US Car Accidents Report

July 15, 2020

Contents

1 Introduction

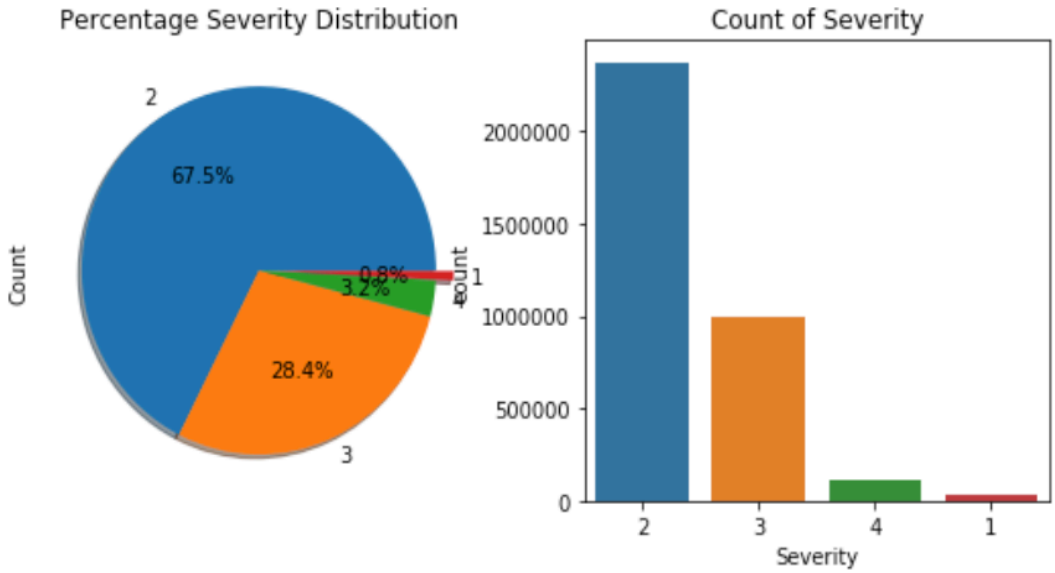
The US Car Accident dataset is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events 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.

In this project, we are going to predict the severity of the car accident from the other given variables. There are many statistical analysis reports available online and for this report, we have regenerated the reports into a comprehensive more. The specific reports that we have used here are attached in the reference section at the end. Prior to this step of the project, we have cleaned and the data of the US Car Accidents 2016-2020.

2 Overview of the Dataset

The response variable that we are interested into is the severity of the car accident. "Severity" is 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)." Here is the distribution of the severity values:

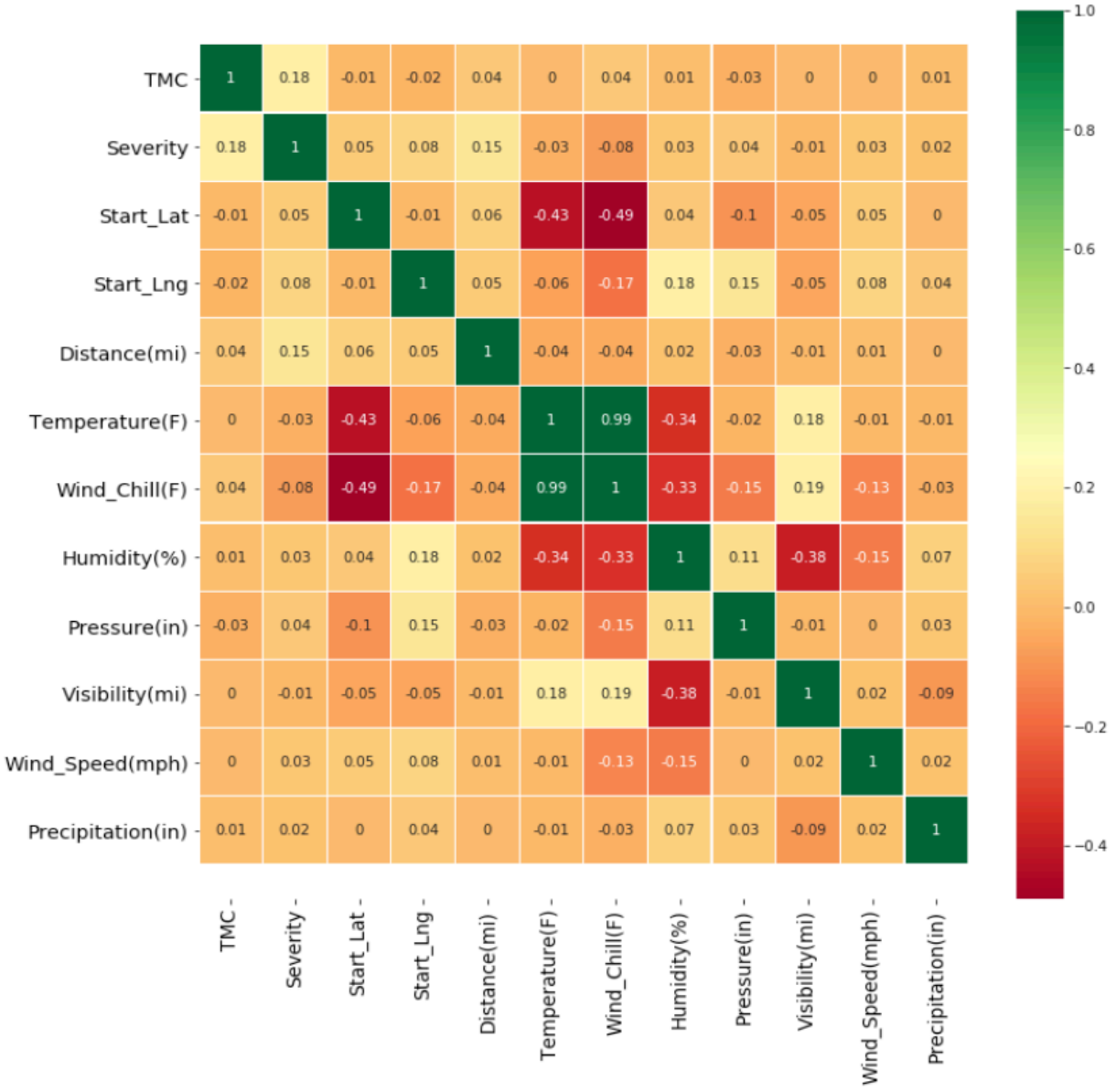
```
2    2373210
3     998913
4     112320
1      29174
Name: Severity, dtype: int64
```



This distribution suggests that the majority of the severities of accidents are those labeled

as 2 or 3.

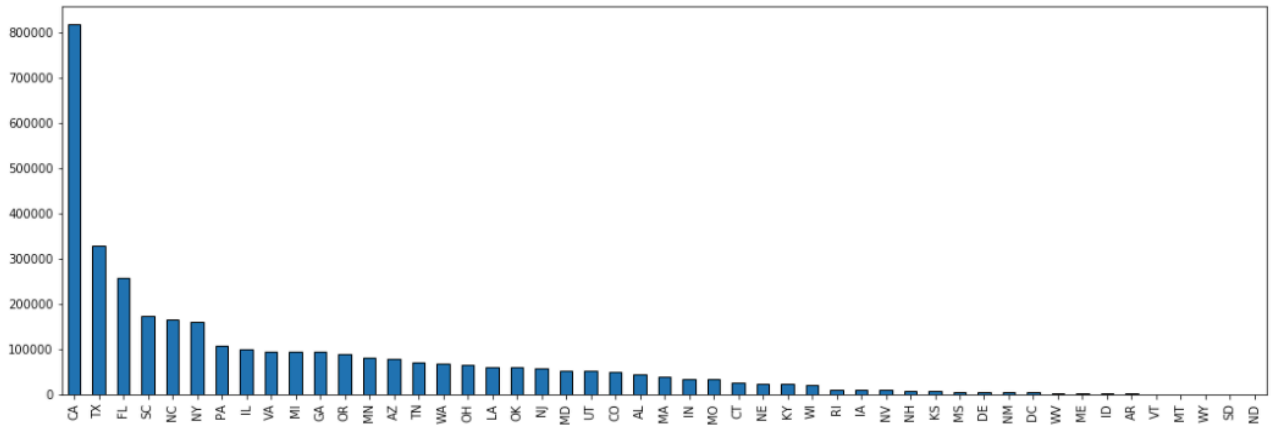
To investigate on the correlations among the variables, we can perform a heat-map to check which of the columns are correlated with the severity the most.



This heat-map does not suggest any strong correlation between a numerical variable and severity of an car accident. This can be seen from the Pearson Coefficients related to "Severity" that all ρ's are far smaller than 0.5. Thus, the severity of the car accidents can be attributed to a combination of factors instead, and further analysis of the dataset is therefore needed.

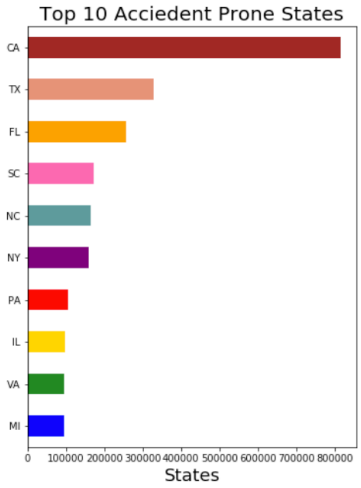
3 Relationship between Location and Severity

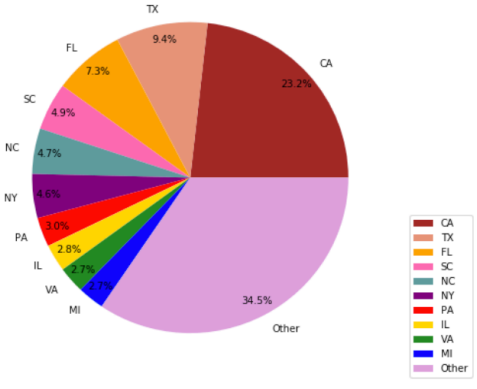
Firstly, let's check the distribution of the accidents over the states.



We can observe from this graph that the number of accidents significantly differ from state to state. However, the other part of the question is whether the severity of an accident is associated with the state or not. In order to answer this question, we will check the distribution of the severities of the accidents in 10 states with the highest number of accidents.

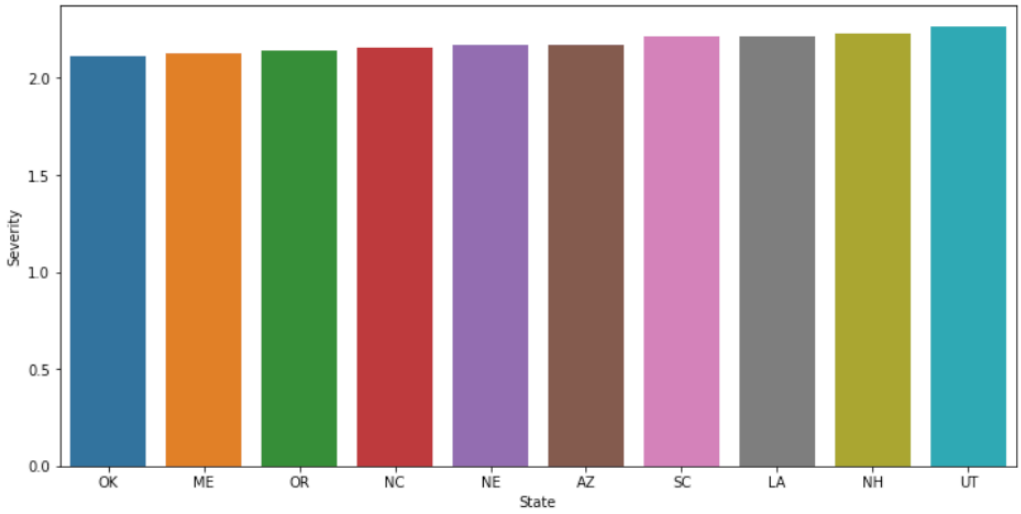
```
CA    816825
TX    329284
FL    258002
SC    173277
NC    165958
NY    160817
PA    106787
IL     99692
VA     96075
MI     95983
Name: State, dtype: int64
```



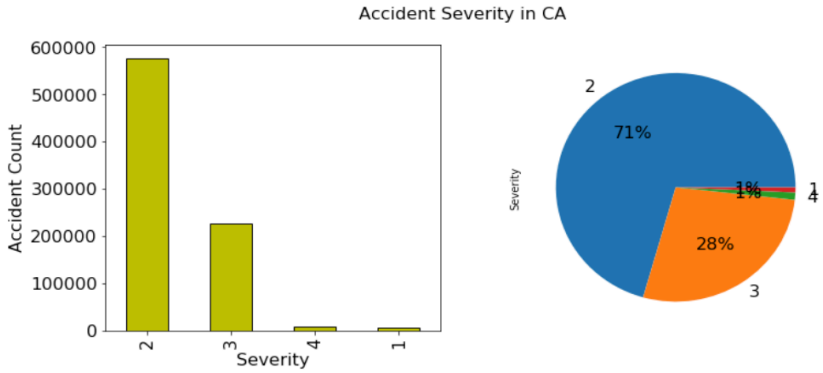


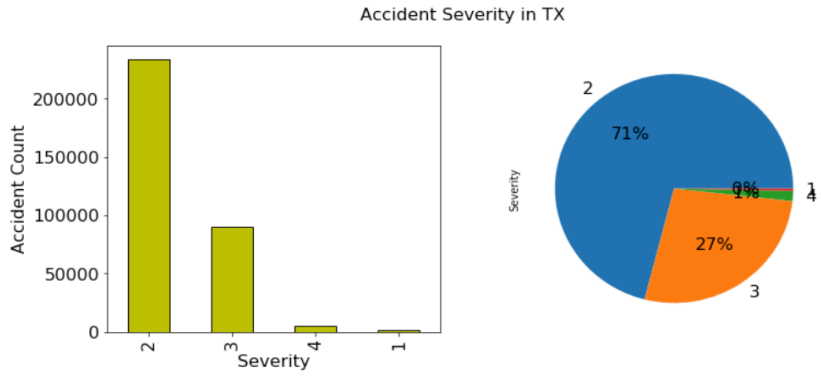
We can see that more accidents are happening in state of California(CA), Texas(TX) and Florida(FL).

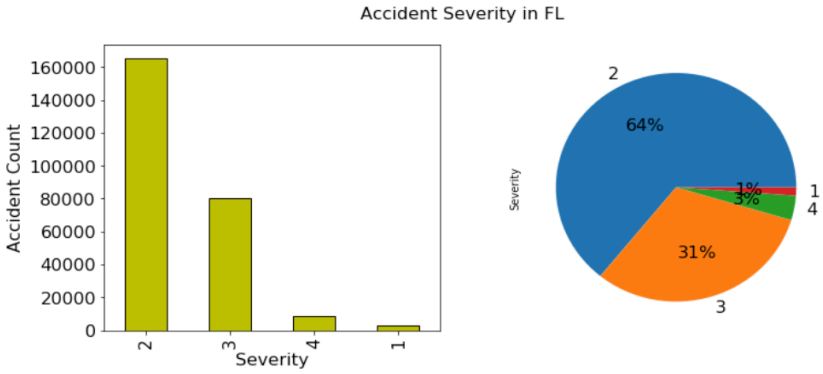
Besides, in terms of the level of mean severity, the top 10 states are shown below:

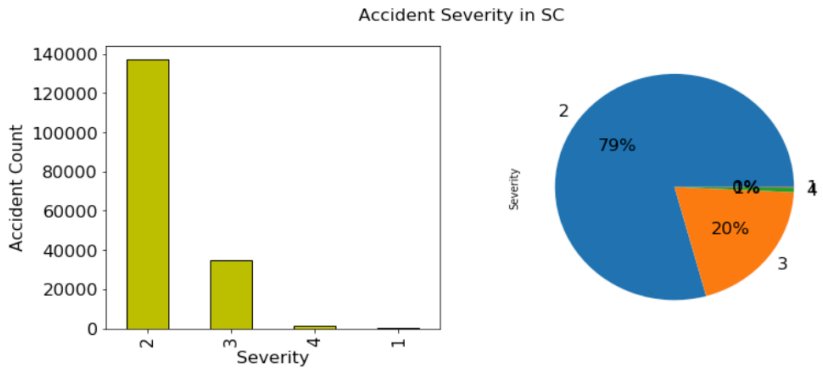


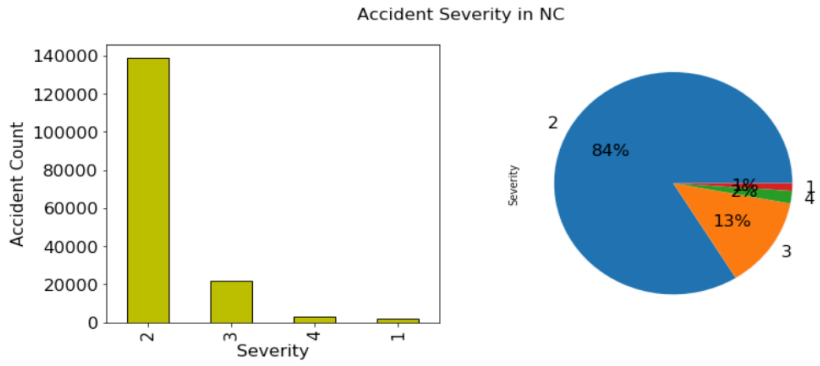
Now, let's take a further look at the top 10 states with the most number of car accidents in the US sorted by the level of severity.

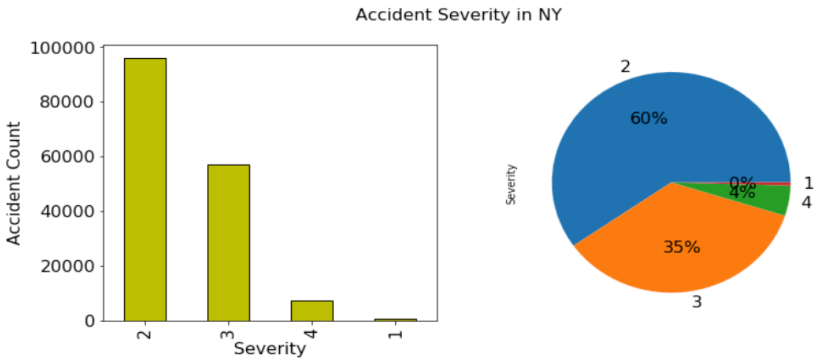


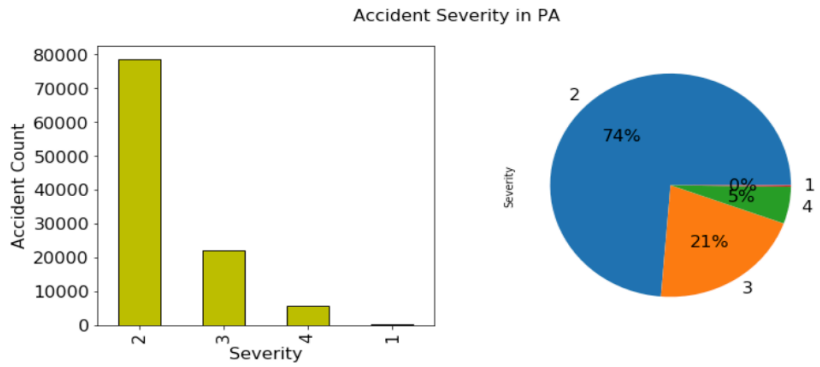

Accident Severity in TX
Accident Count
200000
150000
100000
50000
0
2
3
4
1
Severity
2
71%
27%
3
Severity

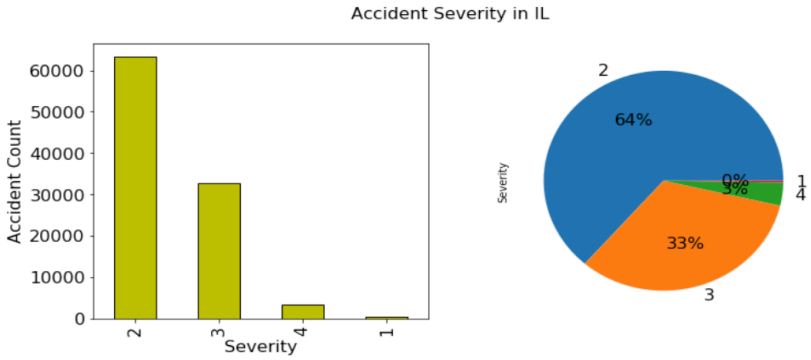

Accident Severity in FL
Accident Count
160000
140000
120000
100000
80000
60000
40000
20000
0
2
3
4
1
Severity
2
64%
31%
3
Severity

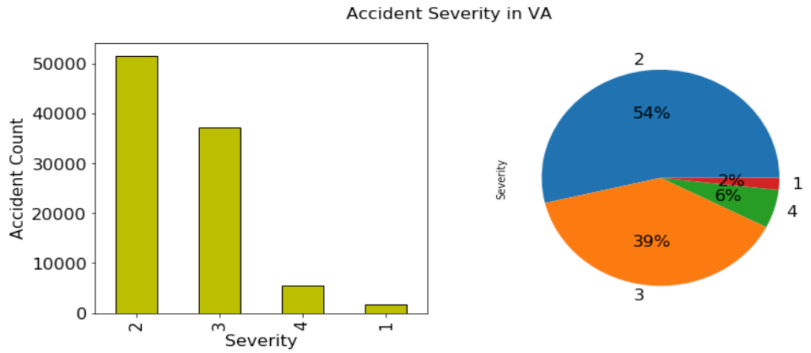

Accident Severity in SC
Accident Count
140000
120000
100000
80000
60000
40000
20000
0
2
3
4
1
Severity
2
79%
20%
3
Severity

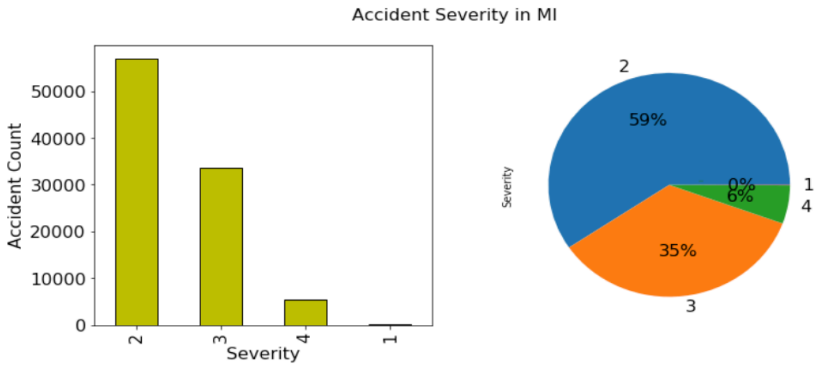

Accident Severity in NC
Accident Count
140000
120000
100000
80000
60000
40000
20000
0
2
3
4
1
Severity
Severity
2
84%
1%
2%
13%
3
1
4


Accident Severity in NY
Accident Count
100000
80000
60000
40000
20000
0
2
3
4
1
Severity
Severity
2
60%
0%
4%
35%
3
1
4


Accident Severity in PA
Accident Count
80000
70000
60000
50000
40000
30000
20000
10000
0
2
3
4
1
Severity
Severity
2
74%
0%
5%
21%
3
1
4

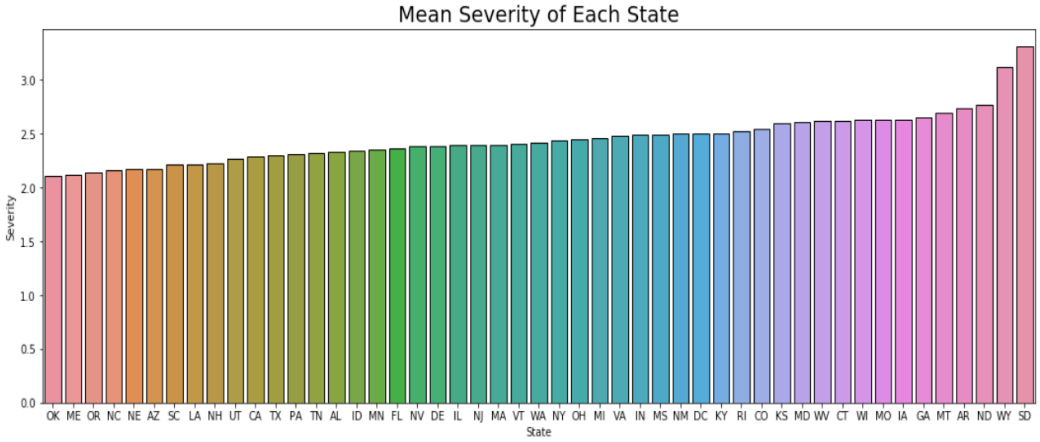




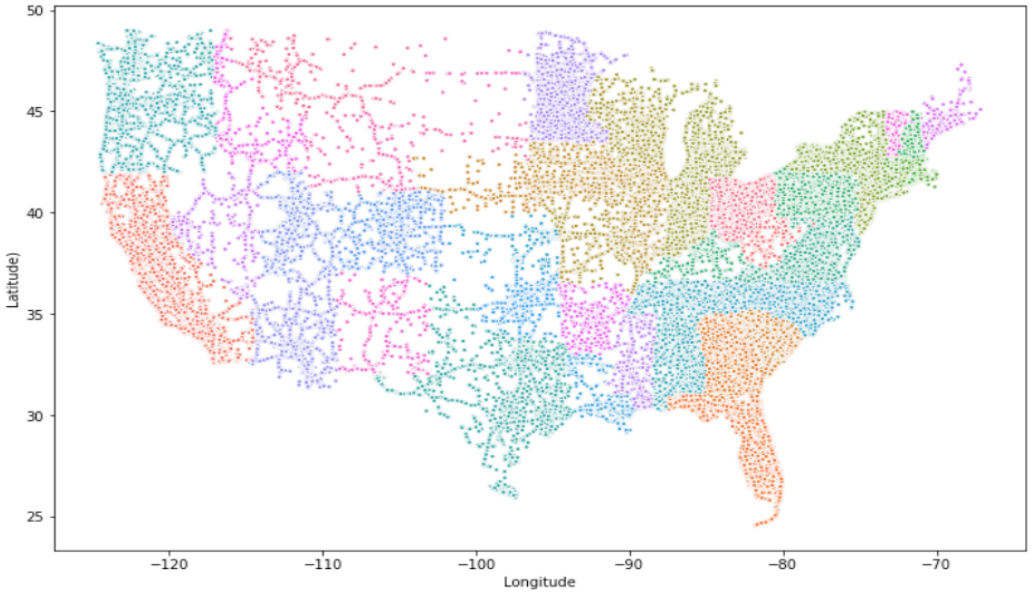


These graphs clearly state that the severity distribution differs from state to state. So state is a variable that may affect the severity of an accident.

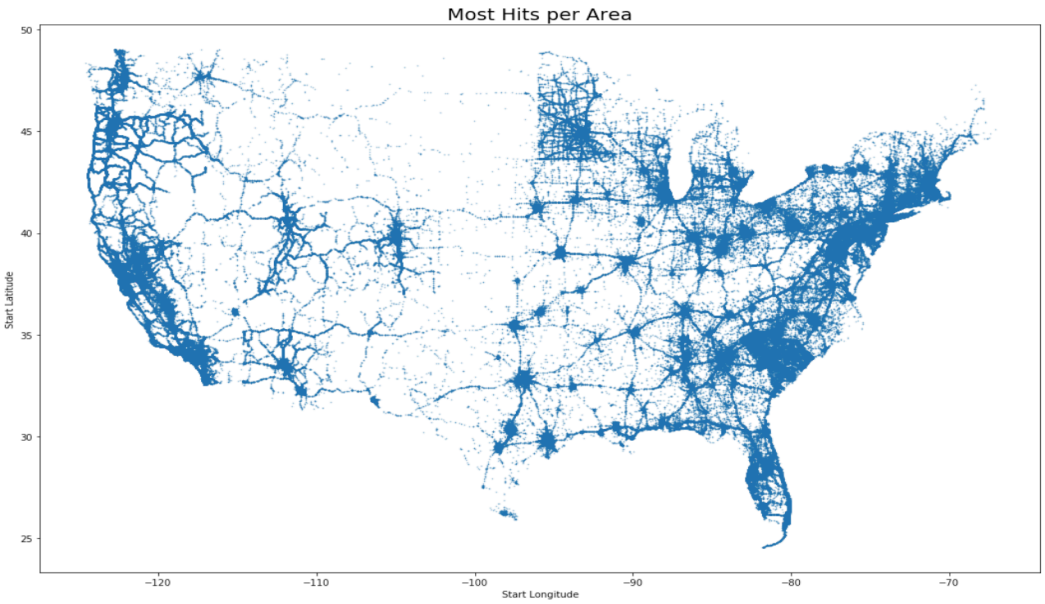
Since severity is a discrete numerical variable in ordinal level, so mean severity may have meaning in the context, we can show the mean severities of each state on a bar graph.



To get a more explicit visualisation on the density of car accidents, let's check the accident data on a map by colouring each state with a different colour.

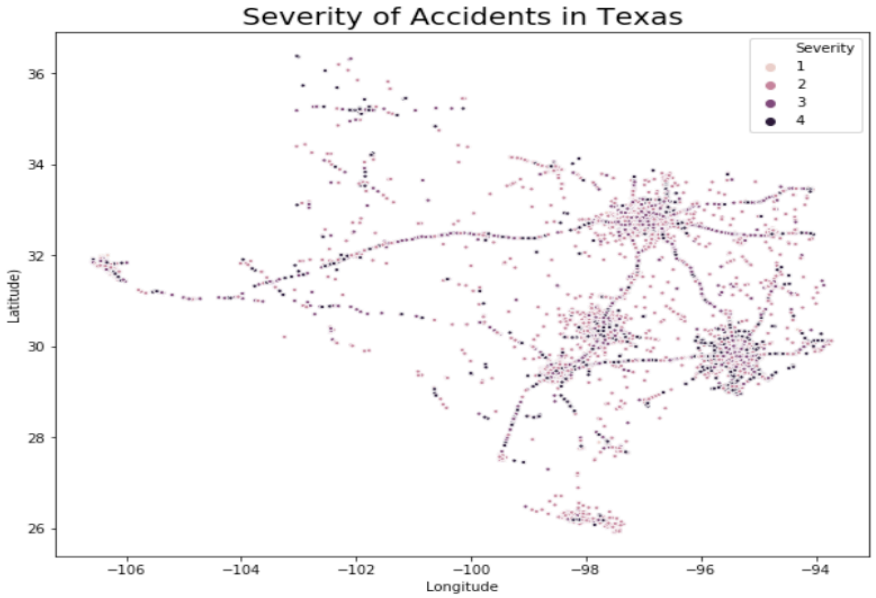


Even though this graph can shows the distribution of the accidents in each state, it does not clearly shows the clusters. If we remove 'hue'=State, we can observe which areas and which roads have more accidents than the others.

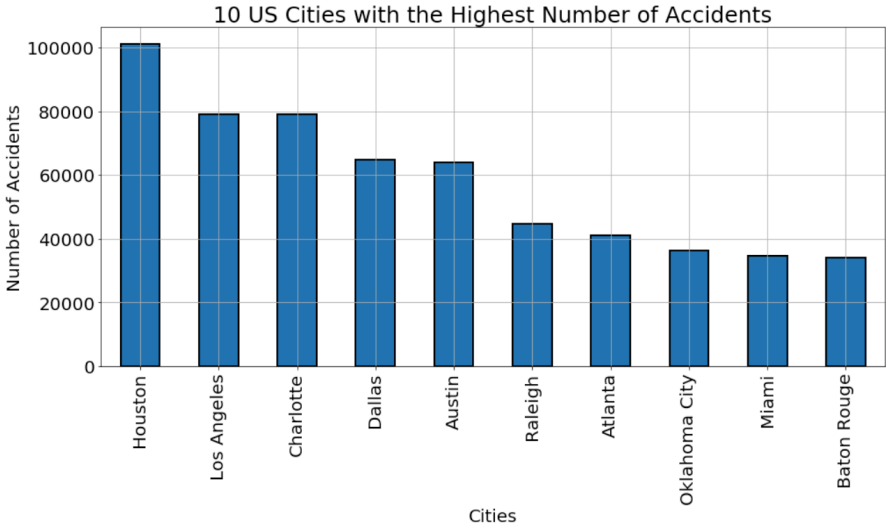


The graph above shows all the accidents in the USA, and we can easily see that the accidents are clustered in metropolitan areas and major highways. However, it does not indicate the severity of accidents.

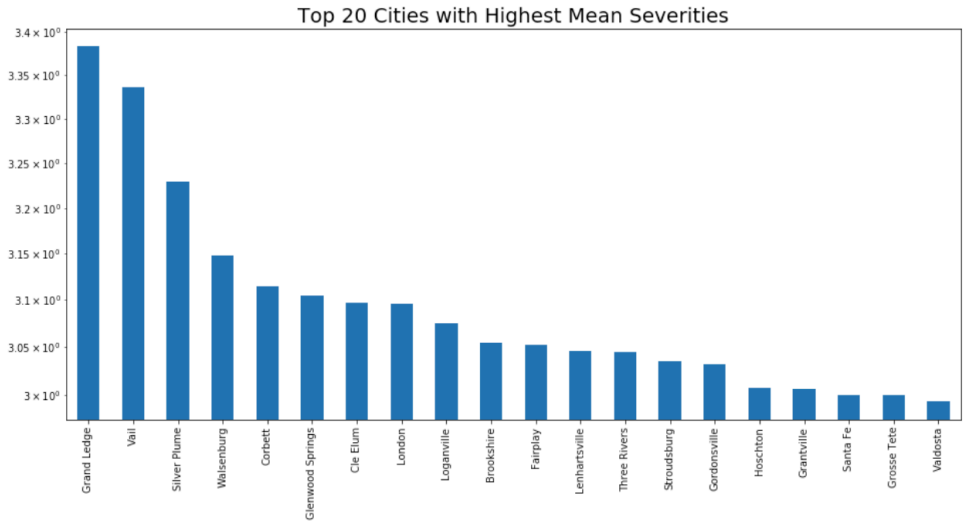
In order to show the severity of an accident, we choose the State of Texas to see how more severe accidents are clustered: We can perform a similar scatter plots to see the distribution



of severities of accidents in each city. I chose Los Angeles, since Los Angeles is a unique name for Los Angeles, CA in this dataset.

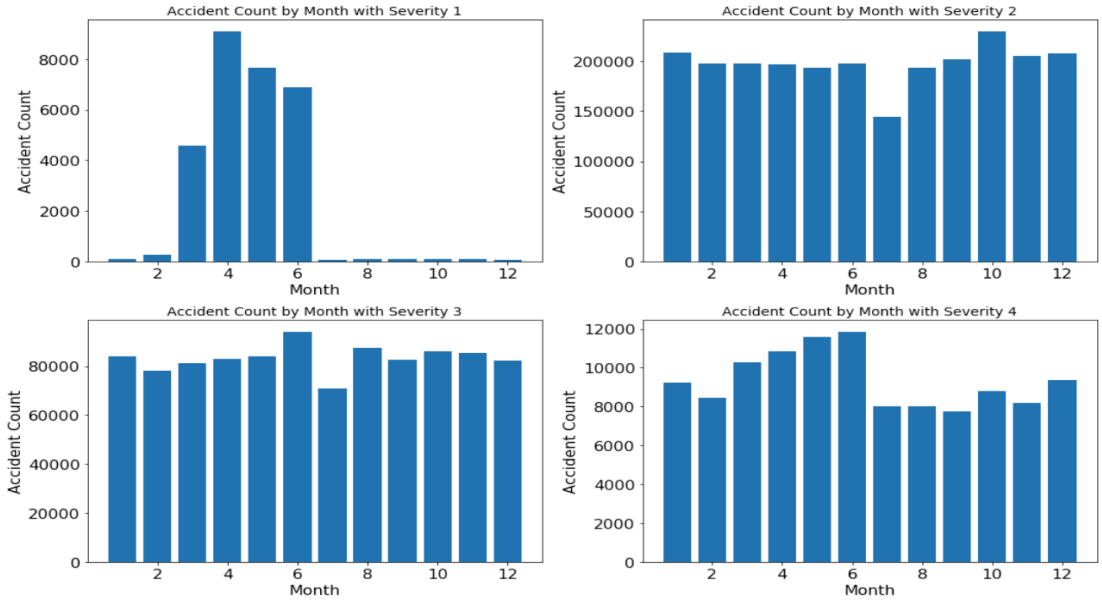


Previously, we have checked on the top states with highest mean severity. Now, we will check which cities have the highest mean of severities. One of the problem is many cities have only one car accident data, and if their accident's severity is 4, their mean will be 4, which will be misleading. We will sort the cities which has more than 100 car accidents to see which cities have a high number of accidents with a high mean severity. Note that we will use log y-scale since the mean values are pretty close to each other.

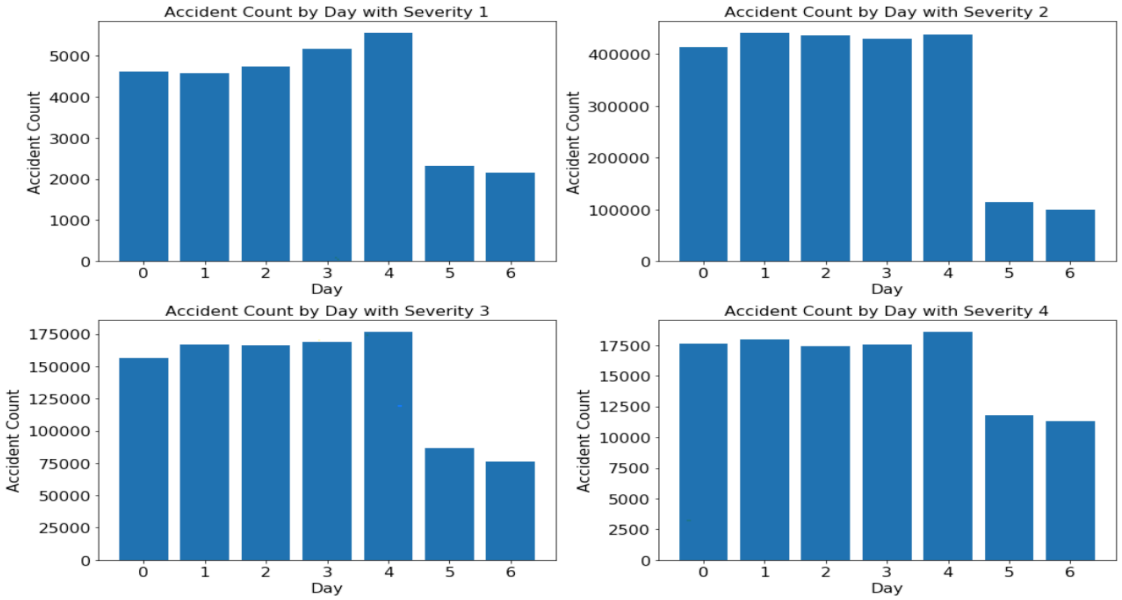


4 Relationship between Time and Severity

4.1 Monthly report



4.2 Daily report



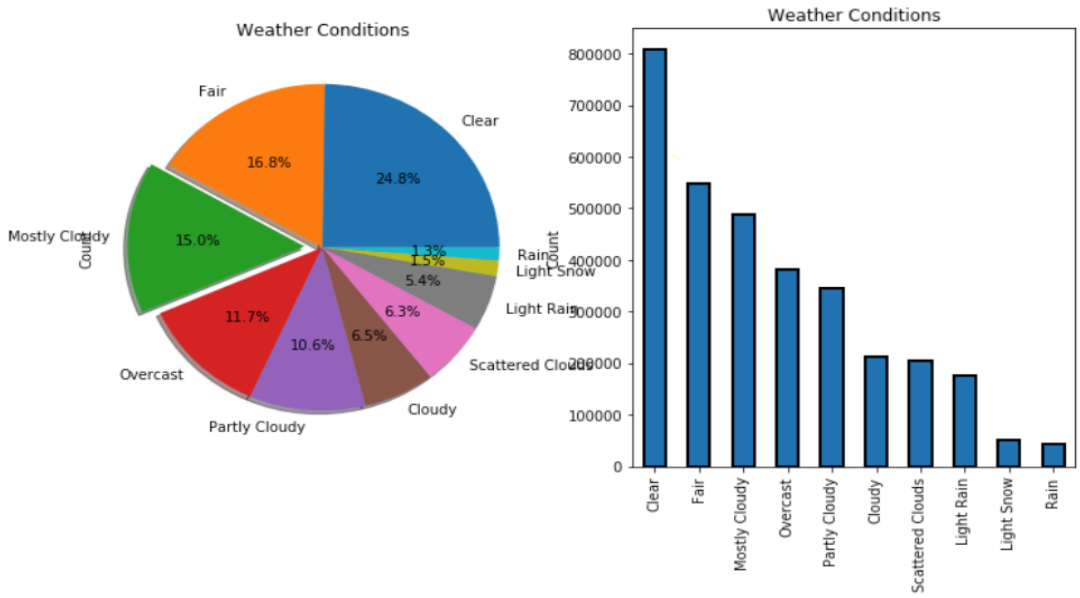
From the Monthly report by severity, we find that the number of accident counts is the smallest in July for all the four levels of severity.Besides, the count pattern for severity of level 2 and 3 are very similar, while the patterns for level 1 and 4, especially level 1, is very different.

From the daily report, we can see that there is a drop in the number of accidents for all severity levels during the weekend. Although, the relative drop for level 3 and 4 is smaller.

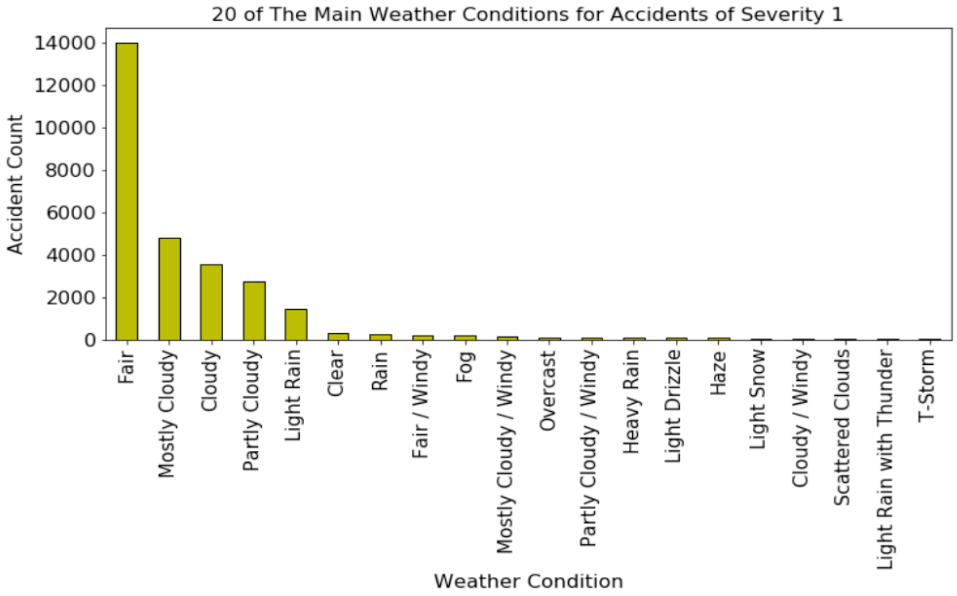
5 Relationship between Weather and Severity

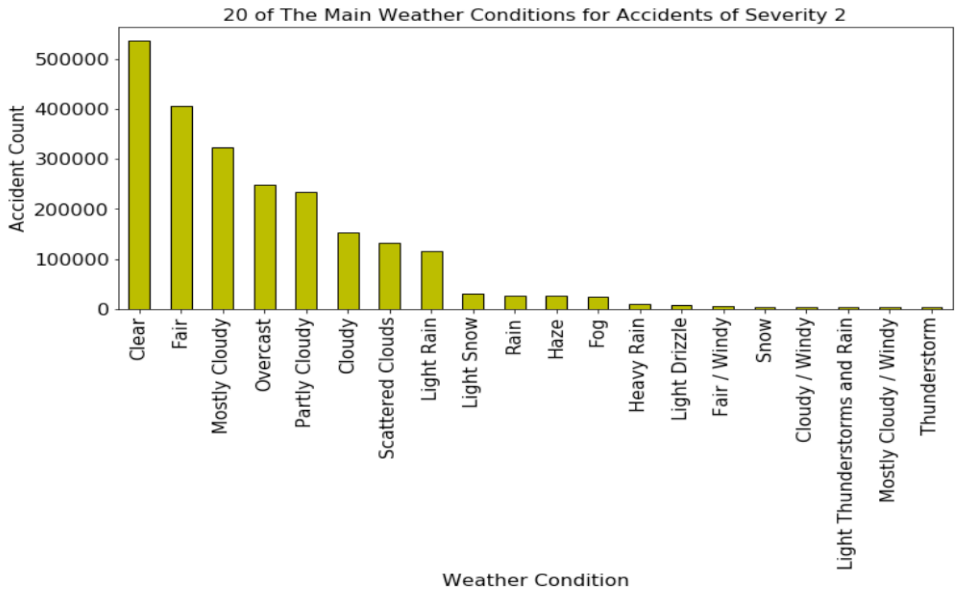
5.1 Most Frequent Weather Conditions

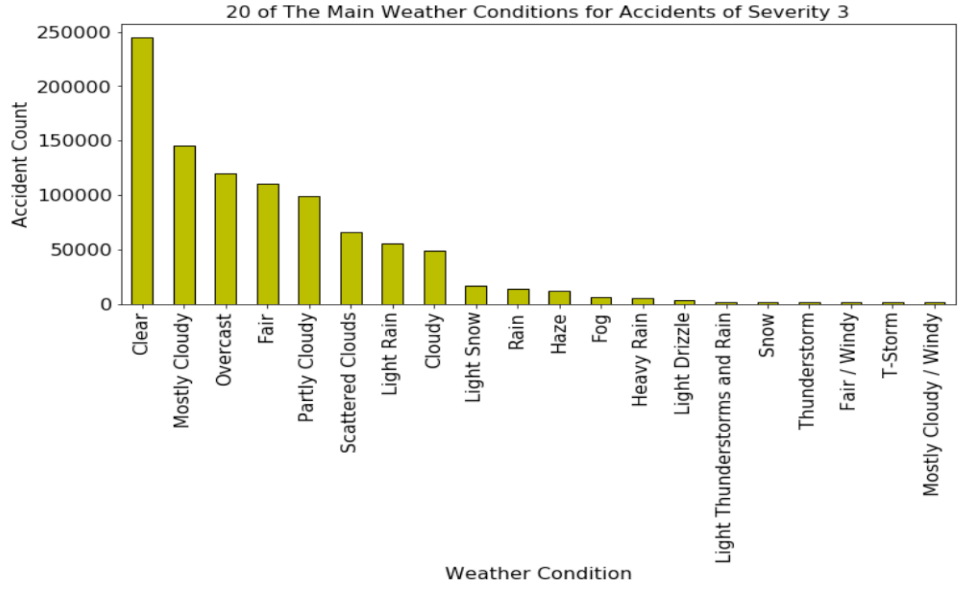
The top 10 accident weather conditions is shown as below:

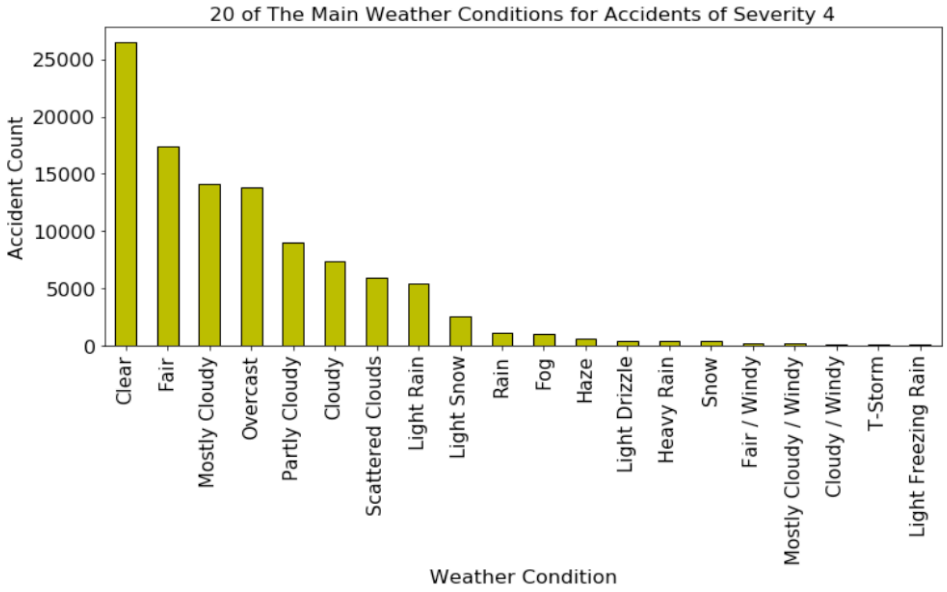


Across all levels of severity, most accidents happen under clear, cloudy, fair or similar weather conditions. These conditions are considered benign compared to rain and snow, Perhaps they are the most frequent conditions. Light rain and light snow are the top adverse weather conditions. Most likely these cause accidents since they can make roads slippery without causing concern in the drivers. Details are shown below:

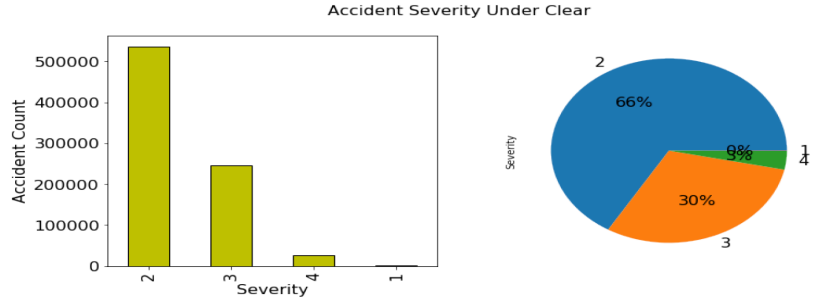


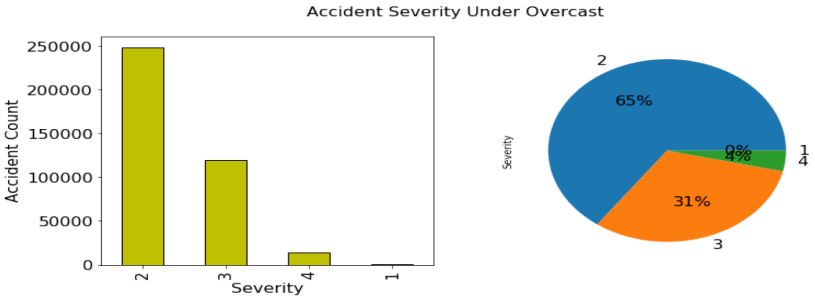

20 of The Main Weather Conditions for Accidents of Severity 2
Accident Count
500000
400000
300000
200000
100000
0
Clear
Fair
Mostly Cloudy
Overcast
Partly Cloudy
Cloudy
Scattered Clouds
Light Rain
Light Snow
Rain
Haze
Fog
Heavy Rain
Light Drizzle
Fair / Windy
Snow
Cloudy / Windy
Light Thunderstorms and Rain
Mostly Cloudy / Windy
Thunderstorm
Weather Condition

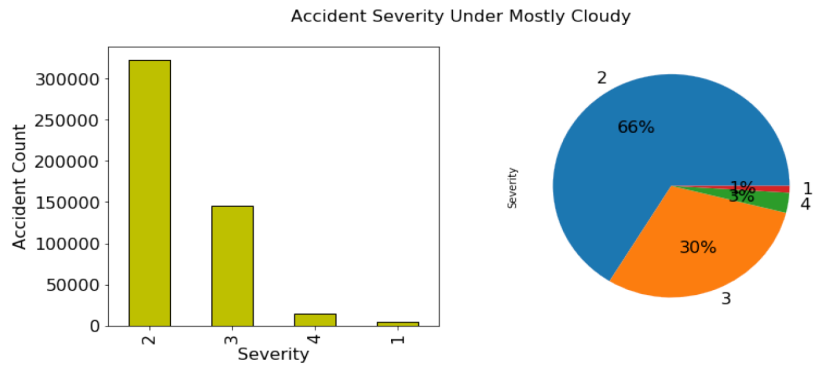

20 of The Main Weather Conditions for Accidents of Severity 3
Accident Count
250000
200000
150000
100000
50000
0
Clear
Mostly Cloudy
Overcast
Fair
Partly Cloudy
Scattered Clouds
Light Rain
Cloudy
Light Snow
Rain
Haze
Fog
Heavy Rain
Light Drizzle
Light Thunderstorms and Rain
Snow
Thunderstorm
Fair / Windy
T-Storm
Mostly Cloudy / Windy
Weather Condition

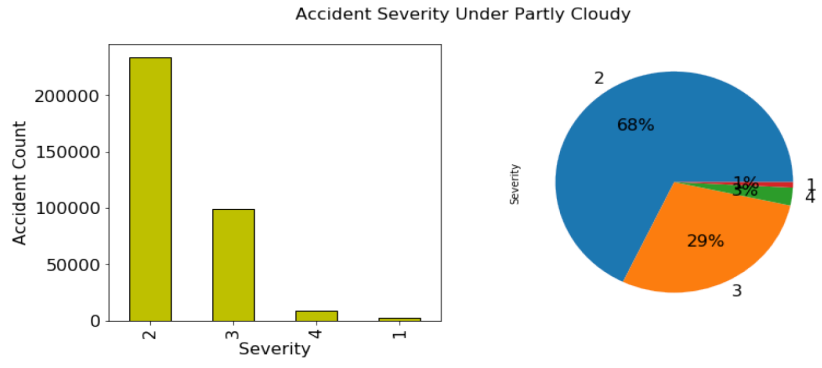


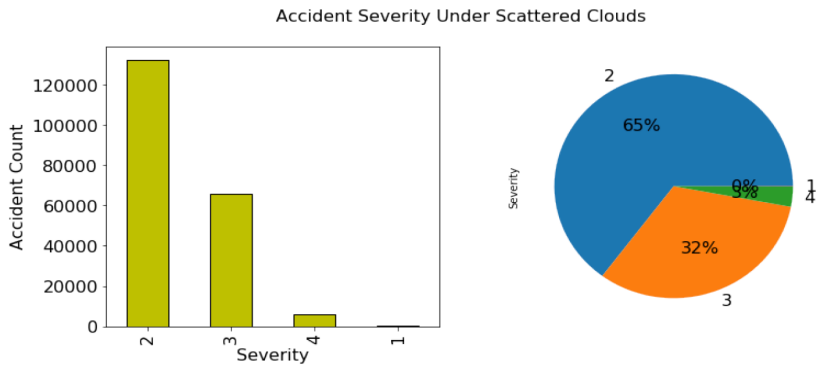
5.2 Severity under each Weather Conditions

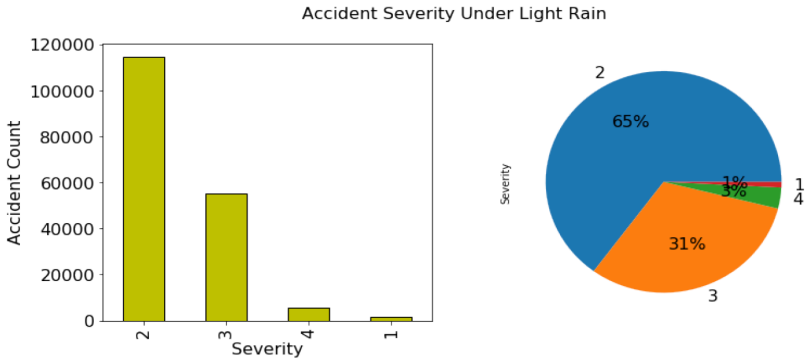


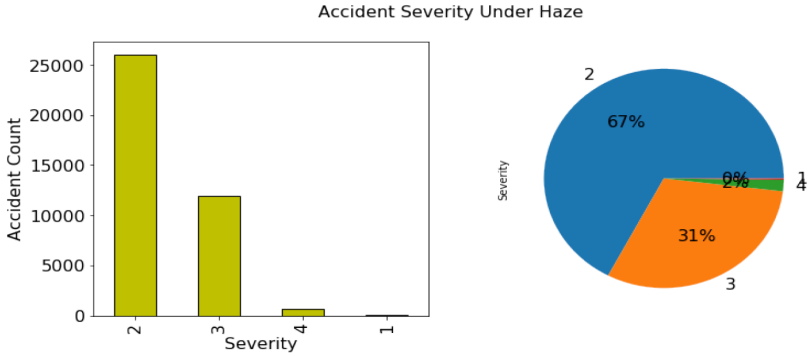


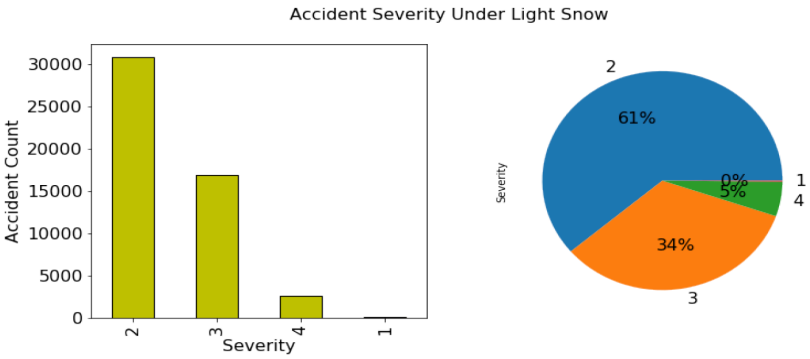

Accident Severity Under Mostly Cloudy
Accident Count
300000
250000
200000
150000
100000
50000
0
2
3
4
1
Severity
2
66%
1%
3%
1
4
30%
3
Severity

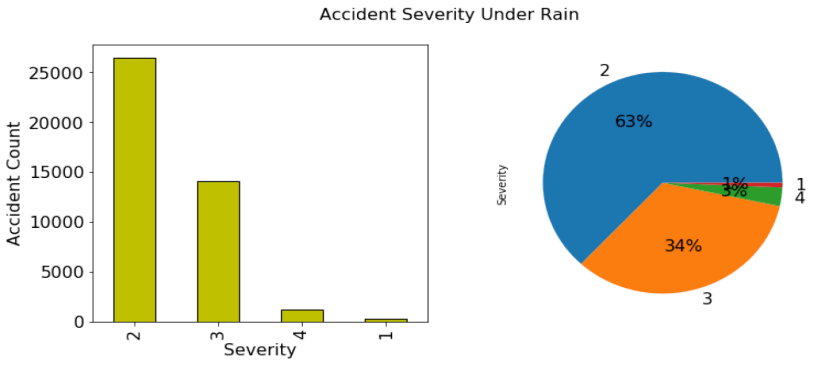

Accident Severity Under Partly Cloudy
Accident Count
200000
150000
100000
50000
0
2
3
4
1
Severity
2
68%
1%
3%
1
4
29%
3
Severity

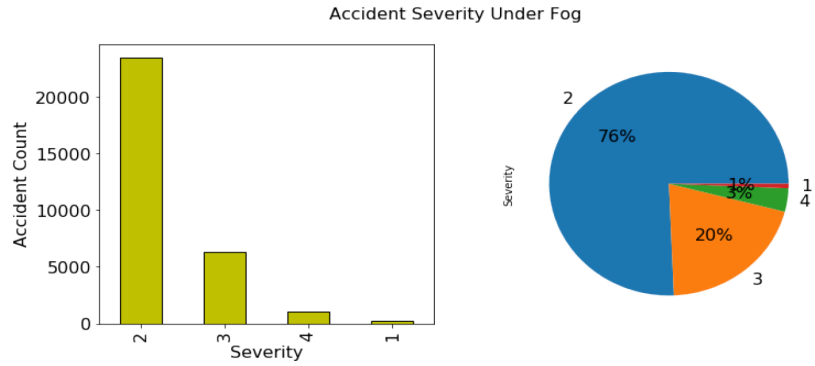

Accident Severity Under Scattered Clouds
Accident Count
120000
100000
80000
60000
40000
20000
0
2
3
4
1
Severity
2
65%
0%
3%
1
4
32%
3
Severity

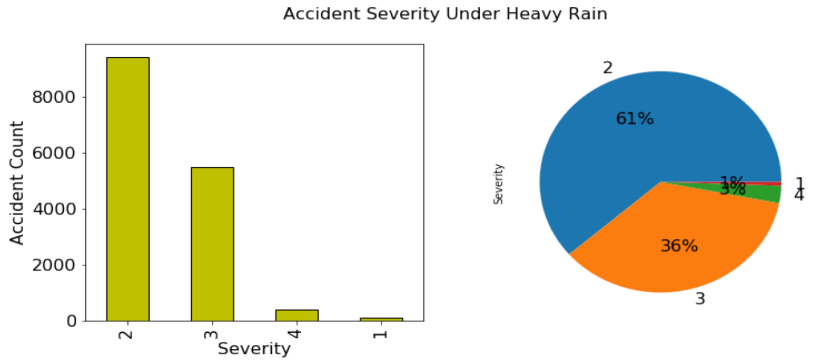

Accident Severity Under Light Rain
Accident Count
120000
100000
80000
60000
40000
20000
0
2
3
4
1
Severity
Severity
2
65%
1%
3%
1
4
31%
3

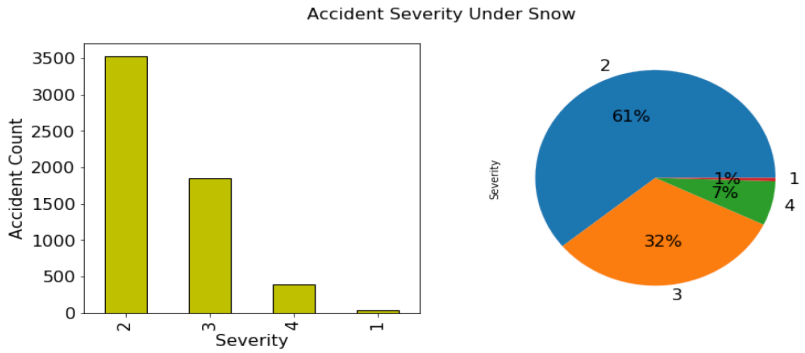

Accident Severity Under Haze
Accident Count
25000
20000
15000
10000
5000
0
2
3
4
1
Severity
Severity
2
67%
0%
2%
1
4
31%
3


Accident Severity Under Light Snow
Accident Count
30000
25000
20000
15000
10000
5000
0
2
3
4
1
Severity
Severity
2
61%
0%
5%
1
4
34%
3


Accident Severity Under Rain
Accident Count
25000
20000
15000
10000
5000
0
2
3
4
1
Severity
Severity
2
63%
1%
3%
1
4
34%
3


Accident Severity Under Fog
Accident Count
20000
15000
10000
5000
0
2
3
4
1
Severity
Severity
2
76%
1%
3%
1
4
20%
3


Accident Severity Under Heavy Rain
Accident Count
8000
6000
4000
2000
0
2
3
4
1
Severity
Severity
2
61%
1%
3%
1
4
36%
3

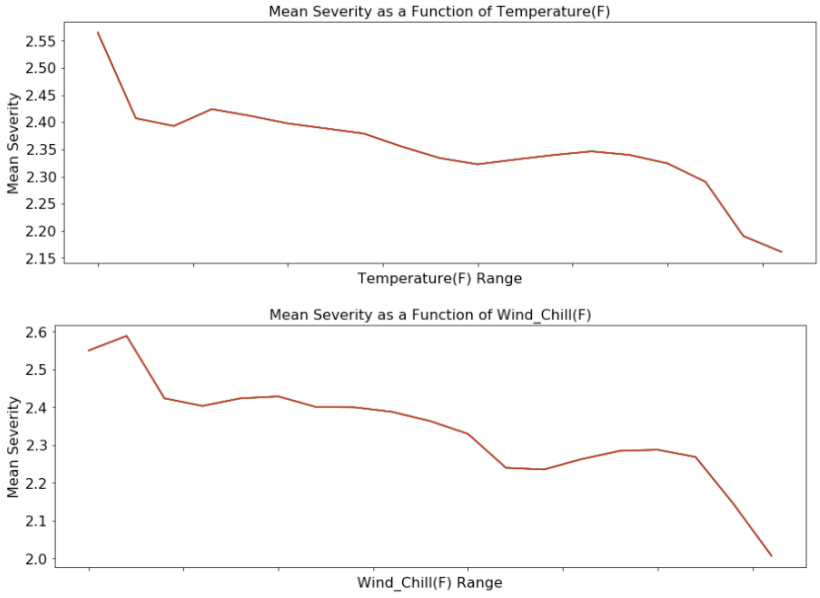


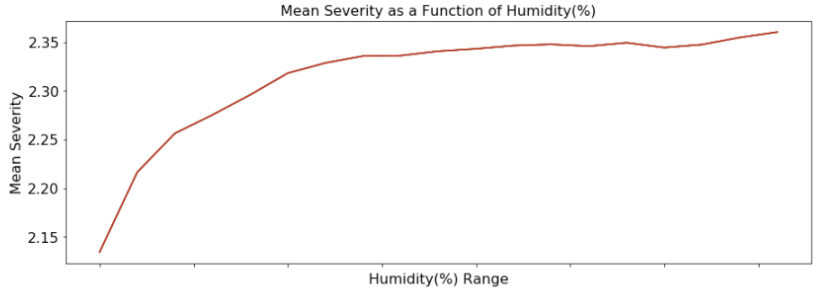
These graphs suggest that the waether conditions have a significant impact on severity of an accident.

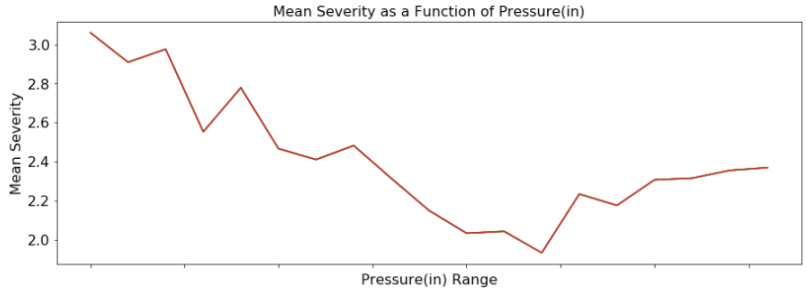
The proportion of level 3 and 4 accidents increases as weather changes from fog (23%) to light rain (34%) to rain (37%) to heavy rain (39%) to snow (39%).

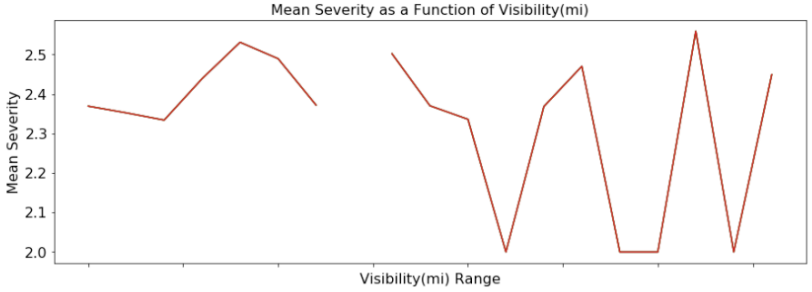
5.3 Other Weather factors

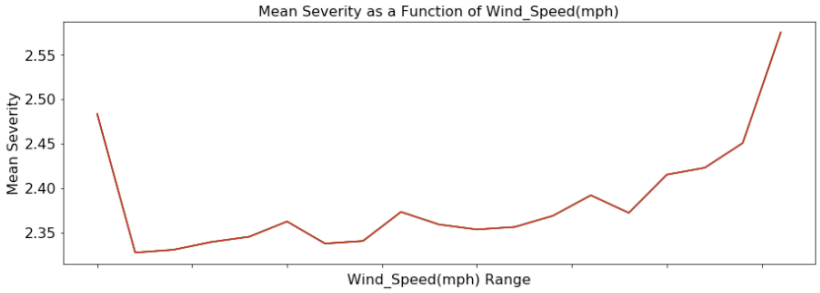
Mean severity increases as conditions for freezing precipitation increase, and as we saw in the previous section rain and snow have higher proportion of level 3 and 4 severity. These conditions include decreasing temperature, wind chill, and air pressure [1] as well as increasing humidity. Severity also increases as a function of wind speed. The data for visibility and precipitation is not complete. (https://sciencing.com/rain-pressure-low-8738476.html) The details are shown below:

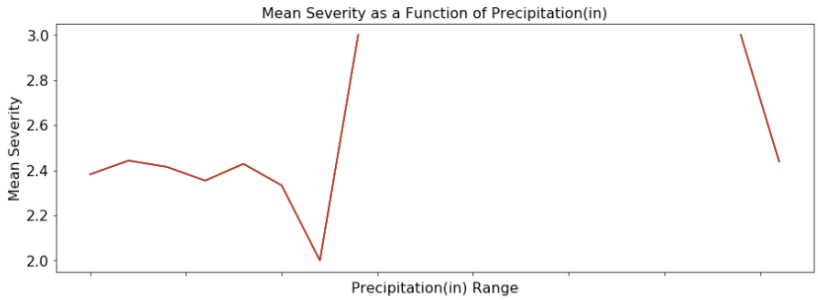



Mean Severity as a Function of Humidity(%)
Mean Severity
2.35
2.30
2.25
2.20
2.15
Humidity(%) Range

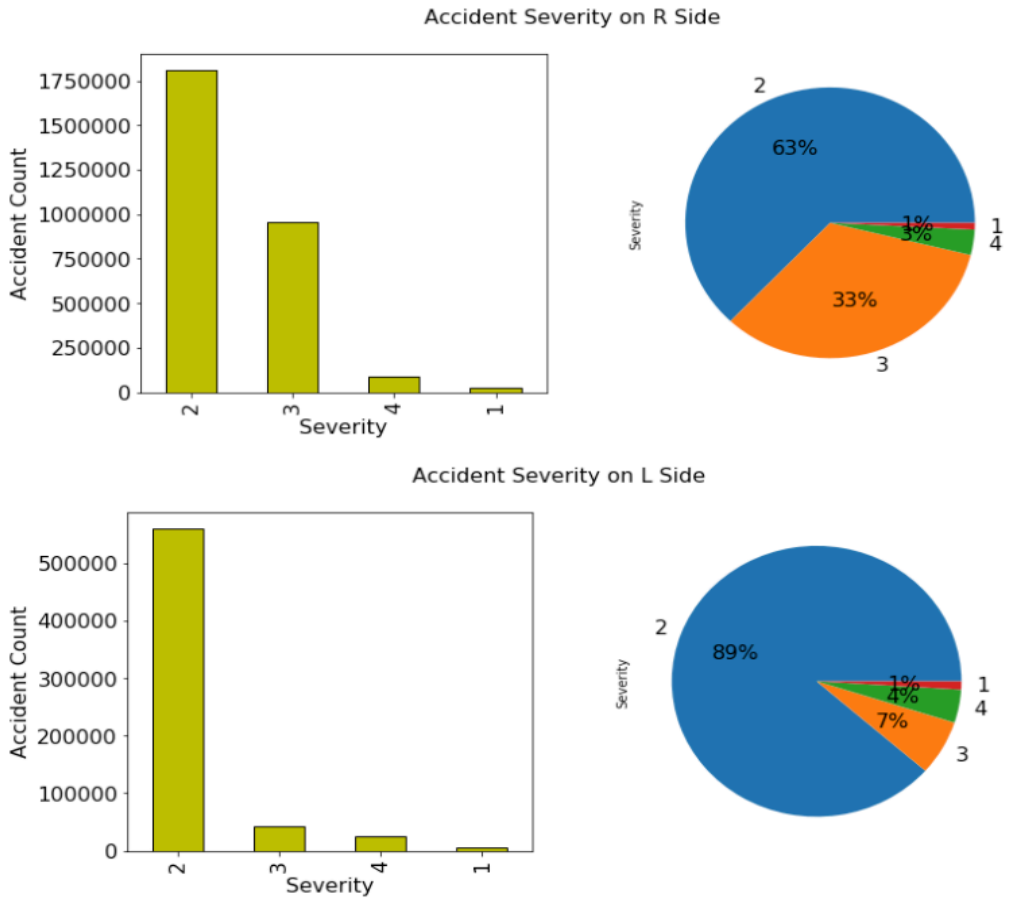

Mean Severity as a Function of Pressure(in)
Mean Severity
3.0
2.8
2.6
2.4
2.2
2.0
Pressure(in) Range


Mean Severity as a Function of Visibility(mi)
Mean Severity
2.5
2.4
2.3
2.2
2.1
2.0
Visibility(mi) Range


Mean Severity as a Function of Wind_Speed(mph)
Mean Severity
2.55
2.50
2.45
2.40
2.35
Wind_Speed(mph) Range

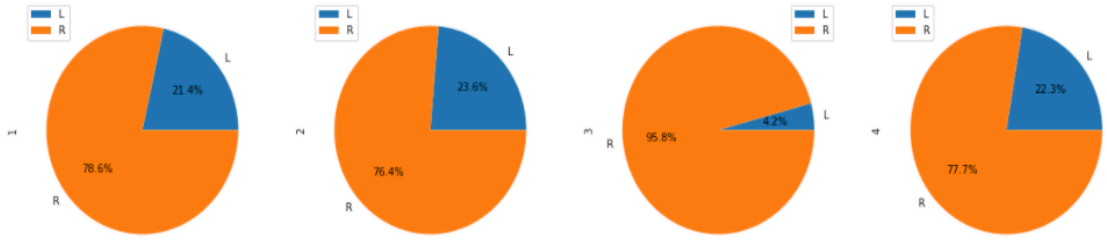


6 Does the Side have an effect on the Severity?



These two graphs suggest that the Left Side accidents are more likely to be less severe.

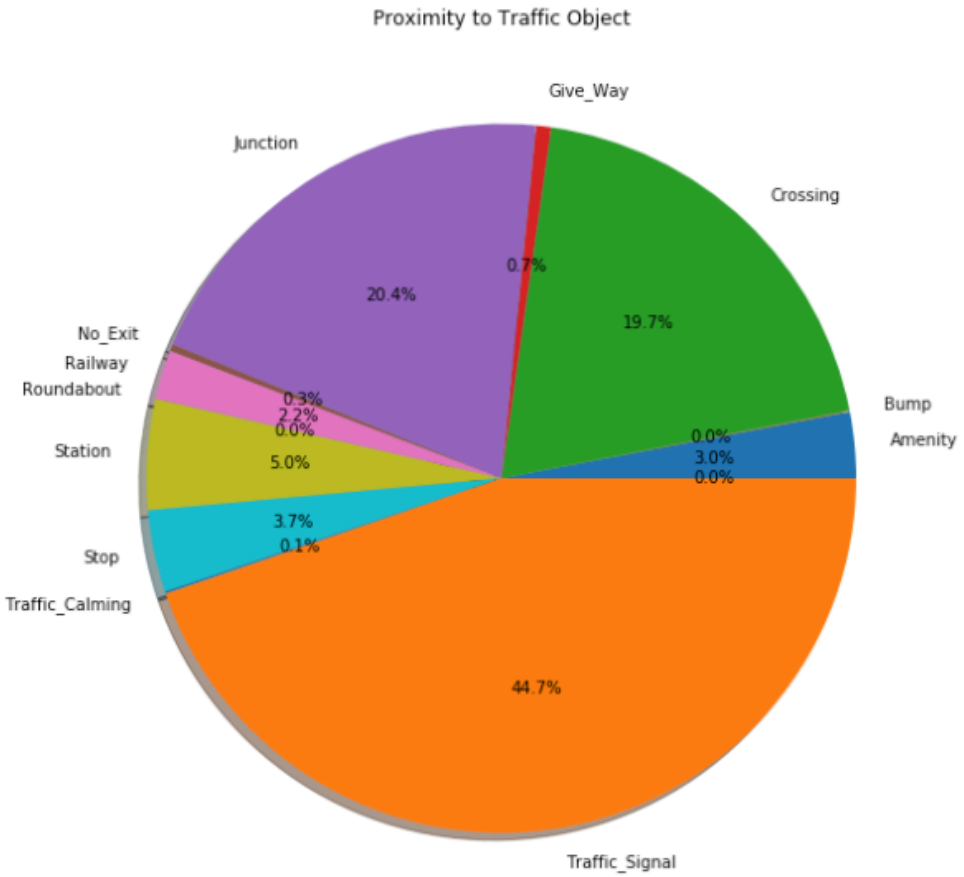
Also, this conclusion can be further proven by classifying using the level of severities. The results are shown below:



7 Relationship between Infrastructure and Severity

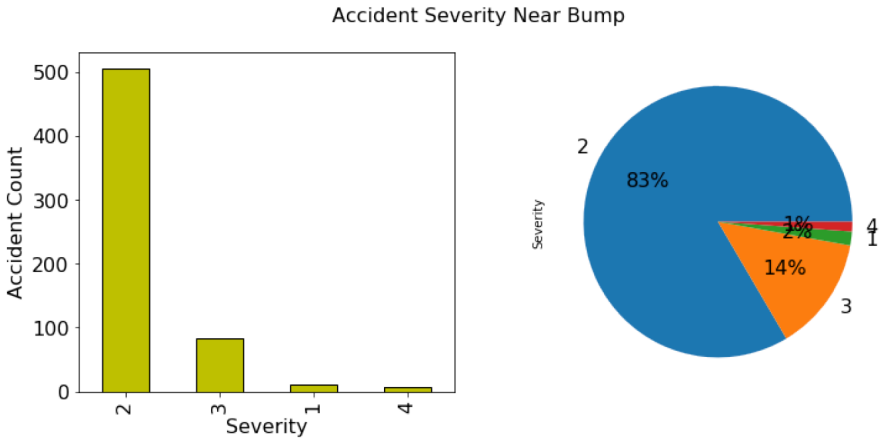
In this dataset, there are some boolean values that determines whether the accident happened near a traffic signal, stop sign, etc. We will See how proximity to an object affects the accident number and severity.

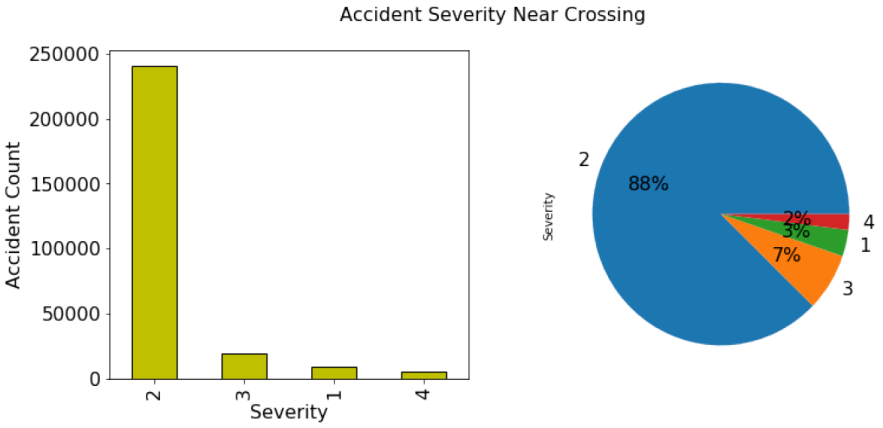
Firstly we will identify columns with boolean datatypes and express with a pie plot.

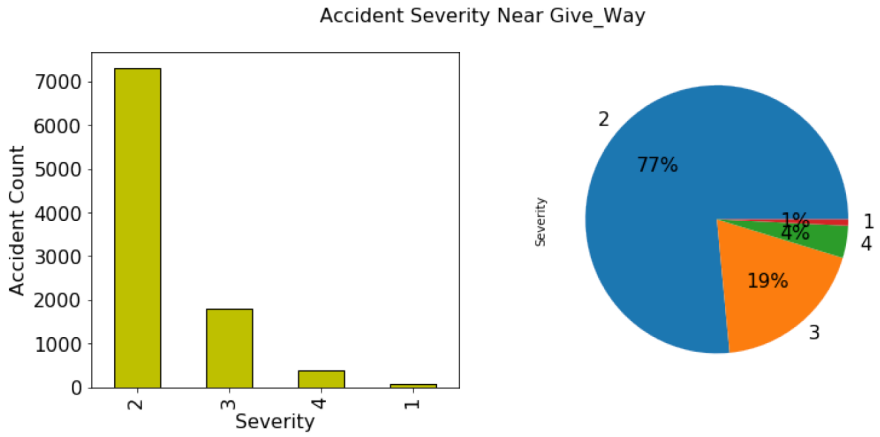


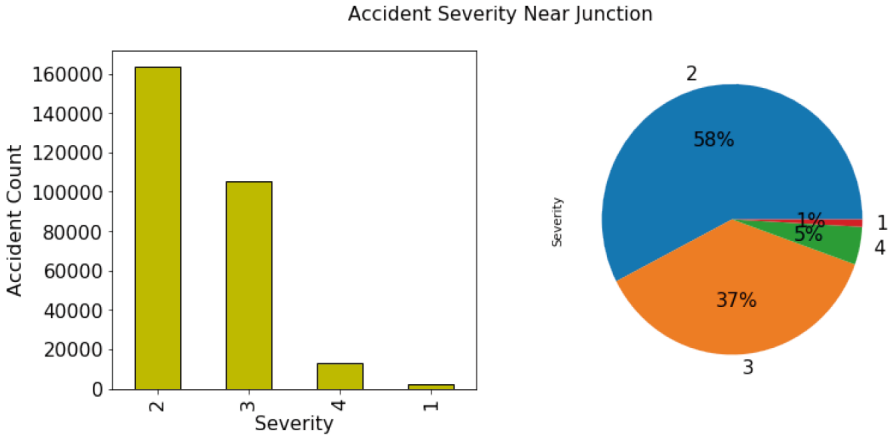
One problem about this pie chart is, some of the accidents may have more than one boolean values.

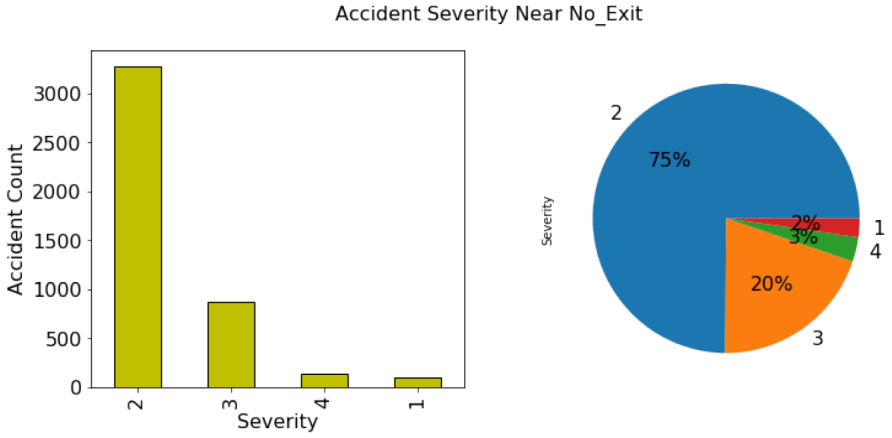
There are 284711 non one hot metadata rows, which are 8.1% of the data. Now we will check on the accident severity near each type of infrastructure. From the results, we can find that junctions, give way, and no exit have the highest proportion of level 3 and level 4 severity accidents. The details are shown below:

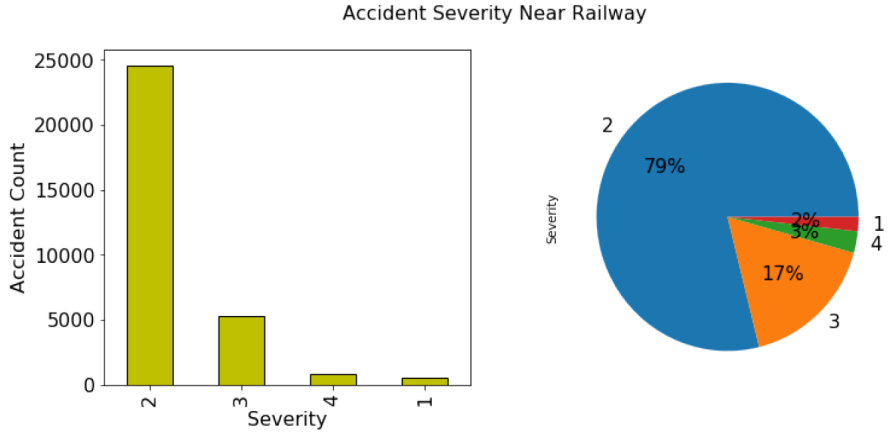


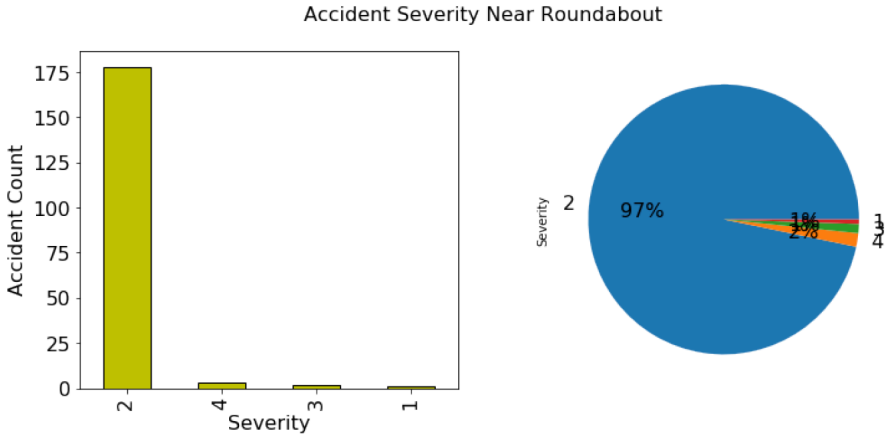


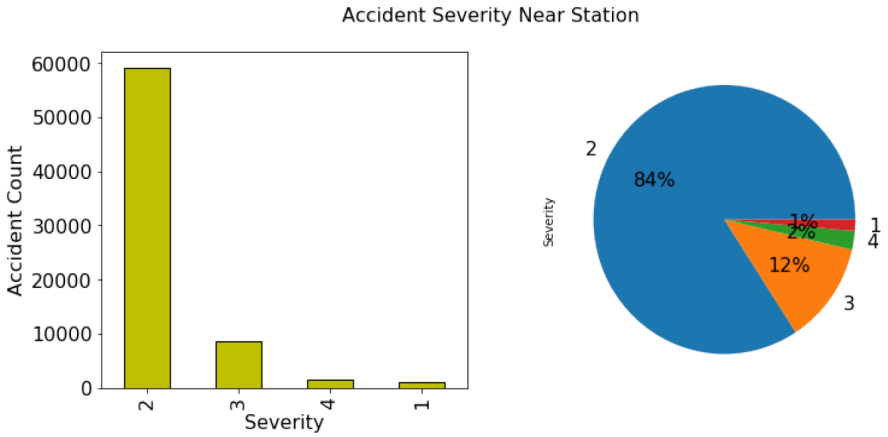

Accident Severity Near Give_Way
Accident Count
7000
6000
5000
4000
3000
2000
1000
0
2
3
4
1
Severity
Severity
2
77%
1%
1
4%
4
19%
3

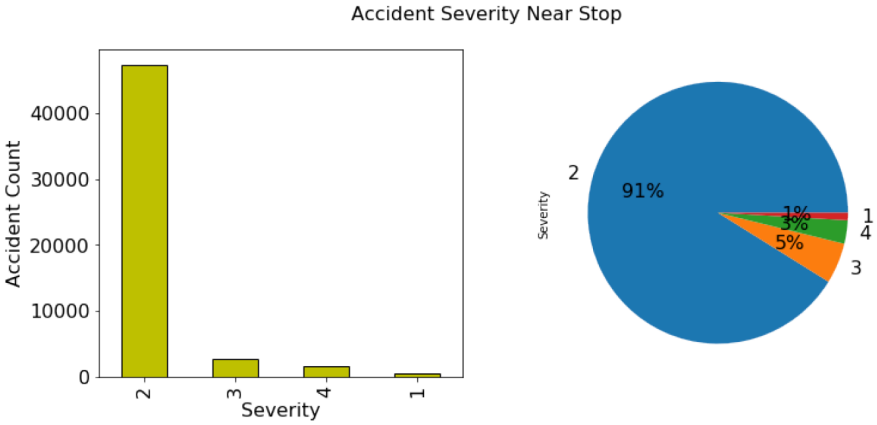

Accident Severity Near Junction
Accident Count
160000
140000
120000
100000
80000
60000
40000
20000
0
2
3
4
1
Severity
Severity
2
58%
1%
1
5%
4
37%
3

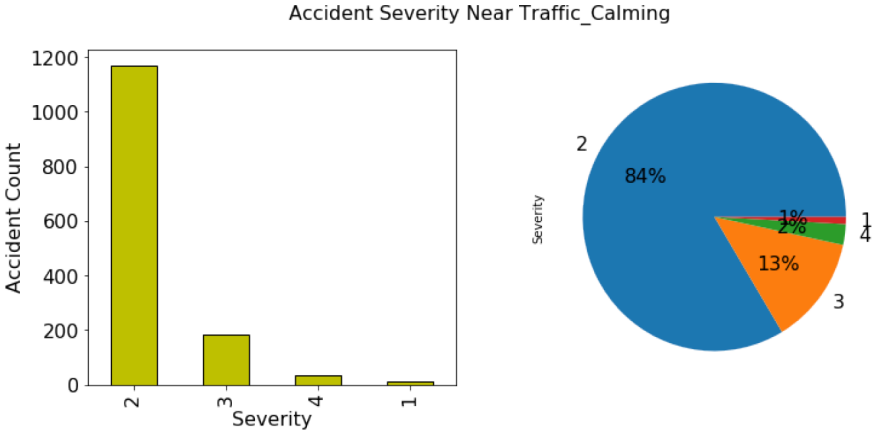

Accident Severity Near No_Exit
Accident Count
3000
2500
2000
1500
1000
500
0
2
3
4
1
Severity
Severity
2
75%
2%
1
3%
4
20%
3

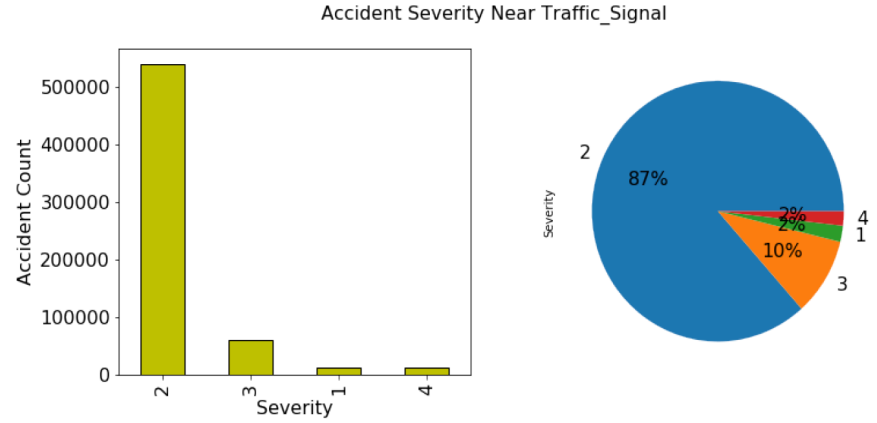

Accident Severity Near Railway
Accident Count
25000
20000
15000
10000
5000
0
2
3
4
1
Severity
Severity
2
79%
17%
3
2%
1
3%
4


Accident Severity Near Roundabout
Accident Count
175
150
125
100
75
50
25
0
2
4
3
1
Severity
Severity
2
97%
1%
2%


Accident Severity Near Station
Accident Count
60000
50000
40000
30000
20000
10000
0
2
3
4
1
Severity
Severity
2
84%
12%
3
1%
1
2%
4


Accident Severity Near Stop
Accident Count
40000
30000
20000
10000
0
2
3
4
1
Severity
Severity
2
91%
1%
3%
5%
1
4
3


Accident Severity Near Traffic_Calming
Accident Count
1200
1000
800
600
400
200
0
2
3
4
1
Severity
Severity
2
84%
1%
2%
13%
1
4
3


Accident Severity Near Traffic_Signal
Accident Count
500000
400000
300000
200000
100000
0
2
3
1
4
Severity
Severity
2
87%
2%
2%
10%
4
1
3