Analyzing the Causes of Road Accidents

Alexios Prodanas

BIT 5524

Jeremy Don Sudweeks

October 19th, 2021

Background information:

Road accidents are a growing problem to our society. Every year we have multiple deaths and injuries which are related to road accidents. To become more precise, on an average there are 11 million car accidents in the U.S every year[1] which are related to driver errors, weather conditions, road conditions and car malfunctions. This is a serious problem because many people lose their lives or they get seriously injured. Every year, more than 1.35 million people lose their lives in the United States because of road accidents and 20-50 million people suffer from non-fatal injuries which are often result in long term disabilities[2]. Car accidents have a collective impact on our society since they can traumatize people and their families. Moreover, studies have shown that a decrease in road accidents can boost income growth. Road traffic crashes affect medium- and long-term growth prospects by removing prime age adults from the workforce, and reducing productivity due to the burden of injuries[3]. Therefore, it is necessary to analyze the data which are related to road accidents and try to find the relationships between the accidents and the conditions that caused them.

Problem Statement:

The increased number of road accidents is growing every year, which also causes the increase of the total number of fatalities and non-fatal injuries. It would be ideal to decrease the number of road accidents that take place every year because it will also minimize the possibility of new deaths and injuries which are related to road accidents.

Analysis Plan:

There are many ways to approach the analysis to meet the solution for our problem statement. One study leveraged junction accident data from over 1,000 accidents and utilized cluster analysis to identify similar crash characteristics and their respective severity. This analysis was published with the intent to improve road safety by fixing junctions with risky attributes.[4.] Another research article used regression analysis to estimate crash safety mechanisms in lightweight cars. This research aimed to identify the components with the most energy (absorption, so cars could become lighter and more efficient while maintaining their safety standards[5]. The analysis of our project will use a variety of analytical methods to assist in answering our problem statement. We can use descriptive statistical methods to understand our road accident dataset, gauge direction and range of accidents, and illustrate better which methods can best assist us. We can use cluster analysis to try to define relationships between high risk factors for fatal accidents. Furthermore, we can use predictive analytics such as regression to try to predict fatal accidents in the future if lower risk techniques are used.

---

[1] https://www.askadamskutner.com/motorcycle-accident/how-do-car-accidents-compare-to-motorcycle-accidents/

[2] https://www.asirt.org/safe-travel/road-safety-facts/

[3]https://www.worldbank.org/en/news/press-release/2018/01/09/road-deaths-and-injuries-hold-back-economic-growth-in-developing-countries#:~:text=Using%20detailed%20data%20on%20deaths,over%20a%2024%2Dyear%20horizon.

[4]https://www.sciencedirect.com/science/article/abs/pii/S0001457517302464

[5]https://www.worldscientific.com/doi/abs/10.1142/S0217979208050851

Expected Results:

Throughout our analysis we expect to find correlations between the potential factors that cause car accidents and the amount of accidents occurring. They may be more prevalent throughout a specific time of the year or day, while different types of weather are occurring such as rain or snow, attributable to car malfunctions, etc. One thing we can expect to find on top of that could be the specific factors that increase fatal injuries. In an analysis that was published to the National Center for Biotechnology Information, factors that increased fatal injuries included, but are not limited to, drinking and driving, impairment by alcohol, not using a seatbelt resulting in ejection from the vehicle, and driving over the speed limit[6]. These will be all interesting factors we can observe and expect to see very similar results to this study. It is not a secret that there are now many more distractions while driving as technology has become more advanced. Cell phones are a distraction for drivers who use them while operating their vehicle and can be a major factor in causing accidents. According to the National Safety Council, cell phone use while driving leads to 1.9 million car accidents a year and 1 out of every 4 in the US is caused by texting and driving[7]. Another expectation could be the result of an analysis done on phone use in the car. Seeing an association between that and car accidents would be no surprise at all for us.

This analysis not only can bring light upon the factors that cause accidents and potential ways to decrease the amount of them, but it can also open our eyes to potential things we may do while driving which create a situation more likely to end in an accident. Everyone has either been in a car accident or knows someone who has been in one. Doing this analysis can educate us on these factors and provide us with the knowledge necessary to understand the risks while driving in specific situations and possible measures we can take in order to prevent an accident from occurring. We expect to be thorough with our analysis, create clean and easy data to use, and conduct relevant research in order to produce impactful results for ourselves and others to learn from.

---

[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1067816/

[7] https://www.edgarsnyder.com/car-accident/cause-of-accident/cell-phone/cell-phone-statistics.html

Data Summarization, Exploration, Results and Findings

Car accidents in our society are all too prevalent and affect so many people nationwide on a daily basis. Either someone you know or you yourself have more than likely been in an accident before which could have resulted in serious injury or even death. Our main goal as a group here is to find out what factors are most prevalent in causing severe accidents on the road. Going through a dataset with multiple factors that cause accidents could provide us with further insight on the main causes for the more severe accidents to hopefully give us a better understanding on what to look out for when going out on the road. The main business goal for this project is to find what factors cause the most accidents along with the highest severity and potentially come up with a solution that could reduce the overall number of road accidents we see on a daily basis in the United States.

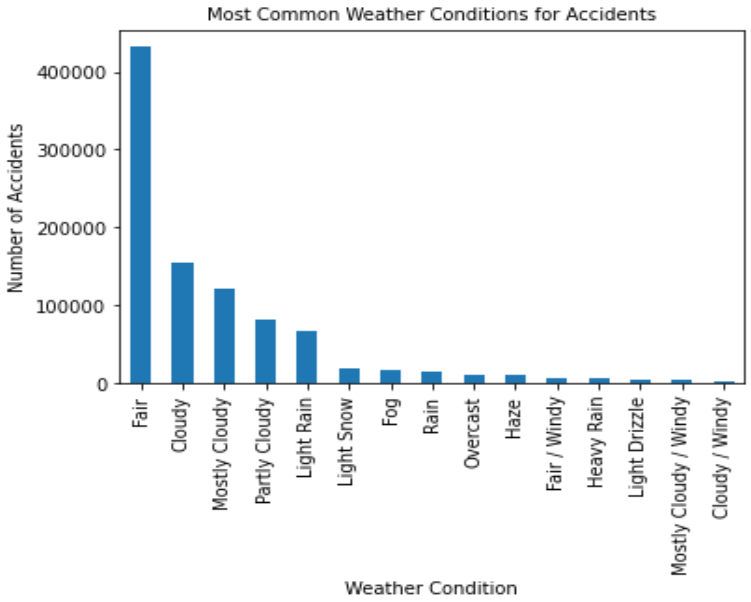
After going through our potential datasets mentioned above we decided to go with the first dataset listed which is titled "US accidents (2016-2020)." This data was collected from the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. It was collected in 48 of the 50 US States from February 2016- December 2020. The dataset was collected in order to compare conditions with the severity of accidents in order to predict what conditions cause the most accidents. In this dataset there were a total of 47 categories ranging from any potential factor that could have an effect on causing an accident, where the accidents happened, what time it occured, weather conditions, the type of roadway in which the accident occured, so on and so forth. These factors can all play a role in car accidents and this dataset will allow us to explore a more in depth analysis for the business intelligence problem at hand. We ended up reducing the number of factors in the dataset since it was incredibly large from 47 to 12 categories. This cleaned version of the dataset was made in order to make it easier to work with as well as use the factors we felt would play the largest role in causing accidents. Since the data was collected from the years 2016-2020 it is relevant to today's society and a potential solution can be immediately implemented in order to reduce accidents as a whole. The data dictionary provided will include only the categories from the cleaned dataset that was used. The 12 categories used were Severity, Start_Time, State, City, Temperature(F), Visibility(mi), Weather_Condition, Pressure(in), Humidity(%), Wind_Speed(mph), Precipitation(in) and one we created called Severity Text. Weather and location are some of the largest known factors to play a role in accidents so looking for patterns with specific weather conditions in certain locations could allow us to narrow down what conditions drivers need to express more caution in.

While cleaning the dataset, as mentioned above we got rid of variables we felt would not have a large impact on the accident itself or would not be of use in the analysis in regards to our business intelligence problem. We also added a variable that provided text information to the severity column in order to make it easier to understand. The severity column had numbers listed from 1-4 with 1 having the least impact on traffic and 4 having the most significant impact on traffic so we added that to the dataset for clarification purposes. The next part of cleaning that was done was checking for null values within the data in order to remove them so they would not adversely affect the results of our analysis. There were null values found in Temperature(F), Visibility(mi), Weather_Condition, Pressure(in), Humidity(%), Wind_Speed(mph), and Precipitation(in) so they were all removed to maintain the integrity of the data and provide us with the best results possible. The next step in cleaning the data was to identify potential outliers in the set so we checked the maximum values within the data and found there to be outliers for the Wind_Speed (mph) category with an entry coming in at 984 miles per hour.

After seeing this we set a threshold for the speed at 100 mph in order to ensure reasonable results. This was the final step in cleaning the data so we were able to move on to actually analyzing the data to search for potential results.

The first step in the data exploration was to find out which weather condition had the most amount of accidents involved. Below you can see the graph showing some surprising results. The weather that the most accidents occured in was fair conditions over 400,000 accidents. This was followed by cloudy and mostly cloudy which each had well over 100,000 accidents. Rain, fog, and light snow which would have been predicted to be higher on the list did not come into the rankings until 8th, 7th, and 6th respectively. This could be a potential indication that weather might not be as much of a factor when looking into car accidents as we have originally thought. It may just be more of the fact that people are driving recklessly or not paying as much attention to the road as they should be which causes their accidents but it is early in the analysis and we have much more to look into.

Figure 1: Number of Accidents vs Weather condition



Data Dictionary

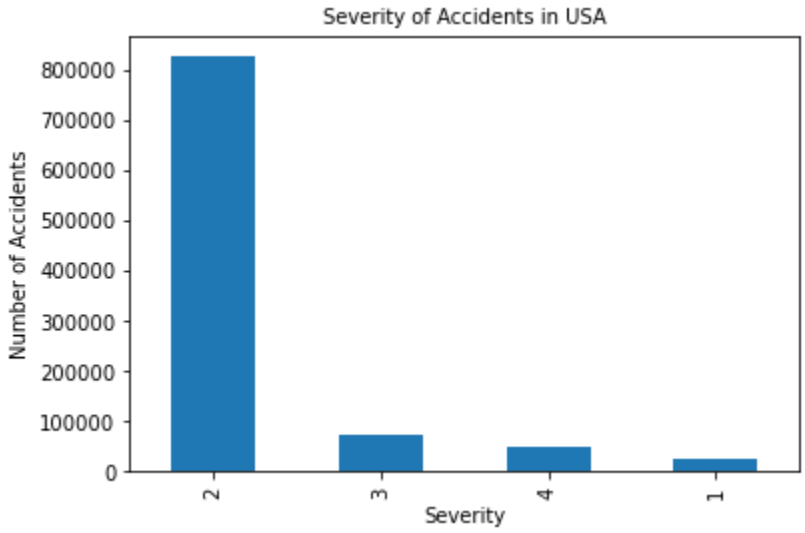
Name of Variable	Variable Definition	Data Type	Range of Values

Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	integer	1-4
Start_Time	Shows start time of the accident in the local time zone.	object	2/8/16 0:37-12/31/20 23:28
State	Shows the state in the address field.	object	AL-WY
City	Shows the city in the address field.	object	Abbeville-Zwingle
Temperature	Shows the temperature (in Fahrenheit).	float	(-89)-113
Visibility	Shows visibility (in miles).	float	0-10
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	object	NA
Humidity	Shows the humidity (in percentage).	float	2-100
Pressure	Shows the air pressure (in inches).	float	0.02-58.04
Wind_Speed	Shows wind speed (in miles per hour).	float	0-98.0
Precipitation	Shows precipitation amount in inches, if there is any.	float	0-24.0

Severity Text	Shows the severity of the accident in relation to the original severity column in terms of a short delay (1), medium delay (2), high delay (3), and heavy delay (4).	object	Short delay-heavy delay

After exploring the dataset and reducing it to one that is ready for analysis, our team was ready to use python to begin to answer our business intelligence problem. Ultimately, our team decided to focus on weather-related independent variables to use on the explanatory variable of severity: in other words, do certain weather conditions affect the severity of a vehicle accident. By graphically illustrating the severity variable, as seen in figure 2, the "2" severity is by far the most common in the country. This, like the weather condition variable, helps to explain the intuitive response that the majority of accidents come in standard situations: fair weather, moderately severe. In a dataset this large, that is expected - but our team is more concerned with exploring the causes for the most severe accidents, and how policy or procedure could help mitigate those accidents.

Figure 2: Severity of accidents frequency chart



The first way our team decided to tackle the relationship between severity and weather condition was by creating a correlation matrix between our independent variables, seen in Figure 3. We can see that the Visibility has a negative relationship with severity, implying that lower visibility is linked to higher levels of severity in vehicle accidents. Furthermore, Wind Speed, Humidity, and Precipitation are positively correlated, implying that higher wind speed, humidity, and precipitation are associated with higher levels of accident severity. These correlations are generally very weak, which is likely a result

of the extremely high frequency of instances in the data, and the disproportionate amount of fair weather accidents. Despite this, there still is a directional relationship between the variables that explain a story between weather and accident severity.

Figure 3: Correlation Matrix

	Severity	Temperature	Visibility	Pressure	Humidity	Wind_Speed	Precipitation
Severity	1.000000	-0.000981	-0.026748	-0.031933	0.055861	0.055158	0.016916
Temperature	-0.000981	1.000000	0.224143	0.122691	-0.398395	0.092173	0.001185
Visibility	-0.026748	0.224143	1.000000	-0.042075	-0.376656	0.004157	-0.104873
Pressure	-0.031933	0.122691	-0.042075	1.000000	0.193511	-0.060204	0.017480
Humidity	0.055861	-0.398395	-0.376656	0.193511	1.000000	-0.145224	0.073881
Wind_Speed	0.055158	0.092173	0.004157	-0.060204	-0.145224	1.000000	0.030227
Precipitation	0.016916	0.001185	-0.104873	0.017480	0.073881	0.030227	1.000000

Another way our team found success in trying to answer this question was to use the groupby pandas function to explore the averages of conditions based on their severity. Our findings are located in Figure 4. The first observation based on these findings is that precipitation seems to trend upward with severity: there is a 131.8% increase in precipitation average from severity 2 accidents to severity 3 accidents. This change can be explained intuitively, as rain can impair drivers vision and lead to, in general, more dangerous conditions for crashes. Furthermore, there was roughly a 30% increase in the average wind speed from severity 2 to severity 3 crashes. This corroborates the precipitation story, as wind speeds are correlated with storm systems that could create more dangerous conditions for drivers on the road. The average temperatures of severities 2, 3, and 4 suggest that the average accident temperature falls around 60 degrees fahrenheit, potentially implying that a greater share of accidents happen in seasons that aren't summer.

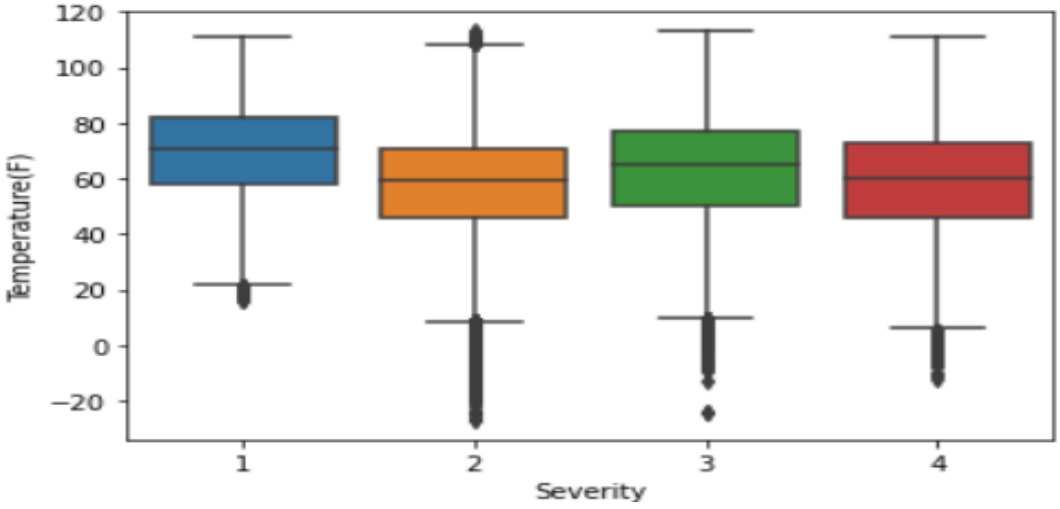
Figure 4: Severity by average conditions

Severity	Temperature	Visibility	Pressure	Humidity	Wind_Speed	Precipitation
1	70.637200	9.510815	29.066384	51.265317	8.353052	0.005413
2	58.011781	8.940652	29.395174	66.365479	6.973345	0.006846
3	62.818627	8.948192	29.093476	65.622921	8.940147	0.015872
4	58.647998	8.697738	29.292784	69.428430	7.802672	0.009748

The next analytics step taken by our team was to explore the conditions that have, on average, the most severe crashes. This could offer insight into where and when to allocate emergency resources, reinforce road safety features, or disperse traffic around to prevent the frequency of severe crashes. The findings are located in Figure 4, with averages of severity 3 and above. One observation from this analysis was that all of the conditions with extremely high averages included a condition of snow or wind in their description, with 'Light Blowing Snow' and 'Freezing rain / Windy' having an average severity of 4. This is also consistent with the analyses above, where precipitation and wind speed seem to be correlated with higher severity. In general, these observations could help inform decision making to increase the presence of emergency vehicles and safety protocols when these weather conditions are present.

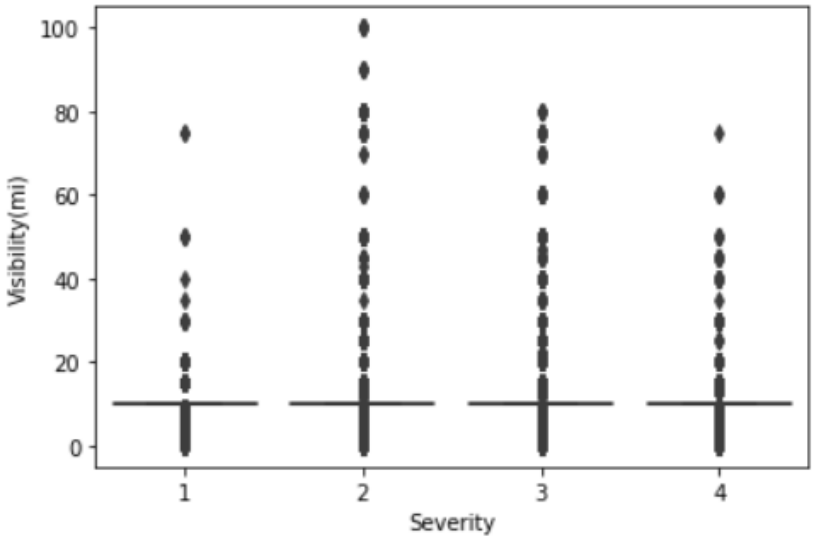
Additionally, we were able to create boxplots for each one of the categories to have a better visualization of the results. On figure 5 we can see the boxplot for Temperature and Severity levels, where we can derive that higher Temperatures usually lead to accidents with type 1 severity. However,we can see that the boxplots of type 2, 3 and 4 severity have a lot of outliers which could imply that extreme cold (Temperatures under 10 Fahrenheit) temperatures could be one of the reasons that lead to an accident of higher severity.

Figure 5: Severity and Temperature BoxPlot



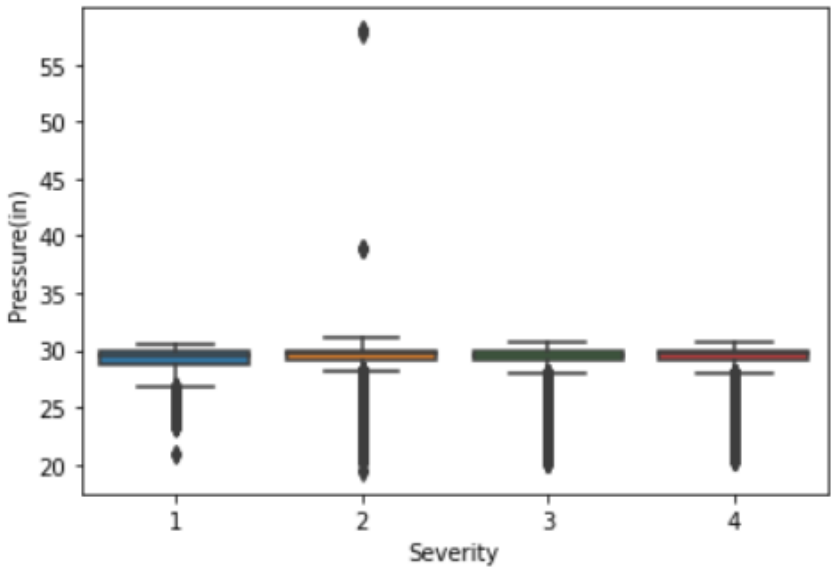
Looking at Figure 6 we can see the boxplot that represents the effects of Visibility to the severity of the accident. It is hard to see it on the boxplot graph but after looking at Figure 4 we can see that Severity 4 has the lowest mean value of 8.6 miles and Severity 1 has the highest mean value of 9.5 miles. This could imply that the lower the visibility of the driver the higher the chances of getting into a more severe accident.

Figure 6: Severity and Visibility BoxPlot



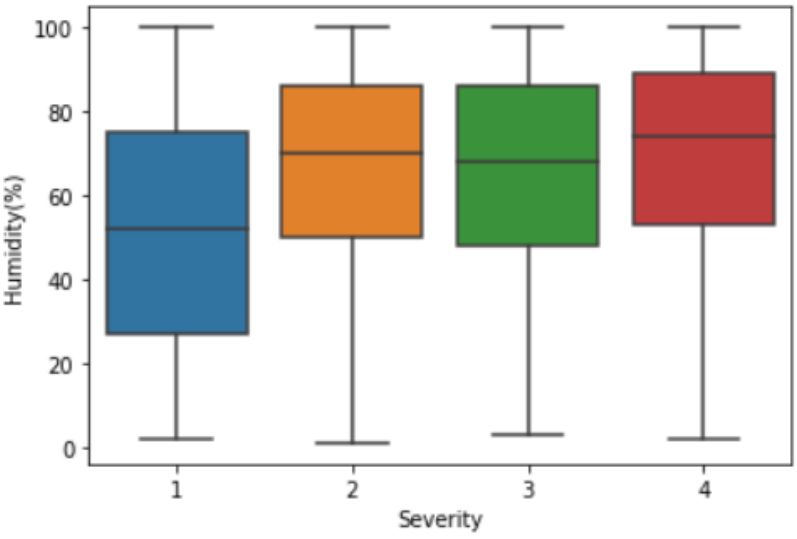
On Figure 7 we are able notice a similar situation with the Severity and Visibility boxplot. Here we can see that the mean values of Pressure in each level of Severity are similar which again it can indicate that Pressure cannot have a big impact on how severe an accident can be.

Figure 7: Severity and Pressure BoxPlot



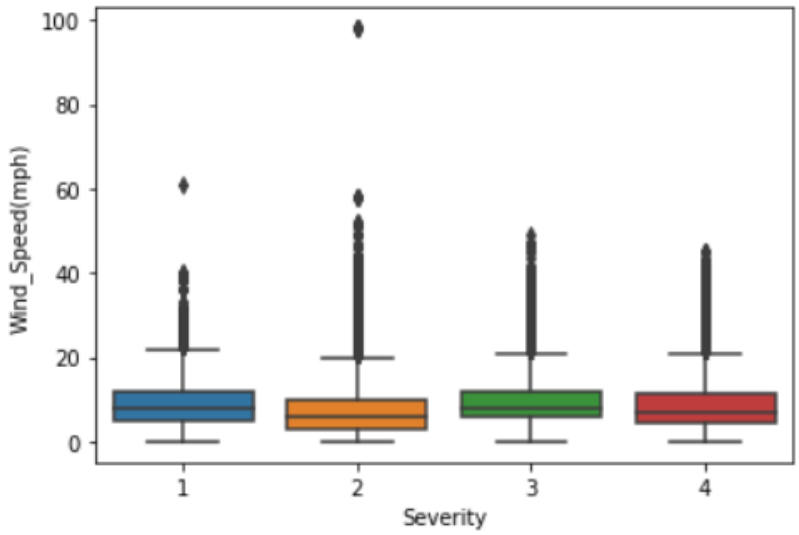
On Figure 8, it is very easy to notice that a lower percentage of humidity can lead to a less severe accident. We can see that the boxplot that represents the type 1 Severity has the lowest mean value of Humidity (51%). However, on the other hand we can see that type 4 Severity has the highest mean value of Humidity (69%). This could imply that rainy weather conditions that cause an increase in the humidity could result in more severe accidents.

Figure 8: Severity and Humidity BoxPlot



On Figure 9 we can see that the lowest mean value of Wind speed belongs to Severity 2 (6.97). On the other hand the highest mean value of Wind speed belongs to Severity 3 (8.9). Such results could imply that there is a possibility to have a more severe accident when wind speed is higher.

Figure 9: Severity and Wind Speed BoxPlot



Finally on figure 10, we can see that the mean value of Precipitation for all four levels of severity have a similar value close to zero. For example, severity 1 has a mean value of 0.005, severity 2 a mean value of 0.006, severity 3 a mean value of 0.015 and severity 4 has a mean value of 0.009. We can see that severity 3 and 4 have higher mean values than severity 1 and 2, therefore we could assume that higher levels of Precipitation could lead to accidents of higher severity.

Figure 10: Severity and Precipitation BoxPlot

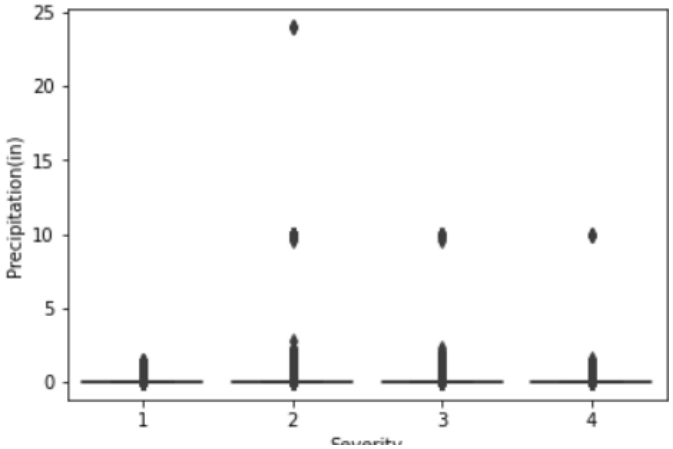


Figure 11: Weather conditions and their subsequent average severity

Weather_Condition	
Light Blowing Snow	4.000000
Freezing Rain / Windy	4.000000
Patches of Fog / Windy	3.600000
Smoke / Windy	3.106383
Light Thunderstorms and Snow	3.000000

Lastly, our team decided to include a multiple regression model to test the relationship between accident severity and these weather conditions. The output of the regression result is seen below in Figure 12. While the significance of the model and features are acceptable, the R-squared value of .01 tells us that these conditions do not explain a high level of the variance in accident severity. Nonetheless, the directionality of the results are consistent with the findings above, as visibility affects severity in a negative way - ie. more visibility, less severity - and humidity, precipitation, and wind speed affect severity in a positive way. The highest coefficient is with the precipitation variable, where a one inch increase in precipitation affects the severity of an accident by almost half of a point value, holding others constant. This continues to corroborate the story that poor conditions have an impact on vehicle crash severity.

Figure 12: Multiple regression

OLS Regression Results

Dep. Variable:	Severity	**R-squared:**	0.010
Model:	OLS	**Adj. R-squared:**	0.010
Method:	Least Squares	**F-statistic:**	1641.
Date:	Fri, 03 Dec 2021	**Prob (F-statistic):**	0.00
Time:	15:46:08	**Log-Likelihood:**	-7.6219e+05
No. Observations:	976622	**AIC:**	1.524e+06
Df Residuals:	976615	**BIC:**	1.524e+06
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.6043	0.014	184.165	0.000	2.577	2.632
Temperature	0.0010	3.4e-05	30.478	0.000	0.001	0.001
Visibility	-0.0007	0.000	-3.121	0.002	-0.001	-0.000
Pressure	-0.0231	0.000	-46.932	0.000	-0.024	-0.022
Humidity	0.0020	2.75e-05	71.182	0.000	0.002	0.002
Wind_Speed	0.0058	9.72e-05	59.945	0.000	0.006	0.006
Precipitation	0.0486	0.005	9.036	0.000	0.038	0.059

Omnibus:	424371.487	**Durbin-Watson:**	1.314
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1974869.139
Skew:	2.134	**Prob(JB):**	0.00
Kurtosis:	8.506	**Cond. No.**	2.52e+03

Conclusion

In conclusion, vehicle accident safety still remains, and will continue to remain a constant and astronomically important aspect to American life. The analysis seeked to begin to paint a picture of vehicle crashes based on just one aspect of driving - weather. While weather seemingly has an impact on the severity of the crash, the picture still remains highly complicated and unclear. This research has provided insight that statistics can prove that there are relationships between weather conditions and accident severity, and the results should be used to improve safety protocols and standards when these conditions arise, especially in extreme circumstances. Wind, rain, and humidity seem to contribute positively to accident severity, and lower visibility seems to contribute as well.

The bottom line remains that there are insights to draw from accident data, and this data can be used to help save lives. This data was only a small sliver into conditions that exist when any person gets behind the wheel, and there is plenty more to be done. While weather plays a role in accidents, so does human error, distraction level, road condition, population density, and countless other factors - all which could be explored and eventually turned into helpful information.