746	436
2	1238	166

Visibility_Poor	0	1
Cluster		
0	4876	164
1	4986	196
2	1335	69

Based on the tables above, one could generally conclude the following personas for each cluster:

Cluster 0: Consist of accidents that lead to injury, occur primarily during the day, least likely to occur during rough water conditions, and least likely to occur during poor visibility.

Cluster 1: Consists of accidents that did not lead to death, and may or may not have lead to injury, occur during the day, do not usually occur during rough water or poor visibility (but are more likely than cluster 0 to occur during rough water and poor visibility conditions).

Cluster 2: Consists of accidents that lead to death, often have injuries, occur primarily during the day but are more likely than the other clusters to occur during the night, more likely than the other clusters to occur during rough water conditions, and more likely than the other clusters to occur during poor visibility.

OLS Regression Model

The OLS regression modeling was conducted to help answer research question one by trying to identify some the leading features that explain the variation in the total cost of damage for accidents. Recall, prior to addressing outliers, there was significant variance between the min and the max of the total damage feature. To help better understand this variation I performed an OLS regression and employed backwards elimination using the P-value approach. I began with the full model and at each step, I removed the least significant feature based on P-value. As long as the Adjusted R squared value did not reduce, I continued the process. My stepwise process concluded on the fourth step, model 4. The general equation for my model is as follows:

"TotalDamage = B0 + B1*NumberDeaths + B2*NumberInjured + B3*NumberVesselsInvolved + B4*DayofWeek + B5*AccidentEvent + B6*OperatorUsingAlcohol + B7*VesselType + B8*Yearbuilt" from the data = trainreg)

This model resulted in a R-squared value of 0.21 the MSE is 502620146.89. Below is the detailed output for the modeling:

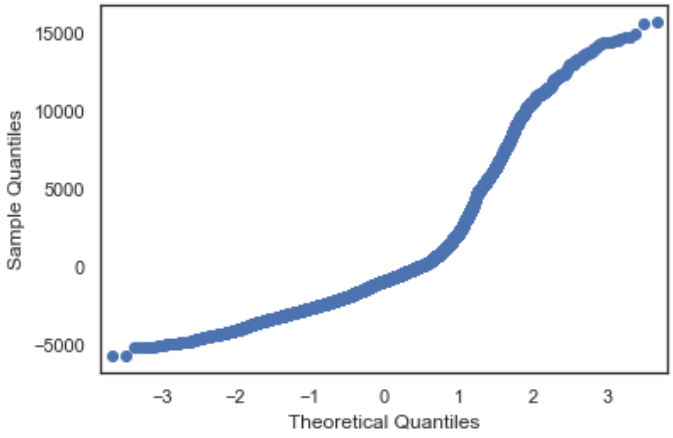
OLS Regression Results

Dep. Variable:	TotalDamage	R-squared:	0.214
Model:	OLS	Adj. R-squared:	0.209
Method:	Least Squares	F-statistic:	43.17
Date:	Thu, 02 Jul 2020	Prob (F-statistic):	0.00
Time:	11:08:55	Log-Likelihood:	-77725.
No. Observations:	8138	AIC:	1.556e+05
Df Residuals:	8086	BIC:	1.559e+05
Df Model:	51		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.36e+04	6916.255	-3.412	0.001	-3.72e+04	-1e+04
DayofWeek[T.Monday]	217.8235	167.487	1.301	0.193	-110.495	546.142
DayofWeek[T.Saturday]	383.4560	133.779	2.866	0.004	121.215	645.697
DayofWeek[T.Sunday]	200.0776	136.535	1.465	0.143	-67.566	467.721
DayofWeek[T.Thursday]	411.5504	179.404	2.294	0.022	59.872	763.229
DayofWeek[T.Tuesday]	250.4134	175.563	1.426	0.154	-93.736	594.562
DayofWeek[T.Wednesday]	218.5010	171.695	1.273	0.203	-118.065	555.067
AccidentEvent[T.Carbon monoxide exposure]	-2491.7208	797.070	-3.126	0.002	-4054.183	-929.259
AccidentEvent[T.Collision with commercial vessel]	1026.4628	482.301	2.128	0.033	81.028	1971.898
AccidentEvent[T.Collision with fixed object]	1278.1359	230.496	5.545	0.000	826.304	1729.968
AccidentEvent[T.Collision with floating object]	747.8207	411.394	1.818	0.069	-58.618	1554.260
AccidentEvent[T.Collision with government vessel]	-1312.6309	858.340	-1.529	0.126	-2995.198	369.936
AccidentEvent[T.Collision with recreational Vessel]	175.5704	1985.413	0.088	0.930	-3716.349	4067.490
AccidentEvent[T.Collision with recreational vessel]	1186.6034	235.483	5.039	0.000	724.995	1648.211
AccidentEvent[T.Collision with submerged object]	1027.7095	318.432	3.227	0.001	403.501	1651.918
AccidentEvent[T.Fall in Vessel]	-3032.2566	2424.213	-1.251	0.211	-7784.338	1719.825
AccidentEvent[T.Fall in vessel]	-2034.2229	310.091	-6.560	0.000	-2642.081	-1426.364
AccidentEvent[T.Falls overboard]	-2051.6347	255.511	-8.030	0.000	-2552.502	-1550.767
AccidentEvent[T.Fire/explosion (fuel)]	1084.4984	295.104	3.675	0.000	506.018	1662.979
AccidentEvent[T.Fire/explosion (non-fuel)]	527.7423	364.457	1.448	0.148	-186.687	1242.171
AccidentEvent[T.Fire/explosion (unknown origin)]	21.8942	440.527	0.050	0.960	-841.653	885.441
AccidentEvent[T.Flooding/swamping]	749.6385	240.044	3.123	0.002	279.090	1220.187
AccidentEvent[T.Grounding]	1030.0970	250.619	4.110	0.000	538.818	1521.376
AccidentEvent[T.Other]	-1381.7008	481.049	-2.872	0.004	-2324.682	-438.720
AccidentEvent[T.Person departed vessel]	-2272.9819	339.897	-6.687	0.000	-2939.267	-1606.697
AccidentEvent[T.Person ejected from vessel]	-1263.2860	282.450	-4.473	0.000	-1816.962	-709.610
AccidentEvent[T.Person struck by propeller]	-2768.0370	526.549	-5.257	0.000	-3800.209	-1735.865
AccidentEvent[T.Person struck by vessel]	-2189.3364	566.958	-3.862	0.000	-3300.721	-1077.952
AccidentEvent[T.Skier mishap]	-2627.7695	282.450	-9.303	0.000	-3181.444	-2074.095
AccidentEvent[T.Sudden medical condition]	-2784.2519	3425.073	-0.813	0.416	-9498.277	3929.774
AccidentEvent[T.falls overboard]	-141.2401	3466.632	-0.041	0.968	-6936.732	6654.251
VesselType[T.Auxiliary sail]	1616.9193	492.798	3.281	0.001	650.908	2582.931
VesselType[T.Cabin motorboat]	800.9020	460.010	1.741	0.082	-100.835	1702.639
VesselType[T.Canoe]	-1861.0406	560.720	-3.319	0.001	-2960.197	-761.884
VesselType[T.Houseboat]	210.8175	572.593	0.368	0.713	-911.611	1333.247
VesselType[T.Inflatable]	-1889.1540	724.345	-2.608	0.009	-3309.056	-469.252
VesselType[T.Kayak]	-1945.4396	514.234	-3.783	0.000	-2953.471	-937.409
VesselType[T.Open motorboat]	789.4167	449.823	1.755	0.079	-92.352	1671.186
VesselType[T.Other]	640.1023	694.399	0.922	0.357	-721.099	2001.304
VesselType[T.Personal watercraft]	-246.9945	457.270	-0.540	0.589	-1143.361	649.372
VesselType[T.Pontoon]	612.1494	474.292	1.291	0.197	-317.585	1541.884
VesselType[T.Rowboat]	-1243.2868	629.308	-1.976	0.048	-2476.892	-9.681
VesselType[T.Sail (only)]	-132.7202	642.501	-0.207	0.836	-1392.187	1126.746
VesselType[T.Sail (unknown propulsion)]	274.5374	965.592	0.284	0.776	-1618.272	2167.347
VesselType[T.Standup Paddleboard]	-348.8437	1177.519	-0.296	0.767	-2657.084	1959.396
VesselType[T.Standup paddleboard]	-1293.5084	1228.535	-1.053	0.292	-3701.753	1114.736
VesselType[T.Unknown]	122.0727	538.995	0.226	0.821	-934.495	1178.641
NumberDeaths	-607.3071	90.765	-6.691	0.000	-785.229	-429.385
NumberInjured	-657.2618	60.892	-10.794	0.000	-776.626	-537.898
NumberVesselsInvolved	293.4113	105.221	2.789	0.005	87.151	499.672
OperatorUsingAlcohol	661.8763	146.259	4.525	0.000	375.172	948.581
Yearbuilt	12.8352	3.444	3.727	0.000	6.084	19.586

Generally, we see that all things being equal, an accident that occurs on Thursday cost the most, and on average $411 more than an accident on Friday (the reference group). With respect to accident events, all things being equal, colliding with a fixed vessel results in the highest cost on average, followed closely by collision with another recreational vessel, while falling in the vessel, naturally, leads to the least amount of cost on average. Additionally, we see that, all thing held constant, for each additional boat involved in the accident, the average cost of damage increases by an average of $293.

A quick inspection of the residuals via a Q-Q plot shows that our assumption of relatively normal residuals is slightly challenged.



Decision Tree Classification Model

The Decision Tree Classification Model was conducted to help answer research question two by developing a model that can accurately predict if an accident is likely to have resulted in death based on a collection of given features. This model is helpful in that if the Coast Guard is dispatched to a marine call where the status of the people onboard is unknow, a few details about the accident and the water conditions may help predict if there is likely to be a casualty.

For this modeling effort I began by creating dummies for the DayofWeek, AccidentEvent, and VesselType features and joining them to the regdf datframe and titling it regdf_dummies. I then dropped the original values for these dummies. Additionally, I dropped the NumberDeaths, NumberInjured, and Injury (1 = Yes) from the dataframe. Failing to remove NumberDeaths would result in an accuracy of 1 since the model would assume, naturally, if there is at least a value of 1 in the NumberDeaths feature, there was a death. I decided to remove NumberInjured, and Injury (1 = Yes) from the data as well to simulate the Coast Guard having as few details about the occupants involved in the accident as possible, but trying to predict if a death will be discovered at the accident when first responders arrive.

After splitting train/test (70/30) and training the model, the regression classification tree performed pretty well at making predictions on the test dataset. Using sklearn.metrics and the confusion matrix class we generated the following confusion matrix:

DTC Confusion Matrix

	No Death	Death
True No Death	2872	188
True Death	172	256

The confusion matrix is a table that categorizes predictions made by the classifier according to whether they matched the actual value recorded in the test data. Confusion matrices are extremely helpful in providing details in the way prediction errors were made, normally referred to as Type I and Type II errors, or False Positive and False Negative errors. As we can see above, there are several instances where the accident was classified death, but was actually no death (False Positive), or a more sensitive situation, when there was a death but it predicted no death (False Negative).

Additionally, using scikit-learn we can also look at the classification report to further understand the results displayed sklearn.metric class classification_report I was able to review other important model performance metrics. This output is seen below.

```
Clasification report:
              precision    recall

           0       0.94      0.94
           1       0.58      0.60

    accuracy
   macro avg       0.76      0.77
weighted avg       0.90      0.90
```

Precision is defined as the proportion of positive predicted examples that are actually positive. This measure relates to the "trust" of the model in making correct predictions. A precise model will only make positive class predictions in cases that are very likely to be positive. A model with low precision would be willing to take more risk (require lower probability of certainty) in making predictions and although it will select a larger pool of positive predictions, the results would be less trustworthy. Recall is a measure of how complete the results are. This is calculated by dividing the number of true positives by the total number of positives. A model with high recall captures a large portion of positive examples with its wide breadth, but it comes at a tradeoff in precision. Generally, this model is pretty good at predicting when deaths do not occur, but suffers slightly at making trustworthy predictions on actual deaths. However, overall the model performs well with an overall accuracy of .90.

CONCLUDING REMARKS

Boating can certainly be a lot of fun, however, this analysis has shown that it can also be dangerous. Although just under half of all boating accidents reported to the Coast Guard result in injury, and only 12% result in death, rough seas, poor visibility, and other boaters can turn a fun day on the water catastrophic. However, this analysis has shown that the power to avoid

boating accidents may in most cases lie in the hands of the operator. They can improve their efforts to pay attention, properly assign lookouts to help keep watch for other boats and obstacles, or they can elect to stay at the dock when weather and visibility are degraded. This analysis has provided details on how certain factors impact the fallout from an accident in terms of total damages ($). Lastly, through Decision Tree Classification, this analysis was successful at producing a model with 90% accuracy on predicting the death/no death outcome of an accident. There is certainly much more to learn from this data, and with access to more longitudinal data, some interesting trend analysis could certainly help show if the boating community is becoming more or less safe over time.

****Presentation with audio voice track provided via separate PowerPoint file*