**Predicting the Severity of Car Accidents**

Reducing the fatal car accident is one of the most critical public safety issues. This project aims to predict the severity of a car accident from the other given variables such as the time of the accident, the location of the accident, the road types, etc.

The primary audience of this problem is the policymakers such as Department of Public Safeties in the US. What policies can be made to decrease the number of cars on the roads when there is a high potential for accidents. What additional warning systems might be generated to have fewer drivers when there is a severe weather condition that correlates with severe car accidents.

**Data Wrangling**

The dataset was obtained from Kaggle <https://www.kaggle.com/sobhanmoosavi/us-accidents> which contains US Accident data from 2016 - 2019. The data was read from a .csv file and includes 49 states (excluding Hawaii) accident data, and there are 1048575 rows and 49 columns. Each row represents a unique incident, and the rows represent the variables associated with each incident.

One of the major issues with this dataframe is two columns: 'End\_Lat', 'End\_Lng' have no information in it, so those two columns were dropped. Then, I check the total number of missing values per column:

Precipitation(in) 927055

Wind\_Chill(F) 875145

Number 651772

Wind\_Speed(mph) 207628

Visibility(mi) 28973

Weather\_Condition 28951

Humidity(%) 25111

Temperature(F) 24059

Pressure(in) 21069

Wind\_Direction 17639

Weather\_Timestamp 17631

Airport\_Code 10290

Timezone 581

Zipcode 140

Nautical\_Twilight 29

Astronomical\_Twilight 29

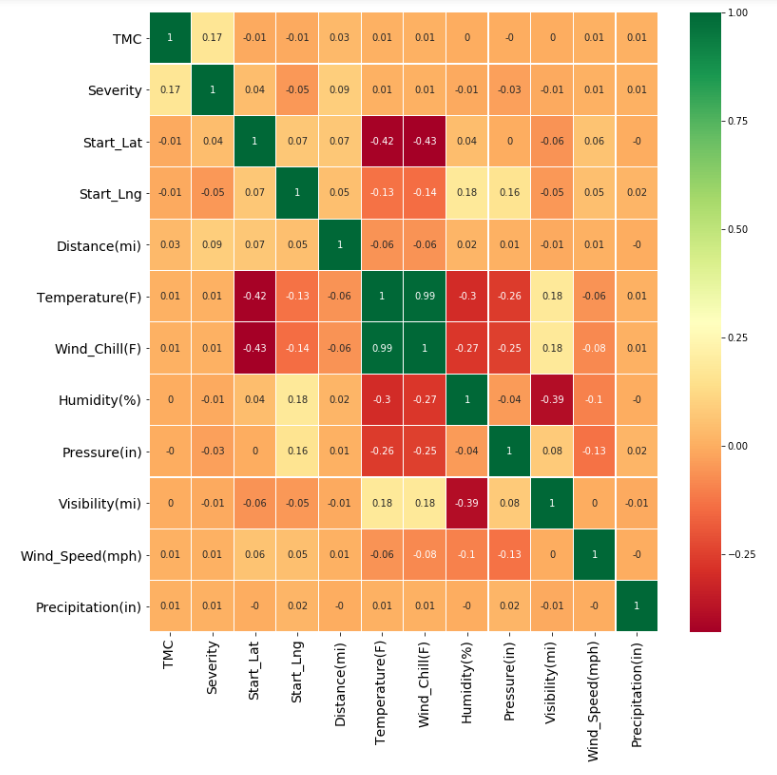
Civil\_Twilight 29

Sunrise\_Sunset 29

City 27

Description 1

Before filling the missing values, I checked the Heat Map from Seaborn library to see whether there is a strong correlation between two given columns:



This Heat Map suggests that Temperature(F) and Wind\_chill(F) columns (r=0.99) strongly correlated.

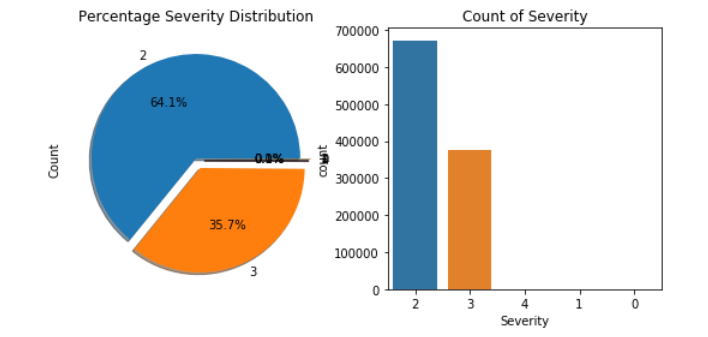
In order to fill the missing values, the following methods were used:

1. Humidity, Precipitation, Pressure, Visibility and Wind Speed values are filled with the mean values of the corresponding column.
2. Temperature(F) column was filled with the average temperature of the corresponding state by using a groupby method.
3. There is a strong correlation between Temperature(F) and Wind\_chill(F) columns (r=0.99) The wind Wind\_chill(F) column was filled with the linear model
4. Weather Condition, Description, Wind Direction and City columns were filled by”NaN’.
5. Zip code and Airport Codes are filled by the most common codes for the corresponding state.
6. The Start Time and End Time are datetime objects, that I extracted the years, days, and hours only from the Start Time column.
7. Nautical\_Twilight, Sunrise\_Sunset, Civil\_Twilight, Astronomical\_Twilight missing values are filled with Day or night depending on the Start Time
8. Weather Timestamp is filled with the corresponding Start Time
9. Number column is filled with firstly average number in the corresponding zip code, then the city, then the state.

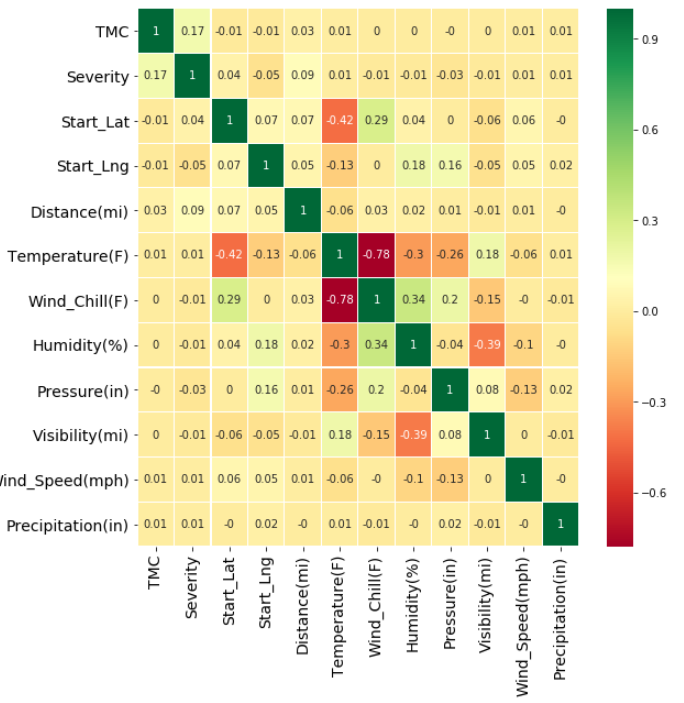
After these steps, the dataframe has no missing values. Now, it is time to select the feature values. In order to select the feature values, I will do some exploratory data analysis to see which factors might affect the target variable: severity the most.

**Exploratory Analysis**

Before observing the feature variables, let’s check how the target variable is distributed. I used subplots with count plot and pie plot from Seaborn library to observe the counts and percentages of each severity type:



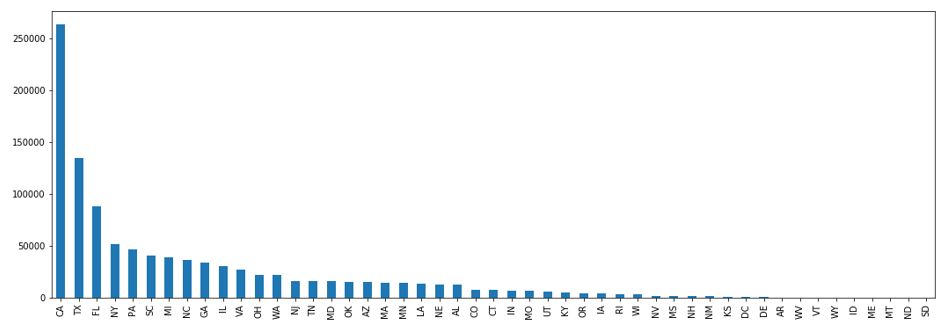
**1. What is the correlation between numerical variables and severity**[**¶**](http://localhost:8888/notebooks/Desktop/Jupyter_Notebooks/Capstone_Milestone.ipynb#1.-What-is-the-correlation-between-numerical-variables-and-severity)



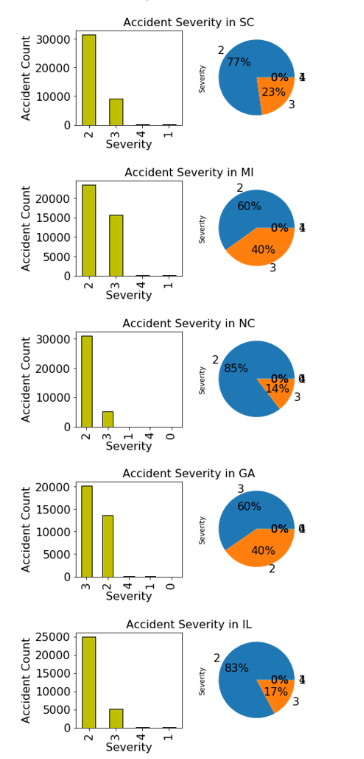
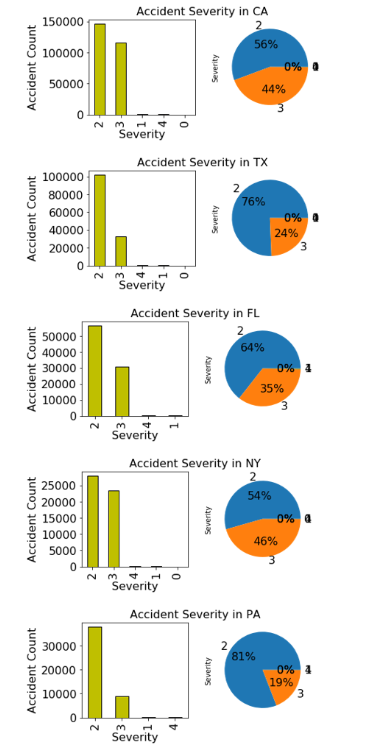
We only used “float64” and “int64” data types that are not coming from dummy variables. According to this heat map, the highest correlation with severity is TMC which is 0.17/. This heatmap does not suggest any strong correlation between a numerical variable and severity of an accident.

### **2. What is the relationship between location and the severity of accidents?**

The second question is how the number and severity of accidents are associated with the location. The first graph we used to see the number of accidents for each state:

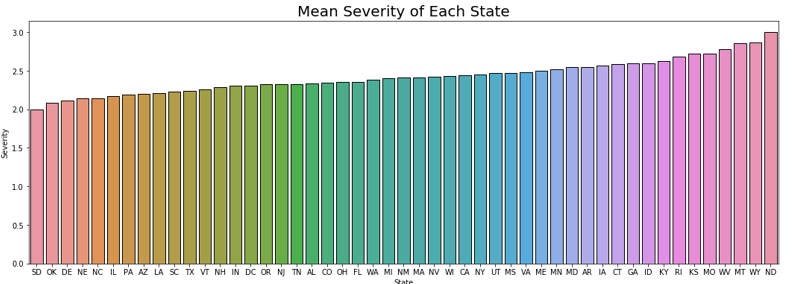


We can observe from this graph that the number of accidents significantly differ from state to state. However, the other part of the question is whether the severity of an accident is associated with the state or not. In order to answer this question, we will check the distribution of the severity of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

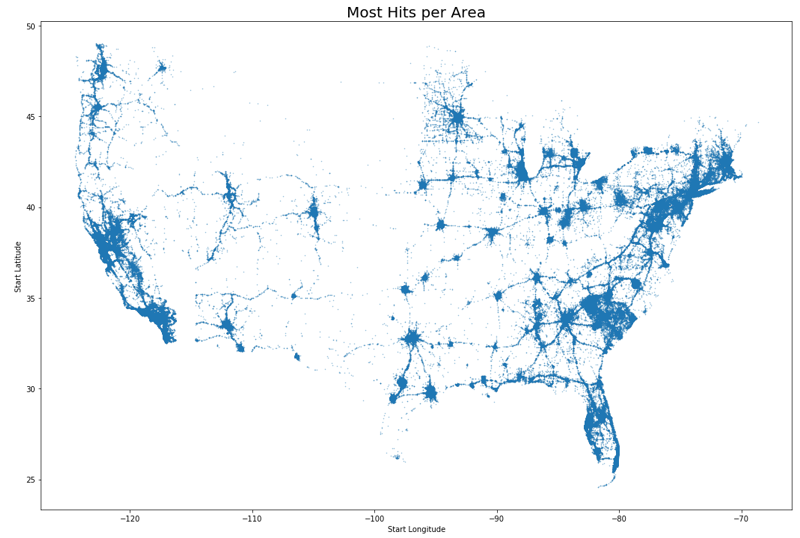


These graphs clearly state that the severity distribution differs from state to state. So state is a variable that may affect the severity of an accident.

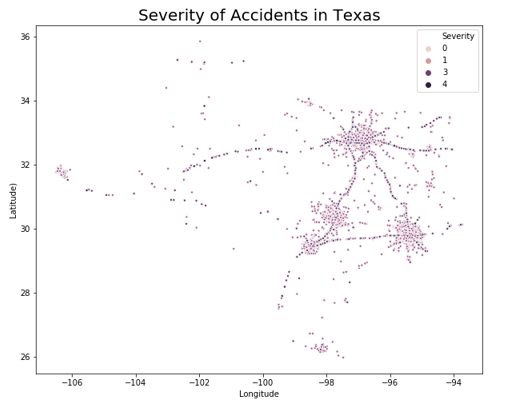
Severity is a discrete numerical variable in ordinal level, so mean severity my have meaning in the context:

****

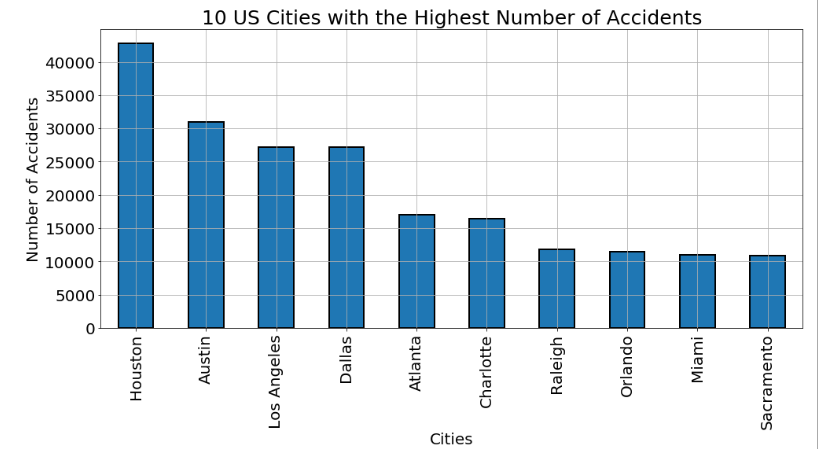
One other question is how the number of accidents is associated with the geographic location. We can show on a map:



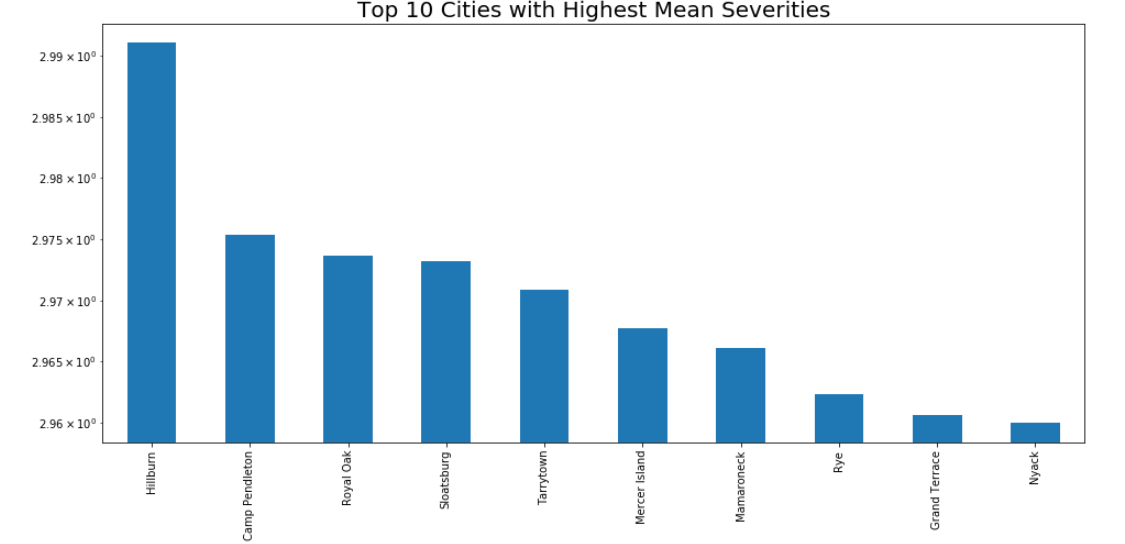
The graph above shows all the accidents in the USA, and we can easily see that the accidents are clustered in metropolitan areas and major highways. However, it does not indicate the severity of accidents. In order to show the severity of an accident, we choose the State of Texas to see how more severe accidents are clustered:



The most severe accidents which are darker are clustered in some certain roads, unfortunately we do not have any data for the type of roads.



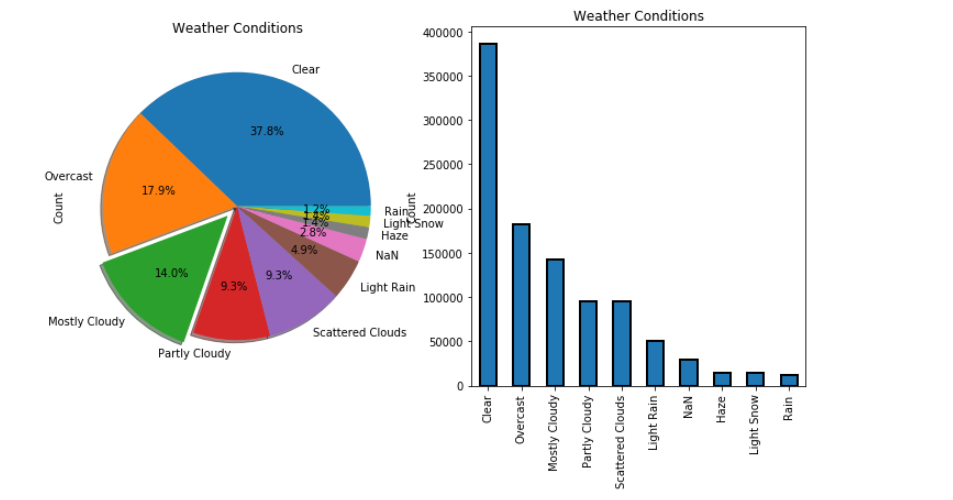
We will check which cities have the highest mean of severities. However, there are some cities which have too few accidents that their mean severity values might be outliers. I set an indicator as “big city” if the number of accidents is more than 100.



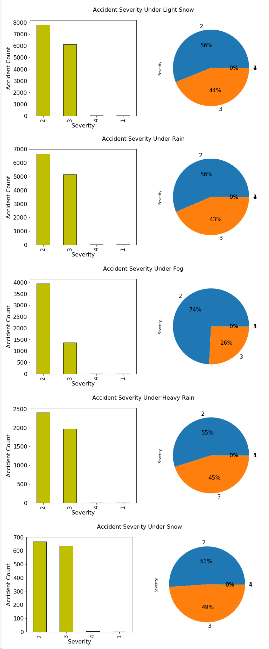
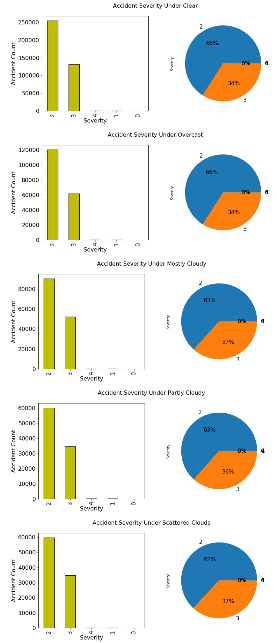
I used a logarithmic scale since the values are very close that it might be hard to see the difference between the values.

**3. Severity of Accidents with different weather conditions**

The other question is how the amount of accidents are distributed in different weather conditions. also, we would like to see how some common weather conditions affect the severity of accidents



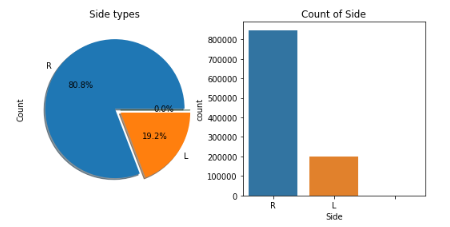
Now, let’s see how the severity is distributed in these 10 weather conditions:



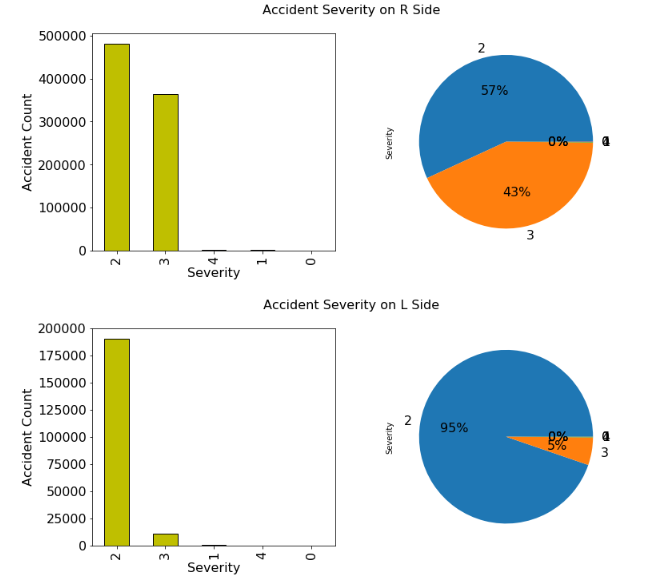
These graphs indicate that the severity is distributed differently in these categories.

**4. Does Side Type Have An Effect on Accidents & Severity?**

The distribution of sides are like this:



And the distribution of severity for each side is:

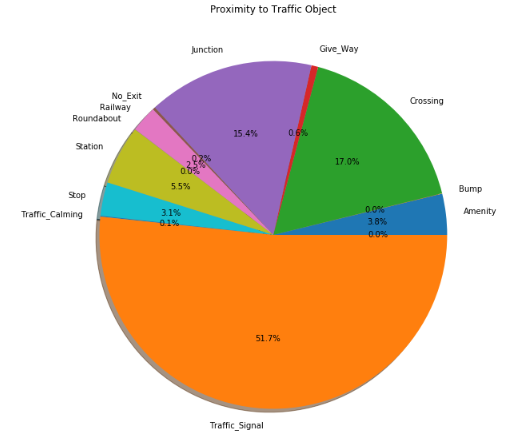


These graphs suggest that the number of accidents on the left sides are significantly lower in amount and tend to be less severe.

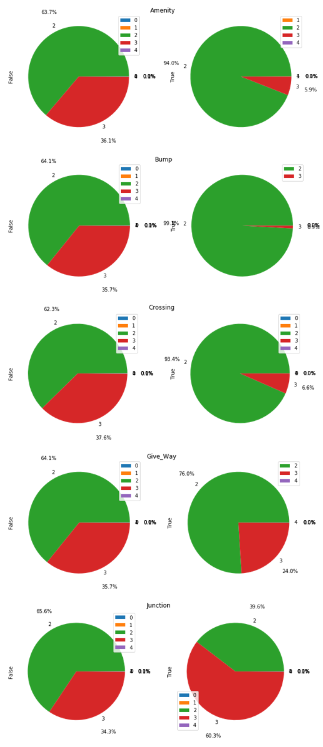
**5. Proximity to Traffic Objects**

In this dataset, there are some boolean values that determine 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.

Many of the accidents happened close to one of the traffic objects

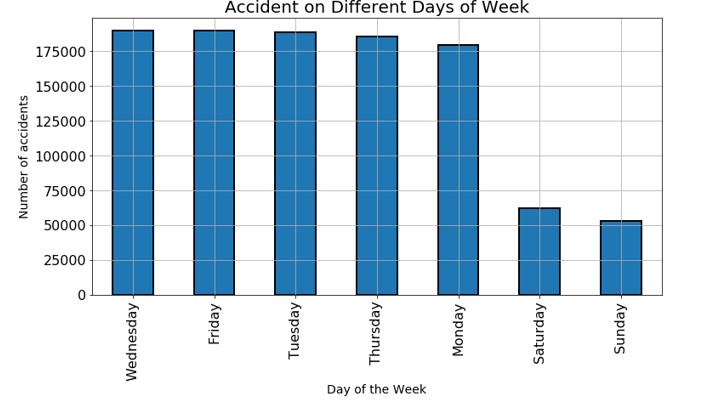


Here are the some examples of severity distributions in some of the amenities:



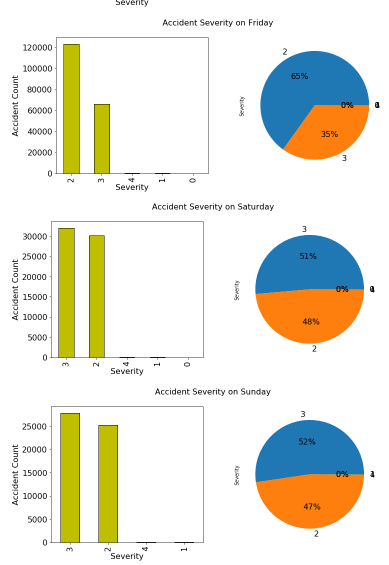
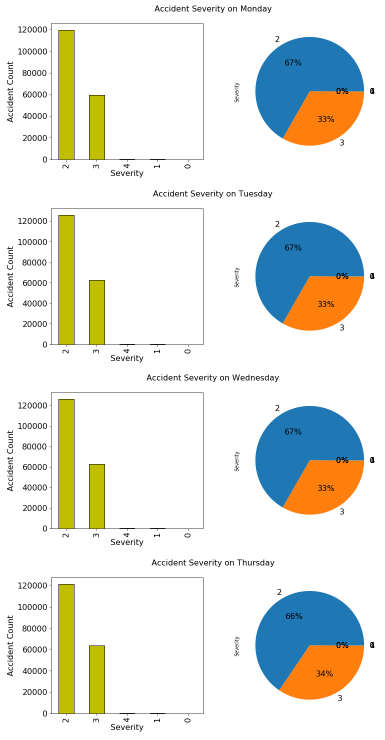
This graphs show that proximity to traffic objects may also affect the accident severity.

**6. How are the Number of Accidents & Severities Associated with Days of Week?**

****

The graph above suggests that the number of accidents on weekdays are pretty close, but there are significantly less accidents on the weekends.

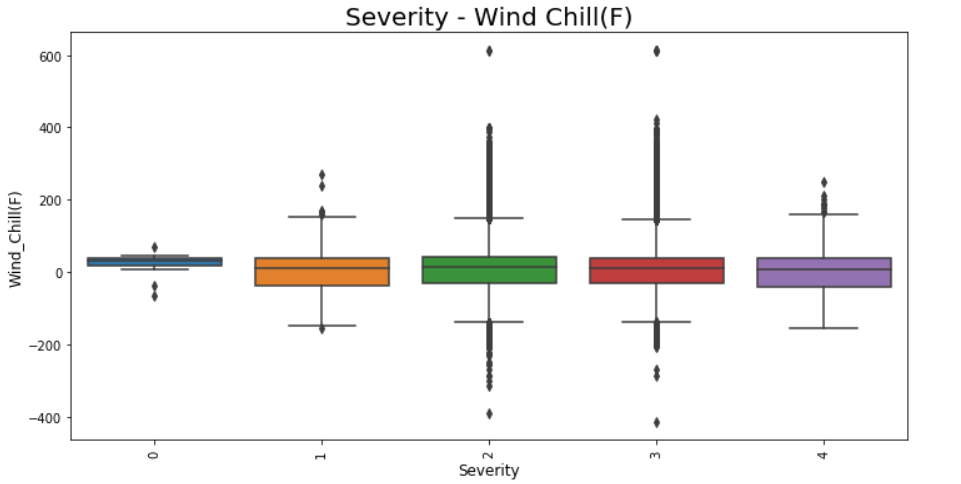
Now, we will check the severity distribution of each day:

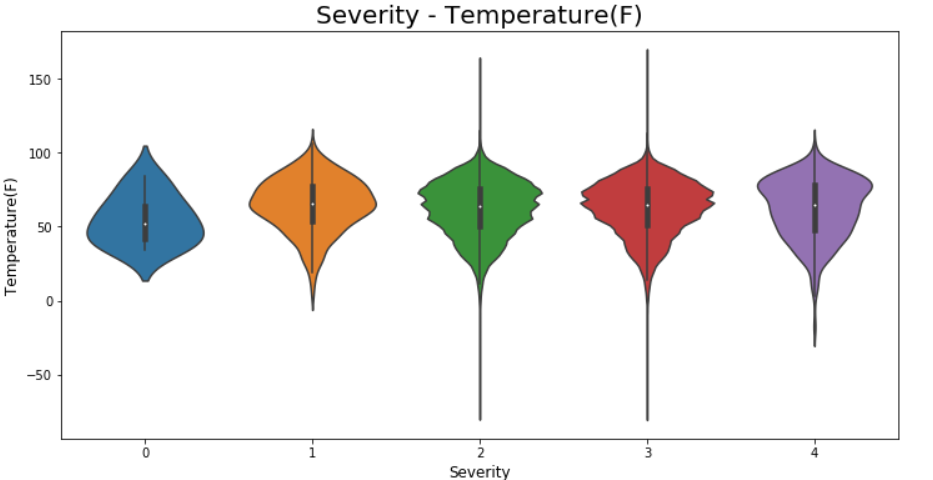


When we split the days as weekdays and weekends into two groups, there is not a big difference observed within the groups. However, both the number of accidents and the severity distribution alters significantly between these two groups.

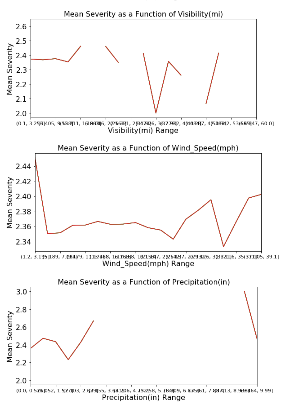
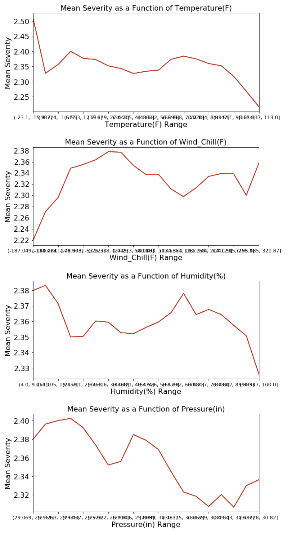
**7. Visualization of Severity vs. Continuous Variables**

In this dataset, there are two continuous variables that we will observe their distribution in each severity categories:

****

****

The median temperatures and wind chills look quite similar. The next graphs are the line plots of mean severity over the continuous variables:



As we can observe from these line plots, the mean severity changes with these continuous variables, but it is hard to determine without further analysis, how each variable changes the severity.

**8. Perform a t-test**

We can perform a t-test to see if the severity changes significantly between two groups. However, the dataset is very large with more than 1 million values, which makes the standard error very small. We expect to get very large t-values. Just to see the significance, I will perform a t-test to see the distribution of severity with a boolean: Junction.

Ttest\_indResult(statistic=123.45381748295537, pvalue=0.0)