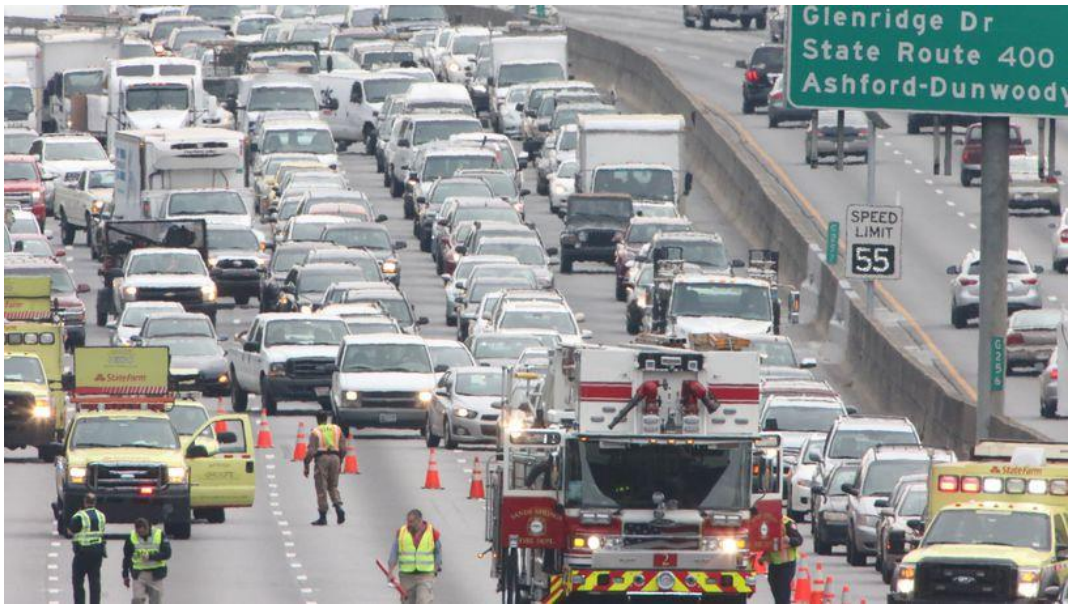
Data wrangling – Project Report 2022-2023 (P3)

Determining traffic accident severity in the USA



Group 2

Project group members

1. Sebastião Rosalino, sxx209
2. Kilian Diederix, kdx300
3. Justine de Jong, jde245
4. Mick van den Boer, mbr633

Professor

Prof. dr. Sandjai Bhulai

Previous note

The jupyter notebook file contains all detailed information and explanation regarding every operation mentioned throughout this report. This report, therefore, does not replace the notebook to fully understand the whole data analysis carried out.

1. Research question

"What are the key factors that determine a higher/lower traffic accident severity in the USA?"

To answer this question, we will examine three sub questions concerning three major factors:

- To what extent do weather conditions impact the severity of traffic accidents?
- To what extent do time conditions impact the severity of traffic accidents?
- To what extent do infrastructure conditions impact the severity of traffic accidents?

Furthermore, we will study the accident by geographical distribution throughout the country by states and cities; by date (months and days of the week) where accidents are prevalent; by duration in hours; and by distance in traffic congestion.

This study will help the authorities to better understand traffic problems and to establish public policies to minimize this critical issue, and for example for insurance companies to define their commercial policy. If we can identify the patterns of how these serious accidents happen and the key factors, we might be able to implement well-informed actions and better allocate financial and human resources.

2. Data sources

Moosavi, Sobhan. (2019). US Accidents(2016 - 2021), Version 6 (Dec 2021). Retrieved 12th of January 2023 from https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?datasetId=199387&sortBy=voteCount&search=severity.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 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.

File name: "US_Accidents_Dec21_updated.csv"
File format: CSV
Last access date: January 25, 2023
Website: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

3. Data wrangling methods

The data consists of 2.845.342 instances of accidents. Every accident has a unique ID so we used that as an index. For every accident there is a severity level. Severity represents a number between 1 and 4, where 1 indicates the least impact on traffic (short delay because of the accident) and 4 indicates a significant impact on traffic (long queue delay). The relation between the severity level and how long the delay is, is not provided by the author of the dataset, because it is not tracked. However, by his own analysis he claims to have found a correlation between the severity and delay, the results of which can be found in Appendix A4.

The other features can be categorized into four groups: **weather conditions, phase of the day**, **infrastructure conditions**, **location attributes**, and **accident characterization** (see Appendix A3).

As soon as we started analyzing the data, we discovered that 89% of the recorded accidents in our dataset have a severity score of 2 (See appendix A1). And most of the accidents with severity score 2 have taken place in 2021 (See appendix A2). Since there was no clear explanation why the severity score is so unevenly distributed between accidents and time, we decided not to reshape the data. Although it will be something to take into consideration for our analysis.

Feature engineering and Data cleaning

We discovered that the features regarding the time of the accident were not proper datetime python objects. Therefore, we converted these features to *datetime64* format. This way we could use the time features to create new columns that we could use for our analysis. For example, we could now aggregate accidents by year, month, weekday, and hour, and calculate the duration of an accident. Also, we made sure that our newly created month feature was chronologically ordered so we could use the values for plotting a graph over the course of a year.

Since Duration was a continuous variable, it is still not in its most meaningful format. It is better suited for analysis if it is converted into a categorical format using bins. Following the same situation, the variable Distance is in continuous (miles) format. It is better suited for analysis if it is converted into a categorical format using bins.

We used the one-hot encoding technique for the Weather Condition, Infrastructure Condition and Phase of the Day feature groups. These operations were revealed as necessary because they split the Weather Condition: Sunrise Sunset and Civil Twilight variables into multiple binary categorical variables to serve for correlation maps for the data exploration section later.

We noticed there was irrelevant/redundant data on many columns, so we decided to drop them: "Start_Lat", "Start_Lng", "End_Lat", "End_Lng", "Description", "Number", "Street", "Side", "County", "Zipcode", "Country", "Airport_Code", "Weather_Timestamp", "Nautical_Twilight", "Astronomical_Twilight".

Handling missing values

Out of over 2.8+ million records, only 137 had no value (NaN) for the variable *City*, such a small subsection of the dataset will be irrelevant for any analysis, so these records were dropped. The variable containing the Distance in miles caused by the accident, will be of utmost analytical importance. Since there are merely 329 missing values in a collection of 2845342 records, we decided to drop those instances. Following the same issue as the above variables, the Timezone variable simply contains 3659 missing values, which is a fairly low number, hence the drop.

The same could not be said for the variables: Humidity (%), Pressure(in), Temperature(F), Wind_Chill(F) & Wind_speed(mph). These variables had nearly 950.000 NaN combined (see Appendix A5 for distribution). While this is still only ~33% of the dataset we did not want to drop these records as they represent important numerical data. We thus decided to input all the weather numerical missing values as the average of every state where the accident took place.

4. Correlations

To understand whether our independent variables had a positive or a negative correlation with the dependent variable Severity (or no correlation at all). We constructed three correlation matrices. Containing the variables in the columns: Weather Conditions, Phase of the Day and Infrastructure Conditions found in A3 of the appendix. The matrix for all three can be found in appendix B1-B3.

5. Linear Regression

To predict the influence our independent variables will have on our dependent variable we used a linear regression model [2]. We have divided our independent variables into 3 groups: variables that have a positive correlation with the dependent variable Severity, variables that have a negative correlation with the dependent variable Severity and lastly, a combination of both. We will only review the last one in this report, the others can be found in the notebook.
Having fit our model on the data, the variables that showed the strongest explanation power over the Severity (hence the higher absolute coefficient), are Weather_Condition_Clear, Junction, Civil_Twilight_Day, Civil_Twilight_Night, Weather_Condition_Fair, Weather_Condition_Cloudy and Crossing. The results obtained can be found in appendix C1.

6. Conclusion

To answer our research question, we needed to find a correlation between the severity of an accident and the conditions that were recorded when an accident took place. We divided all the features in the dataset into five groups of similar conditions. We specifically focused on three kinds of conditions: Weather conditions, Time conditions and Infrastructure conditions.

Our analyses resulted in the following main findings:

- The great majority of the accidents in the dataset lie on a severity value of 2 (~ 89%). And the accident severity mean is approximately 2,137
- Most accidents take between 1 and 3 hours to be solved. On the other hand, accidents that take up more than 15 hours to be solved are residuals
- The great majority of the accidents do not provoke huge problems in terms of queue length, as the 0-1 distance interval is the most frequent situation
- The distance caused by the accident increases the average severity caused by the accident itself increases as well. The data are, therefore, coherent
- The year with greatest number of accidents in the dataset is, by far, 2021, and the year with less accidents is 2016
- The months pertaining to the highest road accident rates are the latest months in year, from September to December
- The highest accident rates take place during the workdays, as Friday topping. The accident numbers decline on the weekend
- The day hour where most of the accidents happen are in the morning from 6 to 8 hours, and, in the evening from 14 to 17 hours
- The states of California and Florida are responsible for nearly half of the total accidents
- The cities of Miami (Florida), Los Angeles (California) and Orlando (Florida) are top-3 most dangerous U.S. cities to drive

In terms of correlation analysis towards severity, we can conclude that the most relevant variables are:

- In positive direction: Humidity (%), Pressure(in), Wind Speed(mph) Weather Condition Clear, Civil Twilight Night, Junction
- In negative direction: Temperature(F), Wind Chill(F), Weather Condition Fair, Weather Condition Cloudy, Civil Twilight Day, Crossing, Station, Traffic Signal

Having fit a linear regression model of the data, the variables that showed strongest explanation power over the Severity are Weather Condition Clear, Junction, Civil Twilight Day, Civil Twilight Night, Weather Condition Fair, Weather Condition Cloudy and Crossing.

Overall response to the research question:

In view of the results achieved, important information can be drawn on the main predictors of accidents in the USA, as well as their impact in terms of severity and consequences for road traffic. Thus, and answering the research question, we can say that the severity of accidents recorded in the USA between 2016 and 2021 is influenced by factors related to the weather situation (mainly cloudiness and wind), road infrastructure (mainly junctions, crossings and traffic signs) and with the phase of day (day or night). In addition, the present work also references the occurrence of accidents by geographic location, identifying the states and cities with the highest accident rates. In this sense, this research will be able to provide important data to the authorities to deepen the diagnosis produced and, consequently, define strategies aimed at solving the identified problems.

7. Improvement opportunities

As further project development ideas, we identify the following:

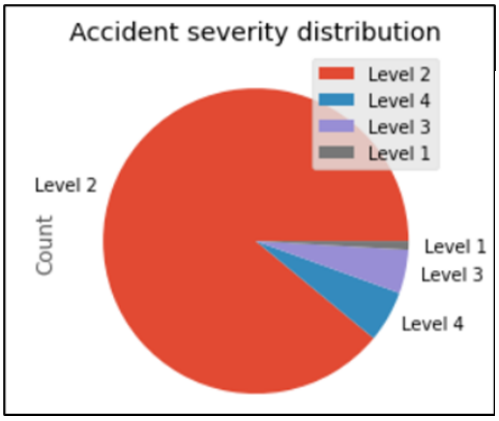
- This work would benefit from accessing other sources of information, providing more characterizing data.
- It would also be useful to divide the dataset into training and test datasets, and later apply more diversified regression models that would allow predicting the occurrence of accidents in real time.
- It will be equally important to try to deepen the study of the relationships between the explanatory variables and the level of severity of accidents, with more explanation by geographic location.

References

[1]: Moosavi, Sobhan. (2019). Correlation between Severity Level and Traffic Delay. Kaggle. Retrieved January 27, 2023, from https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/discussion/152370?search=severity

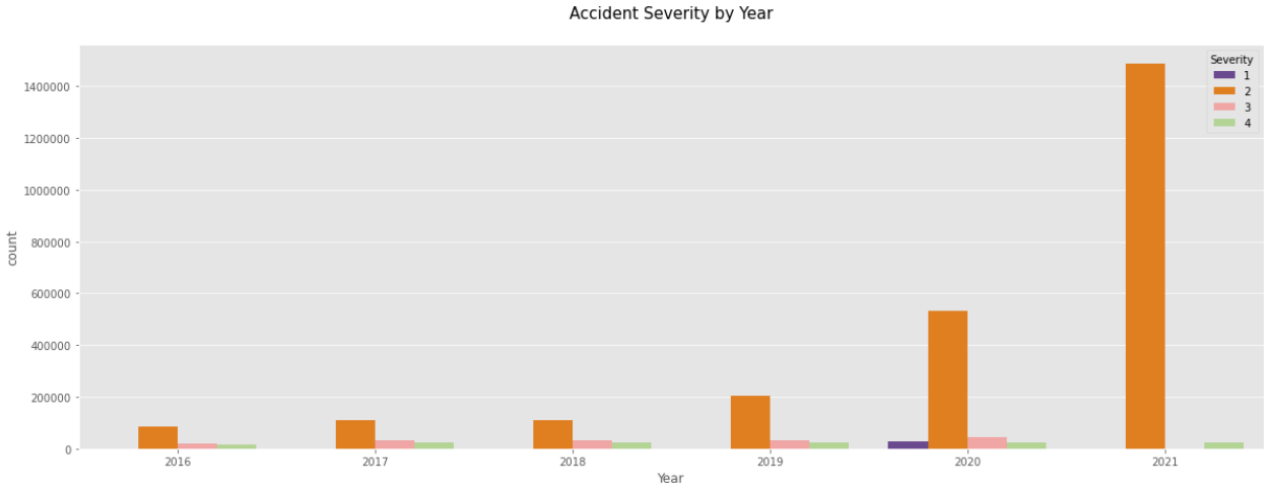
[2]: IBM. (n.d.). About Linear Regression. Netherlands | IBM. Retrieved February 3, 2023, from https://www.ibm.com/nl-en/analytics/learn/linear-regression#:~:text=Linear%20regression%20analysis%20is%20used,is%20called%20the%20independent%20variable

Appendix A



Severity level	Accident Count	Percentage (%)
2	2532991	89.022374
3	155105	5.451190
4	131193	4.610799
1	26053	0.915637

A1. Pie chart showing the severity score distribution over the total amount of accidents recorded in our dataset, including list with percentage.



A2. Bar chart showing the severity score distribution per year.

Weather conditions (11):

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

Time (4):

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

Infrastructures conditions (13):

- Amenity: A Point-Of-Interest (POI) annotation which indicates presence of amenity in a nearby location.
- Bump: A POI annotation which indicates presence of speed bump or hump in a nearby location.
- Crossing: A POI annotation which indicates presence of crossing in a nearby location.
- Give_Way: A POI annotation which indicates presence of give_way sign in a nearby location.
- Junction: A POI annotation which indicates presence of junction in a nearby location.
- No_Exit: A POI annotation which indicates presence of no_exit sign in a nearby location.
- Railway: A POI annotation which indicates presence of railway in a nearby location.
- Roundabout: A POI annotation which indicates presence of roundabout in a nearby location.
- Station: A POI annotation which indicates presence of station (bus, train, etc.) in a nearby location.
- Stop: A POI annotation which indicates presence of stop sign in a nearby location.
- Traffic_Calming: A POI annotation which indicates presence of traffic_calming means in a nearby location.
- Traffic_Signal: A POI annotation which indicates presence of traffic_signal in a nearby location.
- Turning_Loop: A POI annotation which indicates presence of turning_loop in a nearby location.

Geographical distribution

- Number: Shows the street number in address field.
- Street: Shows the street name in address field.
- Side: Shows the relative side of the street (Right/Left) in address field.
- City: Shows the city in address field.
- County: Shows the county in address field.
- State: Shows the state in address field.
- Zipcode: Shows the zipcode in address field.
- Country: Shows the country in address field.
- Timezone: Shows timezone based on the location of the accident (eastern, central, etc.).

The variables that provide characterization about the accident are the following (10):

- ID: This is a unique identifier of the accident record.
- 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 result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
- Start_Time: Shows start time of the accident in local time zone.
- End_Time: Shows end time of the accident in local time zone.
- Start_Lat: Shows latitude in GPS coordinate of the start point.
- Start_Lng: Shows longitude in GPS coordinate of the start point.
- End_Lat: Shows latitude in GPS coordinate of the end point.
- End_Lng: Shows longitude in GPS coordinate of the end point.
- Distance(mi): The length of the road extent affected by the accident.
- Description: Shows natural language description of the accident.

A3: List of all features categorized into 5 groups.

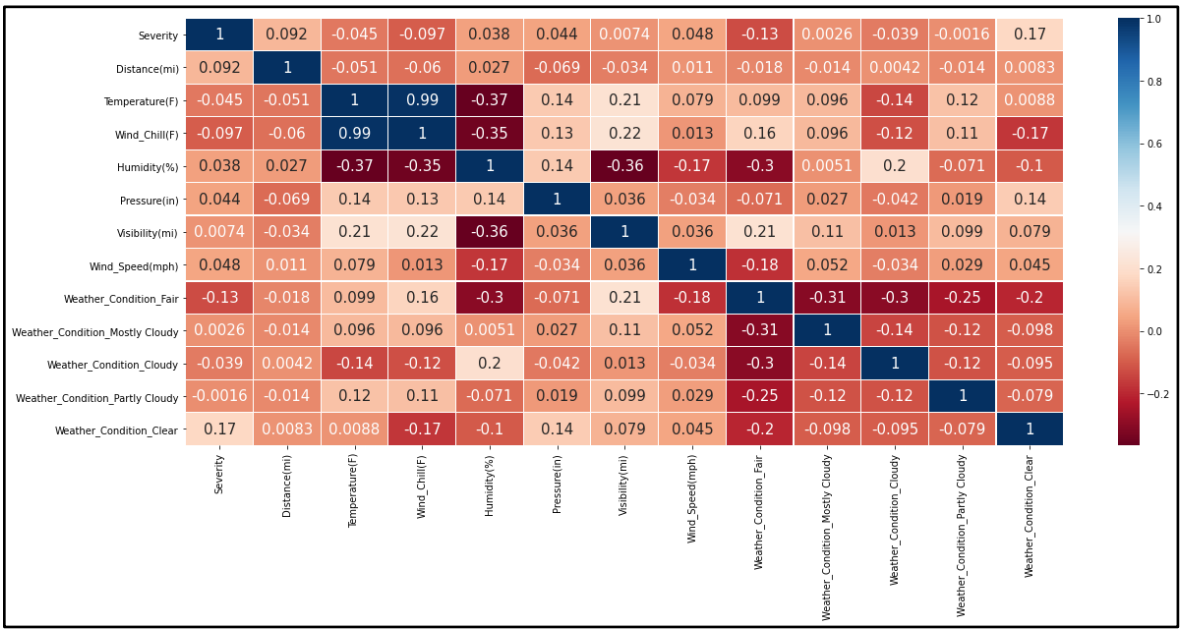
Severity	Delay
1	2 minutes and 30 seconds
2	3 minutes and 15 seconds
3	8 minutes
4	18 minutes

A4: Table showing correlation between severity level and traffic delay [1].

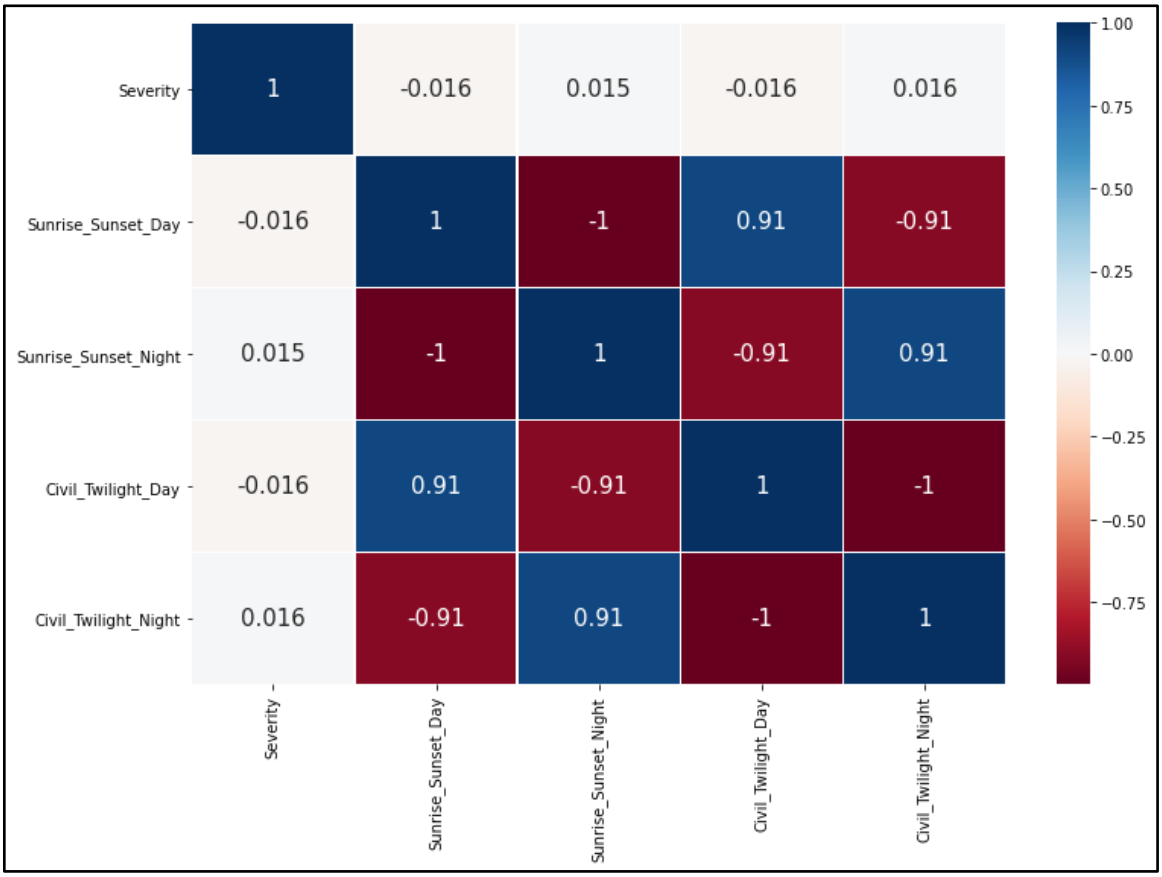
Temperature(F)	69274
Wind_Chill(F)	469643
Humidity(%)	73092
Pressure(in)	59200
Visibility(mi)	70546
Wind_Direction	73775
Wind_Speed(mph)	157944
Precipitation(in)	549458
Weather_Condition	70636

A5: Table of NaN values of the numerical Weather Condition variables

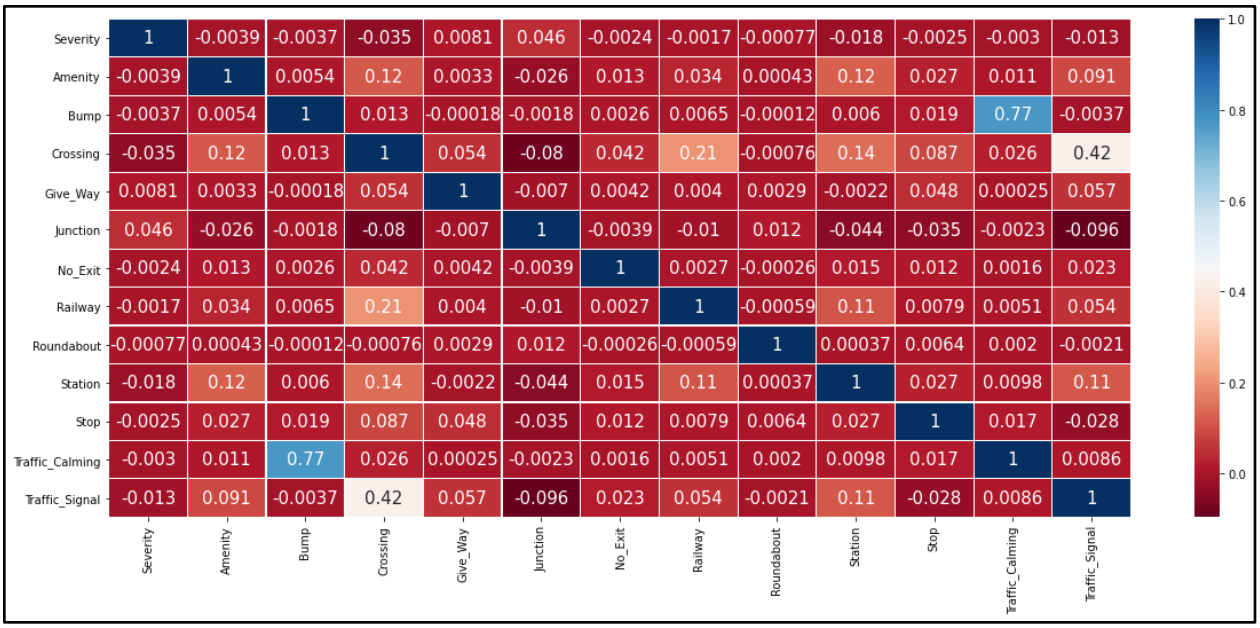
Appendix B



B1: Correlation matrix between independent variables 'Weather Condition' and 'Severity'



B2: Correlation matrix between independent variables 'Phase of the Day' and 'Severity'



B3: Correlation matrix between independent variables 'Infrastructure Condition' and 'Severity' .

Appendix C

```
Weather_Condition_Clear      0.258076
Junction                     0.032912
Traffic_Signal               0.009942
Pressure(in)                 0.008053
Temperature(F)               0.004463
Wind_Speed(mph)              0.001903
Humidity(%)                  0.000499
Wind_Chill(F)               -0.005226
Station                     -0.024783
Crossing                    -0.046829
Weather_Condition_Cloudy    -0.089589
Weather_Condition_Fair      -0.091510
Civil_Twilight_Night        -0.205054
Civil_Twilight_Day          -0.230671
Name: Explainers coefficients, dtype: float64
```

C1: Coefficients obtained on the linear regression model