Predicting Car Accident Severity in USA

Parth Kalbag

September 2, 2020

1. Introduction

1.1 Background

Road accidents have become very common these days. Nearly
1.25 million people die in road crashes each year, on average, 3,287 deaths a day. Moreover, 20–50 million people are injured or disabled annually. Road traffic crashes rank as the 9th leading cause of death and accounts for 2.2% of all deaths globally. Road crashes cost USD 518 billion globally, costing individual countries from 1–2% of their annual GDP. In the USA, over 37,000 people die in road crashes each year, and 2.35 million are injured or disabled. Road crashes cost the U.S. $230.6 billion per year or an average of $820 per person. Road crashes are the single greatest annual cause of death of healthy U.S. citizens travelling abroad.

1.2 Target Audience

1. The Seattle administration: By targeting areas prone to areas to speeding accidents, interventions such as speed bumps, stop signs etc. can be put in place to reduce accidents.

2. Car Insurance Companies: Areas where parked cars are prone to getting damaged. Owners in those localities may be asked to pay more premium on their car insurance.

3. Health-care workers and emergency services in Seattle: By having enough data on the crash one can predict the severity and therefore take action more quickly potentially saving lives

2. Data

2.1 Data Sources

Data that might contribute to determine the accident severity were obtained from Kaggle. This is a countrywide traffic accident dataset which covers 49 states of United States. The data is continuously being collected from various APIs like Bing and MapQuest from last 5 years. 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. This dataset contains about 3.5 million of rows and 49 columns.

2.2 Features of the dataset

Traffic Attributes (12):

1. ID: This is a unique identifier of the accident record
2. Source: Indicates source of the accident report (i.e. the API which reported the accident.)
3. TMC: A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event
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).
5. Start_Time: Shows start time of the accident in local time zone.
6. End_Time: Shows end time of the accident in local time zone.
7. Start_Lat: Shows latitude in GPS coordinate of the start point.
8. Start_Lng: Shows longitude in GPS coordinate of the start point.
9. End_Lat: Shows latitude in GPS coordinate of the end point.
10. End_Lng: Shows longitude in GPS coordinate of the end point.
11. Distance(mi): The length of the road extent affected by the accident.
12. Description: Shows natural language description of the accident

Address Attributes (9):

1. Number: Shows the street number in address field.
2. Street: Shows the street name in address field.
3. Side: Shows the relative side of the street (Right/Left) in address field.
4. City: Shows the city in address field.
5. County: Shows the county in address field.
6. State: Shows the state in address field.
7. Zipcode: Shows the zipcode in address field.
8. Country: Shows the country in address field.

9. Timezone: Shows timezone based on the location of the accident (eastern, central, etc.).

Weather Attributes (11):

1. Airport_Code: Denotes an airport-based weather station which is the closest one to location of the accident.
2. Weather_Timestamp: Shows the time-stamp of weather observation record (in local time).
3. Temperature(F): Shows the temperature (in Fahrenheit).
4. Wind_Chill(F): Shows the wind chill (in Fahrenheit).
5. Humidity(%): Shows the humidity (in percentage).
6. Pressure(in): Shows the air pressure (in inches).
7. Visibility(mi): Shows visibility (in miles).
8. Wind_Direction: Shows wind direction.
9. Wind_Speed(mph): Shows wind speed (in miles per hour).
10. Precipitation(in): Shows precipitation amount in inches, if there is any.
11. Weather_Condition: Shows the weather condition (rain, snow, thunderstorm, fog, etc.).

POI Attributes (13):

1. Amenity: A Point-Of-Interest (POI) annotation which indicates presence of amenity in a nearby location.
2. Bump: A POI annotation which indicates presence of speed bump or hump in a nearby location.
3. Crossing: A POI annotation which indicates presence of crossing in a nearby location.
4. Give_Way: A POI annotation which indicates presence of give_way sign in a nearby location.
5. Junction: A POI annotation which indicates presence of junction in a nearby location.
6. No_Exit: A POI annotation which indicates presence of no_exit sign in a nearby location.
7. Railway: A POI annotation which indicates presence of railway in a nearby location.
8. Roundabout: A POI annotation which indicates presence of roundabout in a nearby location.
9. Station: A POI annotation which indicates presence of station (bus, train, etc.) in a nearby location.
10. Stop: A POI annotation which indicates presence of stop sign in a nearby location.
11. Traffic_Calming: A POI annotation which indicates presence of traffic_calming means in a nearby location.
12. Traffic_Signal: A POI annotation which indicates presence of traffic_signal in a nearby location.

13. Turning_Loop: A POI annotation which indicates presence of turning_loop in a nearby location.

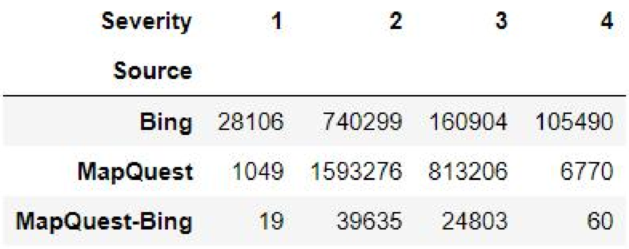
Period-of-Day (4):

1. Sunrise_Sunset: Shows the period of day (i.e. day or night) based on sunrise/sunset.
2. Civil_Twilight: Shows the period of day (i.e. day or night) based on civil twilight.
3. Nautical_Twilight: Shows the period of day (i.e. day or night) based on nautical twilight.
4. Astronomical_Twilight: Shows the period of day (i.e. day or night) based on astronomical twilight.

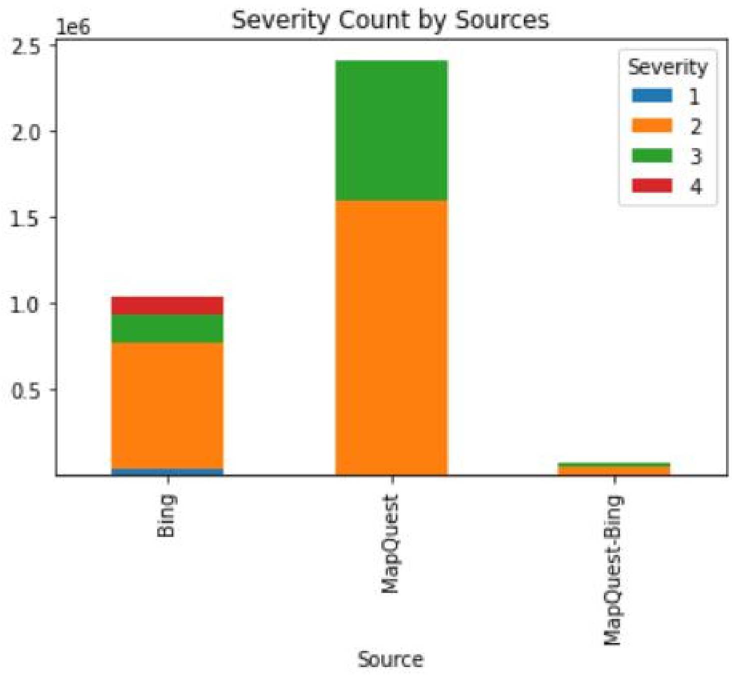
2.3 Data Cleaning

Note that the data has come from two different sources MapQuest and Bing

Since severity is what we really care about in this project, I think it is crucial to figure out the difference.

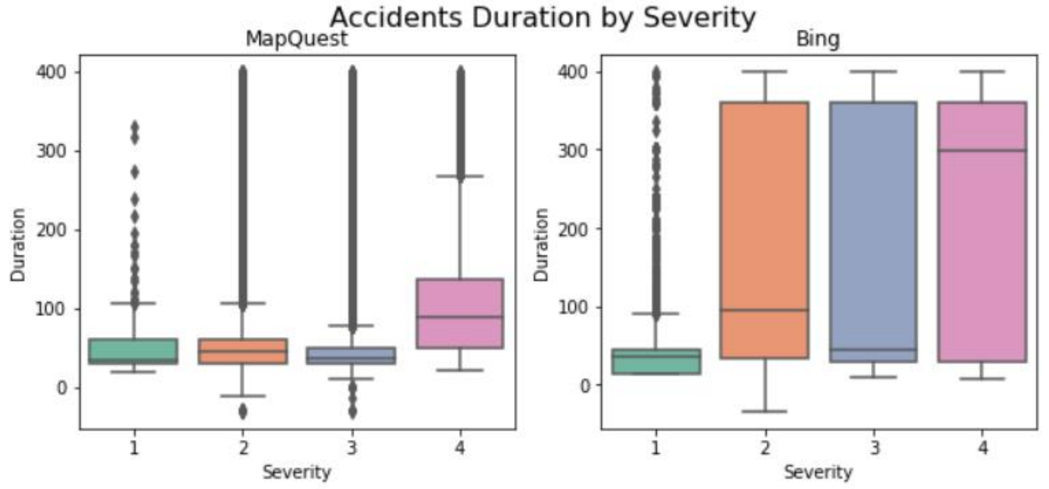


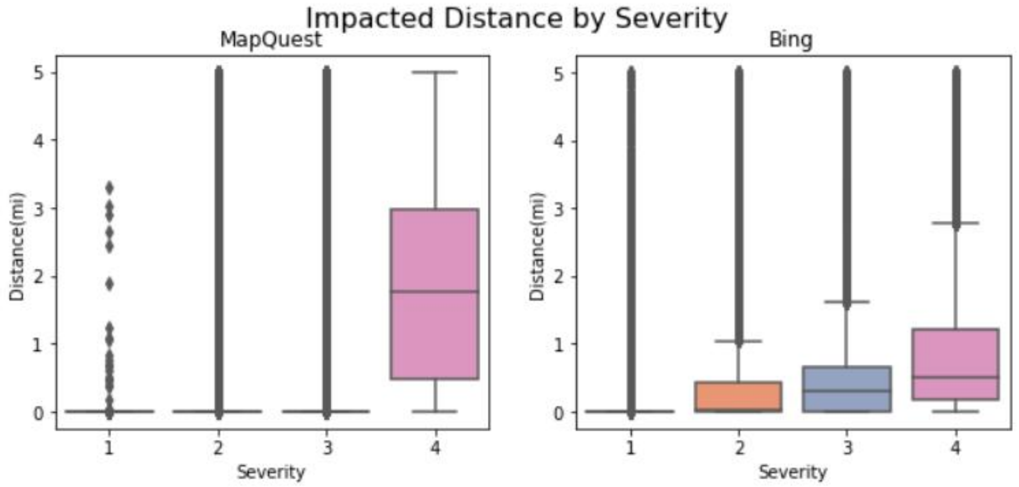
Severity	1	2	3	4
Source				
Bing	28106	740299	160904	105490
MapQuest	1049	1593276	813206	6770
MapQuest-Bing	19	39635	24803	60



The stacked bar chart shows that two data providers reported totally different proportions of accidents of each level. MapQuest reported so rare accidents with severity level 4 which can not even be seen in the plot, whereas Bing reported almost the same number of level 4 accidents as level 2. Meanwhile, MapQuest reported much more level 3 accidents than Bing in terms of proportion. These differences may be due to the different kinds of accidents they tend to collect or the different definitions of severity level, or the combination of them. If the latter is the case, I don't think we can use the data from both of them at the same time. To check it out, we can examine the distribution of accidents with different severity levels across two main measures, Impacted Distance and Duration

2.3.1 Cleaning of some categorical features





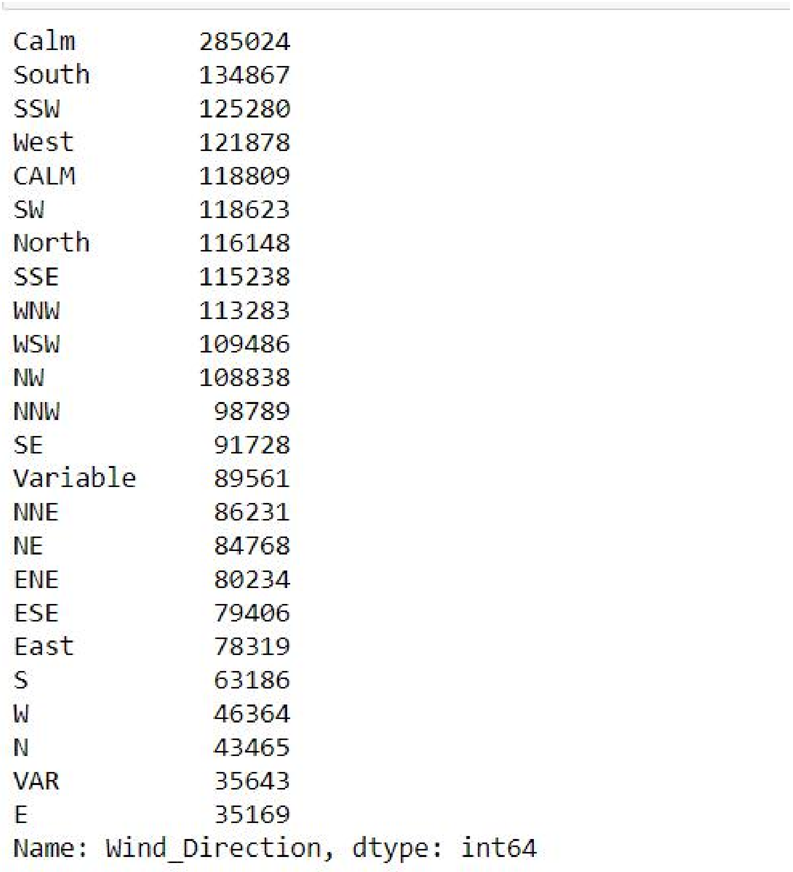
Two differences are obvious in the above plots. The first is that the overall duration and impacted distance of accidents reported by Bing are much longer than those by MapQuest. Second, same severity level holds different meanings for MapQuest and Bing. MapQuest seems to have a clear and strict threshold for severity level 4, cases of which nevertheless only account for a tiny part of the whole dataset. Bing, on the other hand, doesn't seem to have a clear-cut threshold, especially regards duration, but the data is more balanced.

It is hard to choose one and we definitely can't use both. I decided to select MapQuest because serious accidents are we really care about and the sparse data of such accidents is the reality we have to confront.

Finally, drop data reported from Bing and 'Source' column.

2.3.1.1 Removing Useless features

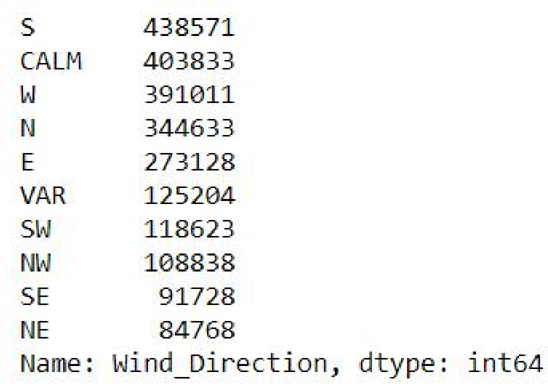
Features 'ID' doesn't provide any useful information about accidents themselves. 'TMC', 'Distance(mi)', 'End_Time' (we have start time), 'Duration', 'End_Lat', 'Country', and 'End_Lng'(we have start location) can be collected only after the accident has already happened and hence cannot be predictors for serious accident prediction. For 'Description', the POI features have already been extracted from it by dataset creators. Let's get rid of these features first



```
Calm        285024
South       134867
SSW         125280
West        121878
CALM        118809
SW          118623
North       116148
SSE         115238
WNW         113283
WSW         109486
NW          108838
NNW          98789
SE           91728
Variable     89561
NNE          86231
NE           84768
ENE          80234
ESE          79406
East         78319
S            63186
W            46364
N            43465
VAR          35643
E            35169
Name: Wind_Direction, dtype: int64
```

2.3.1.2 Cleaning Wind_Direction

Now in the above dataframe the (East and E, ESE, ENE), (South and S,SSW, SSE), (North and N, NNW, NNE), (West and WSW, WNW, W) are basically the same thing so we must convert all of them to a standard variable



```
S       438571
CALM    403833
W       391011
N       344633
E       273128
VAR     125204
SW      118623
NW      108838
SE       91728
NE       84768
Name: Wind_Direction, dtype: int64
```

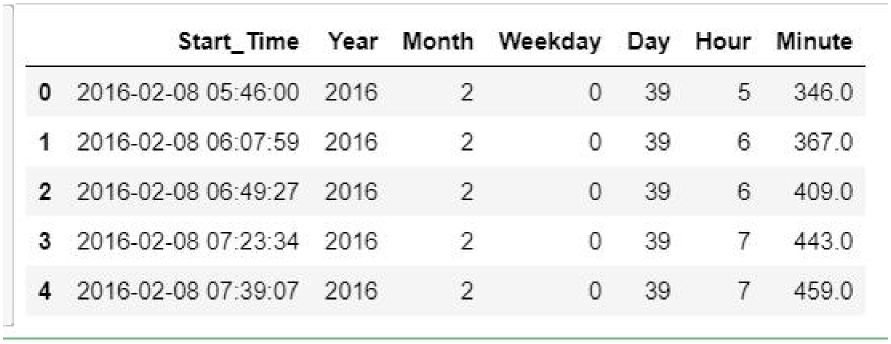
2.3.1.3 Cleaning Weather_Condition:-

```
['Light Rain',
 'Overcast',
 'Mostly Cloudy',
 'Rain',
 'Light Snow',
 'Haze',
 'Scattered Clouds',
 'Partly Cloudy',
 'Clear',
 'Snow',
 'Light Freezing Drizzle',
 'Light Drizzle',
 'Fog',
 'Shallow Fog',
 'Heavy Rain',
 'Light Freezing Rain',
 'Cloudy',
 'Drizzle',
 nan,
 'Light Rain Showers',
 'Mist',
 'Smoke',
 'Patches of Fog',
 'Light Freezing Fog',
 'Light Haze',
 'Light Thunderstorms and Rain',
 'Thunderstorms and Rain',
 'Fair',
 'Volcanic Ash',
 'Blowing Sand',
 'Blowing Dust / Windy',
 'Widespread Dust',
 'Fair / Windy',
 'Rain Showers',
 'Mostly Cloudy / Windy',
 'Light Rain / Windy',
 'Hail',
 'Heavy Drizzle',
 'Showers in the Vicinity',
 'Thunderstorm',
 'Light Rain Shower',
 'Light Rain with Thunder',
 'Partly Cloudy / Windy',
 'Thunder in the Vicinity',
 'T-Storm',
 'Heavy Thunderstorms and Rain',
 'Thunder',
 'Heavy T-Storm',
 'Funnel Cloud',
 'Heavy T-Storm / Windy',
 'Blowing Snow',
 'Light Thunderstorms and Snow',
 'Heavy Snow',
 'Low Drifting Snow',
 'Light Ice Pellets',
 'Ice Pellets',
 'Squalls',
 'N/A Precipitation',
 'Cloudy / Windy',
 'Light Fog',
 'Sand',
 'Snow Grains',
 'Snow Showers',
 'Heavy Thunderstorms and Snow',
 'Rain / Windy',
 'Heavy Rain / Windy',
 'Heavy Ice Pellets',
 'Light Snow / Windy',
 'Heavy Freezing Rain',
 'Small Hail',
 'Heavy Rain Showers',
 'T-Storm / Windy',
 'Patches of Fog / Windy',
 'Drizzle / Windy',
 'Thunder / Windy',
 'Wintry Mix',
 'Squalls / Windy',
 'Rain Shower',
 'Drizzle and Fog',
 'Haze / Windy',
 'Sand / Dust Whirlwinds',
 'Blowing Dust',
 'Fog / Windy',
 'Smoke / Windy',
 'Wintry Mix / Windy',
 'Snow / Windy',
 'Light Rain Shower / Windy',
 'Heavy Snow / Windy',
 'Snow and Sleet',
 'Light Freezing Rain / Windy',
 'Light Drizzle / Windy',
 'Light Snow and Sleet',
 'Partial Fog',
 'Light Snow Shower',
 'Light Snow and Sleet / Windy',
 'Freezing Rain',
 'Blowing Snow / Windy',
 'Freezing Drizzle',
 'Sleet',
 'Light Sleet',
 'Rain and Sleet',
 'Heavy Sleet',
 'Light Snow Grains',
 'Partial Fog / Windy',
 'Light Snow with Thunder',
 'Widespread Dust / Windy',
 'Sand / Dust Whirlwinds / Windy',
 'Tornado',
 'Snow and Thunder',
 'Snow and Sleet / Windy',
 'Heavy Snow with Thunder',
 'Thunder / Wintry Mix / Windy',
 'Light Snow Showers',
 'Heavy Blowing Snow',
 'Light Hail',
 'Heavy Smoke',
 'Heavy Thunderstorms with Small Hail',
 'Light Thunderstorm',
 'Heavy Freezing Drizzle',
 'Light Blowing Snow',
 'Thunderstorms and Snow']
```

The above list represents the weather condition during it so we will select o nly the common weather conditions I.e. ‘Clear’, ‘Cloud’, ‘Rain’, ‘Heavy_Rain’, ‘Snow’, ‘Heavy_Snow’, ‘Fog’

2.3.1.4 Fixing datetime format

As the Weather_Timestamp is basically the same as that of Start_Time and we only need year, month, day, hour, etc. from them we can drop the column.



	Start_Time	Year	Month	Weekday	Day	Hour	Minute
0	2016-02-08 05:46:00	2016	2	0	39	5	346.0
1	2016-02-08 06:07:59	2016	2	0	39	6	367.0
2	2016-02-08 06:49:27	2016	2	0	39	6	409.0
3	2016-02-08 07:23:34	2016	2	0	39	7	443.0
4	2016-02-08 07:39:07	2016	2	0	39	7	459.0

2.3.1.5 Handling Missing data

Following data contains the missing data from the dataset

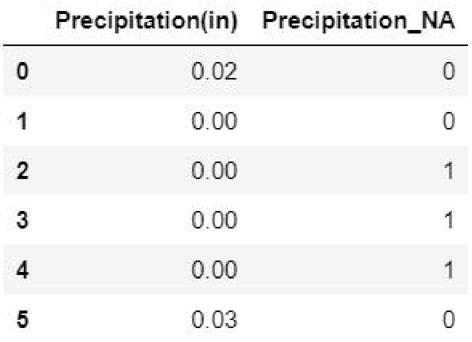
```
Severity
False      2414301
Name: Severity, dtype: int64

Start_Time
False      2414301
Name: Start_Time, dtype: int64

Start_Lat
False      2414301
Name: Start_Lat, dtype: int64

Start_Lng
False      2414301
Name: Start_Lng, dtype: int64

Distance(mi)
False      2414301
Name: Distance(mi), dtype: int64

Number
True       1445664
False       968637
Name: Number, dtype: int64

Street
False      2414301
Name: Street, dtype: int64

Side
False      2414301
Name: Side, dtype: int64
```

```
City
False    2414251
True          50
Name: City, dtype: int64

County
False    2414301
Name: County, dtype: int64

State
False    2414301
Name: State, dtype: int64

Zipcode
False    2413991
True         310
Name: Zipcode, dtype: int64

Timezone
False    2412240
True        2061
Name: Timezone, dtype: int64

Airport_Code
False    2410176
True        4125
Name: Airport_Code, dtype: int64

Temperature(F)
False    2374848
True       39453
Name: Temperature(F), dtype: int64

Wind_Chill(F)
True     1417631
False     996670
Name: Wind_Chill(F), dtype: int64

Humidity(%)
False    2372241
True       42060
Name: Humidity(%), dtype: int64

Pressure(in)
False    2380404
True       33897
Name: Pressure(in), dtype: int64

Visibility(mi)
False    2366747
True       47554
Name: Visibility(mi), dtype: int64

Wind_Direction
False    2380337
True       33964
Name: Wind_Direction, dtype: int64

Wind_Speed(mph)
False    2076023
True      338278
Name: Wind_Speed(mph), dtype: int64

Precipitation(in)
True     1518773
False     895528
Name: Precipitation(in), dtype: int64

Amenity
False    2414301
Name: Amenity, dtype: int64
```

```
Bump
False       2414301
Name:  Bump,  dtype:  int64

Crossing
False       2414301
Name:  Crossing,  dtype:  int64

Give_Way
False       2414301
Name:  Give_Way,  dtype:  int64

Junction
False       2414301
Name:  Junction,  dtype:  int64

No_Exit
False       2414301
Name:  No_Exit,  dtype:  int64

Railway
False       2414301
Name:  Railway,  dtype:  int64

Roundabout
False       2414301
Name:  Roundabout,  dtype:  int64

Station
False       2414301
Name:  Station,  dtype:  int64

Stop
False       2414301
Name:  Stop,  dtype:  int64

Traffic_Calming
False       2414301
Name:  Traffic_Calming,  dtype:  int64

Traffic_Signal
False       2414301
Name:  Traffic_Signal,  dtype:  int64

Turning_Loop
False       2414301
Name:  Turning_Loop,  dtype:  int64

Sunrise_Sunset
False       2414248
True             53
Name:  Sunrise_Sunset,  dtype:  int64

Civil_Twilight
False       2414248
True             53
Name:  Civil_Twilight,  dtype:  int64

Nautical_Twilight
False       2414248
True             53
Name:  Nautical_Twilight,  dtype:  int64

Astronomical_Twilight
False       2414248
True             53
Name:  Astronomical_Twilight,  dtype:  int64

Clear
False       2366785
```

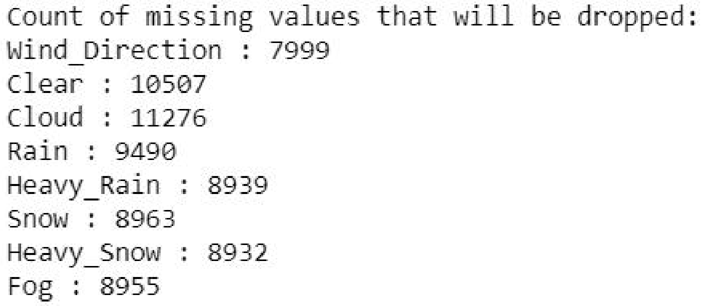
```
True          47516
Name: Clear, dtype: int64

Cloud
False     2366785
True          47516
Name: Cloud, dtype: int64

Rain
False     2366785
True          47516
Name: Rain, dtype: int64

Heavy_Rain
False     2366785
True          47516
Name: Heavy_Rain, dtype: int64

Snow
False     2366785
True          47516
Name: Snow, dtype: int64

Heavy_Snow
False     2366785
True          47516
Name: Heavy_Snow, dtype: int64

Fog
False     2366785
True          47516
Name: Fog, dtype: int64

Year
False     2414301
Name: Year, dtype: int64

Month
False     2414301
Name: Month, dtype: int64

Weekday
False     2414301
Name: Weekday, dtype: int64

Day
False     2414301
Name: Day, dtype: int64

Hour
False     2414301
Name: Hour, dtype: int64

Minute
False     2414301
Name: Minute, dtype: int64
```

Now as far as the Number, Wind_Chill(F), and Precipitation are concerned more than 70 % of the data is missing and we can't just dropna as we would lose that data. 'Number' and 'Wind_Chill(F)' will be dropped because they are not highly related to severity according to previous research, whereas 'Precipitation(in)' could be a useful predictor and hence can be handled by separating feature



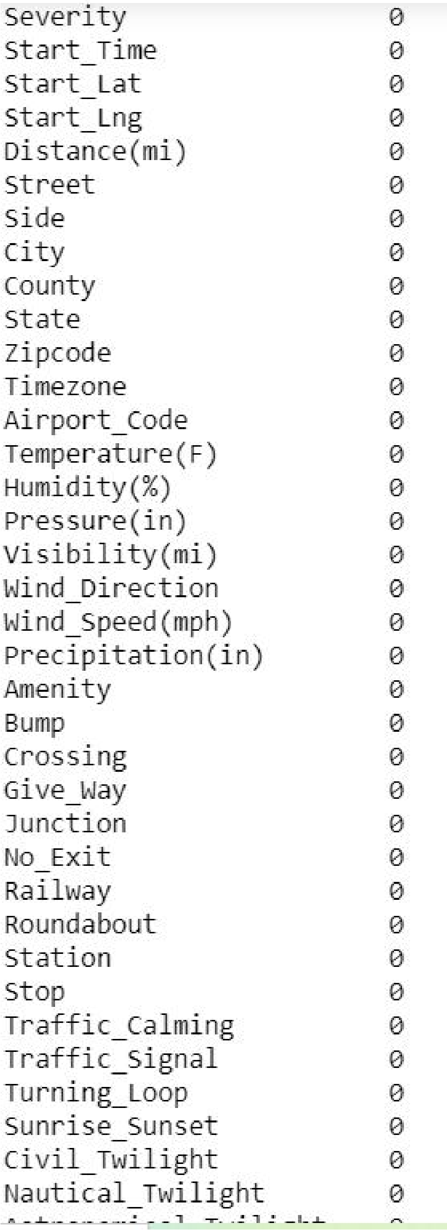
	Precipitation(in)	Precipitation_NA
0	0.02	0
1	0.00	0
2	0.00	1
3	0.00	1
4	0.00	1
5	0.03	0

And as for the remaining values we can't substitute anything regarding it so we can drop the rows containing the values. There are some missing values in weather data but they are very less so we can drop those. Now for weather features mode will better rather than median

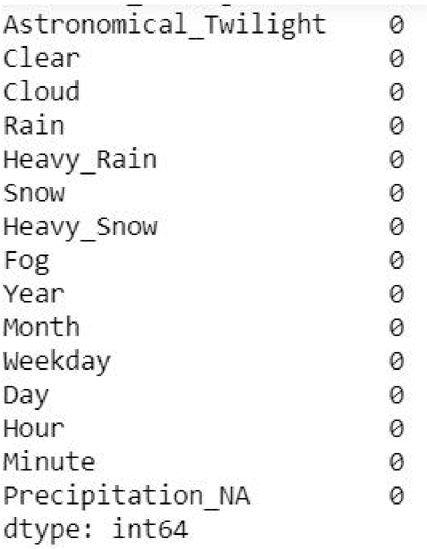


```
Count of missing values that will be dropped:
Wind_Direction : 7999
Clear : 10507
Cloud : 11276
Rain : 9490
Heavy_Rain : 8939
Snow : 8963
Heavy_Snow : 8932
Fog : 8955
```

Hence we have cleaned the data of all the missing values.



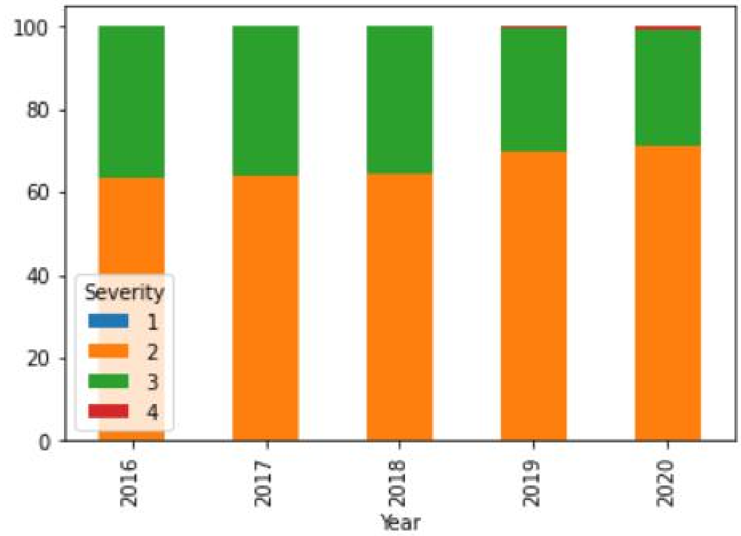
```
Severity                 0
Start_Time               0
Start_Lat                0
Start_Lng                0
Distance(mi)             0
Street                   0
Side                     0
City                     0
County                   0
State                    0
Zipcode                  0
Timezone                 0
Airport_Code             0
Temperature(F)           0
Humidity(%)              0
Pressure(in)             0
Visibility(mi)           0
Wind_Direction           0
Wind_Speed(mph)          0
Precipitation(in)        0
Amenity                  0
Bump                     0
Crossing                 0
Give_Way                 0
Junction                 0
No_Exit                  0
Railway                  0
Roundabout               0
Station                  0
Stop                     0
Traffic_Calming          0
Traffic_Signal           0
Turning_Loop             0
Sunrise_Sunset           0
Civil_Twilight           0
Nautical_Twilight        0
```



```
Astronomical_Twilight    0
Clear                    0
Cloud                    0
Rain                     0
Heavy_Rain               0
Snow                     0
Heavy_Snow               0
Fog                      0
Year                     0
Month                    0
Weekday                  0
Day                      0
Hour                     0
Minute                   0
Precipitation_NA         0
dtype: int64
```

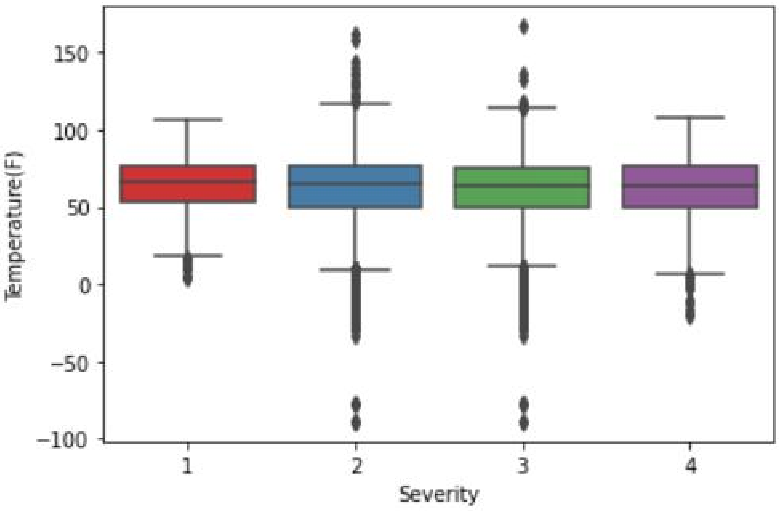
2.3.2 Finding correlation between the data and target column

Let’s find the correlation between Severity and Year:-



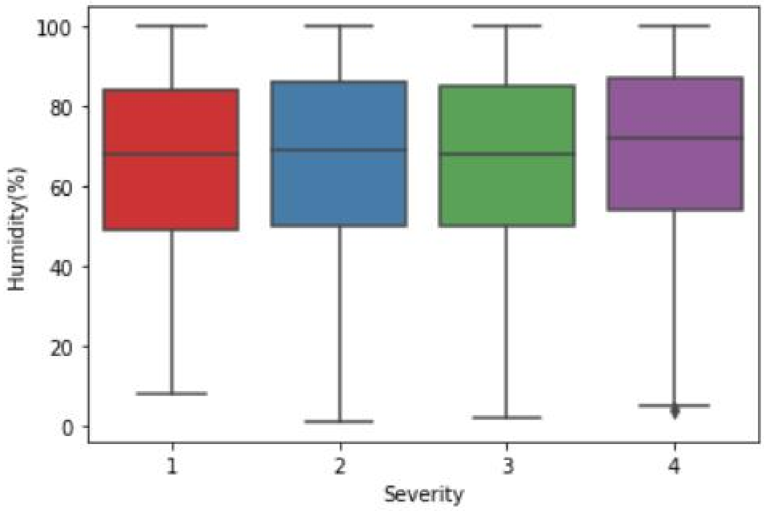
Hence, we can see that year has almost no impact on the severity

Let’s find the correlation between temperature and severity:-

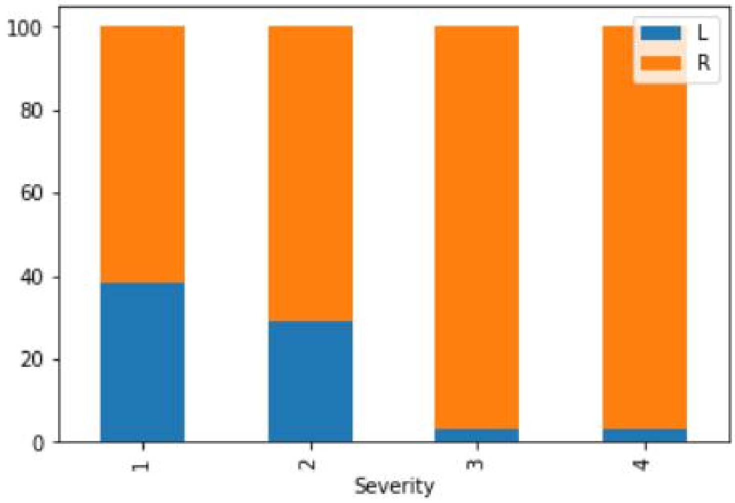


looks like the more severe accident has lower temperature

Let’s find the correlation between humidity and severity:-

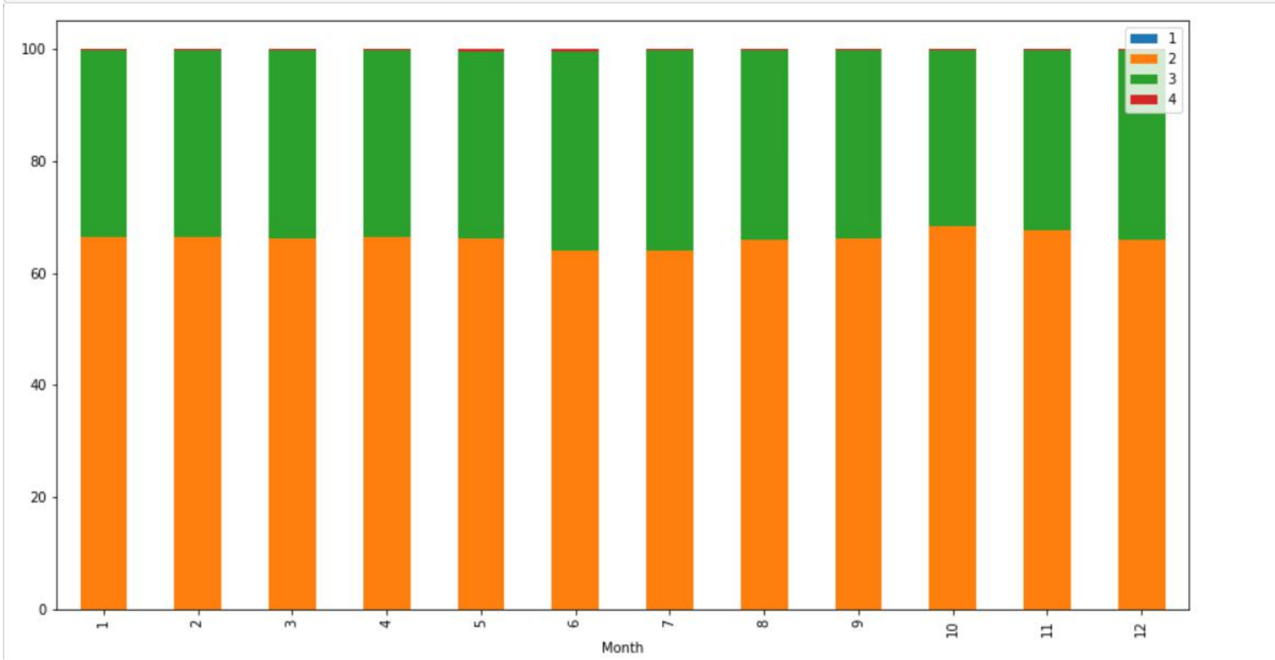


Let’s find the correlation between Side and Severity:-



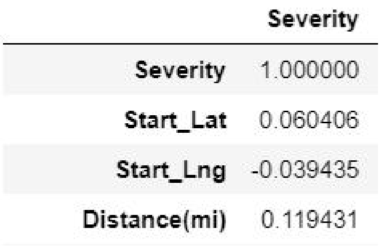
Hence from the above stacked bar chart we can see that most accidents have occurred on the right side of the road.

Let’s find the correlation between Month and Severity:-



Hence, we can see that most of the accidents occurred on the right side of the road.

Let’s find a correlation between distance and Severity:-



	Severity
Severity	1.000000
Start_Lat	0.060406
Start_Lng	-0.039435
Distance(mi)	0.119431

From the correlation between table we can see that the correlation between Severity and Distance(mi) is about 0.1, hence they are strongly correlated also same goes for Start_Lat and Severity.

Let’s find a correlation between Wind_Speed and Severity:-

We can see that at very high wind speed the severity of the accidents have increased so we can say that wind_speed and Severity are correlated.

3. **Modeling**

We can see that the accident_severity has only 4 values i.e. 1, 2, 3, 4 hence we have to consider a classification based algorithm. I used various classification based algorithms like Random Forest, Decision Tree and Logistic Regression. I have also used the jaccard score and f1-score to predict the accuracy of the model based on the data and future data so that to predict how would it perform in real world.

We have obtained all the necessary variables to predict the accident severity and they are :-

- Distance(mi)
- Start_Time and Start_Lat as majority of accidents are occured in CA
- All the weather terms
- Precipitation
- Side i.e. 'L' and 'R' columns
- Visibility(mi)
- Temperature(F)
- Wind_speed(mph)
- Junction

So we will now see how our models perform with these variables.

	Random Forest	Decision Tree	Logistic Regression
Accuracy Score	0.8233	0.7757	0.6989
Jaccard Score	0.7031	0.6412	0.5339
F1-Score	0.8213	0.7761	0.6778

Hence, we can see that Random Forest Model has worked the based producing an accuracy of 82.33 % .

4. Conclusion

In this study, I analyzed the relationship between Severity of the Accidents and the various independent variables like Wind_Speed, Start_Lat, etc. to predict the accident severity using a classification model (i.e. Random Forest). This would help U.S. government to improve the traffic conditions from the relationships described in this report.