Summary:

This report represents the descriptive and predictive statistical analysis on dataset that is a twelve-month period of accidents in New York from 2019-2020. We focused on key variables, including ‘time and day of week’, ‘city’, ‘temperature’, ‘POI’, ‘season’, and ‘severity’ of accidents.

1. Introduction:

1.5 million deaths occurred in children and adolescents age group (aged 5-19 years) globally in 2019, nearly all from preventable causes. To better focus the attention of the global community on improving survival of children and adolescents and to guide effective policy and programs, sound and timely cause of death data are crucial, but often scarce [1].

Road accident which involves preventable factors are the cause of approximately 1.3 million deaths each year [2]. Reducing traffic accidents is an important public safety challenge, therefore, accident analysis and prediction has been a topic of much research over the past few decades [3].

In this report, we employed a US-accidents dataset with the aim of predicting rare accident events and minimising the risk of accident, the ninth leading cause of death and disability globally. The dataset involves information on the location, weather conditions, nearby points-of-interest, and the time length of each accident.

Write about some findings …

2. Data description:

We removed some attributes ‘number’ (i.e., the street number of the address of the incident) from the dataset, which had the highest number of NA (i.e. missing) values (27692 missing values) and it is not relevant to our analysis.

2.1 Data pre-processing

We removed the features which were not relevant to our analysis. These are including ‘Turning_Loop’, ‘Nautical_Twilight’, ‘Civil_Twilight’, ‘number’, ‘Airport_Code’, ‘Timezone’, ‘Country’, ‘State’, ‘, ‘, ‘Astronomical_Twilight’.

We used degree Celsius, millimeters, and kilometers instead of Fahrenheit degree, inches, and miles.

3. Descriptives

We conducted an analysis to determine if there is a dependency on the impact of accident and the severity of the traffic. As depicted in Fig. 1, 8% of the accidents caused a significant impact on the traffic by causing long delays (very high severity), 12% of the accidents caused major delays (high severity), 78% of the accidents caused medium-long delays in traffic (medium severity). In conclusion, we investigate if the accident with the lowest severity has the least impact on traffic.

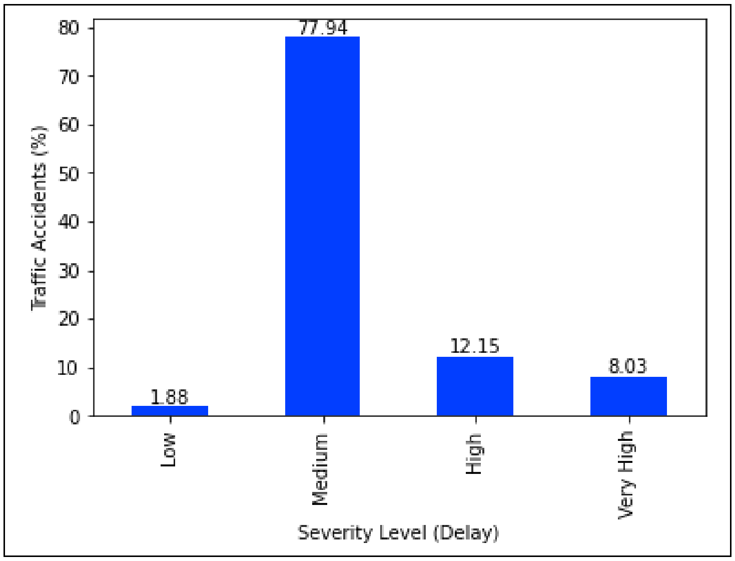


Fig. 1: Word-cloud of 'Description' attribute

Fig. 2 compares the distributions of traffic accident severity across two different years, 2019 and 2020, using a violin plot. We employed a Wilcoxon-Mann-Whitney U test, a non-parametric test to compare the distributions of two independent groups (i.e., both two years). The test examines if the medians of two variables (two-year categories) are equal or not. This test can be used for non-normally distributed data. Unlike t-test, U-test works with the ranks of the observations rather than the actual observations. We used stats.mannwhitneyu() from scipy to find U and p-value. With U = 136937762 and p-value of 0.001, we conclude that the median of the attribute Severity in 2019 and the median of the attribute Severity in 2020 are statistically significantly different. As depicted below, the year 2019 has more proportion of traffic accidents with a rating of 3 or 4 in comparison with the year 2020, while having no traffic accidents with a severity rating of 1.

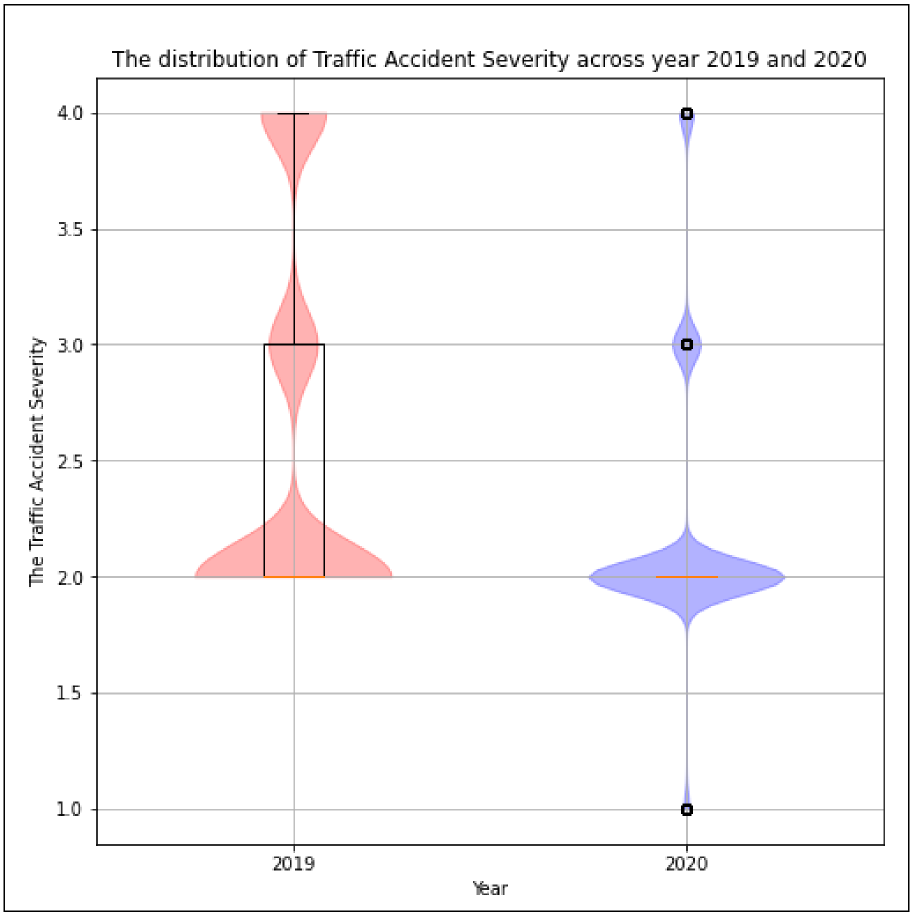
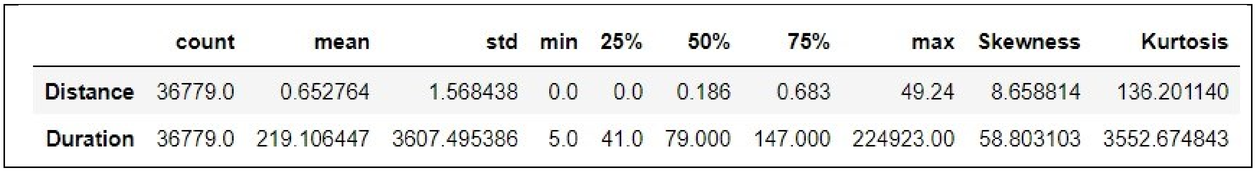


Fig. 2: Distribution of traffic accident severity across the year 2019 and 2020

We also conducted a statistical analysis on the distance attribute to determine the relationship between the length of road and the occurrence of the accident. As listed in Table 1, for a wide range of values (0km – 49.24km), the mean and the median values are 0.65km and 0.19km, respectively. The Interquartile Range (IQR) is Q3 – Q1 = 75% - 25% = 0.68 – 0 = 0.68, that is because of the high occurrence of accident events where the distance is close to zero.

Table 1: Statistics of the Distance attribute



	count	mean	std	min	25%	50%	75%	max	Skewness	Kurtosis
Distance	36779.0	0.652764	1.568438	0.0	0.0	0.186	0.683	49.24	8.658814	136.201140
Duration	36779.0	219.106447	3607.495386	5.0	41.0	79.000	147.000	224923.00	58.803103	3552.674843

Fig. 3*Fig. 3* visualises the distribution of the distance attribute using a box plot. We renamed ‘Distance.mi.’ as to ‘Distance’ attribute. As shown by the box plot, the distribution of distance is positively skewed left with a long right-hand tail. The skewness value was 8.66. The strip charts shows that approximately 12% of data are between 1 and 2 km.

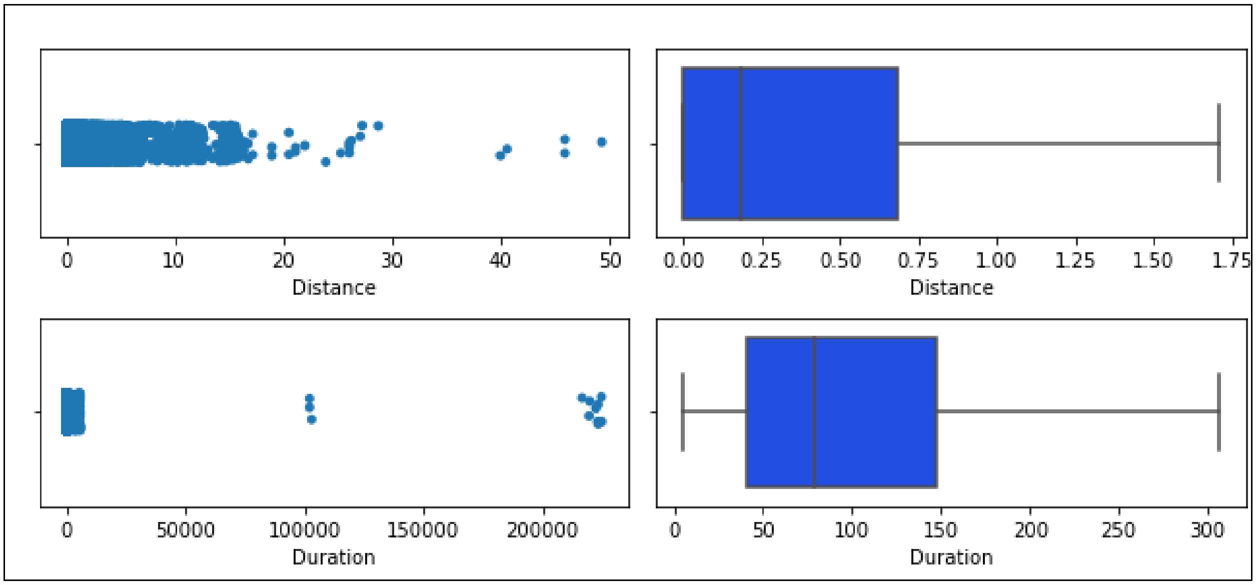


Fig. 3: Strip chart and Box plot of distance and duration attributes

A link was found to exist between Distance and incident Duration as discovered in Fig. 3 on average an incident Duration will be 50 minutes longer for an accident that falls into the first quantile (Q1, 25%) for the Distance attribute. The Fig. 4 indicates that approximately 76% of all observations for the attribute distance have a value of less than 1 km.

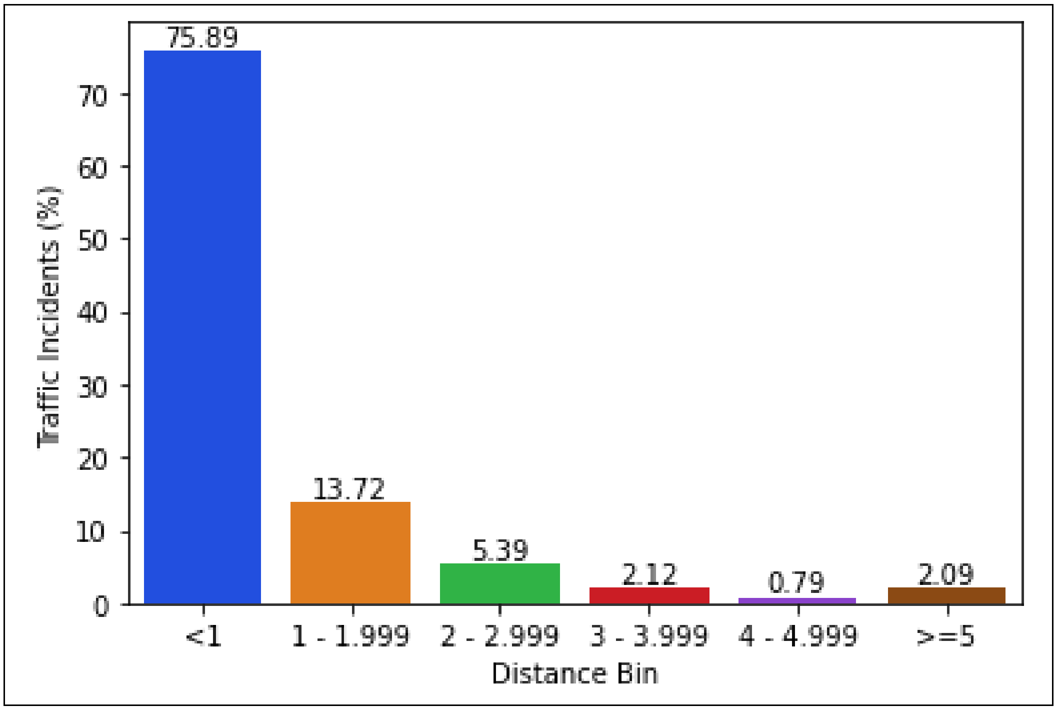
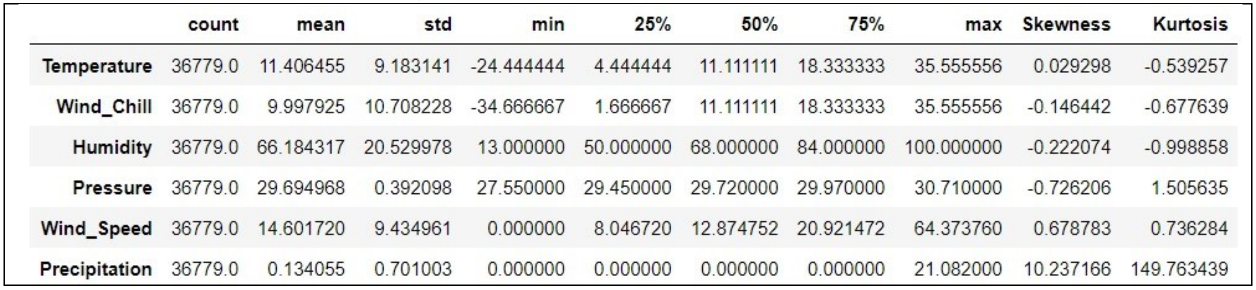


Fig. 4: The percentage of Traffic incidents across the distances

Moosavi et al. suggested to include these weather associated attributes for description of each accident event [4]. Weather attributes such as temperature, winds, and precipitation affect driver capabilities and vehicle performance, which contributes to traffic congestion and accident [5]. Thus, we also conducted a descriptive analysis on the statistics of weather attributes and listed them in Table 2. To do so, first we renamed Temperature.F. to Temperature, Humidity... to Humidity, Pressure.in. to Pressure, Wind_Speed.mph. to Wind_Speed, Precipitation.in. to Precipitation, Wind_Chill.F. to Wind_Chill.

Table 2: Descriptive statistics of weather associated attributes



	count	mean	std	min	25%	50%	75%	max	Skewness	Kurtosis
Temperature	36779.0	11.406455	9.183141	-24.444444	4.444444	11.111111	18.333333	35.555556	0.029298	-0.539257
Wind_Chill	36779.0	9.997925	10.708228	-34.666667	1.666667	11.111111	18.333333	35.555556	-0.146442	-0.677639
Humidity	36779.0	66.184317	20.529978	13.000000	50.000000	68.000000	84.000000	100.000000	-0.222074	-0.998858
Pressure	36779.0	29.694968	0.392098	27.550000	29.450000	29.720000	29.970000	30.710000	-0.726206	1.505635
Wind_Speed	36779.0	14.601720	9.434961	0.000000	8.046720	12.874752	20.921472	64.373760	0.678783	0.736284
Precipitation	36779.0	0.134055	0.701003	0.000000	0.000000	0.000000	0.000000	21.082000	10.237166	149.763439

As listed by the Table 2, Temperature attribute has mean of 11.41, median of 11.11, skewness of around 0.03 (close to zero). These characteristics reveal a normal distribution. Wind_Speed attribute has the mean of 14.6 and median of 12.87. These characteristics reveal that it has a distribution with a positive skewness of around 0.68. As represented by Fig. 5, the attribute Temperature shows a normal distribution. Most traffic incidents occurred within temperatures of 0°C to 9°C, humidity of 80 to 94, pressure of 29.5 to 30, and wind speeds of 10-19 km/h. However, the lowest traffic incidents occurred at temperatures of 30°C and above, humidity of 0 to 20, pressure of 27.5 to 28.5, wind speeds of 30 km/h and above.

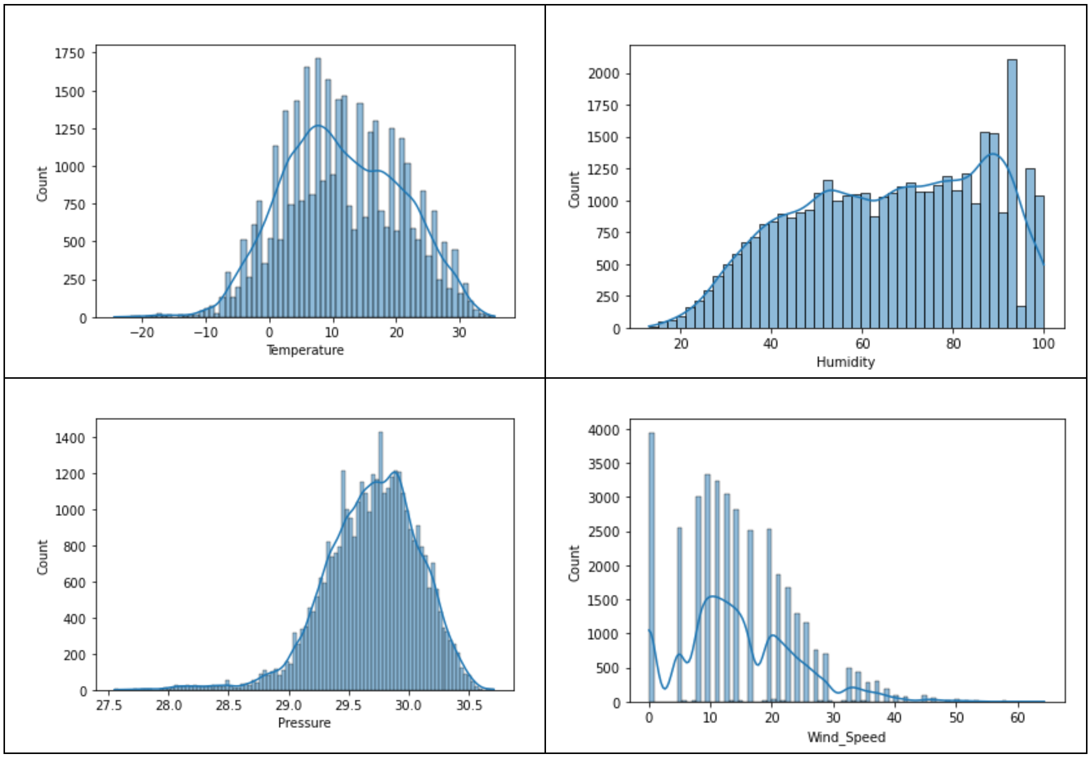


Fig. 5: Data distribution of temperature, humidity, pressure, and wind speed

We also analysed the number of accidents by weather conditions, year, season, month, weekday, time of day, hour, and location:

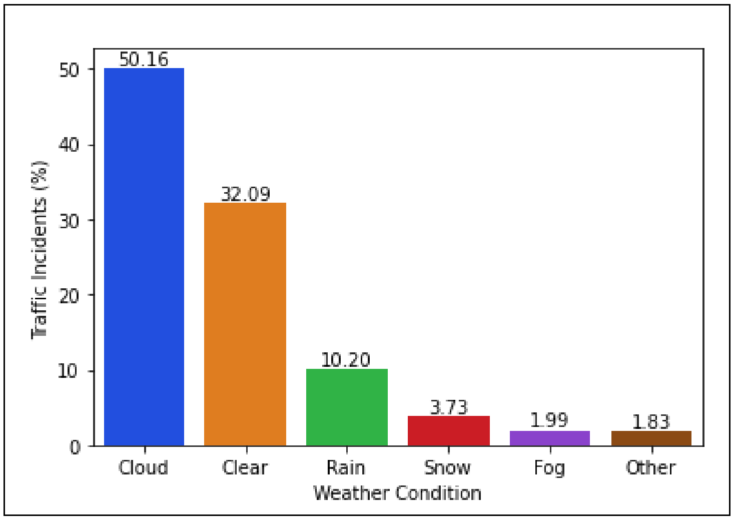
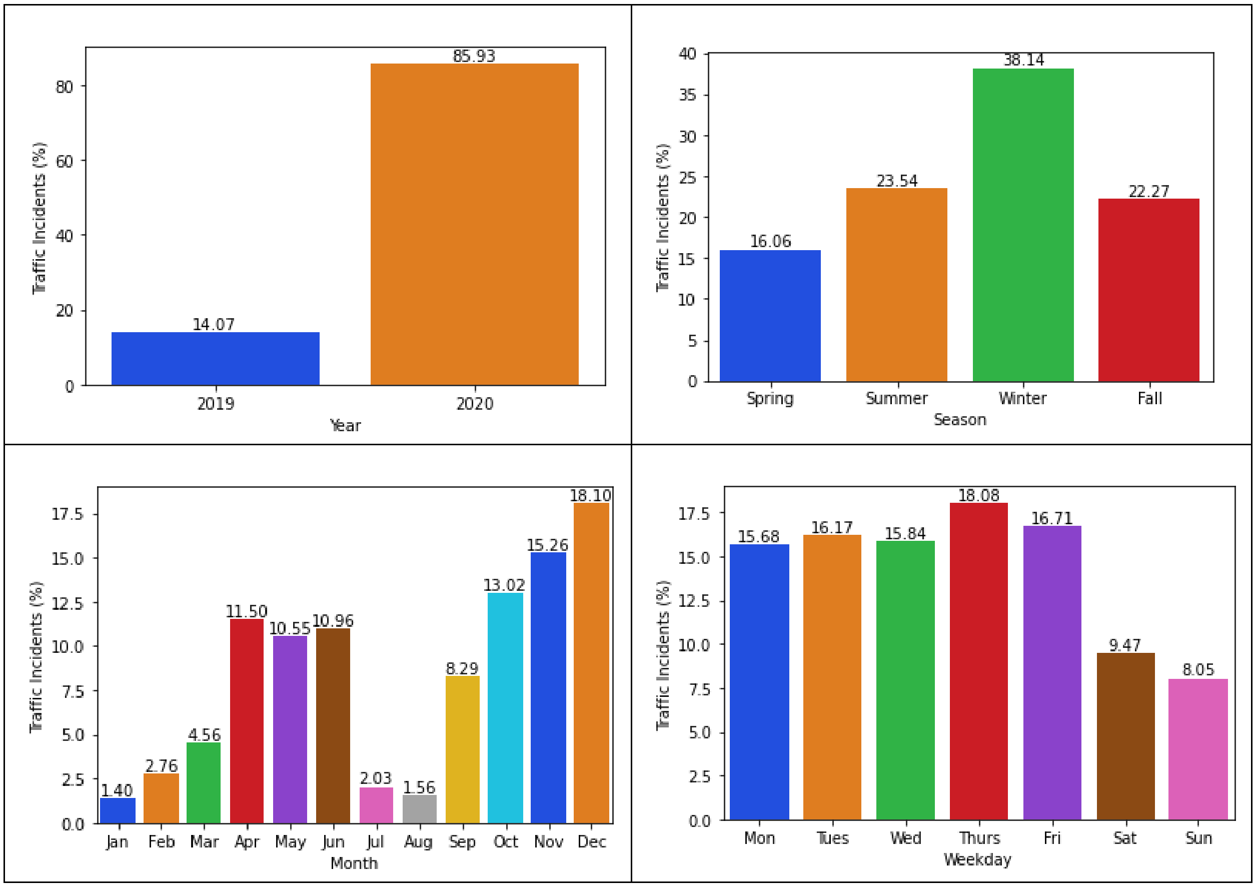


Fig. 6: The percentage of accidents across the weather conditions

As shown by Fig. 6, the cloudy weather contributes to the highest percentage of accidents (around 50%), followed by clear weather (32%), rainy weather (10%), snowing weather (around 4%), foggy weather (around 2%) and other conditions (less than 2%).



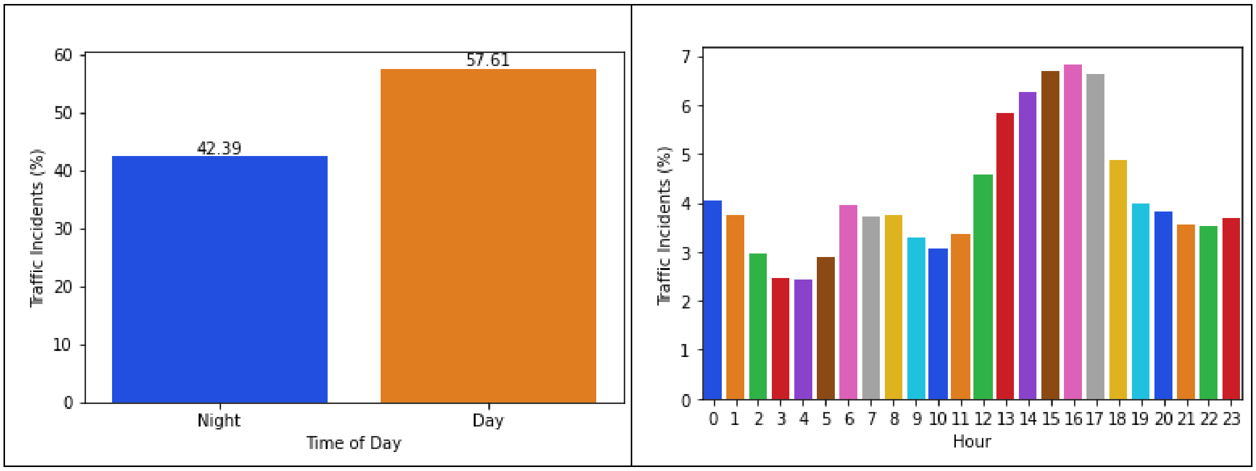


Fig. 7: The percentage of traffic incidents across the year, season, month, weekday, time of day (day/night), and hour

As depicted by Fig. 7, the percentage of traffic incidents across the year 2020 is six times higher than those across the year 2019. This is sensible, because according to the Hedges Company, there were 286.9 million registered cars in the US in 2020. That's 0.84% more than 2019's 284.5 million units [6]. With the increasing number of registered cars every year, it is more vulnerable to observe more traffic incidents. The most percentage of accidents happened in Winter season, this is supported by the distribution of incident percentages across the different months in the USA. December is when the highest percentage of accidents occurred (18%). Because on that month, and that season, the weather conditions (fog, ice, and snow) affect the visibility. Thus, during the winter season (Oct, Nov, and December) the number of traffic incidents is higher than during other seasons. During the weekdays when people are in hurry to commute for their job, the number of accidents is higher than during the weekends. From the time-of-day attribute, it is obvious that most accidents occurred during the daytime than the night-time, which is supported by the distribution of accidents across different hours. As is evident in Fig. 7, most of accidents happened within the time 15-17 (pm time), which is sensible, because most people finish their job and may drive to their home, and the traffic congestion is higher than the other time of day.

Next, we investigate the distributions of number of incident events across the different points of interests (POIs) and top five cities:

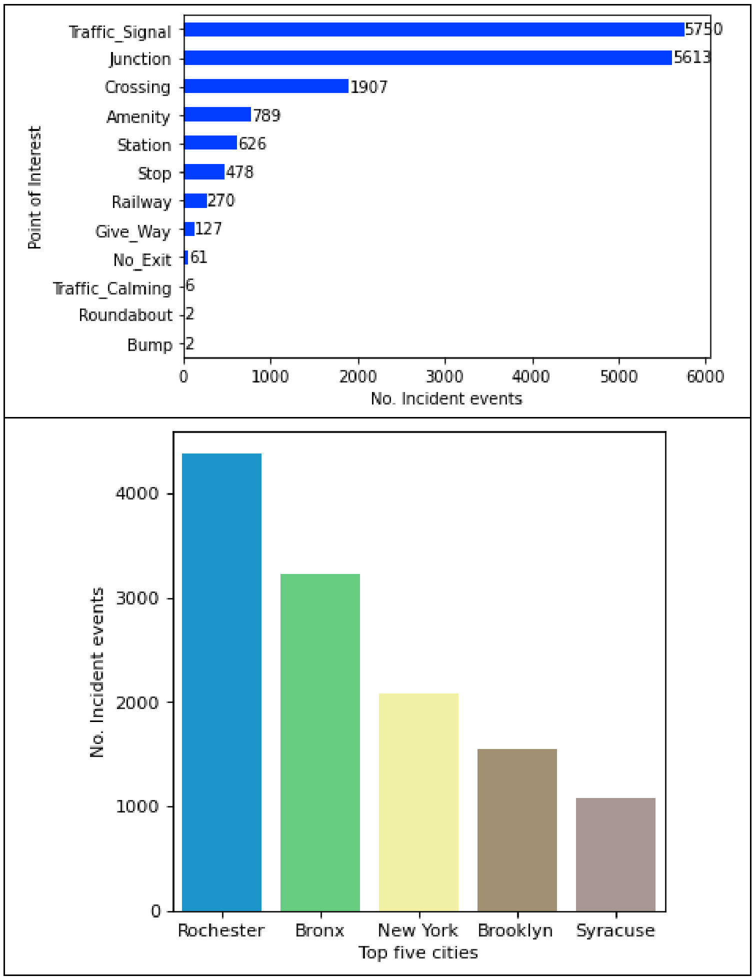


Fig. 8: Distribution of accident cases across different POIs with top five cities

As shown by Fig. 8, Traffic signal has the highest number of incident observations (5750), followed by the junction (5613 observations) and crossing (1907 observations). Other POIs are showing less than 1000 observations. It is also visible that just 5 out of 828 cities are responsible for around 33% accidental events.

We also provided with a correlation matrix, which reveals a comparative level of association between attributes. In particular, we employed a heat map to depict the relationships between different observations. As shown by Fig. 9, Temperature and Wind Chill, Severity and Start Time of Day, Traffic Calming Device and Speed Bump have been shown with a strong correlation.

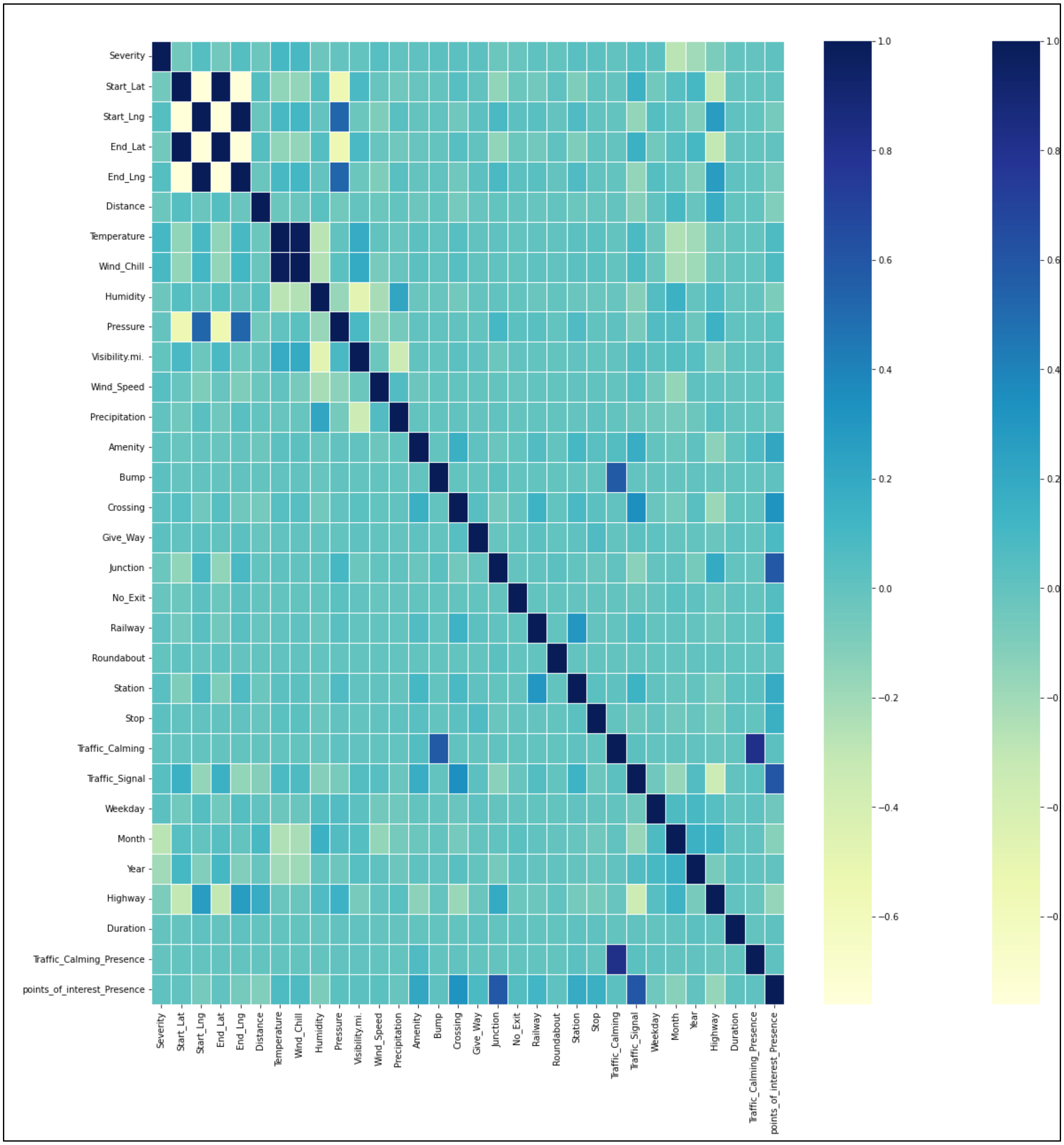


Fig. 9: Correlation Matrix based on the heat map

In conclusion, our descriptive analysis suggests people to plan to use their private vehicle during the safe time; for example, in daytime between 9 a.m. to 11 a.m. or in night-time after 7 p.m. Also, it suggests people to plan to drive in spring and summer, while being careful about the POIs with the highest traffic incidents (i.e., traffic signals, junctions, and crossings). When the visibility is better, and ice and snow are less likely to be on the road. It also suggests people to avoid driving at night when hazards rapidly multiply. If travelling long distances specially in winter or autumn, during the weekdays, between 1 p.m. and 5 p.m., make sure you are well rested and plan where to have a break. Share the driving if possible or allow for stops every two hours or use public transport.

Appendix A:

Put Python code here…

References:

1. Liu, L., et al., *National, regional, and global causes of mortality in 5–19-year-olds from 2000 to 2019: a systematic analysis.* The Lancet Global Health, 2022. **10**(3): p. e337-e347.
2. Organization, W.H., *World health statistics 2022: monitoring health for the SDGs, sustainable development goals.* 2022.
3. Moosavi, S., et al. *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*. 2019.
4. Moosavi, S., et al., *A countrywide traffic accident dataset.* arXiv preprint arXiv:1906.05409, 2019.
5. Elkhazindar, A., M. Hafez, and K. Ksaibati, *Incorporating Pavement Friction Management into Pavement Asset Management Systems: State Department of Transportation Experience.* CivilEng, 2022. **3**(2): p. 541-561.
6. GoodCarBadCar, *Light vehicle sales in the United States between January and December of 2019 and 2020*. 2021: by manufacturer. Retrieved from Statista.