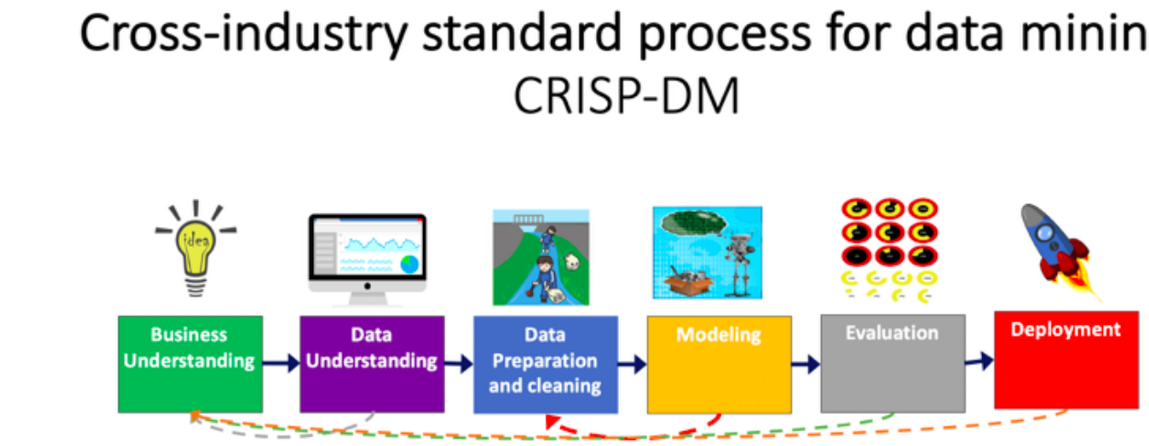
SEVERITY IN CAR ACCIDENT

1. Description of the problem or Business Understanding

This case of study will be concentrate in collecting data about accident in USA and applicate every knowledge that we acquire in this course for develop a model that predict accident, in this case we will walk through every step of CRISP-DM model for complete the work.



When we think about the severity of an accident, specifically in cars accidents, there are many circumstances that come in our minds, like: meteorology conditions, time of the day, state of the highway, velocity, type of cars, age of the conductor, level of drugs like alcohol, etc. For this study we will work with information dataset from https://smoosavi.org/datasets/us_accidents that contain 3.5 millions of records about accidents across 49 States, this information

provide 49 attributes that will help us to prepare our model, this attributes contain information about weather conditions, geographic location and particularities about the locations, like junctions, railway, etc.

For the scope of this case of study we will use the information for develop an model that predict the severity of the accident based in weather an location condition, traffic and locations and we will work at this level of the investigation because we don't have access to the information about other attributes or circumstances of the accident that we willing to have.

For the preparation an understanding the dataset, in the nexts steps we must to reduce or choose the attributes that will be more relevant for the study and select the best data for the model that will give us the better accuracy for the model.

2. Data Understanding

In order to understand the information we found information about concentration of the accident, in this case of the dataset, there are three State that shows high frequency, California, Texas y Florida. Principally the accidents concentrate in Junctions, High-Speed ways, whith 60º (f) and about first hours of the days.

#	Attribute	Description	Nullable
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

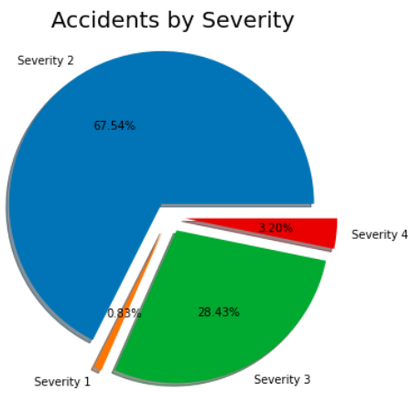
Acknowledgments

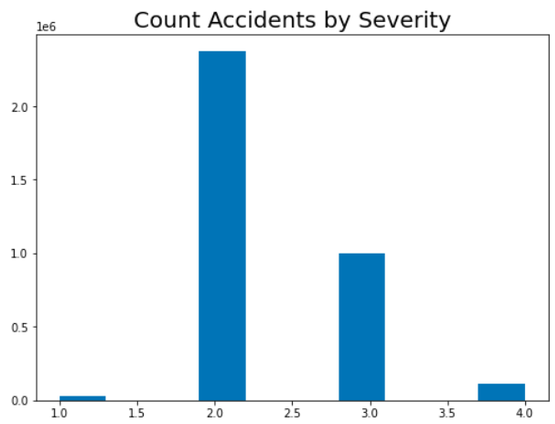
Please cite the following papers if you use this dataset:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (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.

About Severity distribution in dataset

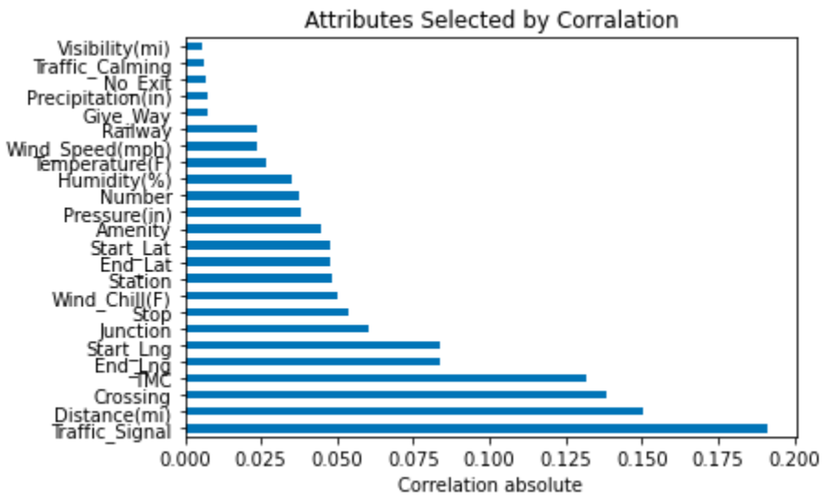
In the next chart we can see that 67% of the dataset refer to Severity 2 and 28.43% refer to severity 3 in data set, en histogram of de right side we can see the frequency of each severity in terms of numbers the case in dataset.





Correlation of the dataset

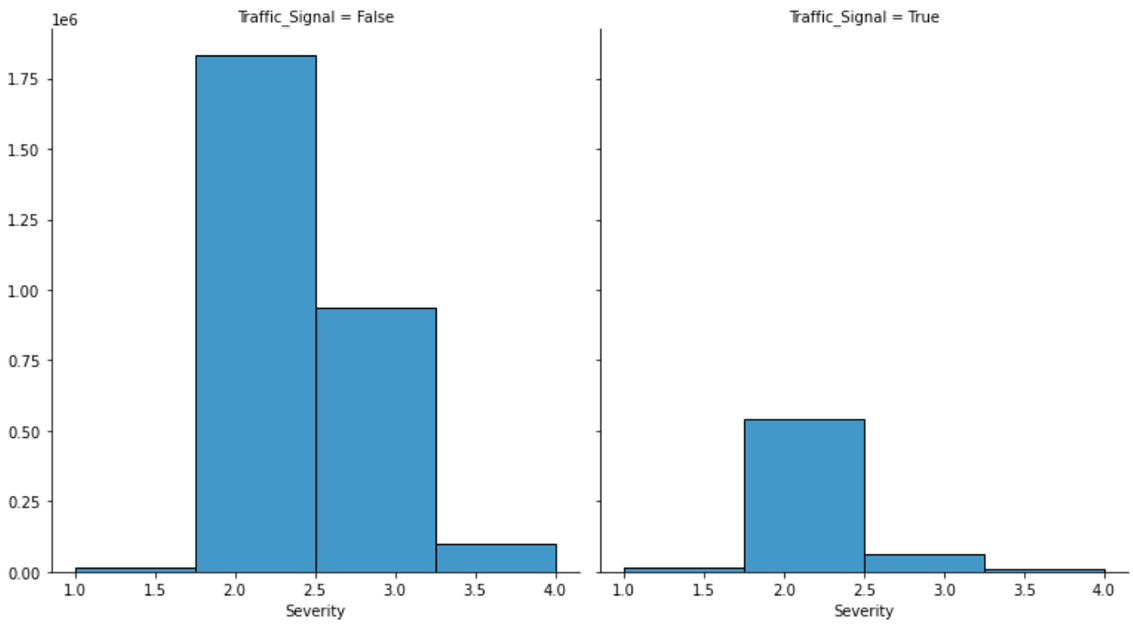
In terms of correlation in dataset we can see that Traffic Signal is the attribute with best correlation with Severity accident with 0.19 and the next is Distance (mi), and Crossing, this attributes are very significative correlated with Severity and this in particular case of cars is something very logic because there are many of the car accident related with this attribute.



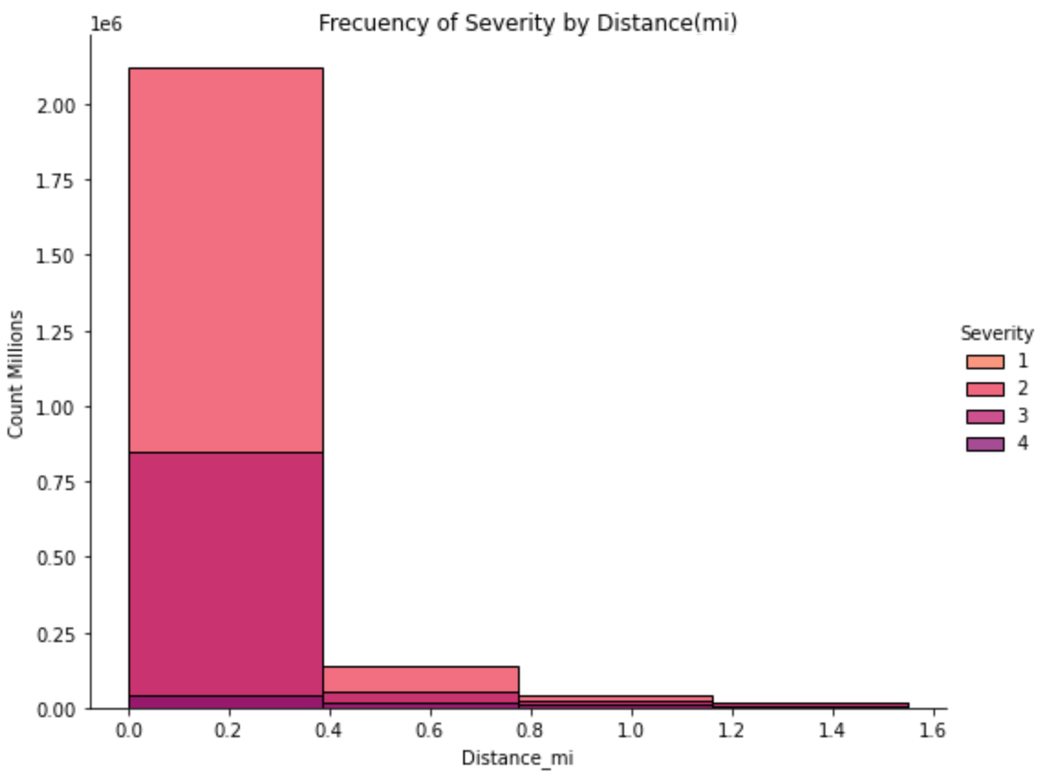

```
Traffic_Signal       0.191531
Distance(mi)         0.150326
Crossing             0.138368
TMC                  0.131999
End_Lng              0.083706
Start_Lng            0.083705
Junction             0.060086
Stop                 0.053500
Wind_Chill(F)        0.049886
Station              0.048260
End_Lat              0.047623
Start_Lat            0.047617
Amenity              0.044494
Pressure(in)         0.038368
Number               0.037541
Humidity(%)          0.035378
Temperature(F)       0.026489
Wind_Speed(mph)      0.023700
Railway              0.023492
Give_Way             0.007747
Precipitation(in)    0.007436
No_Exit              0.006705
Traffic_Calming      0.006073
Visibility(mi)       0.005540
Name: Severity, dtype: float64
```

Traffic Signal, Distance and Crossing description

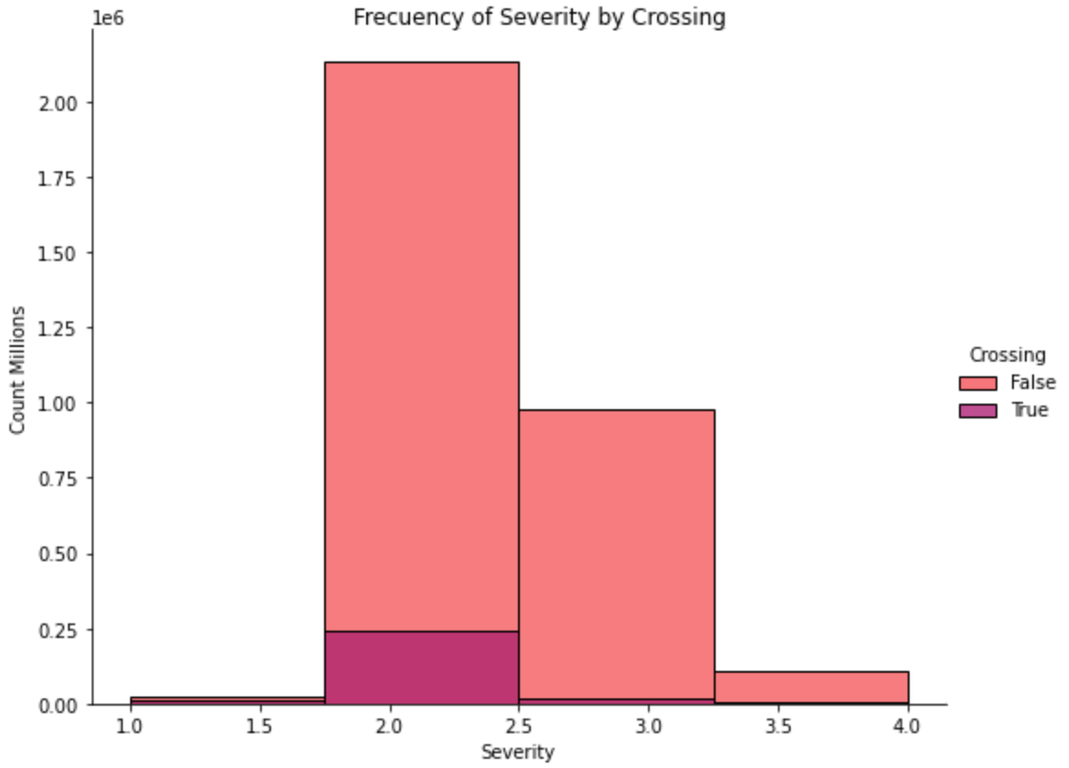
In the next four chart we can see the relation between Severity an Traffic Signal in this case in histogram of frequency we can see much more concentration in case False.



En this case the frequency of the distance of severity is concentrate in less of 0.4 mi with severity two and thee.

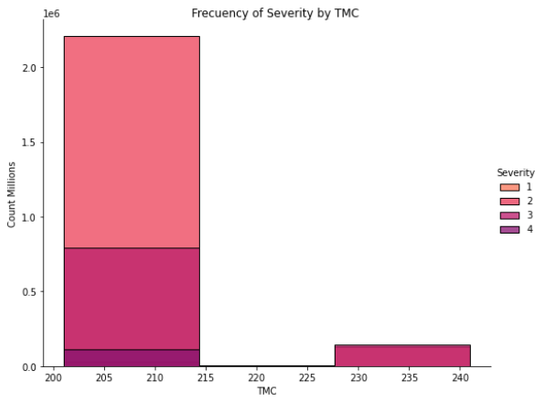


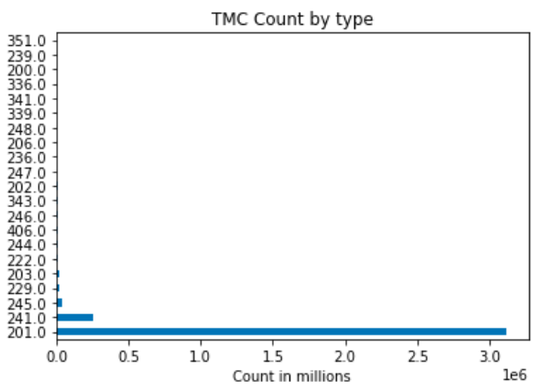
In case of Severity frequency by crossing we can see much more frequency en false case principally in severity two and three



Other Attribute Analyzes

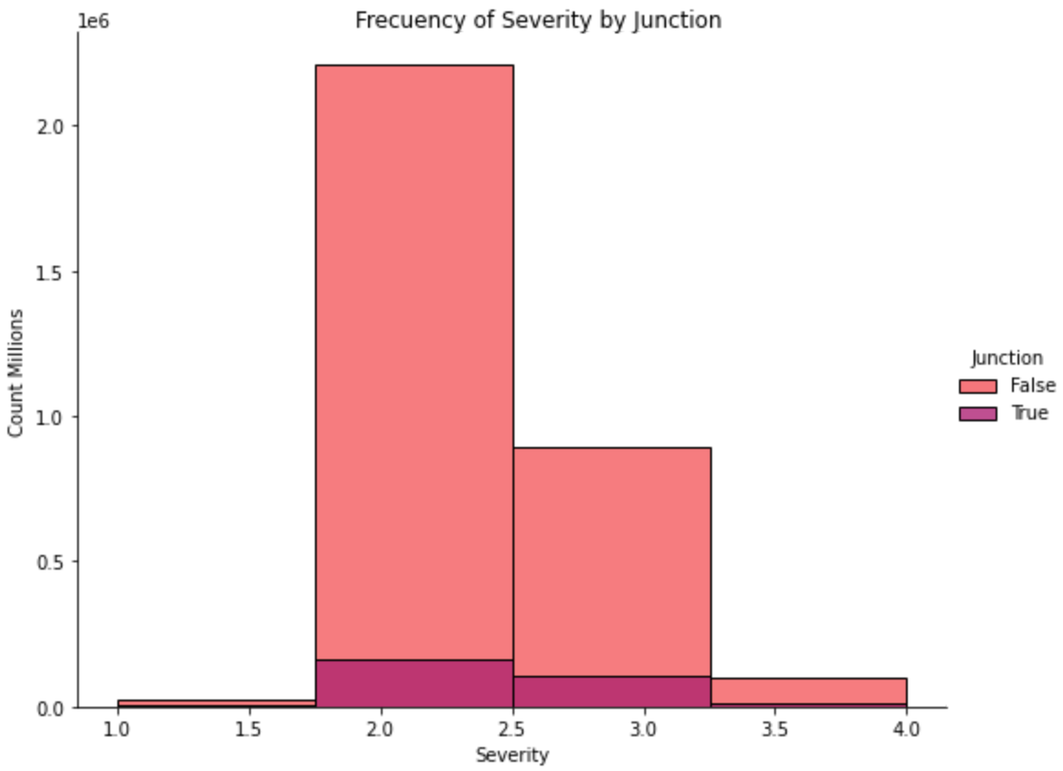
TMC





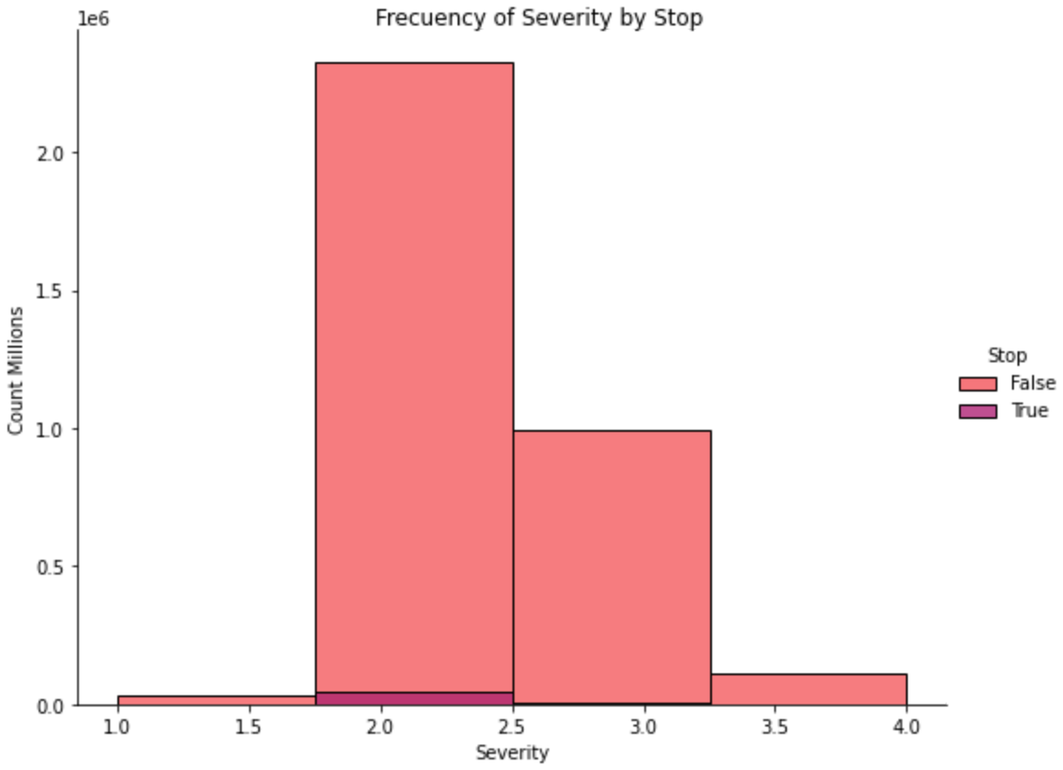
In case of TMC we can see that the common is 201.0 and is principally severity two an tree

Junction Vs Severity



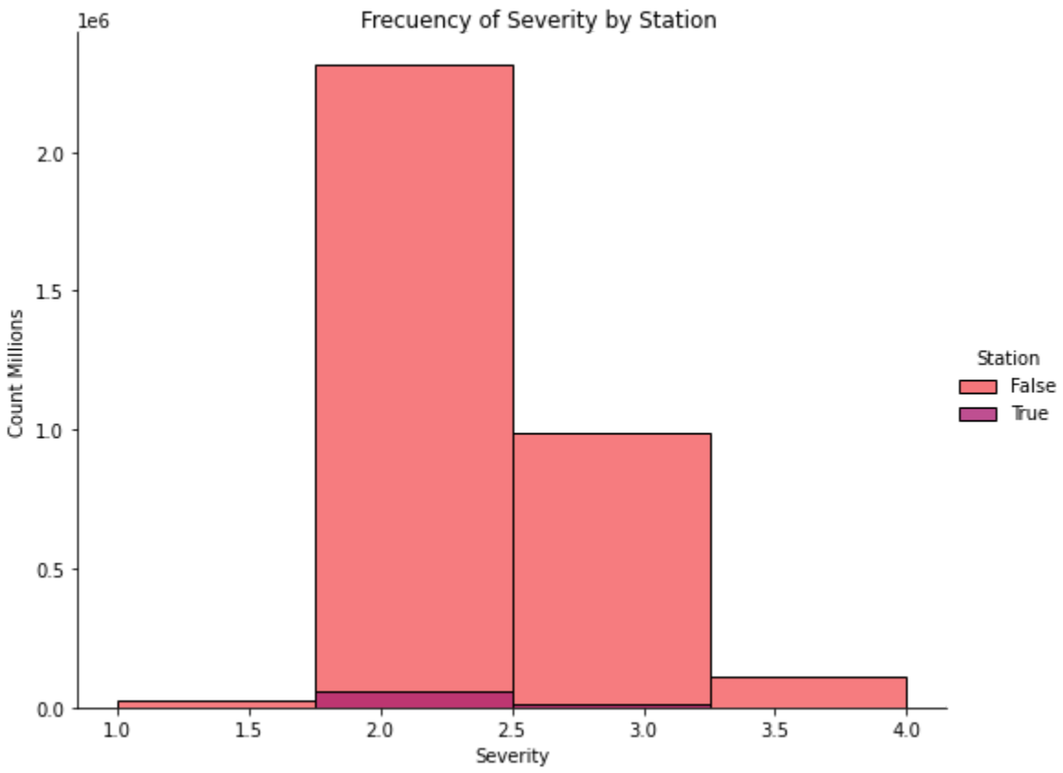
The junction most common is False for the severity two and three

Stop vs Severity



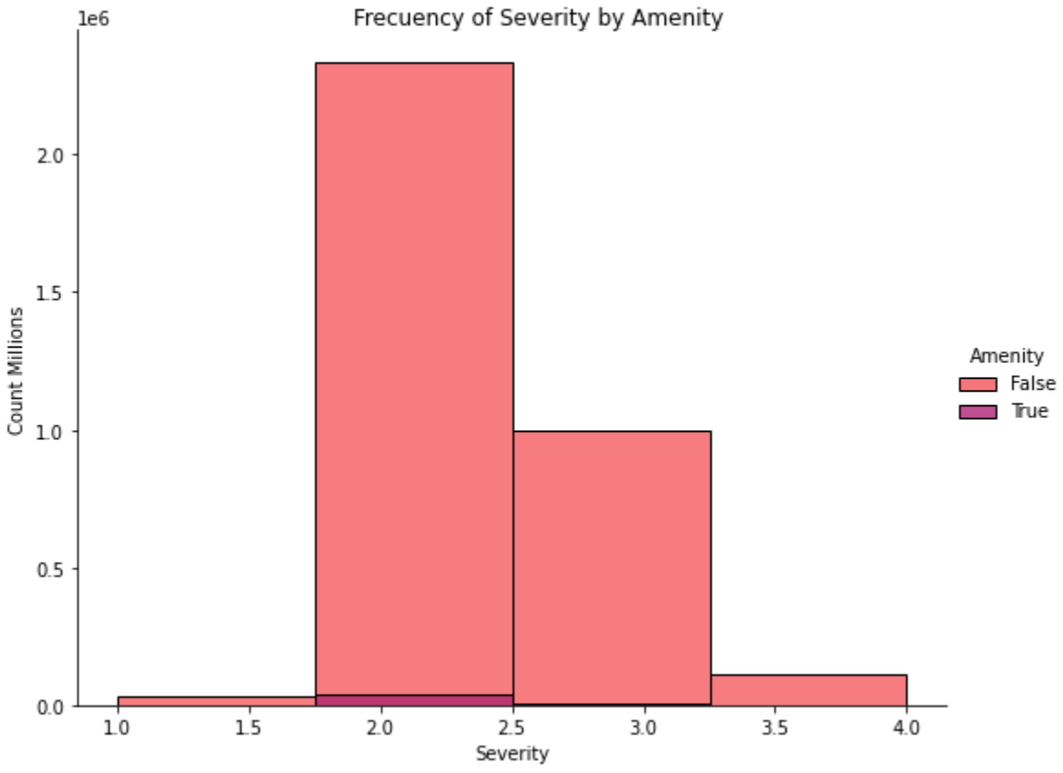
The common Stop is false for the severity two an three

Station vs Severity



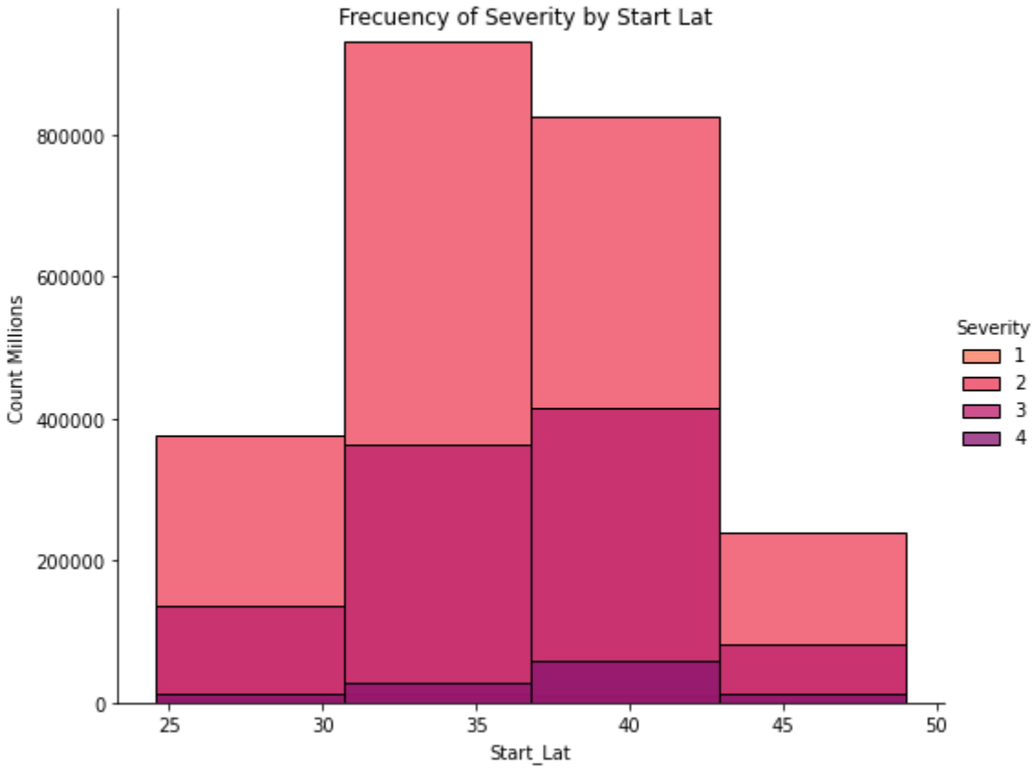
The most common value in Station es False

Amenity vs Severity



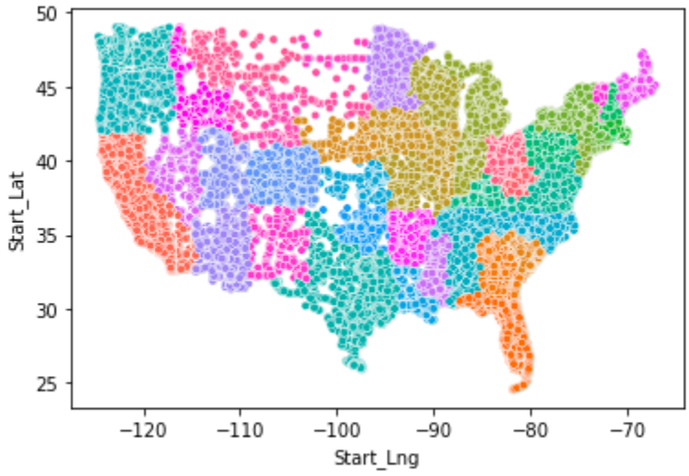
The most common value in Amenity is false

Start Lat vs Severity



Plotting Distribution in Latitude and Longitude

we use basic plotting chart to see the concentration of point in a very basic map, in this case we can see concentration in each cost and less concentration in central and north of the maps.



Statistic Descriptive Basic:

We use statistic descriptive basic for evaluate each attribute for the model

	ID	Source	TMC	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng
count	3513617	3513617	3.513617e+06	3.513617e+06	3513617	3513617	3.513617e+06	3.513617e+06	3.513617e+06	3.513617e+06
unique	3513617	3	NaN	NaN	3200042	3246120	NaN	NaN	NaN	NaN
top	A-1511253	MapQuest	NaN	NaN	2017-05-15 09:22:55	2017-05-15 15:22:55	NaN	NaN	NaN	NaN
freq	1	2414301	NaN	NaN	74	73	NaN	NaN	NaN	NaN
mean	NaN	NaN	2.059544e+02	2.339929e+00	NaN	NaN	3.654195e+01	-9.579151e+01	3.654194e+01	-9.579147e+01
std	NaN	NaN	1.773360e+01	5.521935e-01	NaN	NaN	4.883520e+00	1.736877e+01	4.883503e+00	1.736876e+01
min	NaN	NaN	2.000000e+02	1.000000e+00	NaN	NaN	2.455527e+01	-1.246238e+02	2.455527e+01	-1.246238e+02
25%	NaN	NaN	2.010000e+02	2.000000e+00	NaN	NaN	3.363784e+01	-1.174418e+02	3.363784e+01	-1.174419e+02
50%	NaN	NaN	2.010000e+02	2.000000e+00	NaN	NaN	3.591687e+01	-9.102601e+01	3.591684e+01	-9.102598e+01
75%	NaN	NaN	2.010000e+02	3.000000e+00	NaN	NaN	4.032217e+01	-8.093299e+01	4.032217e+01	-8.093288e+01
max	NaN	NaN	4.060000e+02	4.000000e+00	NaN	NaN	4.900220e+01	-6.711317e+01	4.907500e+01	-6.710924e+01

3. Data Preparation

We search nulls, missing and outliers data for preparation before modeling and replace every data that we will impact in the attribute. in this case we use the median an max value principally for replace the nulls value in data set.

This is the situation initial before the replace:

	ID	Source	TMC	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	...	Roundabout	Station	Stop
0	False	False	False	False	False	False	False	False	True	True	...	False	False	False
1	False	False	False	False	False	False	False	False	True	True	...	False	False	False
2	False	False	False	False	False	False	False	False	True	True	...	False	False	False
3	False	False	False	False	False	False	False	False	True	True	...	False	False	False
4	False	False	False	False	False	False	False	False	True	True	...	False	False	False

Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop	Sunrise_Sunset	Civil_Twilight	Nautical_Twilight	Astronomical_Twilight
False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False	False

And this is before with 49 attribute:

	ID	Source	TMC	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	...	Roundabout	Station	Stop
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False

Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop	Sunrise_Sunset	Civil_Twilight	Nautical_Twilight	Astronomical_Twilight
False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False

3. Modeling, Evaluating and Deployment

We decide to use four models for the data set KNN, Decision Tree, Logistic Regression and SVC but we has trouble deploying KNN an SVC principally for the size of data set, this two model do not perform well with large data set, in this only with knn we use three day computing the result and this is not well.

In summary we use two models with this result...

Algorithm	Jaccard	F1_score	LogLoss	Accuracy
Decision Tree	0.527681	0.677101	NA	0.682432
Logistic Regression	0.507619	0.626959	0.697687	0.698150

Conclusion

With all information gathering in this case and every attribute analyzed we conclude that the best model in this case es Logistic Regression with level of accuracy of 0,6981 that work well determining the level of severity in car accident