FINAL PROJECT CAPSTONE

Car Accident Severity

A case for New York city

Contents

Introduction

The effective treatment of road accidents and thus the enhancement of road safety is a major concern to societies due to the losses in human lives and the economic and social costs. Tremendous efforts have been dedicated by transportation researchers and practitioners to improve road safety. In European Union there has been a consistent reduction in fatalities. In 1991, 76,230 fatalities were recorded in EU, whilst in 2013 the total number of fatalities was 25,938 (CARE, 2015). Greece faces the same significant reduction in fatalities as well. When the type of vehicle.

is examined, recent reports in Greece (ERSO, 2012), show that in 2009, 786 car occupant fatalities and 426 motorcyclist fatalities occurred. Accident severity is therefore an issue which has gained the attention of many researchers so far (Chimba and Sando, 2009; Sze and Wong 2007; Yamamoto and Shankar 2004; Yau, 2004; Chang and Wang 2006; Milton et al. 2008; Quddus et al. 2002; Savolainen and Mannering 2007; Theofilatos et al., 2012). Al-Ghamdi (2002), used logistic regression and found that crash location (intersection, non-intersection) and cause of accident (speed, running a red light, wrong way violation, etc.) significantly influence road accident severity. Ulfarsson and Mannering (2004), explored the differences in injury severity levels between male and female occupants in sport utility vehicles, minivans, pickups and passenger cars. The authors state that significant differences between male and female drivers were found to exist. Valent et al. (2002) applied logistic regression to evaluate the association of driver characteristics and accident severity in Italy. The results indicated that males are more likely to be engaged in fatal accidents and that car drivers are less likely to be fatally injured than killed than motorcyclists. Chang and Wang (2006) carried out a study in Taipei, Taiwan and argue that the most important factor that affected crash severity was the type of vehicle. De Lapparent (2007) found that females and riders of motorcycles with high engine size were more likely to be severely injured. Pai (2009), utilized UK national crash data from 1991 to 2004 and analyzed the severity of two or more vehicle-crashes which have occurred at T-junctions involving at least one motorcycle. The author applied binary logistic regression and found that some factors which seem to increase severity are rider age (over 60 years old), high engine size, fine weather, right-of-way violation and involvement of heavy goods vehicle.

Although various research studies examining road accident severity have been found in international literature, only a few were carried out in Greece (Yannis et al., 2005; Theofilatos et al., 2012). Moreover, previous studies have shown the difference in severities between different types of vehicles. Consequently, the present study aims to contribute to existing knowledge by investigating road accident severity in Greece with particular focus on vehicle type, in order to identify the critical risk factors.

Problem

Traffic accidents are an endemic problem in the US. In 2018 alone, 12 million vehicles were involved in a car crash. Individuals take the road, often out of necessity, not thinking twice about the conditions they will face, and how much they will put themselves and others at risk. What if we could predict the severity of an accident, based on weather condition and time environmental data, so we can provide that information to people who can help to reduce the frequency given advices or controlling the situation on those days that are not very appropriate to drive.

Interest

Obviously, the police will be interested in what type of car accident will be happening during a specific day, depending on the weather condition and other variables like temperature, humidity and pressure.

Data

This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about **3.5 million** accident records in this dataset. Check the below descriptions for more detailed information.

For this project, it is necessary to considerate less states because there is to much data for working. That is way we are going to work with only New York state.

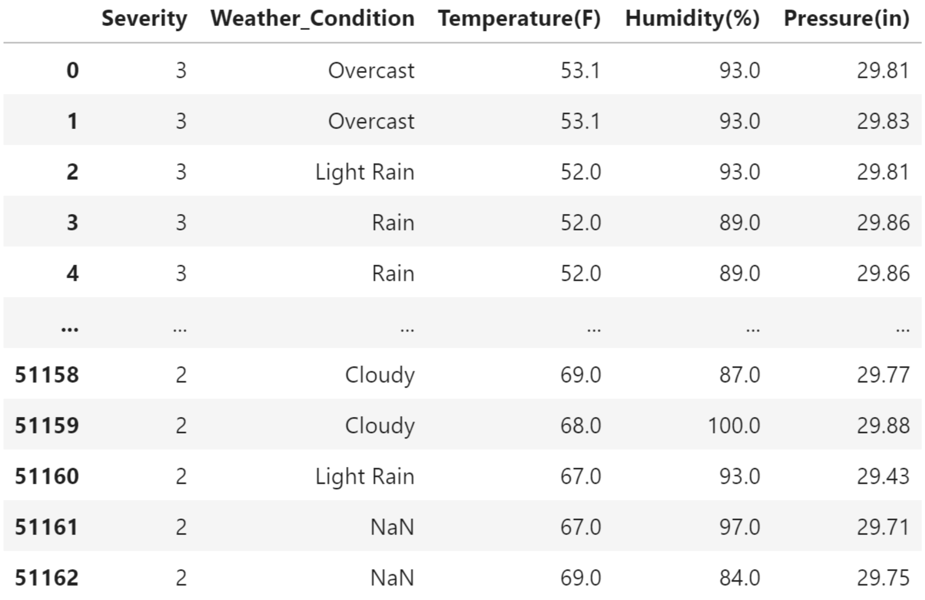
Data Understanding

The first step in analyzing the data is to get an understanding of the attributes. Reading the Metadata pdf, we can allocate the attributes in the following buckets:

- **ID**: This is a unique identifier of the accident record.
- **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).
- **Start_time**: Shows start time of the accident in local time zone.
- **End_time**: Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.
- **Temperature(F)**: Shows the temperature (in Fahrenheit).
- **Humidity (%)**: Shows the humidity (in percentage).
- **Pressure(in)**: Shows the air pressure (in inches).
- **Weather_Condition:** Shows the weather condition (rain, snow, thunderstorm, fog, etc.)

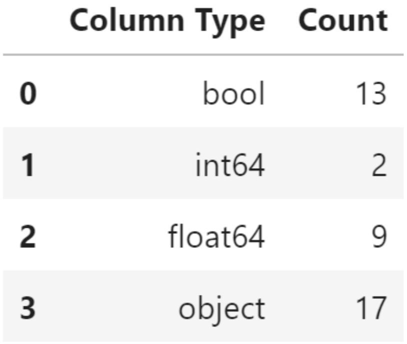
To answer our problem, we are interested in the predictive the severity of the accident depending on the attributes and will therefore discard all other attributes in the data preparation stage.

We print the table to see how the data frame looks .



	Severity	Weather_Condition	Temperature(F)	Humidity(%)	Pressure(in)
0	3	Overcast	53.1	93.0	29.81
1	3	Overcast	53.1	93.0	29.83
2	3	Light Rain	52.0	93.0	29.81
3	3	Rain	52.0	89.0	29.86
4	3	Rain	52.0	89.0	29.86
...	...	...	...	...	...
51158	2	Cloudy	69.0	87.0	29.77
51159	2	Cloudy	68.0	100.0	29.88
51160	2	Light Rain	67.0	93.0	29.43
51161	2	NaN	67.0	97.0	29.71
51162	2	NaN	69.0	84.0	29.75

We describe the data.



	Column Type	Count
0	bool	13
1	int64	2
2	float64	9
3	object	17

As we can see, we are going to try to predict the severity of new car accidents depending on the features which help to explain the dependent variable.

Another thing we can see in the figure 2 there is all types of variables that the data contains. We have 13 variables as bool types, 2 are int types, 9 are float types and 17 object types.

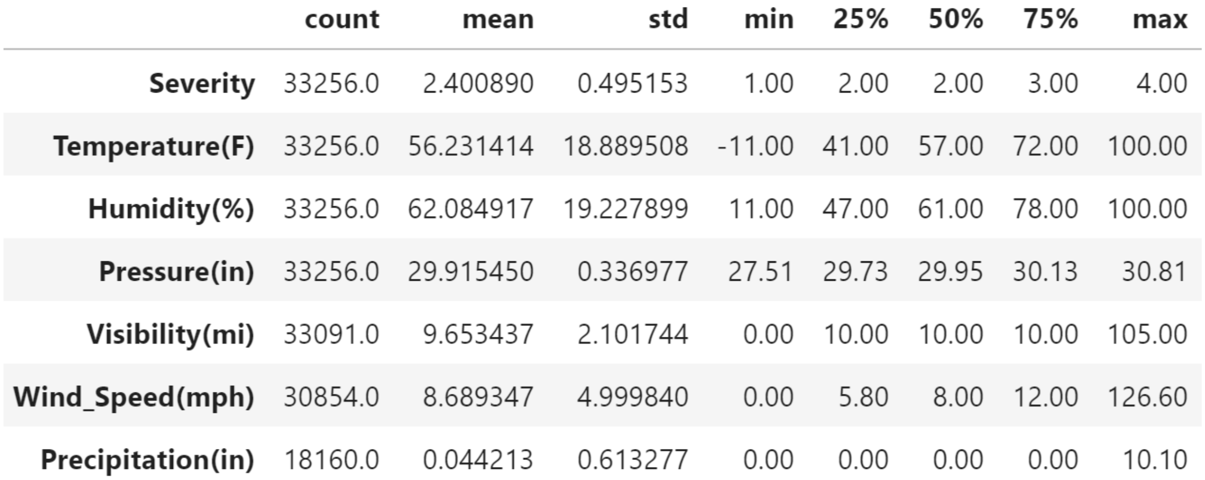
The next step is checking the quality of the data starting with missing values. Checking the number of rows with at least one null reveal that less than 3% of them contain missing values. We will fill the missing data with the mean of the variable we need. For example, we have missing values like Temperature, Humidity and Pressure have missing values, so it is important to fill those values with the mean values for each one.

Data Preparation

Before we ingest the data frame into the different models, we perform the following operations:

- Change the types of all the attributes from Object to:
    - datetime for the Time/Date attributes
    - category for the Environmental attributes
- Change weather condition to dummies
- Change the severity variable into a binary for estimation
- Replace Unknown/Other entries by n/a’s
- Fill n/a’s with most mean values

Here we have a figure that show us how is the data after we fill the non-values

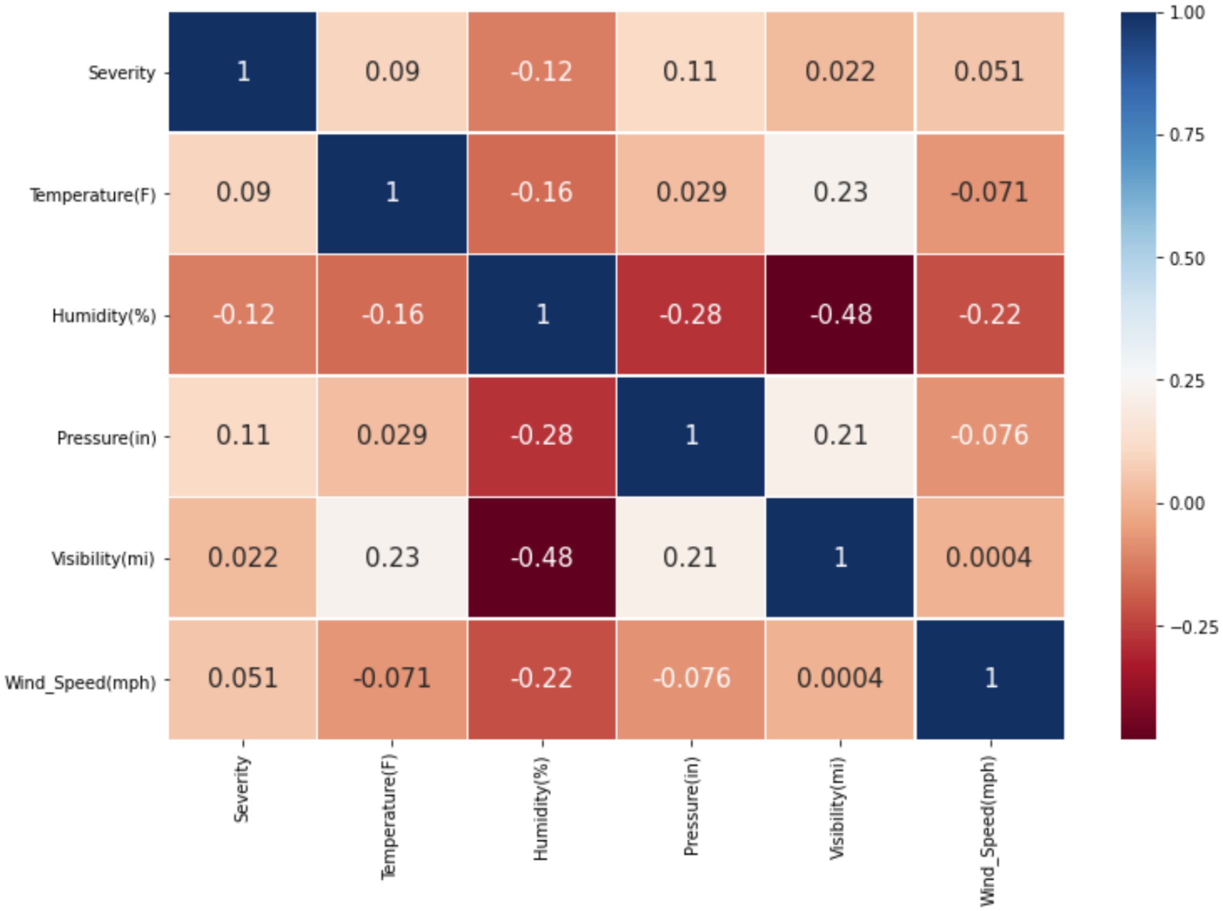


	count	mean	std	min	25%	50%	75%	max
Severity	33256.0	2.400890	0.495153	1.00	2.00	2.00	3.00	4.00
Temperature(F)	33256.0	56.231414	18.889508	-11.00	41.00	57.00	72.00	100.00
Humidity(%)	33256.0	62.084917	19.227899	11.00	47.00	61.00	78.00	100.00
Pressure(in)	33256.0	29.915450	0.336977	27.51	29.73	29.95	30.13	30.81
Visibility(mi)	33091.0	9.653437	2.101744	0.00	10.00	10.00	10.00	105.00
Wind_Speed(mph)	30854.0	8.689347	4.999840	0.00	5.80	8.00	12.00	126.60
Precipitation(in)	18160.0	0.044213	0.613277	0.00	0.00	0.00	0.00	10.10

As we can see, we are working with 33256 values and we do not have missing values for the variables we are using, so now we are rready to work on our project.

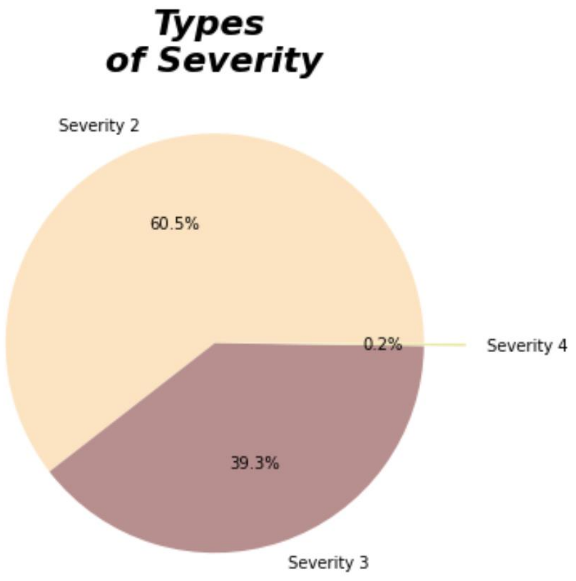
Exploratory Data Analysis

it is always important to see the correlation of the variables, so here we have a graph for a better understanding



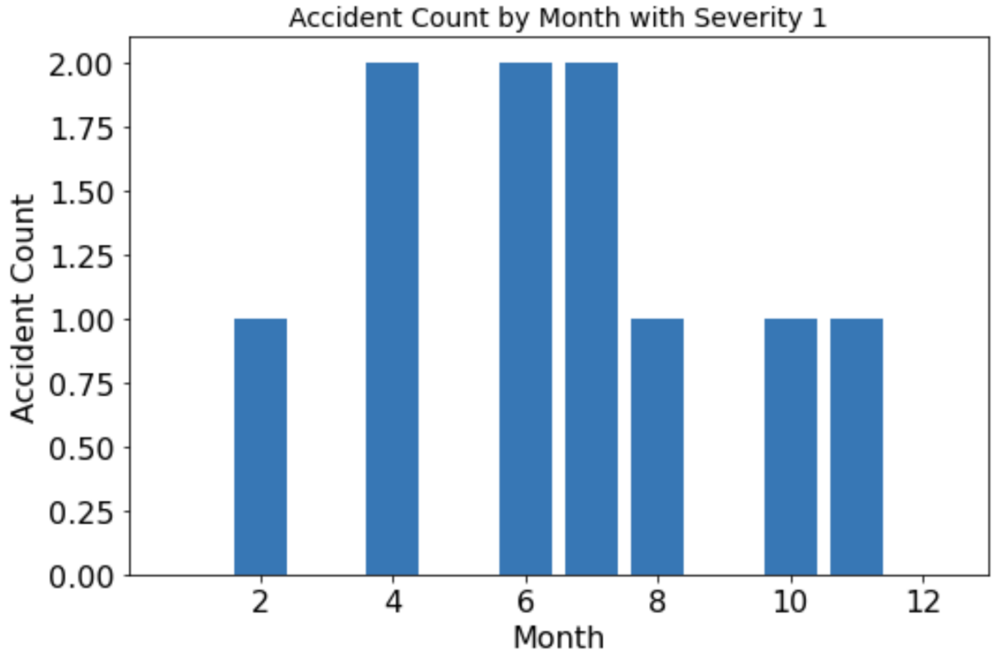
	Severity	Temperature(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)
Severity	1	0.09	-0.12	0.11	0.022	0.051
Temperature(F)	0.09	1	-0.16	0.029	0.23	-0.071
Humidity(%)	-0.12	-0.16	1	-0.28	-0.48	-0.22
Pressure(in)	0.11	0.029	-0.28	1	0.21	-0.076
Visibility(mi)	0.022	0.23	-0.48	0.21	1	0.0004
Wind_Speed(mph)	0.051	-0.071	-0.22	-0.076	0.0004	1

As the temperature increases the severity also does. On the other hand, the more humidity it is, the severity tend to be lower. For pressure is the same as the temperature but higher impact.

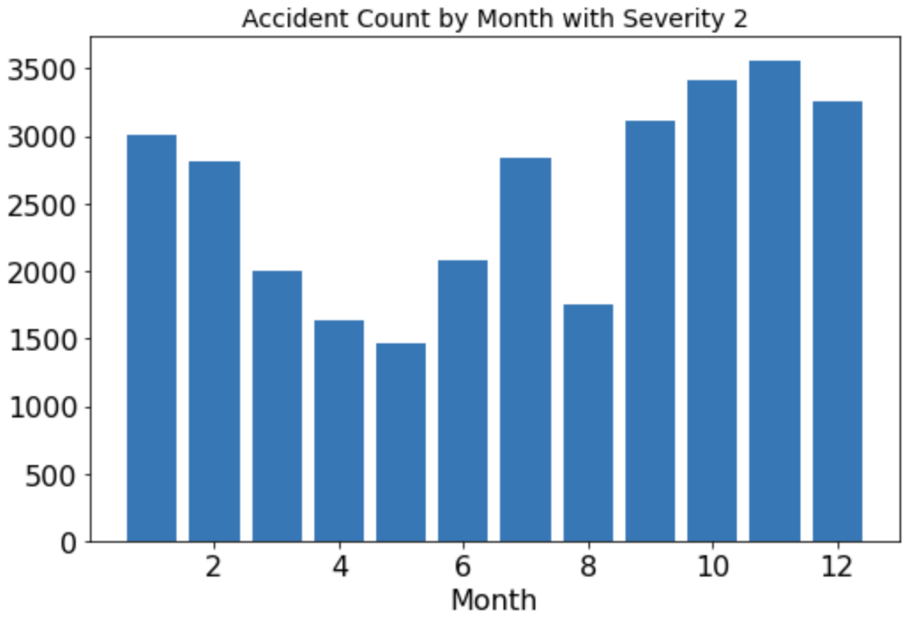


If we look types of severity we see that the most common is car accidents with a severity of 2 with 60.5%. Accidents with severity as 3, there is 39.3%. and only the 0.2% it is severity with type of 4. For those car accidents with severity of 1 it does not show the graph because it is too small proportion which is only 10 car accidents.

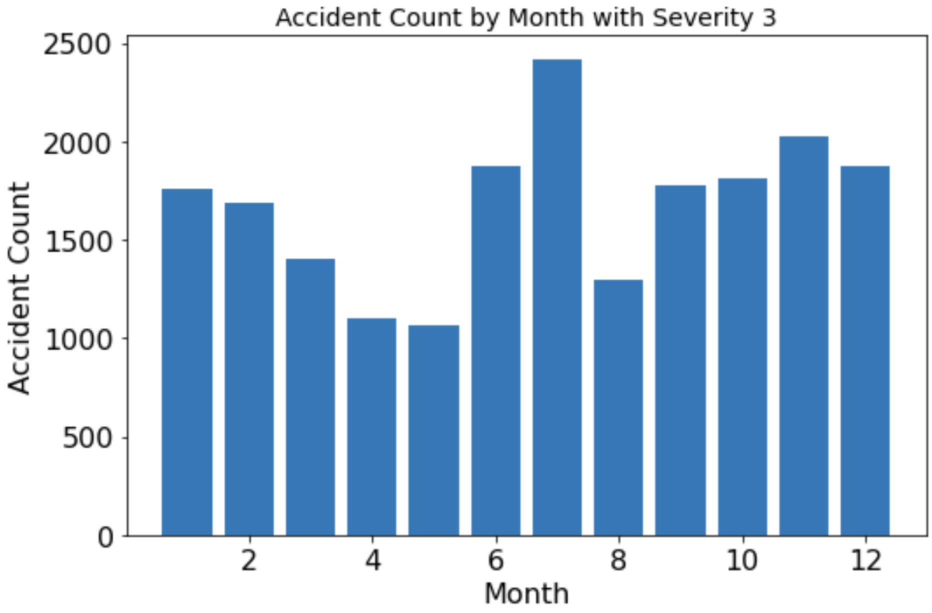
Let's see cases of severity by moth during the year



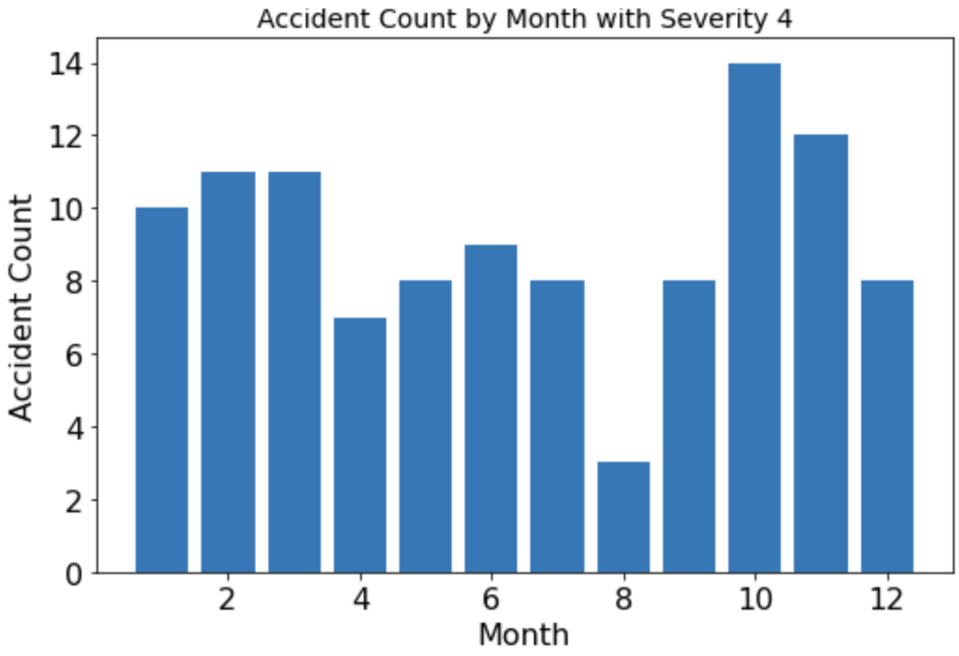
We see thar on April, June and July occurred car accidents with low severity



The most common with severity of 2 occurs on November, octuber and December

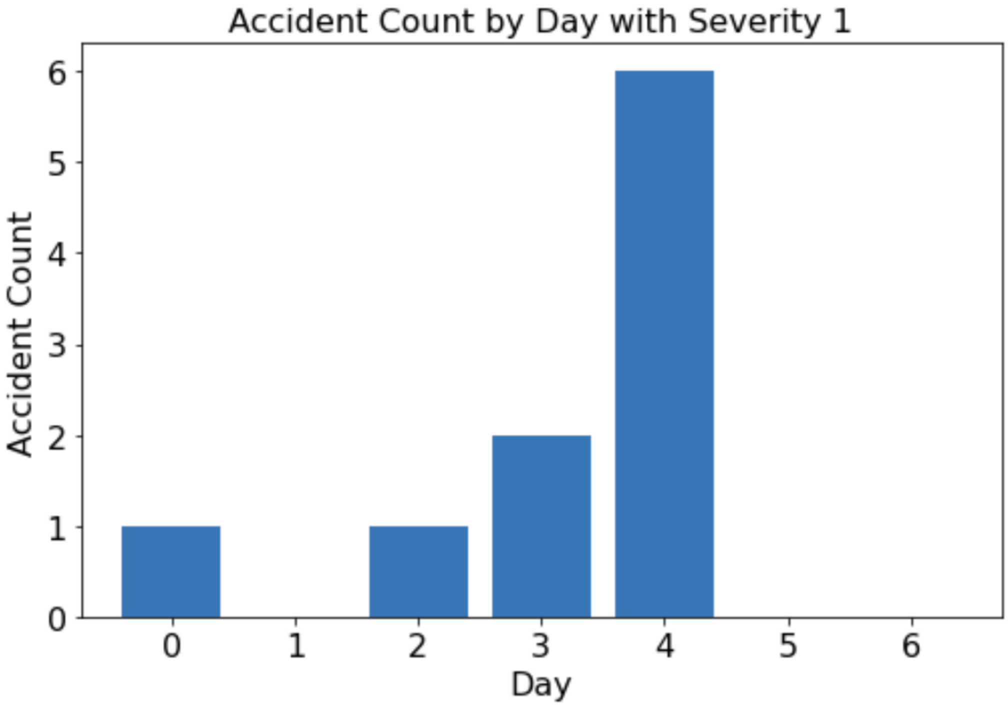


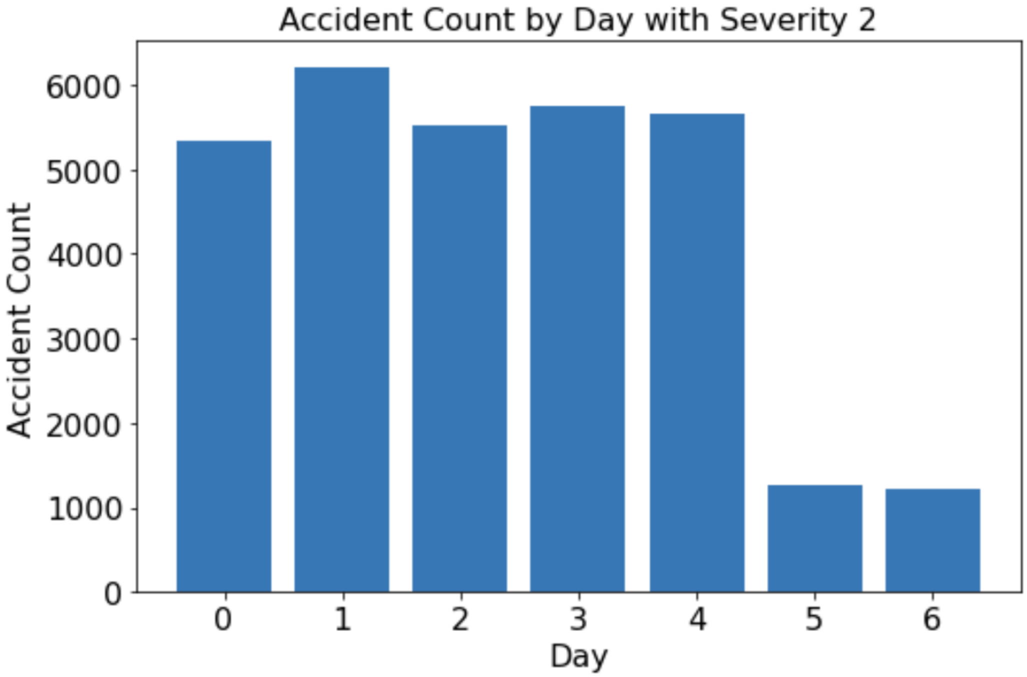
The most common with severity of 3 occurs on July

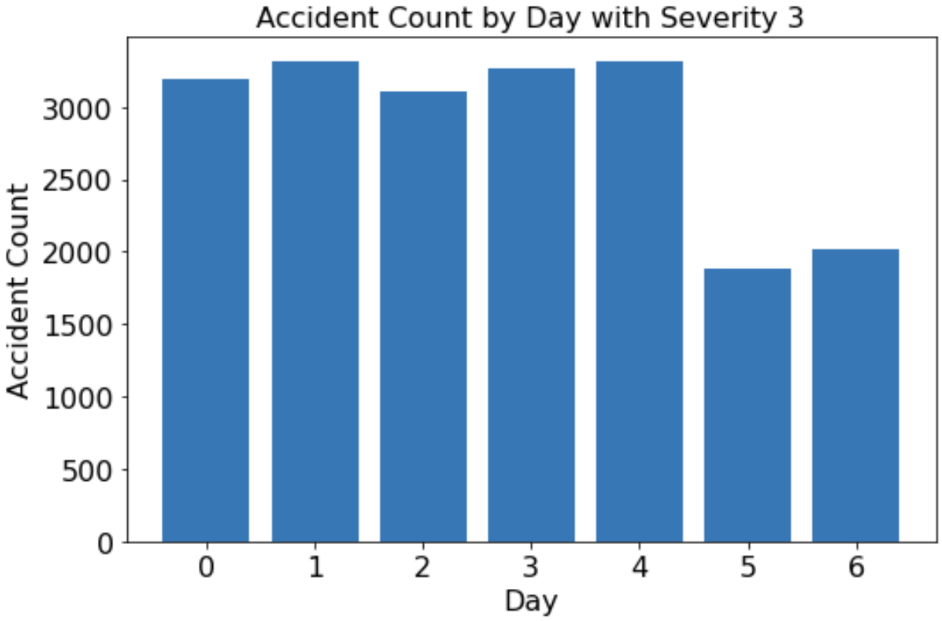


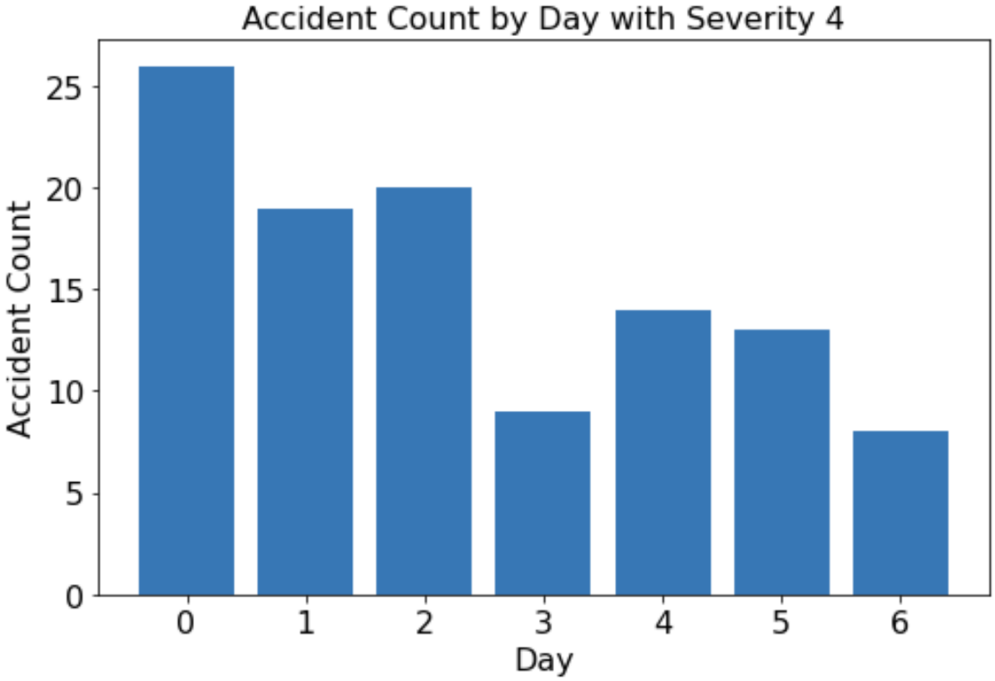
The most common with severity of 4 occurs on October

Now, we consider that doing the analysis by day is also going to provide us interesting information



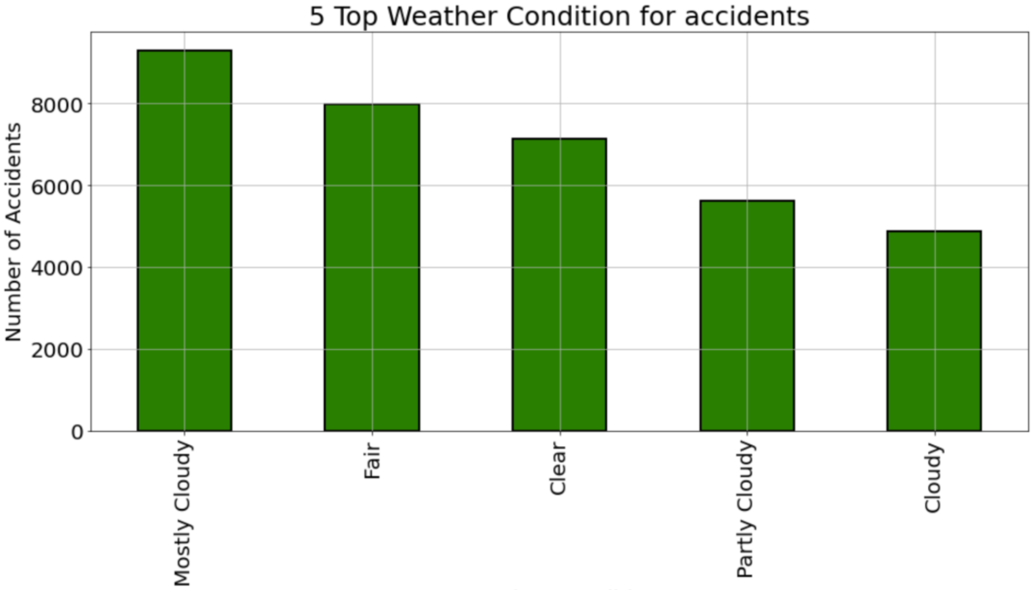

Accident Count by Day with Severity 2
Accident Count
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
Day


Accident Count by Day with Severity 3
Accident Count
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
Day

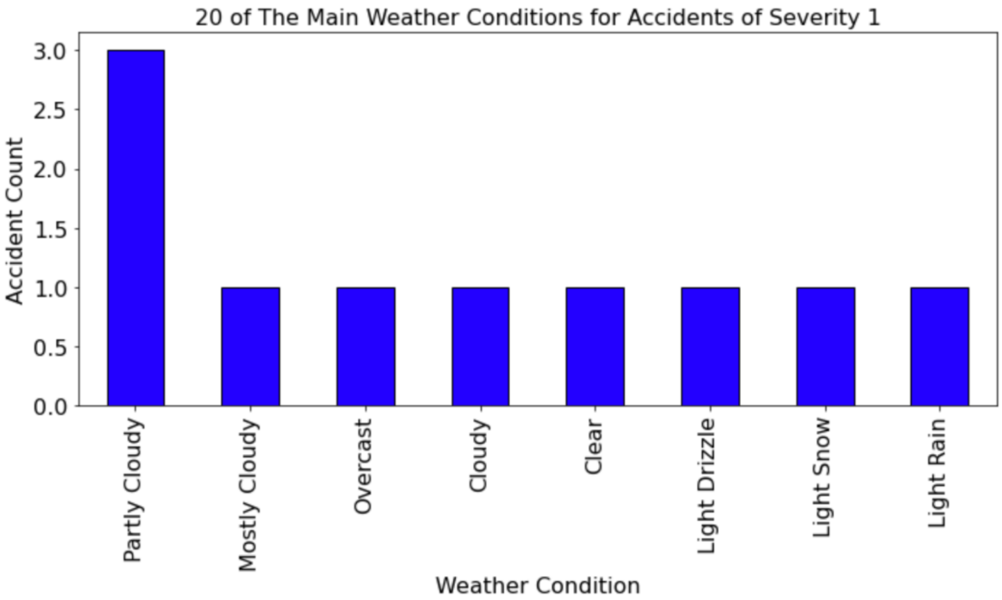


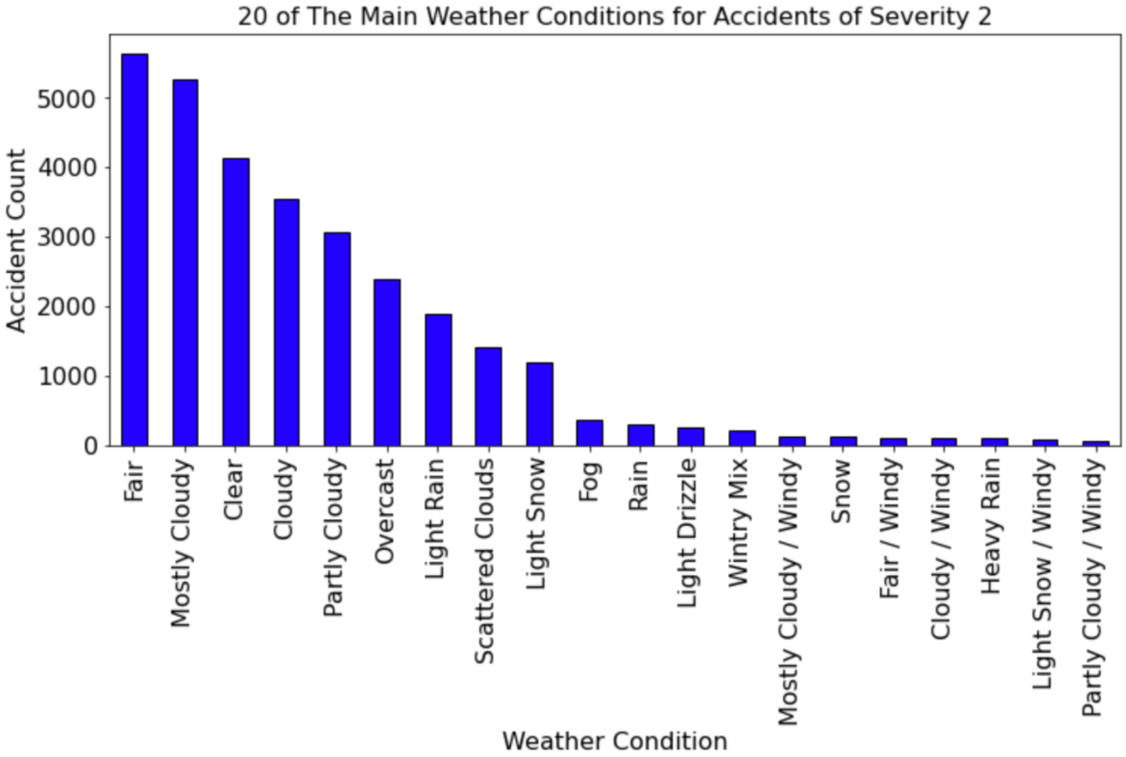
Most of the car accidents happens at the beginning of the week.

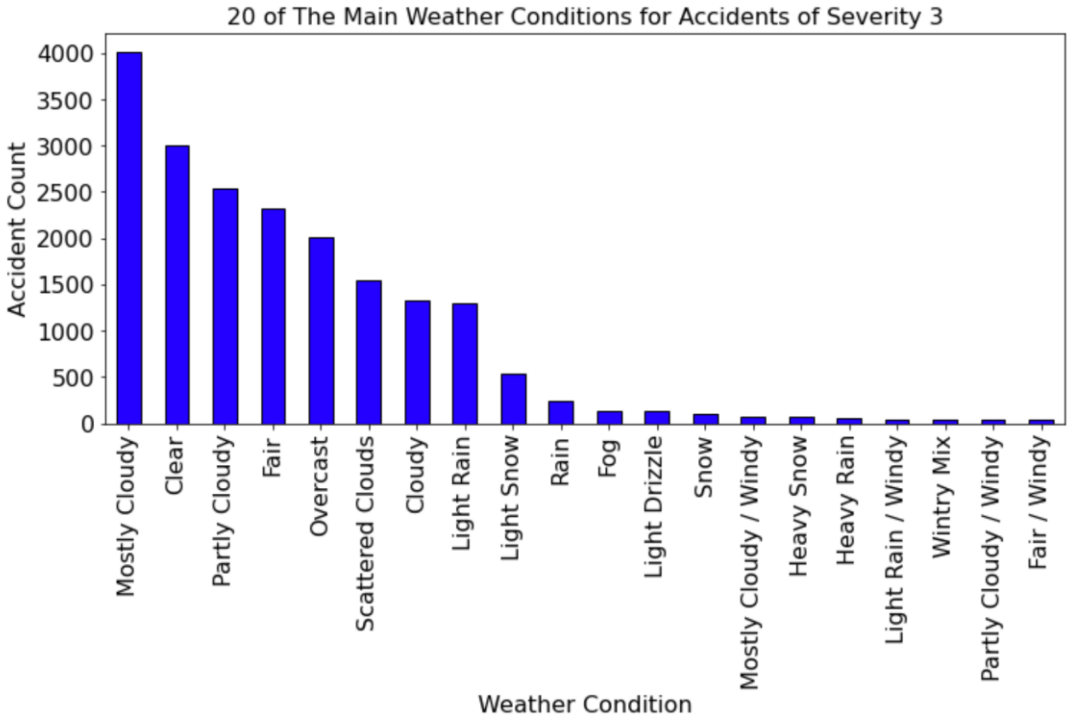
Another variable that we think is key for explaining the severity is the weather condition so here I show the top five

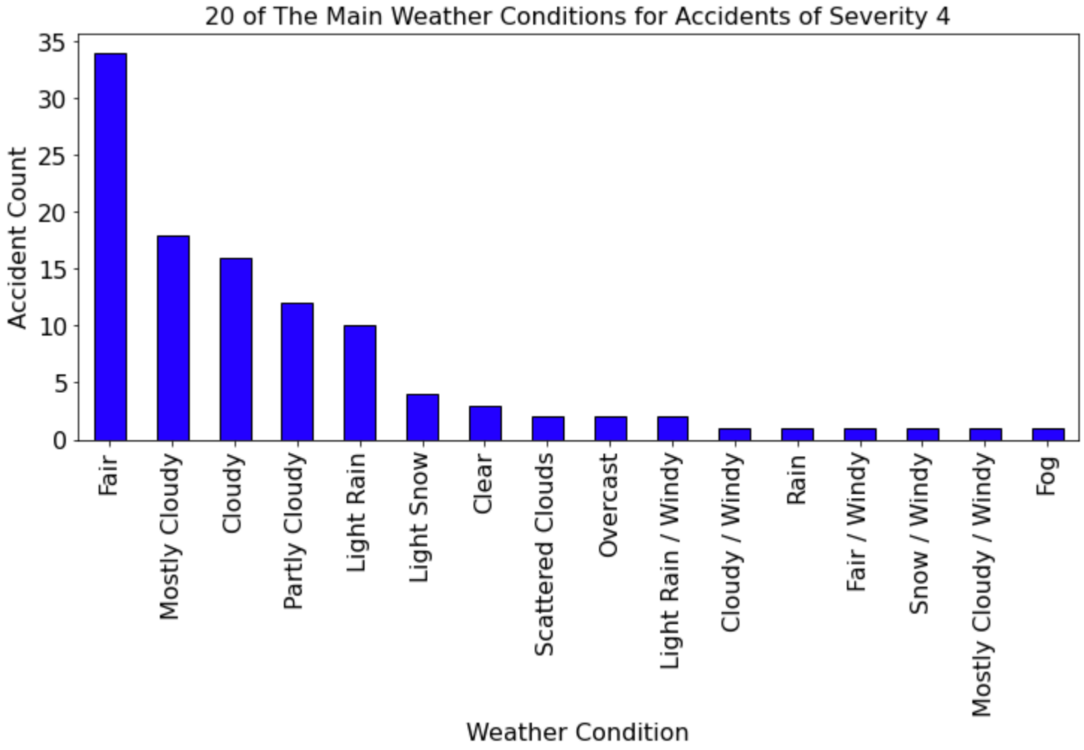


But I also provide an analysis of the weather by severity

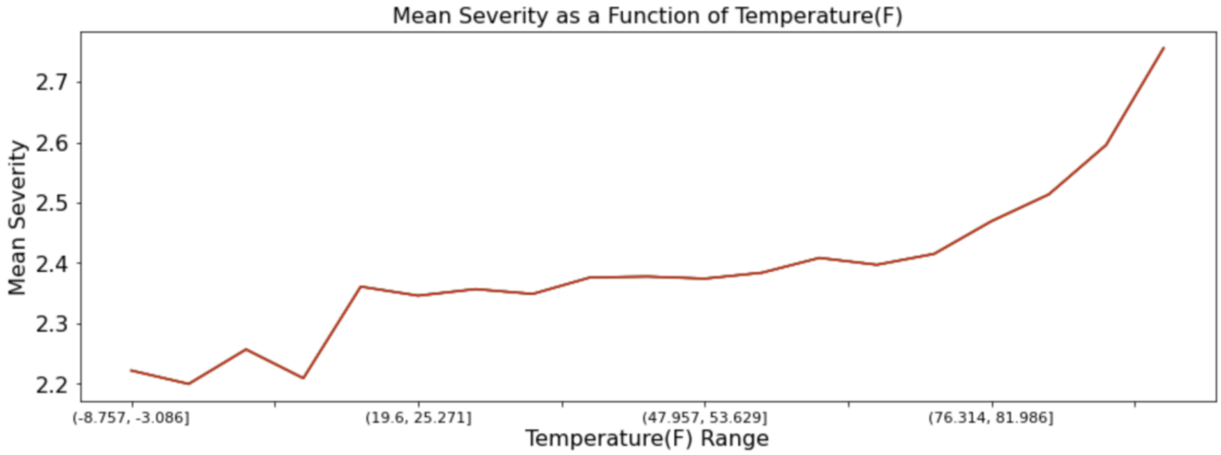

20 of The Main Weather Conditions for Accidents of Severity 1
Accident Count
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Partly Cloudy
Mostly Cloudy
Overcast
Cloudy
Clear
Light Drizzle
Light Snow
Light Rain
Weather Condition


20 of The Main Weather Conditions for Accidents of Severity 2
Accident Count
5000
4000
3000
2000
1000
0
Fair
Mostly Cloudy
Clear
Cloudy
Partly Cloudy
Overcast
Light Rain
Scattered Clouds
Light Snow
Fog
Rain
Light Drizzle
Wintry Mix
Mostly Cloudy / Windy
Snow
Fair / Windy
Cloudy / Windy
Heavy Rain
Light Snow / Windy
Partly Cloudy / Windy
Weather Condition

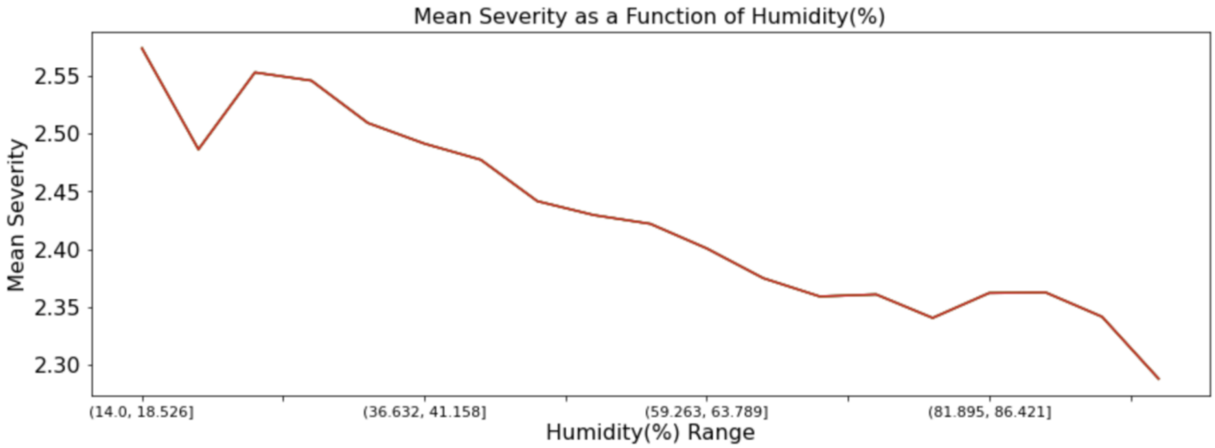

20 of The Main Weather Conditions for Accidents of Severity 3
Accident Count
0
500
1000
1500
2000
2500
3000
3500
4000
Mostly Cloudy
Clear
Partly Cloudy
Fair
Overcast
Scattered Clouds
Cloudy
Light Rain
Light Snow
Rain
Fog
Light Drizzle
Snow
Mostly Cloudy / Windy
Heavy Snow
Heavy Rain
Light Rain / Windy
Wintry Mix
Partly Cloudy / Windy
Fair / Windy
Weather Condition


20 of The Main Weather Conditions for Accidents of Severity 4
Accident Count
0
5
10
15
20
25
30
35
Fair
Mostly Cloudy
Cloudy
Partly Cloudy
Light Rain
Light Snow
Clear
Scattered Clouds
Overcast
Light Rain / Windy
Cloudy / Windy
Rain
Fair / Windy
Snow / Windy
Mostly Cloudy / Windy
Fog
Weather Condition

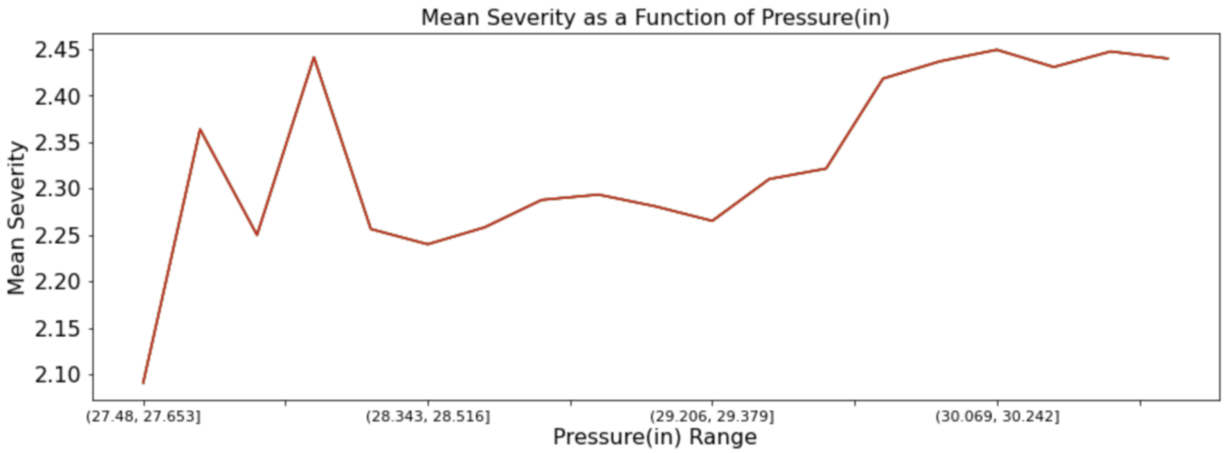
Now, we are going to see what the relation between the severity and the type of environment for climate is. It is important to say that for the severity I used a mean



We see the positive relation between these 2 variables. It is not good a high temperature because increases the severity



Here we have a negative relation. It is desirable to have less humidity because decresess the probability to have a high severity



Finally, if we look at the pressure of the environment, we see that al low inches it is very ambiguous, but the more increases, it also increases the chance to have a car accident with high severity

Modeling

As this is a classification problem, we will compare the performance of some common Machine learning algorithms:

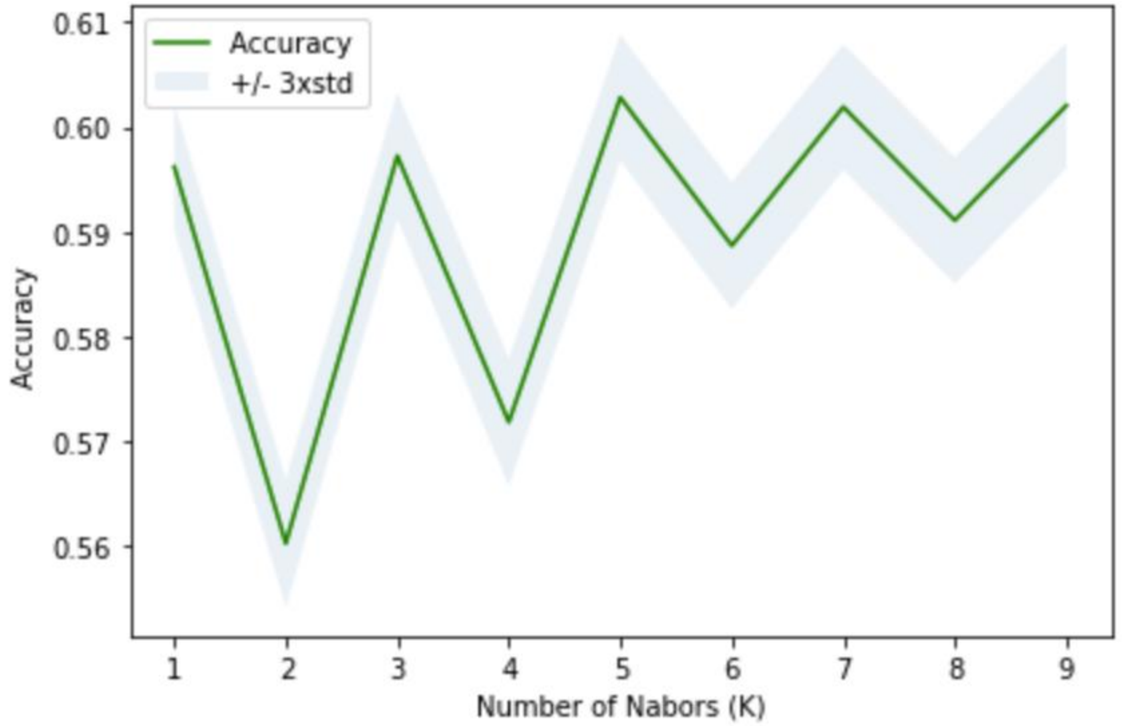
- K Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

. We train-test split our data, leaving 20% for testing. And before we run the the models, we normalize the data

Results

To evaluate the performance of the models, we use accuracy, the F1-score and the Jaccard-score. In the table below we present the scores.

For the KNN model we also calculate the optimum value of k



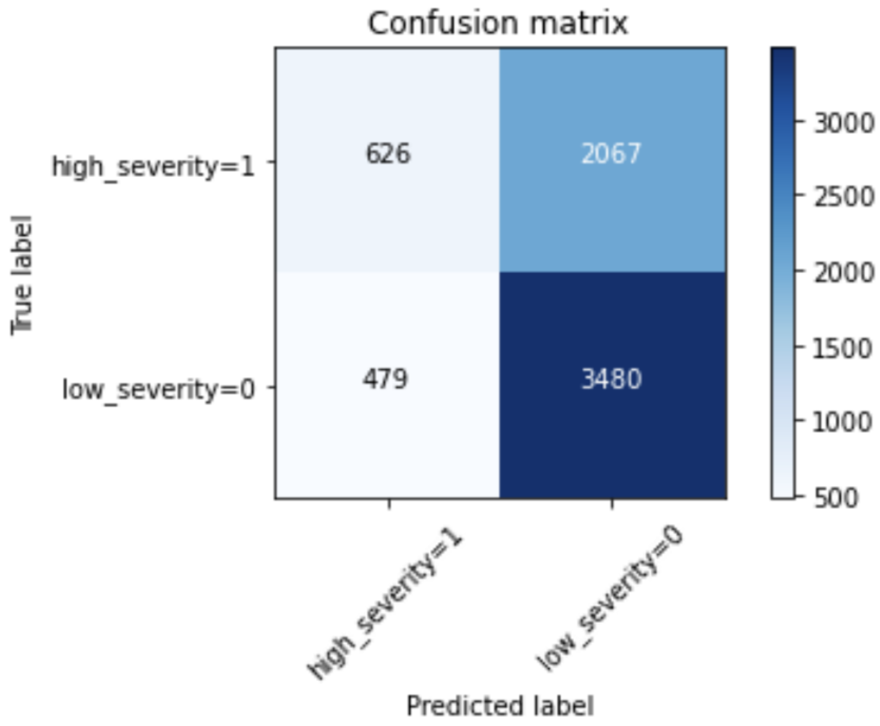
```
The best accuracy was with 0.6028262176788936 with k= 5
```

Algorithm	**Jaccard**	**F1-Score**	**Log-Loss**
KNN	0.6	0.59	
Decision Tree	0.59	0.59	
SVM	0.62	0.56	
Logistic Regression	0.62	0.57	0.65

This are the scores obtained for all models. This indicates that they have a moderate index for all the models. As we see on the table, if we would like to select the best model, that would be the

Logistic Regression model because it has the higher value, so we can do prediction using that model.

Let's see the confusion matrix to see the predictive capacity.



Discussion

At this stage it is worth asking whether the scores can be improved with the features selected. But anyway, we can see that the prediction for a high severity is only 626 cases where the model predicted true and the real value is also true. On the other hand, we see that predicting the cases with low severity is better with 3480 values that are correct.

It is helpful to provide this kind of information because we can be aware of the probability to have a high severity car accident depending on the conditions.

Conclusion

Our original question, “What if we could predict the severity of an accident, if one were to occur, based on weather condition and environmental , and thereby provide information that could reduce their frequency?” we can see that when we have bad weather conditions it is necessary to drive safe because it increases the chance to have an accident with high severity,