IBM Datascience Certificate Program - Capstone Project

Khalid Sayyed

09/13/2020

Predicting Accident Severity

Business Problem:

In the United State of America, more than 38,000 people die every year in crashes on roadways.
The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants.

An additional 4.4 million are injured seriously enough to require medical attention.
Road crashes are the leading cause of death in the U.S. for people aged 1-54.

The economic and societal impact of road crashes costs U.S. citizens 871 billion dollars.
Road crashes cost the U.S. more than $380 million in direct medical costs.

The U.S. suffers the most road crash deaths of any high-income country, about 50% higher than similar countries in Western Europe, Canada, Australia and Japan.

In this study, we will be analyzing US accidents severity data in order to predict how dangerous driving conditions are.

The target audience for this study will be general public/citizens of United States of America who are driving on roads.

The study will analyze set of conditions that lead to high severity accidents.

Based on a given set of conditions on a particular day at a particular place, the algorithm will warn drivers of potential dangerous conditions.

Data:

We will be using the US Accidents Severity data available on Kaggle.com.
The data was collected between February 2016 to March 2019, 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. The dataset contains 2,243,939(2.24 million) rows and 49 features.
Following are each of the 49 features of this dataset and their description/details.

#	Attribute	Description	Nullab
1:	ID	This is a unique identifier of the accident record.	No
2:	Source	Indicates source of the accident report (i.e. the API which reported the accident.).	No
3:	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.	Yes
4:	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).	No
5:	Start_Time	Shows start time of the accident in local time zone.	No
6:	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.	No
7:	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
8:	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
9:	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
10:	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
11:	Distance(mi)	The length of the road extent affected by the accident.	No
12:	Description	Shows natural language description of the accident.	No
13:	Number	Shows the street number in address field.	Yes
14:	Street	Shows the street name in address field.	Yes
15:	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
16:	City	Shows the city in address field.	Yes
17:	County	Shows the county in address field.	Yes
18:	State	Shows the state in address field.	Yes
19:	Zipcode	Shows the zipcode in address field.	Yes
20:	Country	Shows the country in address field.	Yes
21:	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
22:	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
23:	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
24:	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
25:	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
26:	Humidity(%)	Shows the humidity (in percentage).	Yes

27:	Pressure(in)	Shows the air pressure (in inches).	Yes
28:	Visibility(mi)	Shows visibility (in miles).	Yes
29:	Wind_Direction	Shows wind direction.	Yes
30:	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
31:	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
32:	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
33:	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
34:	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
35:	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
36:	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	No
37:	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
38:	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No
39:	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
40:	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
41:	Station	A POI annotation which indicates presence of station in a nearby location.	No
42:	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
43:	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	No
44:	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	No
45:	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	No
46:	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
47:	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Yes
48:	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Yes
49:	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Yes

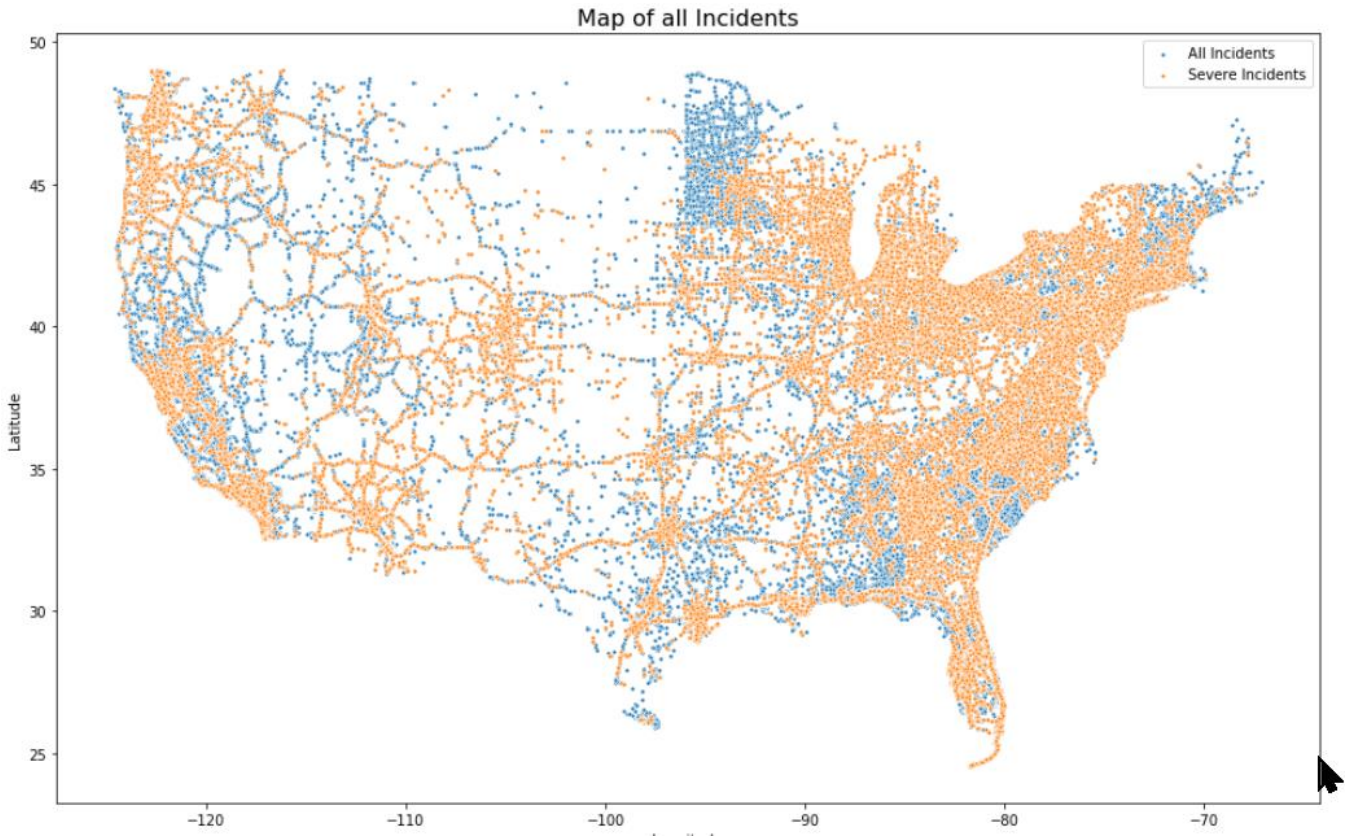
Methodology:

Data Exploration and Manipulation

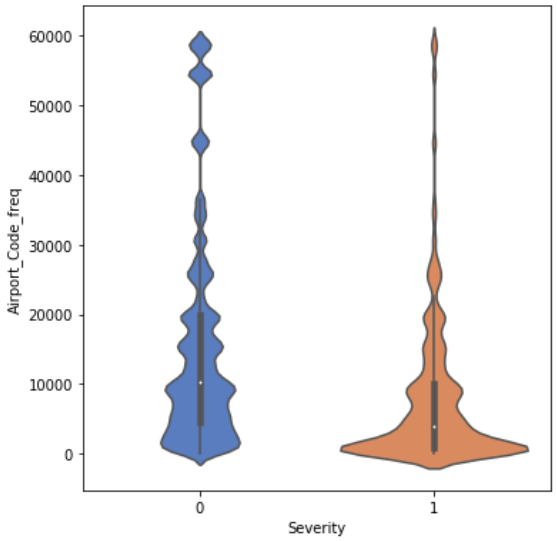
From the map of accidents across latitudes and longitudes, it can be seen that majority of the accidents occur in east and west coast.

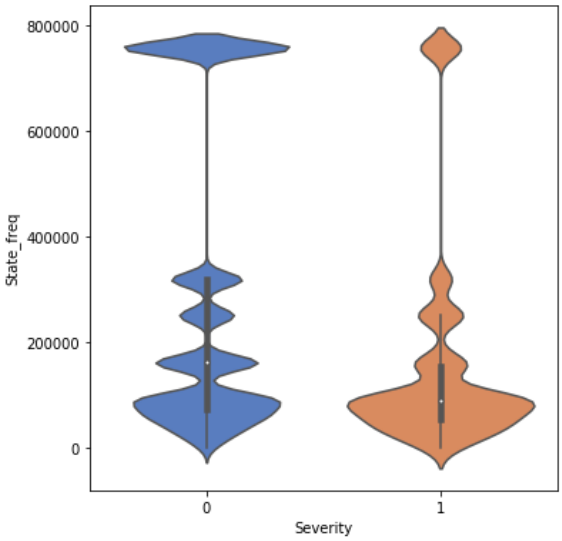
This makes sense because of dense population in these regions of the US.

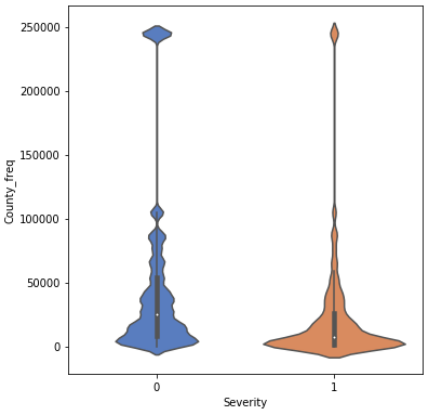
Map (Especially at central USA) shows higher accidents along some lines, indicating interstate highways and cluster of higher accidents around major cities.

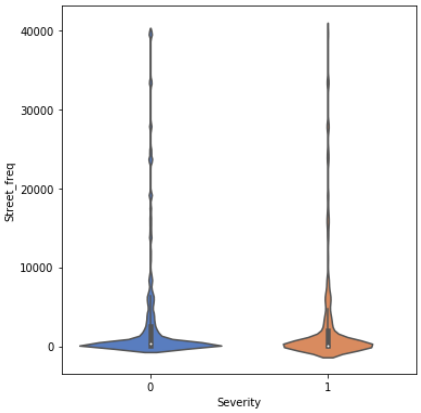


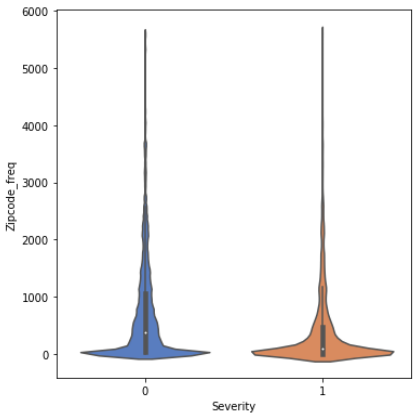
From violin plots of frequency of States, Counties, Streets, Zip Codes and Airport Codes, it can be seen that accidents are higher at some of these places than others.



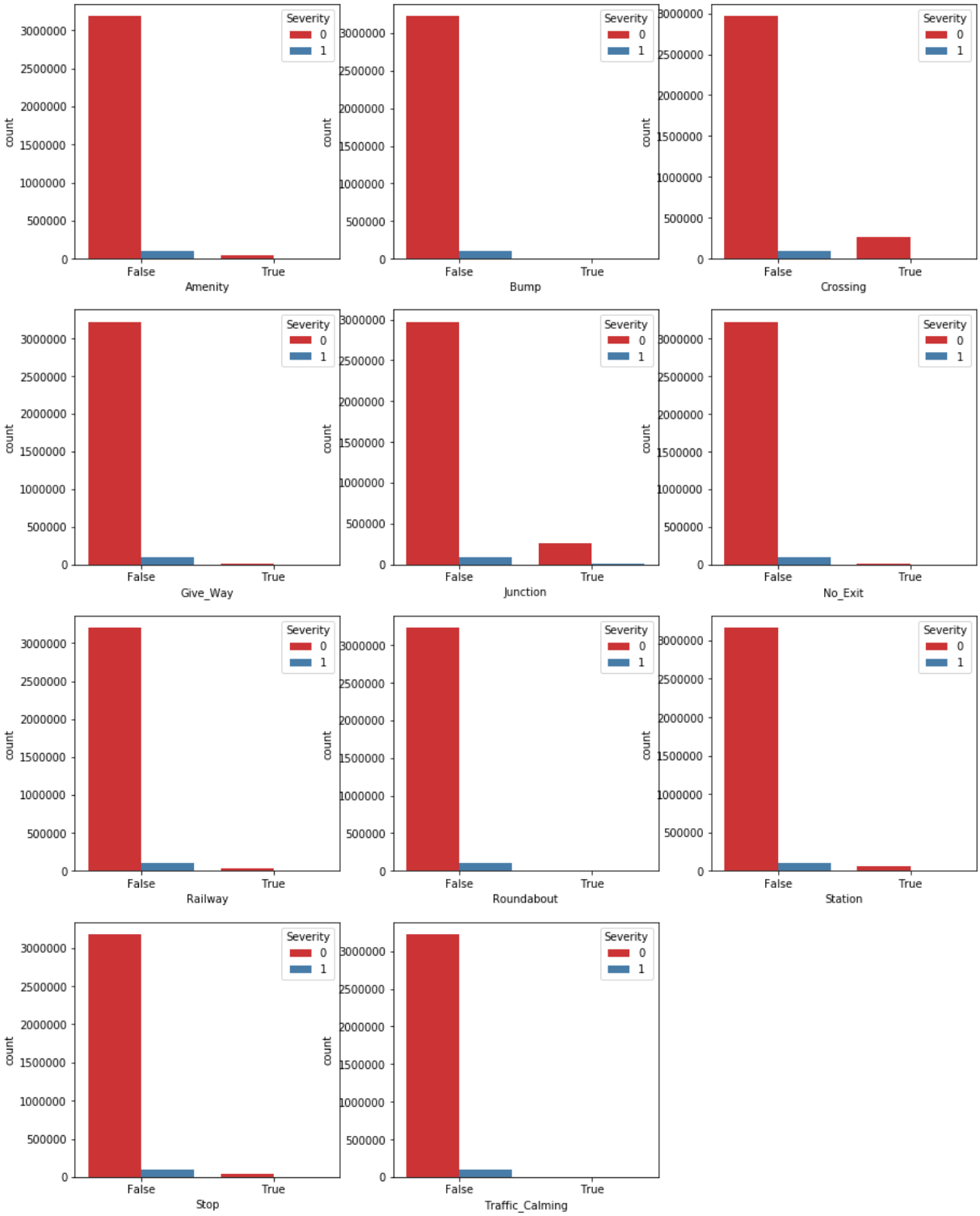




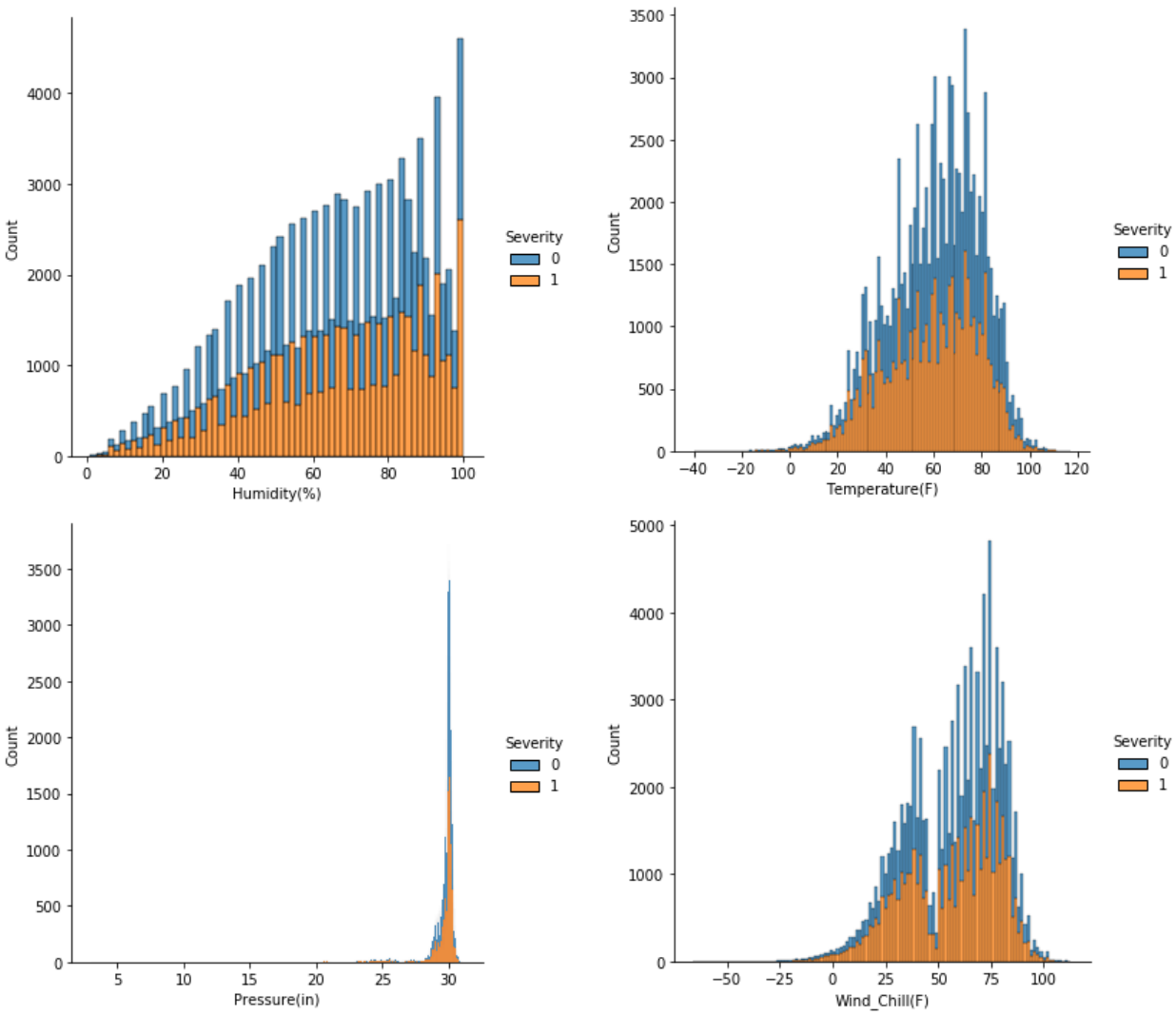




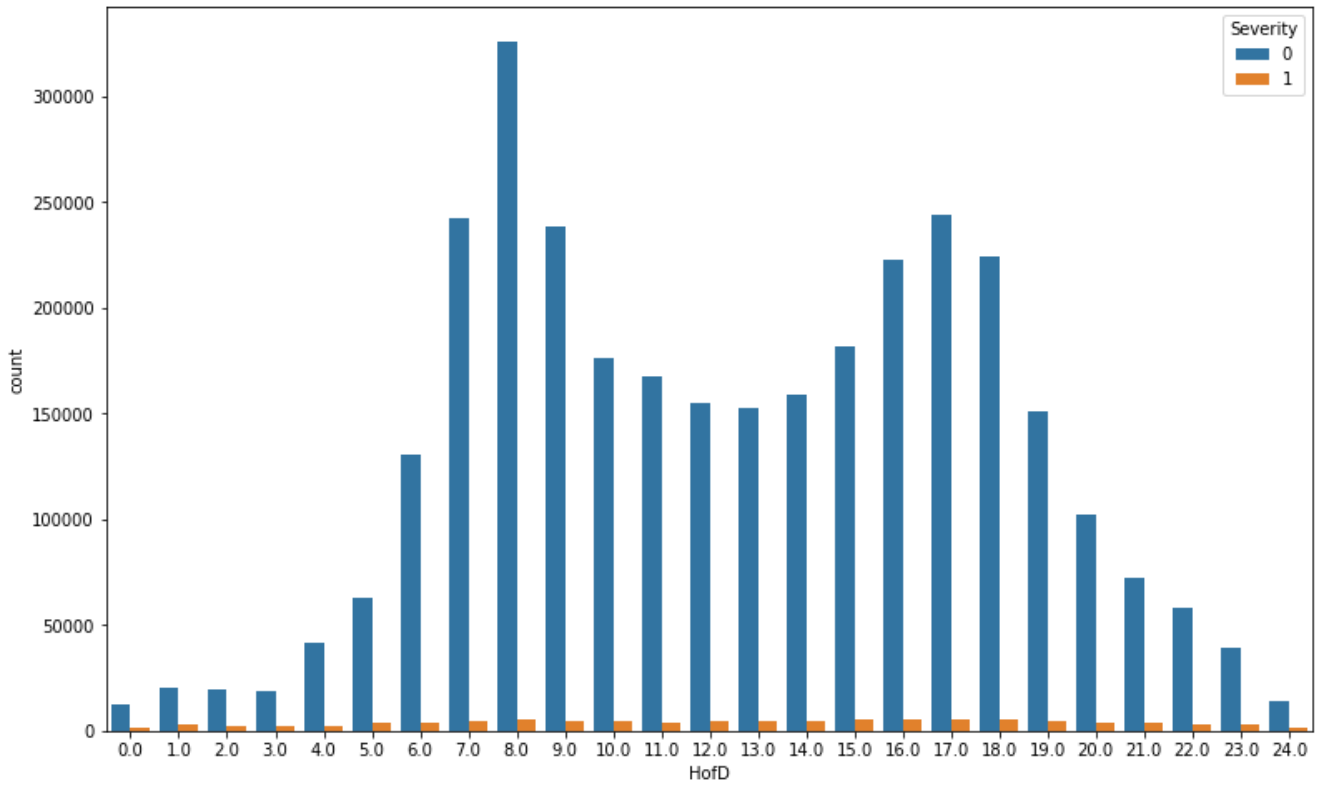
From bar charts of points of interest data, it can been seen the accidents are higher and sever at certain points of interest than others.



Weather conditions distribution plots show that accidents are higher and sever at some weather conditions such as during high humidity (probably during rain).



Hour of the day vs accidents plot shows higher accidents during morning and evening busy hours.



Results:

Machine Learning

1. From the map of accidents across latitudes and longitudes, it can be seen that majority of the accidents occur in east and west coast. This makes sense because of dense population in these regions of the US.

2. From violin plots of frequency of States, Counties, Streets, Zip Codes and Airport Codes, it can be seen that accidents are higher at some of these places than others.

3. From bar charts of points of interest data, it can been seen the accidents are higher and sever at certain points of interest than others.

4. Weather conditions distribution plots show that accidents are higher and sever at some weather conditions such as during high humidity (probably during rain).

5. Hour of the day vs accidents plot shows higher accidents during morning and evening busy hours.

6. Finally, machine learning decision trees shows an accuracy of 92% and 85% on train and test data to predict severe accidents.

Hence we can conclude that the dataset can be used in predicting accidents severity and warning drivers of potential driving hazards.