Car Accident Severity Final Report

IBM Data Science Capstone Project

I. Introduction

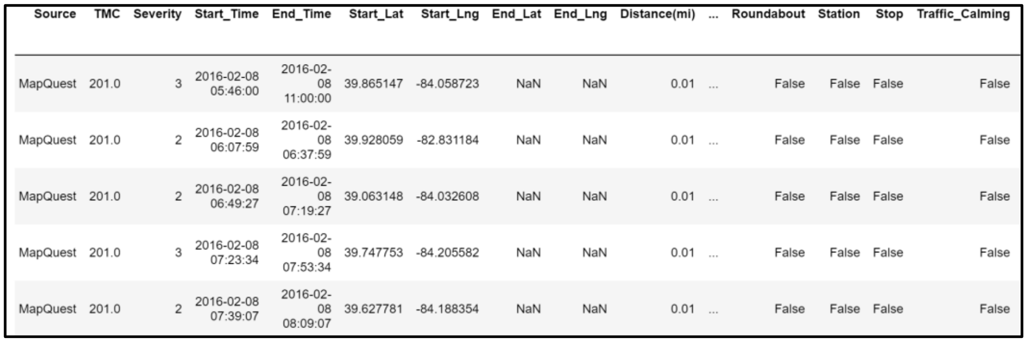
In the United States, an average of 6 million car crashes occur annually. This results in over 37,000 deaths each year, making it the leading cause of fatality for healthy U.S. citizens. In addition, auto accidents cost the U.S. over 200 billion dollars per year. These tragedies are mostly preventable; over 90 percent occur because of driver carelessness. Government officials have worked to remedy this issue for many decades. Despite their efforts, this problem persists today.

For this project, I will build a machine learning model to predict the severity of a car accident based on various attributes. This model can be used to alert drivers of dangerous road conditions and encourage them to drive more cautiously based on present risk. While this approach is unlikely to eliminate the issue altogether, by understanding the factors that contribute to the severity of a car accident, local officials can concentrate their efforts on building a data driven solution to the problem.

II. Data

The data being used for this project is US Accidents (3.5 million records) by Sobhan Moosavi. It contains information on United States car accidents during the years 2016-2020. The data has been sourced from a range of API's, including MapQuest and Bing. There are over 3 million entries and 49 columns. Some of the attributes that will be used for this project include:

- **Location:** Street, City, County, Zip Code, State
- **Weather:** Temperature, Wind Chill, Humidity, Wind Speed, Precipitation
- **Road Type:** Crossing, Junction, Railway, Roundabout
- **Astrological:** Sunrise/Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight



Source	TMC	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	...	Roundabout	Station	Stop	Traffic_Calming
MapQuest	201.0	3	2016-02-08 05:46:00	2016-02-08 11:00:00	39.865147	-84.058723	NaN	NaN	0.01	...	False	False	False	False
MapQuest	201.0	2	2016-02-08 06:07:59	2016-02-08 06:37:59	39.928059	-82.831184	NaN	NaN	0.01	...	False	False	False	False
MapQuest	201.0	2	2016-02-08 06:49:27	2016-02-08 07:19:27	39.063148	-84.032608	NaN	NaN	0.01	...	False	False	False	False
MapQuest	201.0	3	2016-02-08 07:23:34	2016-02-08 07:53:34	39.747753	-84.205582	NaN	NaN	0.01	...	False	False	False	False
MapQuest	201.0	2	2016-02-08 07:39:07	2016-02-08 08:09:07	39.627781	-84.188354	NaN	NaN	0.01	...	False	False	False	False

III. Methodology

a. Feature Selection

Each source reports accident severities differently, so only Bing will be selected to keep things standard. Bing provides a wide distribution of severity classifications and contains enough data to analyze (about 1,000,000 entries), making it a good source to use for this project. To save computation time, a random sample of 100,000 entries will be taken. Finally, any columns that cannot be used to predict the severity of a car accident will be dropped. This means that any columns describing a crash after it occurred will be removed. The features being dropped are:

- Source
- TMC
- End Time
- End Latitude
- End Longitude
- Distance
- Description
- Side
- Country

The data will be split into a train and test set. Twenty percent of the data will be set aside for the test set.

b. Missing Values

The dataset has missing values for several columns, in 3 main groups. These include:

- **Location:** Number, City, Zip Code, Time Zone, Airport Code
- **Weather:** Weather Timestamp, Temperature, Wind Chill, Humidity, Pressure, Visibility, Wind Direction, Wind Speed, Precipitation, Weather Condition
- **Astrological:** Sunrise/Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight

About 75% of values for the Number column are missing. This feature describes the street number where the accident takes place. The name of each street and location of each accident are already provided, so the street number probably won't give any more predictive power. This column will be dropped, as so many values are missing.

Airport Code, Time Zone, and City also have many missing values. Every accident location must be in a time zone, have a closes city and have a closest airport. The longitude and latitude of each accident is known, so this can be used to determine these missing values. KNNImputer will be utilized to fill in the missing values for Airport Code, Time Zone, and City based on information from the nearest accident, in terms of longitude and latitude.

Next up is the weather-related columns. To fill in these missing values, 3 different strategies will be used:

1. **Weather Condition and Wind Direction:** Missing values will be filled in as 'None'. A missing value in these columns could mean that no weather condition or wind was reported at the time of the accident, so trying to predict a value would not make sense. Values not being reported for these columns may have predictive power on its own because of this.
2. **Temperature, Wind Chill, Humidity, Pressure, Visibility, Wind Speed, Precipitation:** Missing values will be filled based on the closest data point, by longitude and latitude, that took place within the same month as the accident with the missing value.
3. **Weather Timestamp:** Missing values will be filled as 0. This feature represents the time that the weather condition present during the accident was observed. As such, when Weather Condition is missing, Weather Timestamp is also missing. This is a continuous variable, so there is no good way to impute this.

The last columns to fill in missing values for are the astrological features - Sunrise/Sunset, Civil Twilight, Nautical Twilight, and Astronomical Twilight. Very few values for these columns are missing. These rare cases will be filled based on the mode of a group. The grouping will be based on month and hour of the day.

c. Feature Engineering

The first column that will be focused on for feature engineering is Street. There are too many unique values in this column for the machine learning models to handle, so key information will be extracted from this feature to describe each street as best as possible. Here are the transformations that will be performed:

- Remove any extraneous whitespace from the street name
- Parse compass direction from the street name - only N, S, W, and E will be included, as opposed to North, South, West and East. This is because the single character often appears in highway names to represent direction of traffic, whereas the full word is often used for street names
- Classify highway type (US-xx = United States Highway, I-xx = Interstate Highway, etc.)
- Categorize street names based on common abbreviations (Rd., Dr., Fwy.)

The next feature that will be created is the distance of the car accident from its nearest city. To do this, the latitude and longitude where the car accident was reported will be compared with the latitude and longitude of the nearest city. In a separate script, I wrote a function to create a DataFrame with the latitude and longitude for each city in the dataset. This will be used to compute the distance for each accident.

The population of each location where the car accidents occurred will likely be relevant to the task at hand. A new column will be created that is a count of the number of accidents that occurred at that location. This will provide a general estimate for how large each location is, relative to the others in the same category. The columns this operation will be performed on are City, County, State, Zip Code, Airport Code and Time Zone.

The Weather Timestamp feature represents the time the weather condition present during the car accident was first observed. A new column will be created, Time Elapsed Weather, which will be the difference in time between the car accident (Start Time), and observance of the weather condition (Weather Timestamp). Some of these values are negative, as the weather condition was observed after the accident. Even though the task is to predict the severity of an accident before it occurs, negative values will still be kept as it indicates an incoming weather condition. After this, an additional column will be created, Holiday, which indicates if the car accident happened during a holiday. This will be determined by the date of the car accident.

Using the Start_Lat, and Start_Lng features, we will compute the X, Y, and Z coordinates of the car accident.

Finally, new features will be created to indicate the direction of the wind in comparison to the direction of traffic where the car accident occurred. This will be computed by comparing the North, South, West and East columns (extracted from street earlier in the feature engineering phase), with the Wind Direction column. The new columns that will be created will be Headwind, Tailwind, and 'Sidewind'. They will be determined by the primary and secondary compass directions of the Wind Direction column. This means that in some instances, an accident will be classified as two of these (i.e. a north facing road and north west facing wind direction will result in a tailwind and a sidewind).

d. Machine Learning

Before fitting the data to the machine learning algorithms, a few final adjustments need to be made. First, dimensionality must be reduced to save computation time. Many of the location features had very large cardinality. For instance, Street has over 16,000 unique values and Zip Code has over 24,000. These columns would need to be dummied later, meaning the size of the dataset would grow immensely. Due to this, City, Zip Code, Airport Code, and Street will be dropped. The predictive power lost due to this should be minimal, for a few reasons. First, much task-relevant information related to these columns have been extracted during the feature engineering phase. Secondly, keeping all these features would probably be overkill. For instance, accidents in the same county would likely have the same Airport Code and Zip Code.

After this, continuous columns will be scaled, and categorical columns dummied. This will be necessary to help the machine learning algorithms process the data. Then, Principal Component Analysis will be applied to the continuous features, as many of them exhibit multi-collinearity. Five Principal Components will be used for prediction.

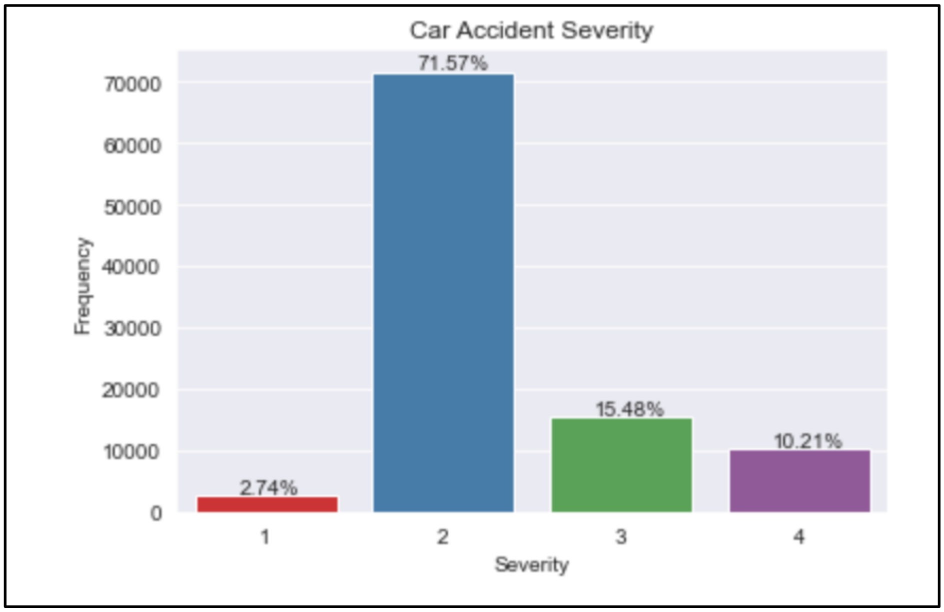
The scoring parameter for this task will be f1 macro, as the task is a multi-class classification problem with imbalanced classes. Grid Search will be used to find the optimal hyperparameters for the machine learning models. The three models that will be used for this task are:

- Logistic Regression (SGD)
- Linear SVC (SGD)
- Random Forest Classification

IV. Analysis

a. Severity

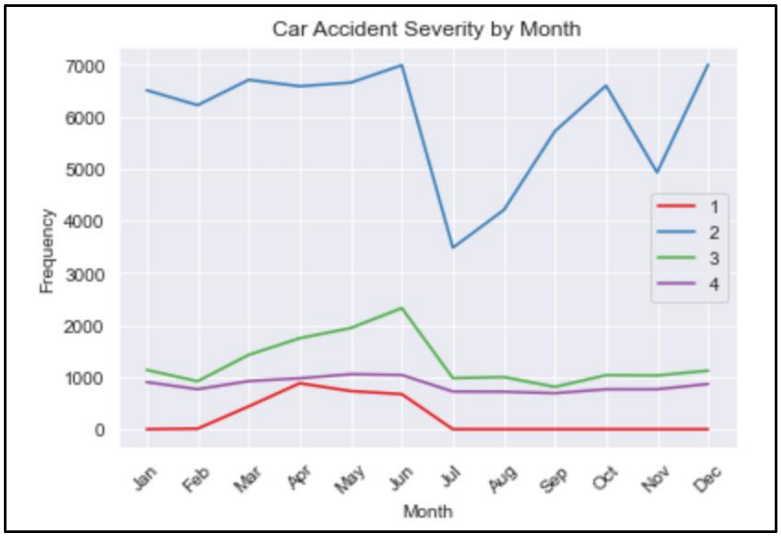
Severity is classified as an integer ranging from 1 to 4. This number represents the amount of traffic delay the accident caused. The chart shows a vast majority of accidents having a severity of 2, and only a very small amount of accidents having a severity of 1. The classifications are heavily imbalanced.



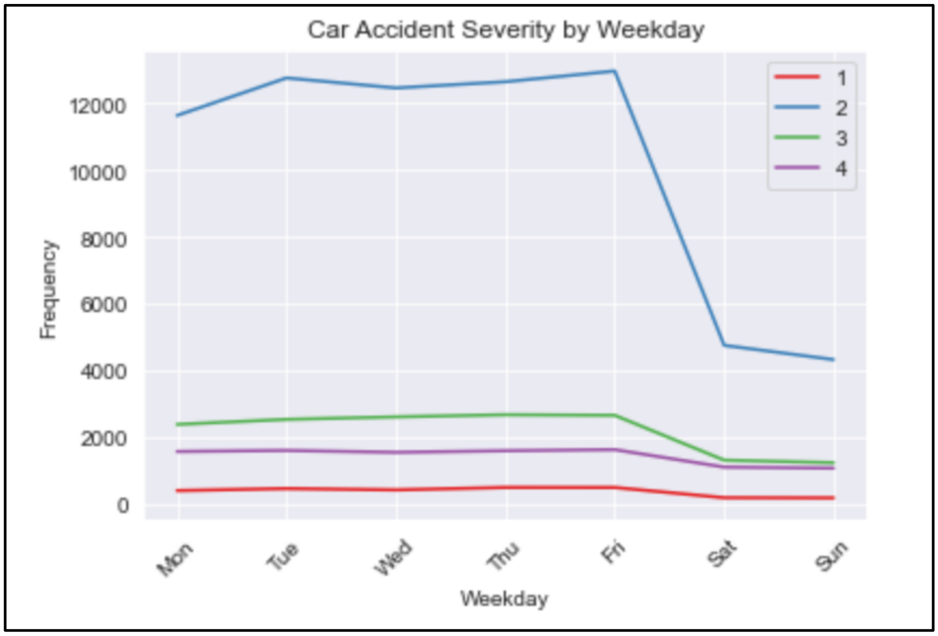
b. Time

Car accidents occur year-round. By viewing the number of car accidents by month, it can be observed that:

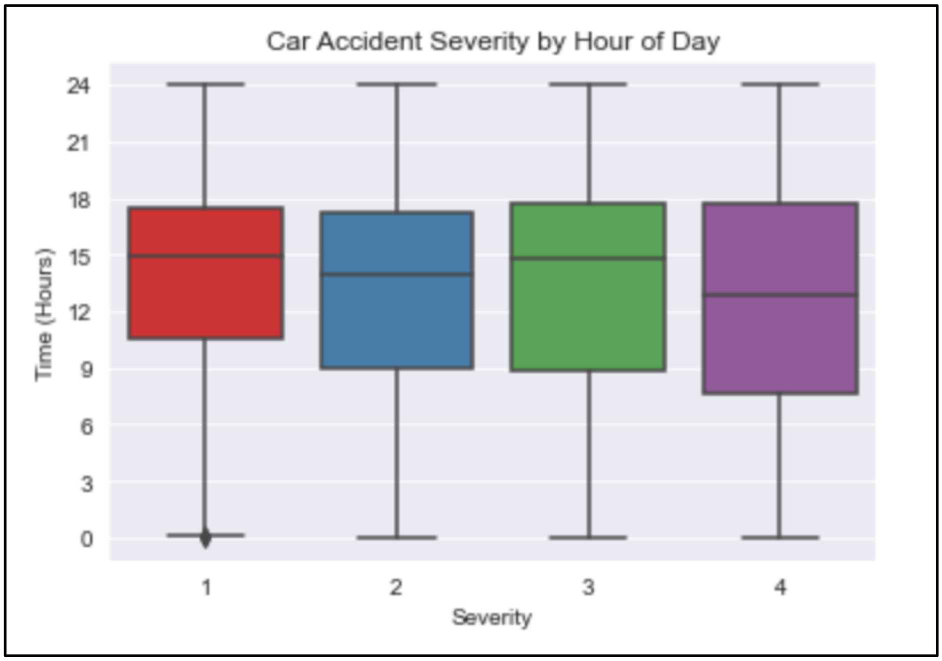
- Severity 1 accidents peak during the spring, and by mid-summer, the number of cases is minimal
- Severity 2 accidents seem to decrease during the summer, and climb their way back by mid-fall
- Severity 3 accidents rise during the spring, fall during the beginning of the summer, and remain relatively constant for the rest of the year
- Severity 4 accidents remain relatively constant throughout the year
- The safest time to drive seems to be mid-summer
- The most dangerous time to drive seems to be late spring to early summer



Weekday has a significant impact on the number of car accidents. Starting with Monday, accidents gradually increase until Friday, and plummet during the weekend. This trend is present for all levels of accident severity.

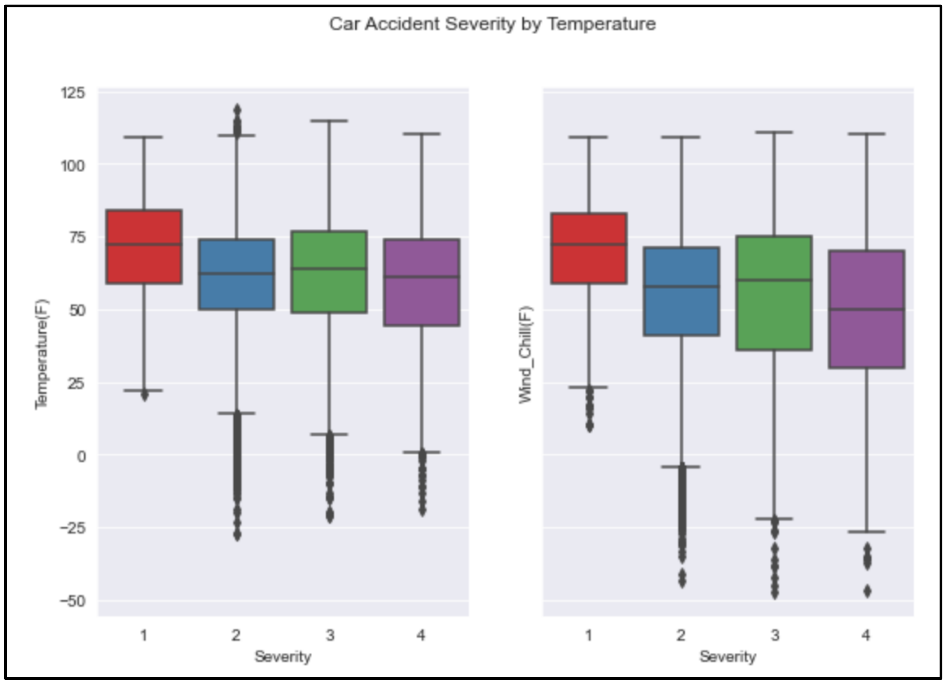


Car accidents of each severity span every hour of the day. The median time of a car accident is during the afternoon, around 3 o'clock. Severity 4 car accidents have the lowest median time, which is closer to noon. Severity 4 car accidents also have the largest interquartile range.

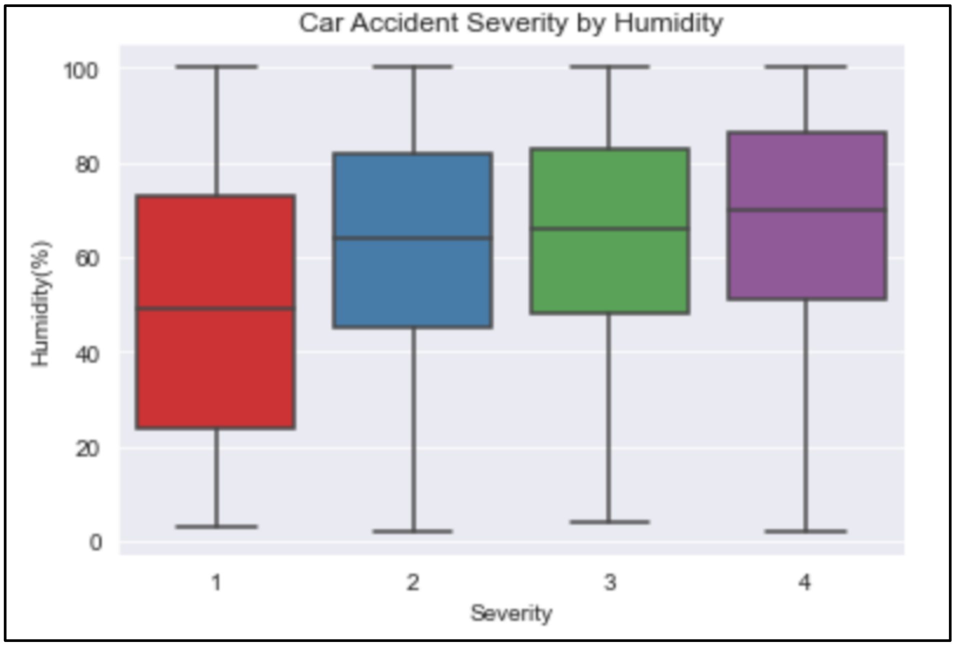


c. Weather

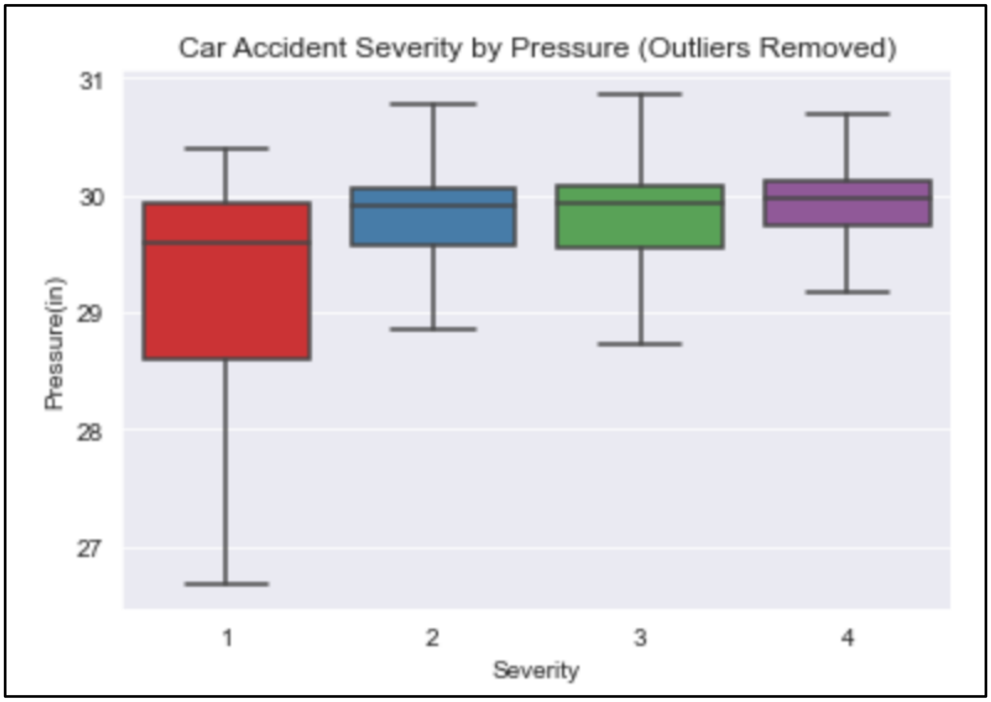
Weather is the next category of features to be examined, beginning with Temperature and Wind Chill. As temperature and wind chill decrease, accident severity also appears to decrease. Wind Chill has roughly larger interquartile ranges for each of the severities, compared to temperature.



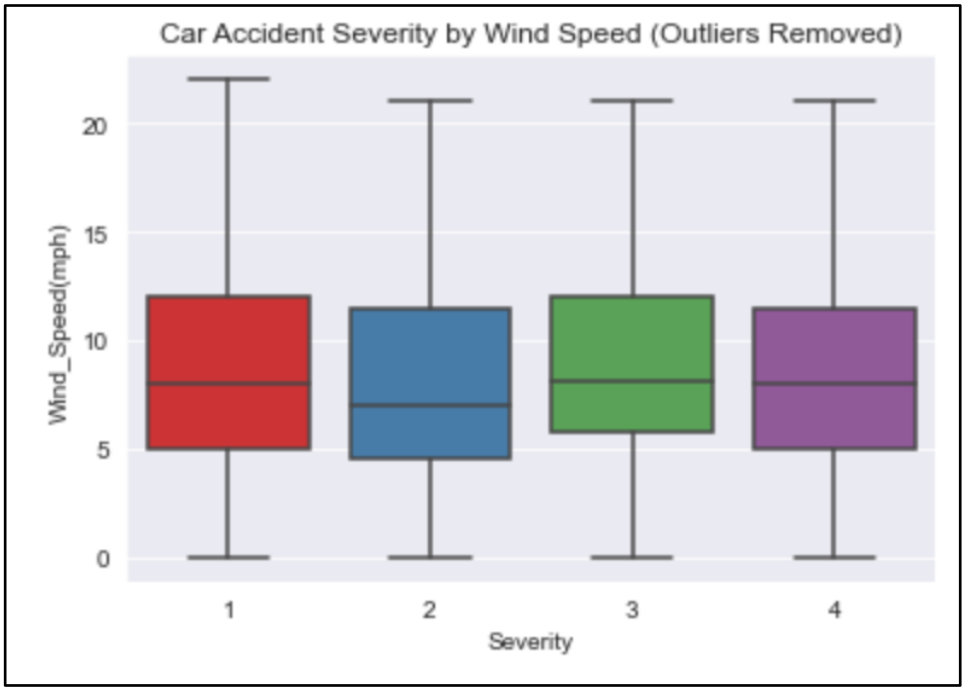
There is a clear relation between Severity and Humidity. The higher the accident severity, the higher the median humidity. Severity 1 accidents have the largest interquartile range.



Pressure has a similar effect on severity. The higher the accident severity, the higher the median pressure. Severity 1 accidents have the largest interquartile range. It is important to note that the outliers have been removed from this chart to make it readable.

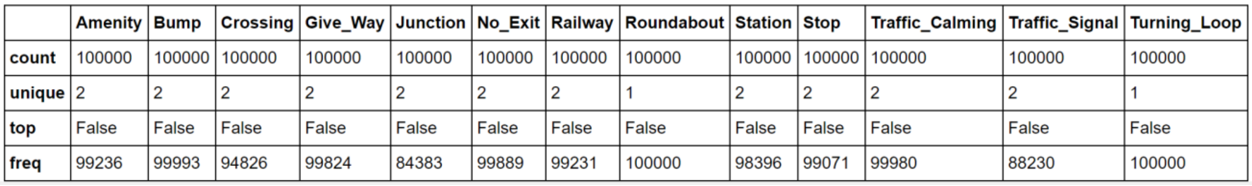


Wind Speed is the final weather-related feature to examine. There is no clear relation between Wind Speed and Severity. Most accidents occurred with a wind speed between 5 and 12, regardless of accident severity. The outliers have been removed from this chart to make it readable.



d. Location

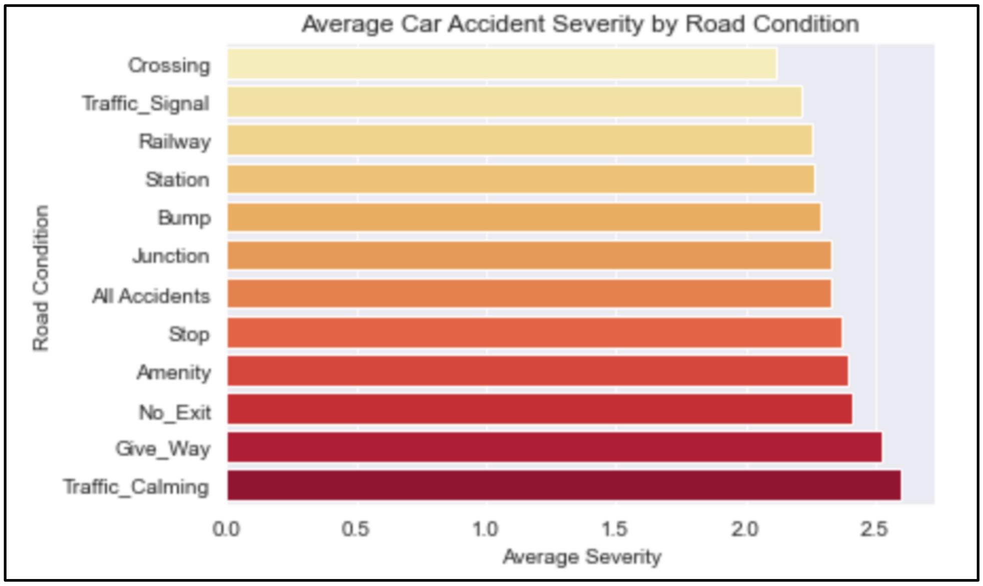
The dataset contains columns of Boolean values signifying the presence of a certain road condition - such as a crossing or a junction. Below are the summary statistics for these columns. The values for each column are mostly false, meaning the road condition is not present. In fact, no car accidents in the sample had a turning loop or roundabout present. Traffic signals are the most common road condition during a car accident.



	Amenity	Bump	Crossing	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop
count	100000	100000	100000	100000	100000	100000	100000	100000	100000	100000	100000	100000	100000
unique	2	2	2	2	2	2	2	1	2	2	2	2	1
top	False	False	False	False	False	False	False	False	False	False	False	False	False
freq	99236	99993	94826	99824	84383	99889	99231	100000	98396	99071	99980	88230	100000

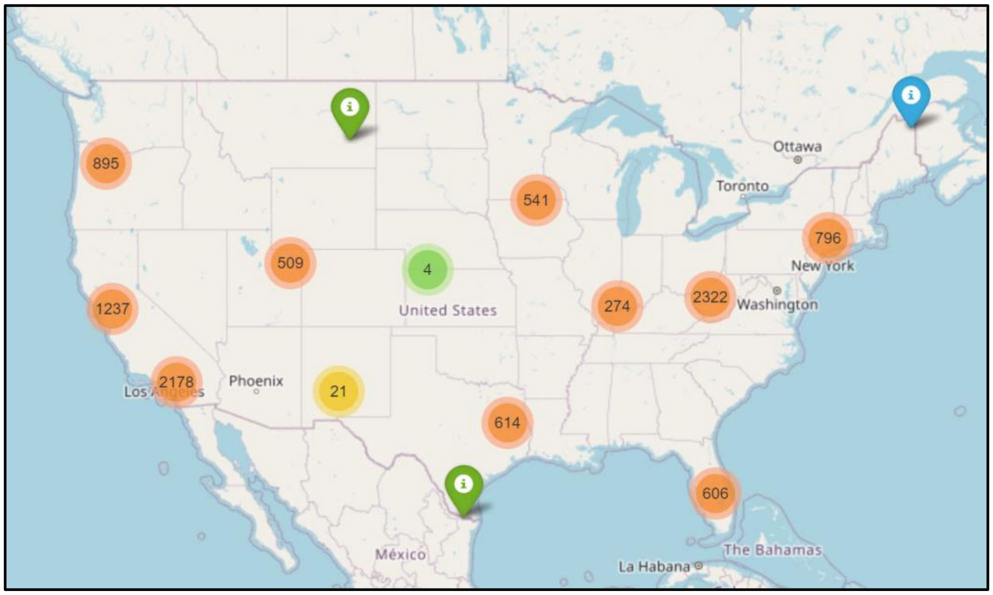
Up next is a plot of the average accident severities for each road condition. Here are the key takeaways:

- Traffic Calming’s have the highest average accident severity - however, only 20 accidents took place with this condition
- Crossings have the lowest average accident severity, but had one of the highest numbers of reported accidents
- Conditions with fewer reported accidents tend to have higher average accident severities



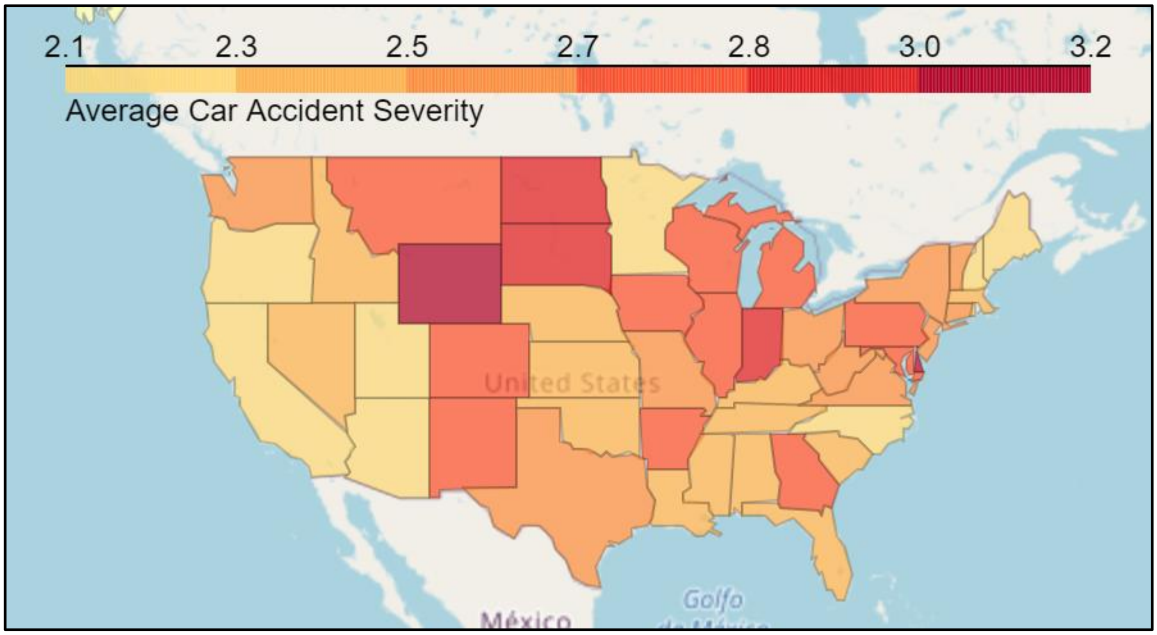
Each entry in the dataset contains the longitude and latitude where the accident occurred. Below is a map of a random sample of 10,000 accidents. The map in the notebook is interactive, whereas the map here is just a snapshot. After some investigating, here are some conclusions that can be made:

- Accidents seem to be clustered near population centers
- The west coast seems to have more accidents than the east coast
- To see where each severity of accident is most likely to occur, more exploration will be needed



Certain states have a higher share of severe accidents than others. The map below shows the average accident severity for each state. It is important to note that states with the highest average severity are not necessarily the most dangerous, as this map does not indicate the total number of accidents, only the average severity. Wyoming and

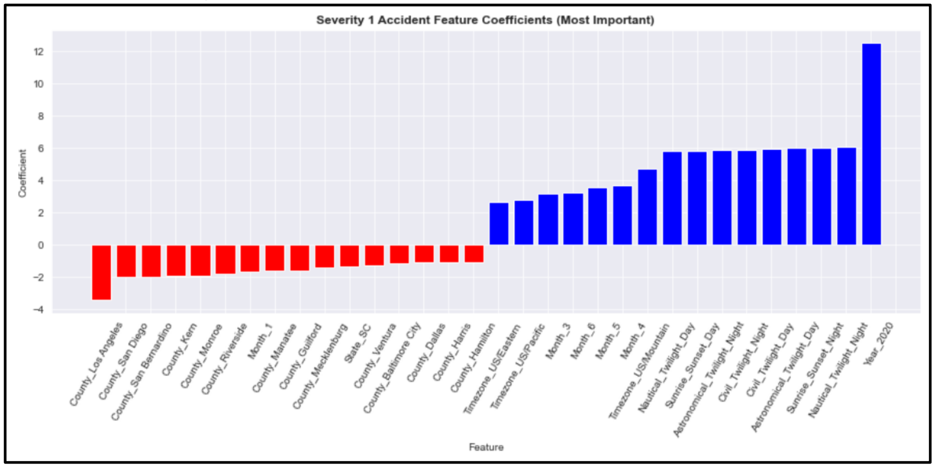
Delaware have the highest averages, but these states are of low population, meaning the finding may be insignificant.

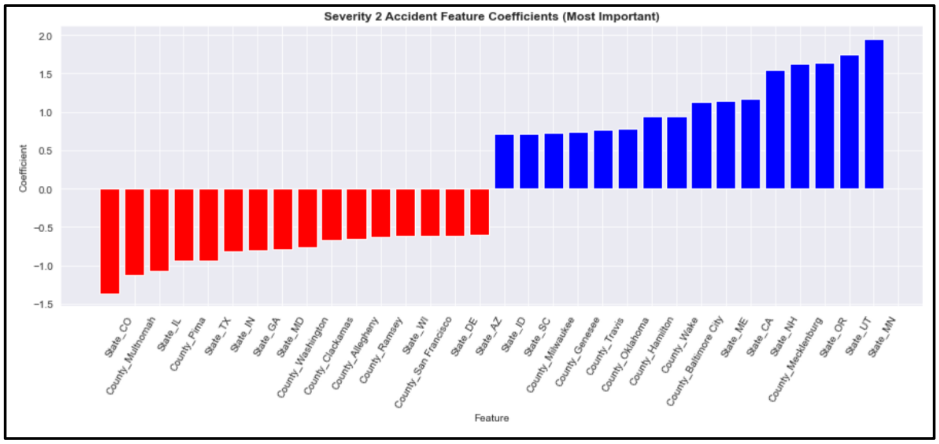


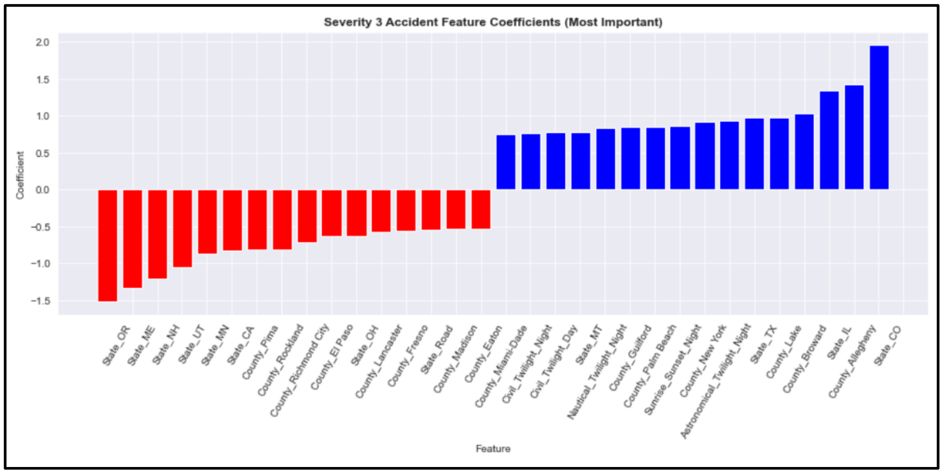
V. Results

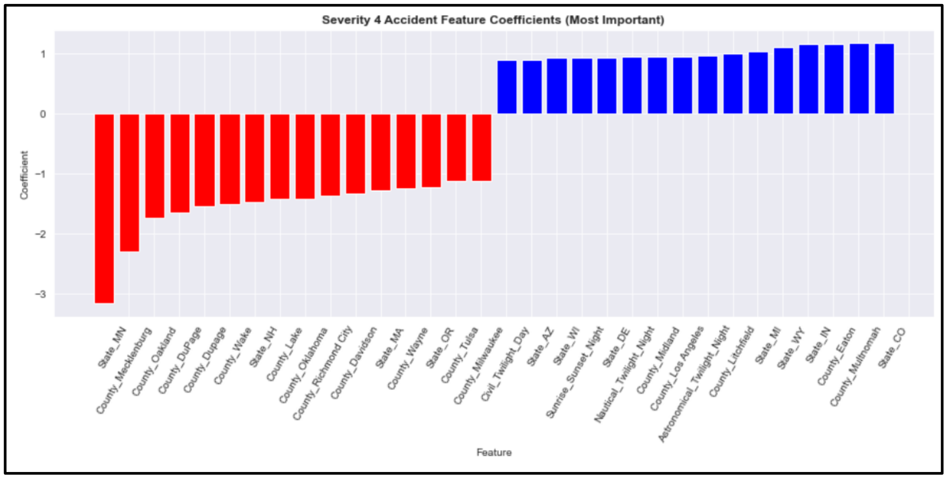
The models made a tradeoff between accuracy and f1 Macro scoring. Essentially, the models were able to correctly identify more rare severities by sacrificing some overall accuracy. The best model for this task is SGD Logistic Regression, with a macro average of .50 and an overall accuracy of .74 for the test set. While performing slightly worse than Random Forest Classification, Logistic Regression fit much better to the data, with a smaller difference between the train score and test score.

The most important factors in classifying the severity of a car accident were time and location. The coefficients were largest for state and county columns, with many of the twilight columns also have big values. A chart of the most important feature coefficients for each severity are shown below:









A big constraint for this project was the size of the dataset. A sample of only 100,000 rows was used for machine learning. While this should have been enough for the given task, the model would have been more robust if all samples were included. This would have allowed for obscure locations or other conditions to have been represented as part of the machine learning models.

If computational cost were not an issue, more high cardinality columns would have been able to be used for machine learning – such as Zip Code and Street. In addition, high dimensionality meant long computation times for the machine learning models, so only simpler models were fit. More complex models could have been used for this project, such as Gradient Boosting Classifier, or non-linear Support Vector Machines, if this were not an issue.

VI. Conclusions

I hope that the results of this project will yield valuable information that can contribute to life saving policies and technologies. To expand upon this project, I believe more accurate results could be produced if information regarding speed limits and peak traffic hours were included in the dataset. In addition, the use of a powerful cloud computing system to aid in the machine learning process would have been of great assistance. A powerful computing platform would have allowed the entire dataset to be used, yielding even stronger results. To accomplish the goal of saving lives from car accidents, the next step would be to present the results of this project to government officials. With the insights drawn, an alert system could be implemented in automobiles to warn the drivers of unsafe conditions.

VII. Appendix

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.
- https://www.kaggle.com/sobhanmoosavi/us-accidents
- https://www.thewanderingrv.com/car-accident-statistics/#:~:text=Annual%20United%20States%20Car%20Crash%20Statistics,-This%20section%20of&text=On%20average%2C%20there%20are%206,22%2C471%20caused%20only%20property%20damage.
- https://aneesha.medium.com/visualising-top-features-in-linear-svm-with-scikit-learn-and-matplotlib-3454ab18a14d