Trend Analysis of motor vehicle accidents involving fatalities and serious injuries by cause of accident and time.*

Aggressive driving is the leading cause of fatal accidents and even though alcohol-related or speeding-related collisions may have higher fatality rates.

Shivik Arora

01 May 2022

Abstract

In this paper I analyze the factors increasing the chances of motor vehicle accidents being fatal. Approximately 13,300 people died or sustained serious injuries in Canada in 2010 due to motor vehicle collision related incidents. Are most of these collisions due to driving under influence? Or nighttime visibility? Are the number of traffic accident fatalities increasing or are cars becoming safer? Analyzing a dataset picked from the Open Data Toronto on traffic collisions involving fatalities or seriously injured persons from 2006-2020. We find that aggravating factors for such incidents like speeding, aggressive driving or driving under influence do not have a significantly different fatality rate. Aggressive driving is the leading cause for accidents involving killed or seriously injured persons.

1 Introduction

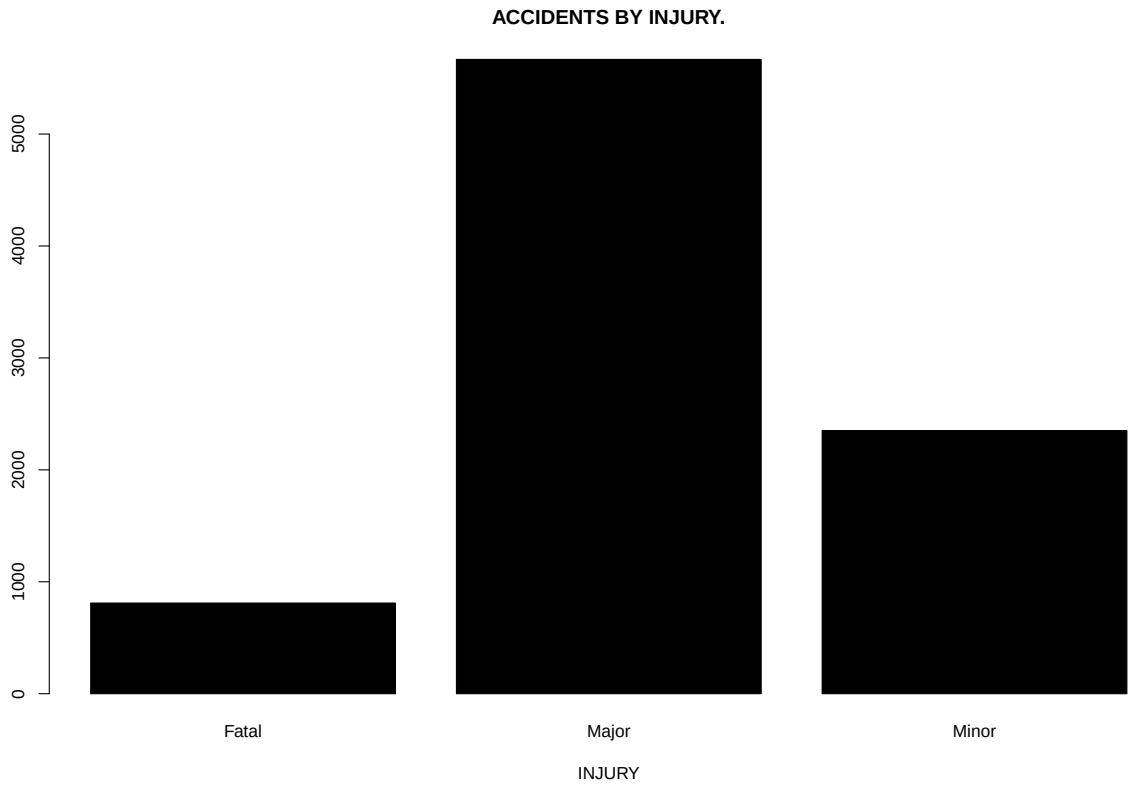
The most valuable gift you have is being alive. If avoidable in any circumstance, even one death a year is too many. Deaths due to traffic collisions are usually completely avoidable. Other than the rare cases of mechanical failures, collisions are almost always the result of human error. If we understand what factors increase the chance of motor vehicle accidents, we can specifically try to put measures in place to reduce them. It is true that mistakes do happen. One can lose control of the car on a turn or something else unexpected can happen, but a large proportion of collisions are due to mistakes in the form of aggressive driving, over speeding or driving under influence. Sometimes these mistakes can be aggravated due to infrastructural errors such as steeper than necessary turns or malfunctioning signals. My aim in this paper is to back these intuitions with data and develop them further. We analyze data from the Toronto Police Service available at https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/ about motor vehicle collisions involving killed or seriously injured persons. On analyzing our dataset after cleaning it, we find that the leading cause for accidents involving killed or seriously injured person is aggressive driving. Factors such as light and visibility do not play a significant role in determining the severity of injuries caused in an accident.

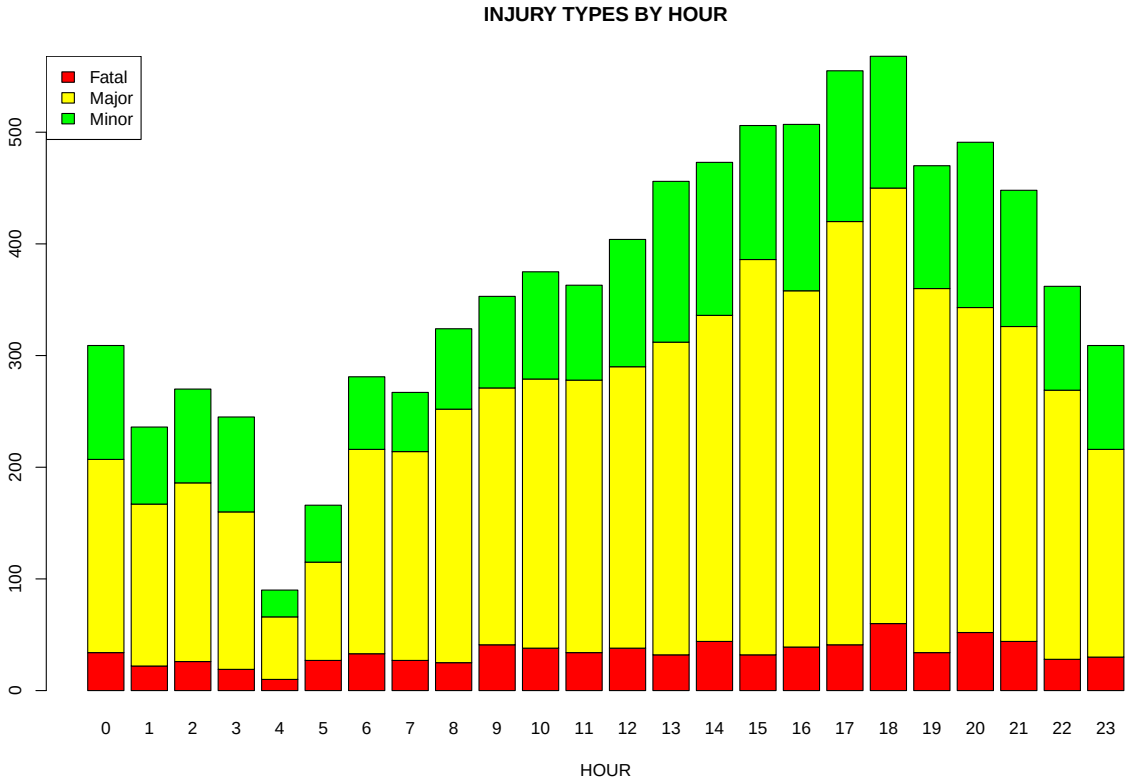
2 Data

The dataset I chose for analysis in this project is from the City of Toronto, the 'Motor Vehicle Collisions involving Killed or Seriously Injured Persons'. It is available on https://open.toronto.ca/dataset/motor-

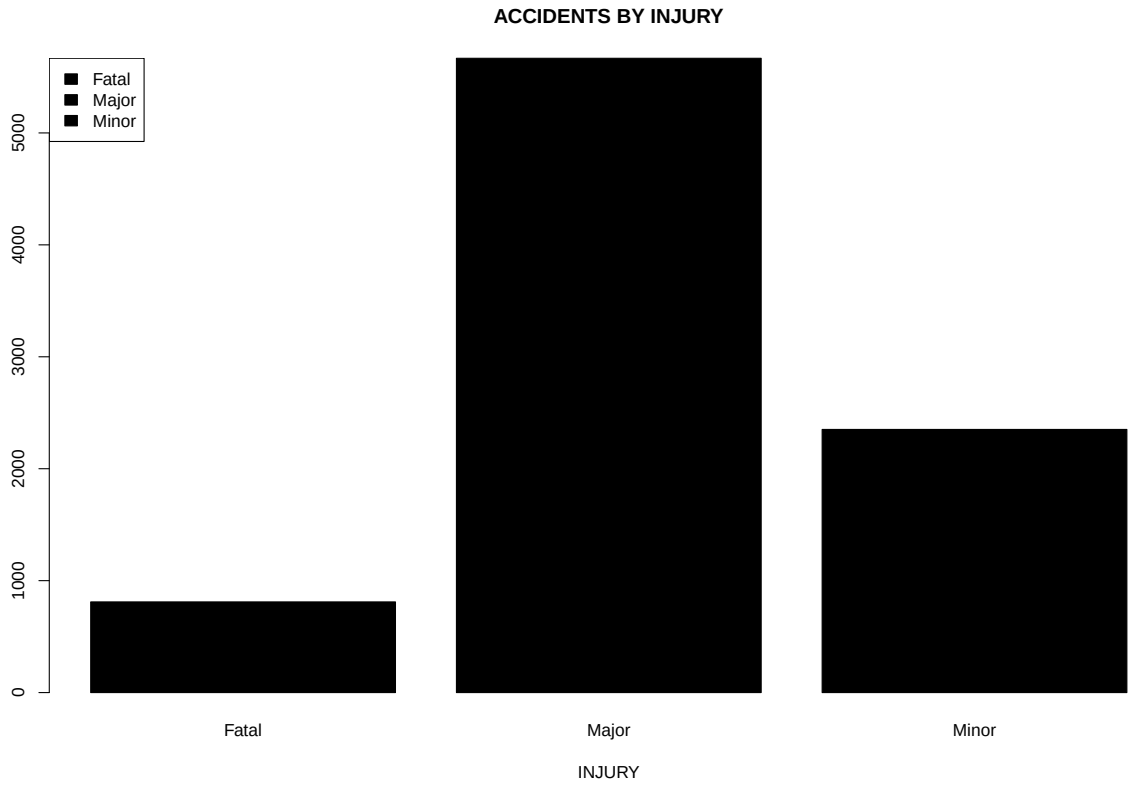
*https://github.com/shivikarora/-304-Traffic

vehicle-collisions-involving-killed-or-seriously-injured-persons/. In the original dataset, we had 16,860 observations of accidents from 2006 to 2020 which involved killed or seriously injured persons with data on location, year, date, time, hour, streets, road classification, visibility and environment condition, light condition, road surface condition, injury severity, condition and actions of persons involved, and the cause which could be speeding related, alcohol related, aggressive driving related, disability related, neighborhood. In total there were 57 data points for each collision. I did not need all these variables and was only interested in the main causes of the accidents. I decided to drop the variables on location since the locations were offset to the nearest node to protect the privacy of the involved individuals. I also dropped date and time since the exact date and time are not of importance. The 'hour' variable gives us enough information in respect to these. As we would expect, most accidents occur in peak traffic hours from 1600 to 1800 hours. However, it seems like the number of fatal accidents are almost uniformly distributed across all hours.



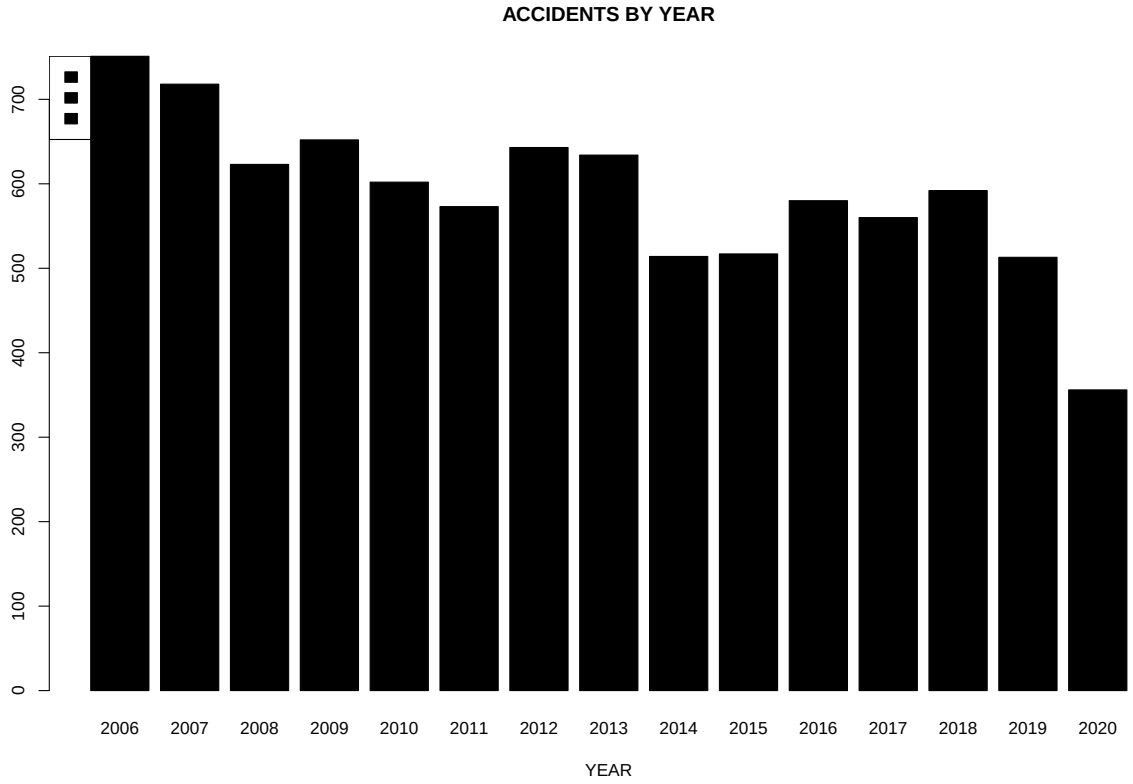


Another main variable of interest is injury type. This told us the severity of the injuries to the persons involved. Entries which did not have datapoints on this were removed. In our cleaned data, we classified any injury which was marked minimal or minor, as minor. This left us with three types of injury: fatal, major, and minor. Of the total accidents, 9%, 65%, and 26% were fatal, major, and minor respectively as shown in Figure 2.



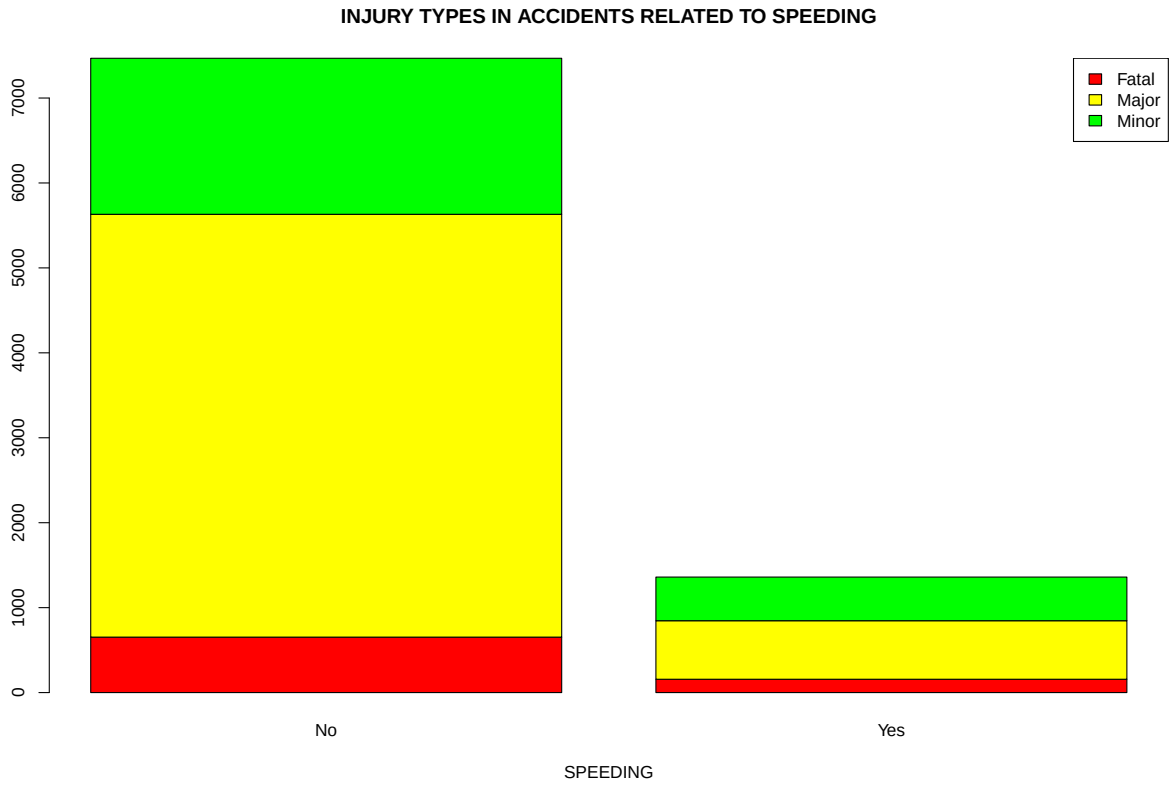
In this paper, we will see if these figures are the same for accidents which are caused due to alcohol related incidents, or speeding related incidents, or aggressive driving incidents and try to see which is the most likely to result in fatalities, or if they make no difference and have no significant comparative difference. Another thing we may be interested in seeing is the trends of accidents across the years as shown in Figure 3 below. Accidents in our dataset do have a declining trend over the years. This could be attributed to the advancements in car safety technologies, such as improved tire qualities, airbags, lane control systems and ABS systems. Throughout the report, I used R to conduct my analysis (R Core Team (2020)). R packages tidyverse (Hadley Wickham [aut 2021), knitr (Yihui Xie ORCID iD [aut 2021), reshape2 (Wickham 2020), janitor (Sam Firke [aut 2021), and kableExtra (Hao Zhu ORCID iD [aut 2021) are used for data cleaning, analysis, and discussion.

The dataset will be downloaded, processed, and analyzed in `R` (R Core Team 2020) primarily using the `dplyr` (Wickham et al. 2021) and `Tidyverse` (Wickham et al. 2019) packages. The packages `knitr` (Xie 2014) are used to generate the R markdown report.

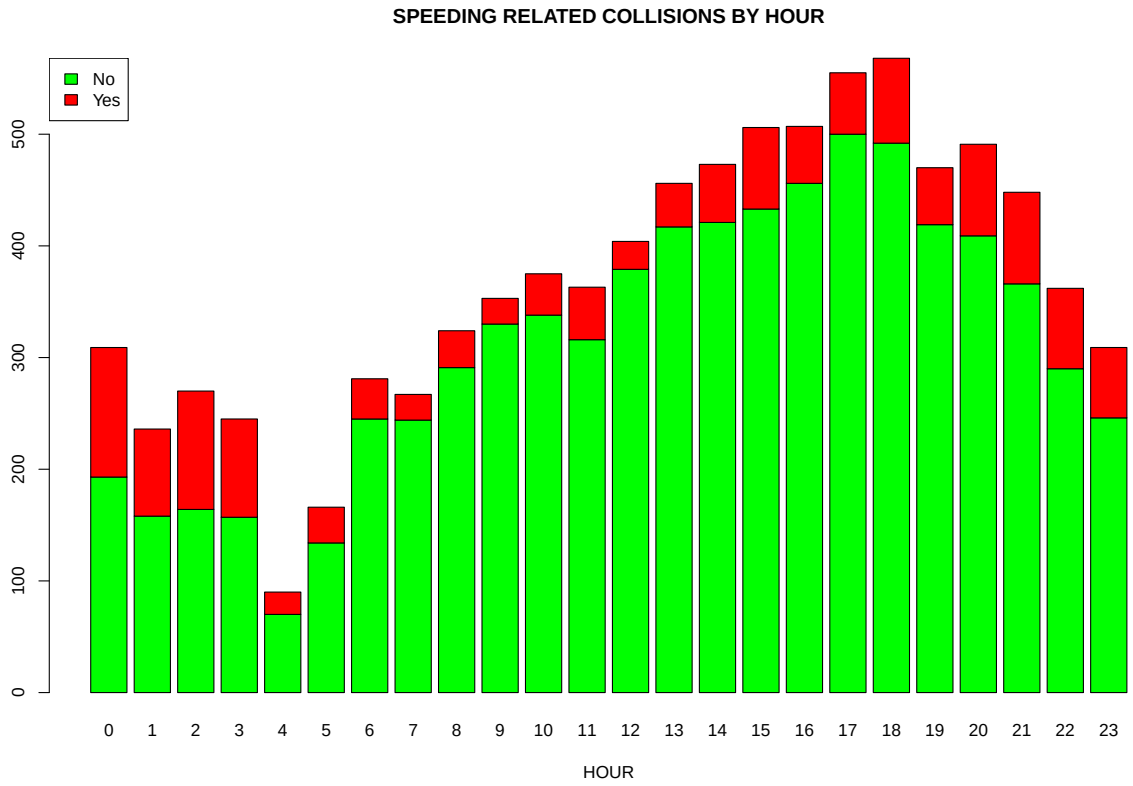


3 Analysis and Results

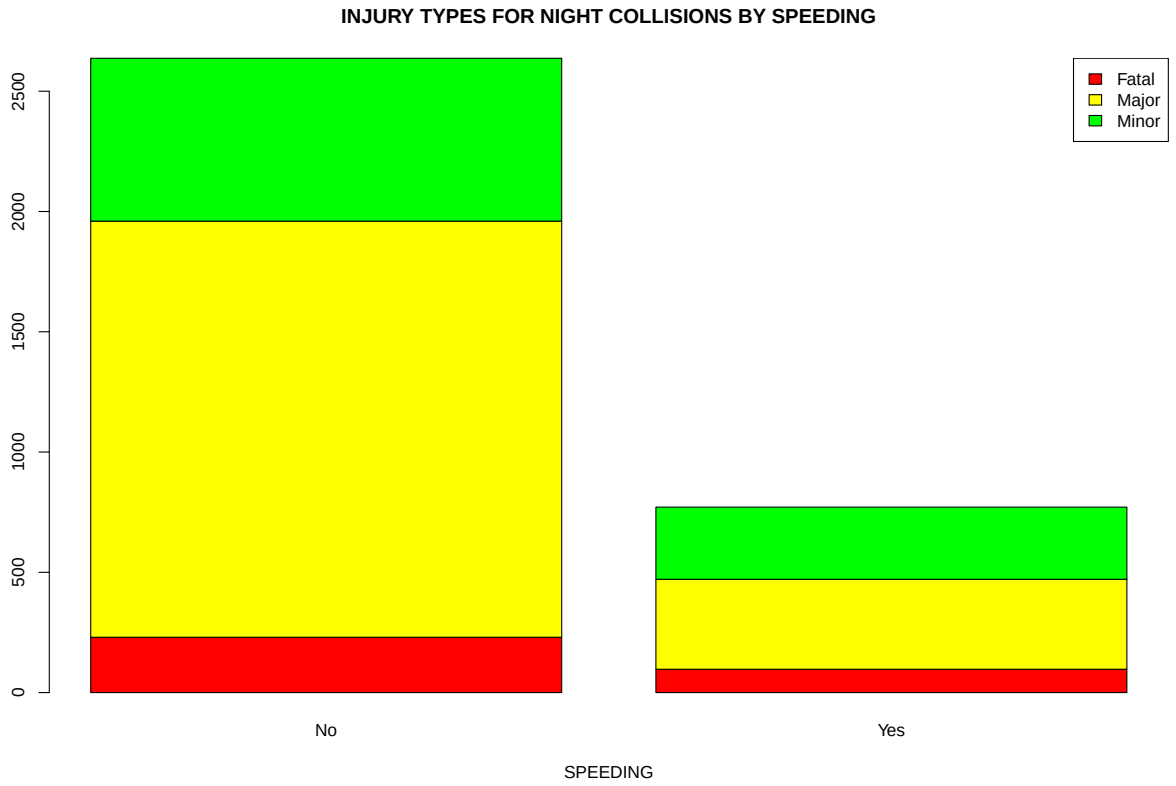
The three main causes of accidents were as we would intuitively expect aggressive driving, speeding, and alcohol related. We wanted to see if the rates of fatality in incidents due to these causes have a significant impact on fatality rates. In Figure 4, we can see the injury types of involved persons in speeding related accidents and non-speeding related accidents. In the speeding related accidents, the percentage of fatal, major, and minor injuries are 11.54%, 50.66% and 37.8% respectively whereas in non-speeding related incidents, these figures stand at 8.7%, 66.65% and 25.65% respectively. There seem to be less fatalities and minor injuries in non-speeding related incidents. Unintuitively, there's a higher proportion of accidents which led to major injuries.



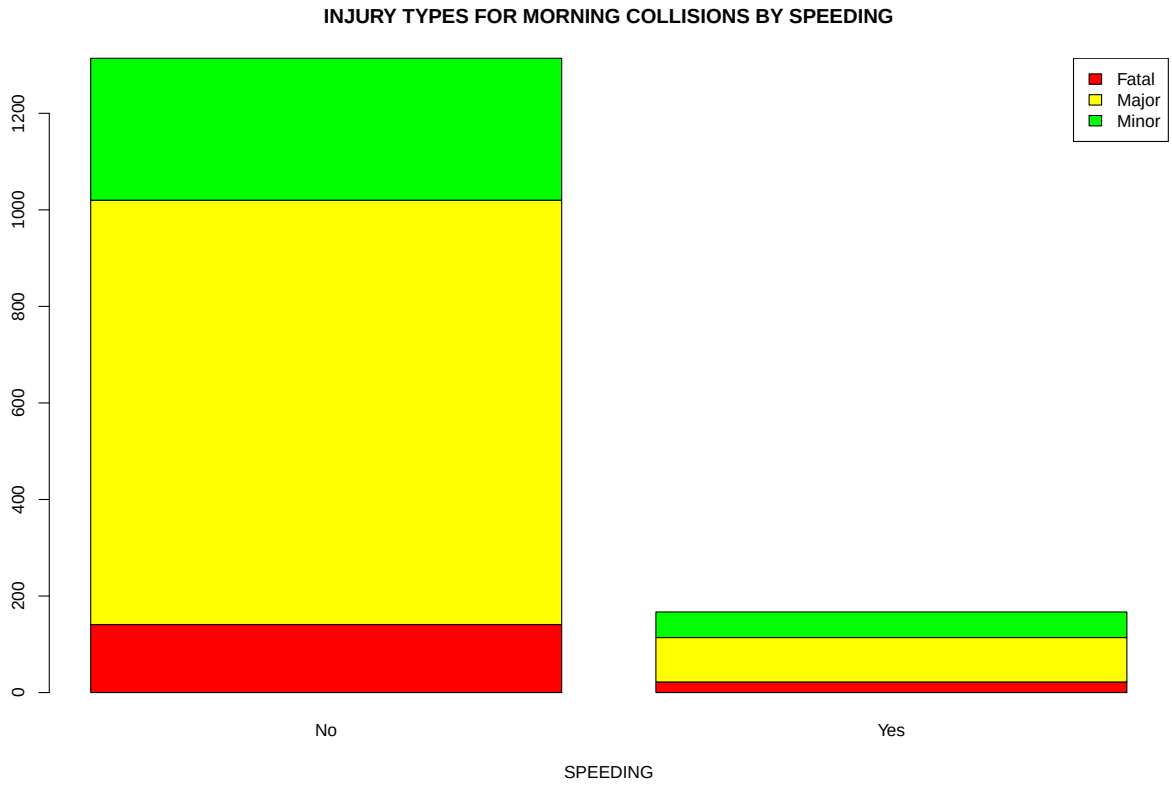
To further analyze this, as we saw above that accidents and injury types do differ by the time of day, I decide to explore what happens when we compare speeding and non-speeding accidents in different periods of the day. For this we divide the day into morning, afternoon/evening, and night. The respective times allotted were from 0400 – 1200 hours, 1200-2000hours, and 2000-0400 hours.



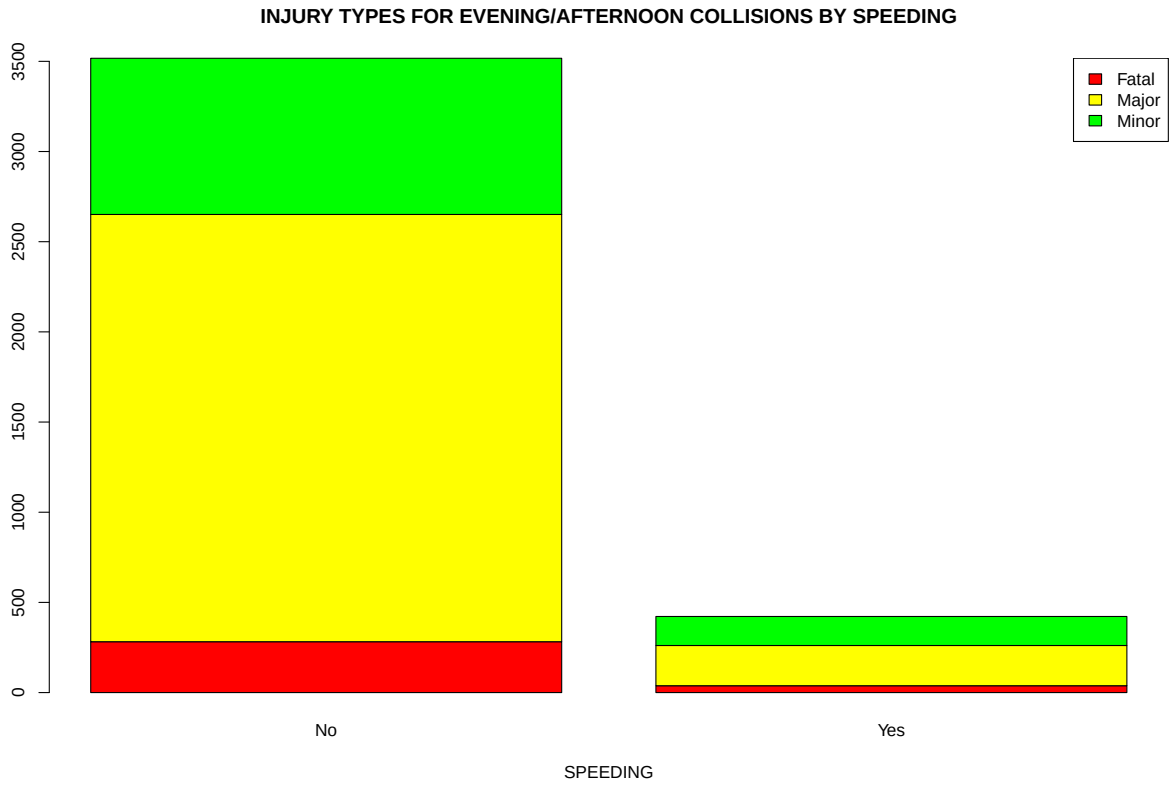
We can see that the highest number and proportion of speeding related accidents occur from 2000-0400 hours. This is what we would expect as there is minimal traffic during these hours and less patrol. If we look at injury types of the driving incidents during night-time which are speeding related and non-speeding related, we get the data shown in Figure 6.



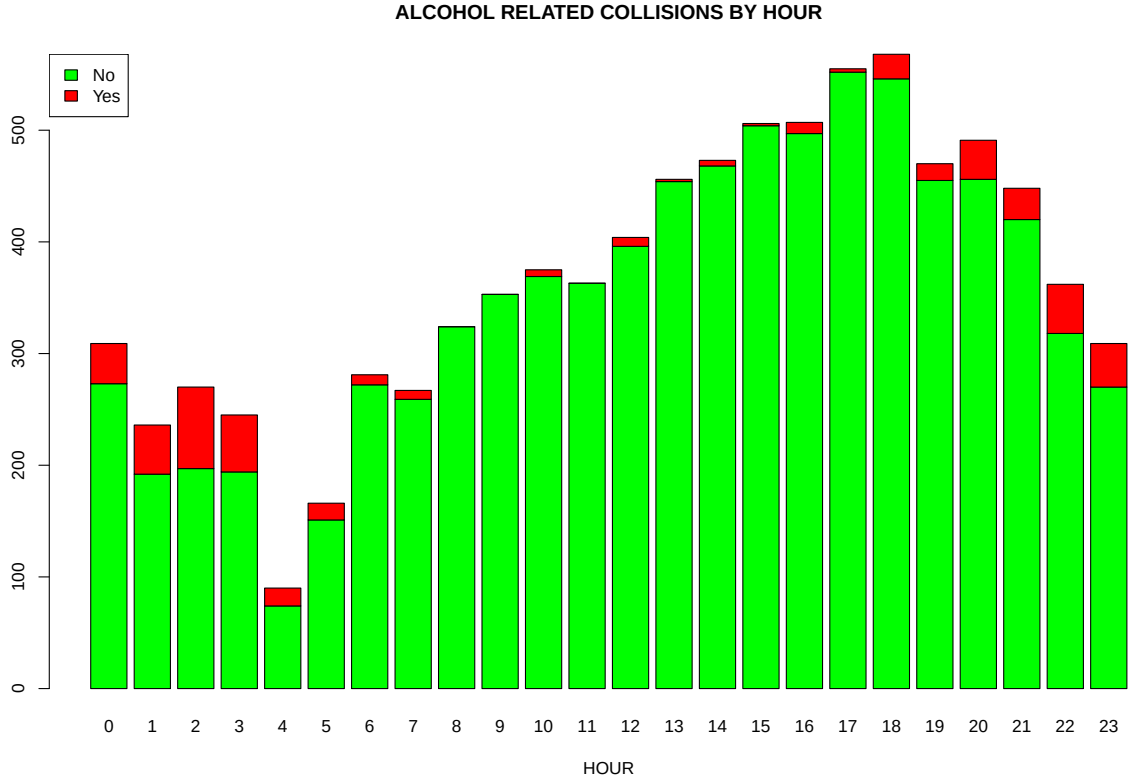
Here, we see that in night-time speeding incidents, 12.6% are fatal, 48.5% cause serious injuries, and 38.9% cause minor injuries. On the other hand, for night-time non-speeding related incidents these rates stand at 8.7%, 65% and 26.3% respectively.



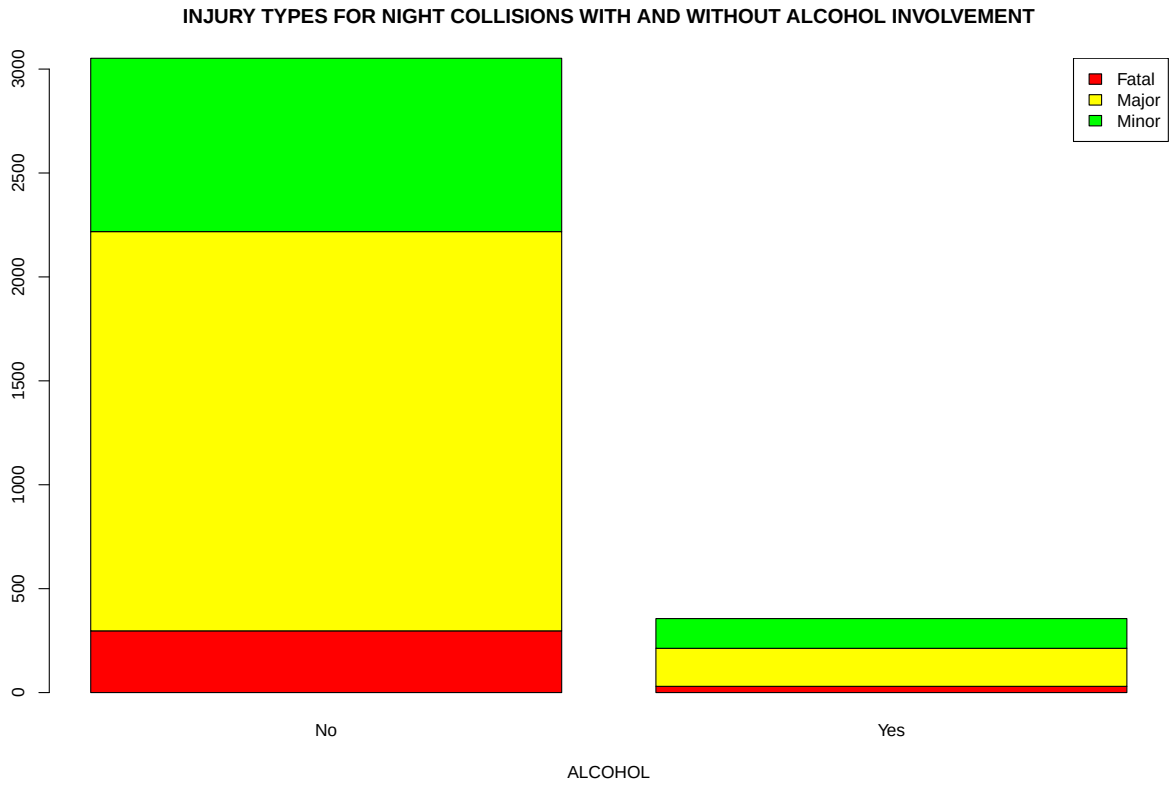
We see a similar trend in the morning with the rates standing at 13%, 55% and 32% for speeding incidents and 10.7%, 66% and 22% respectively. In the evening, a similar trend is seen with the rates standing at 9%, 53% and 38% for speeding incidents and 8%, 67% and 25% respectively.



We conduct similar analysis for the injury types of accidents caused due alcohol related incidents. By checking the occurrences of alcohol-related incidents by hour, our suspicion is confirmed that most of these incidents are at night between 2000 hours and 0400 hours. There are mostly minimal and often even 0 alcohol-related incidents between the hours of 0600 – 1700.

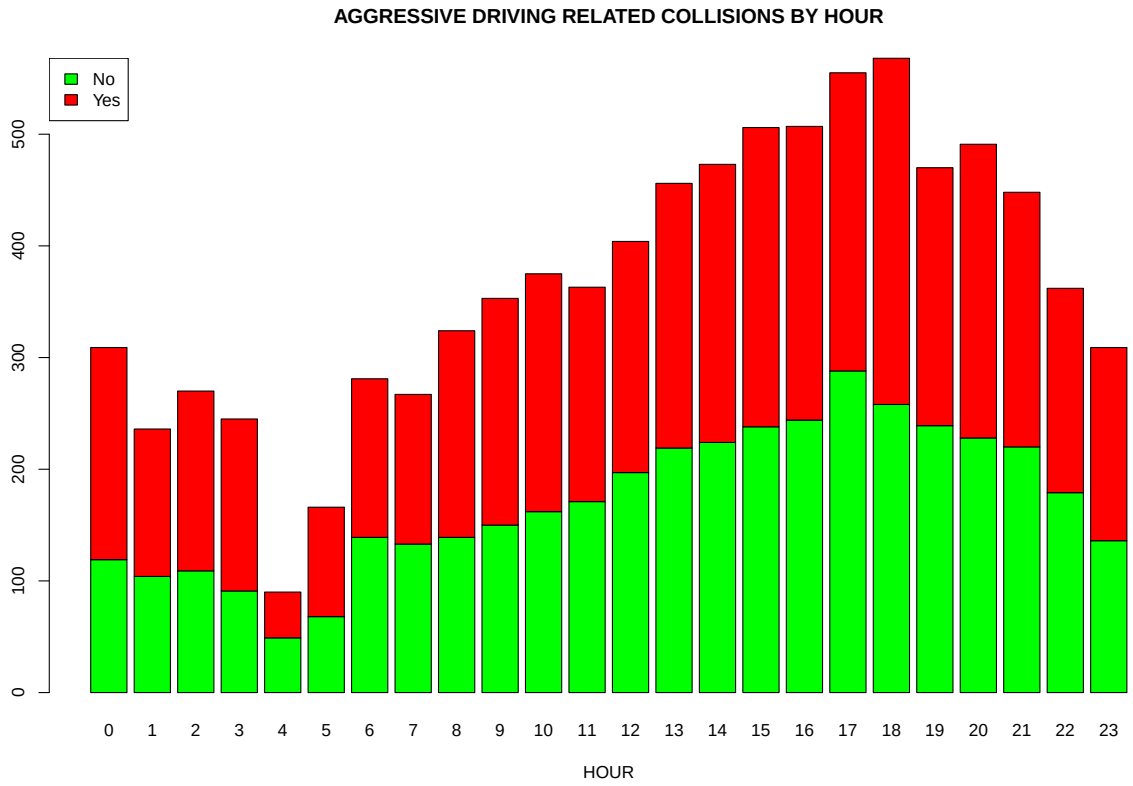


Since most of these incidents occur at night, for alcohol related incidents we only check the difference in injury occurrences of accidents caused at night due to alcohol related and alcohol independent incidents. We do not analyze the alcohol related incidents in the morning and evening as there is a very small sample size which is less than 100 incidents.

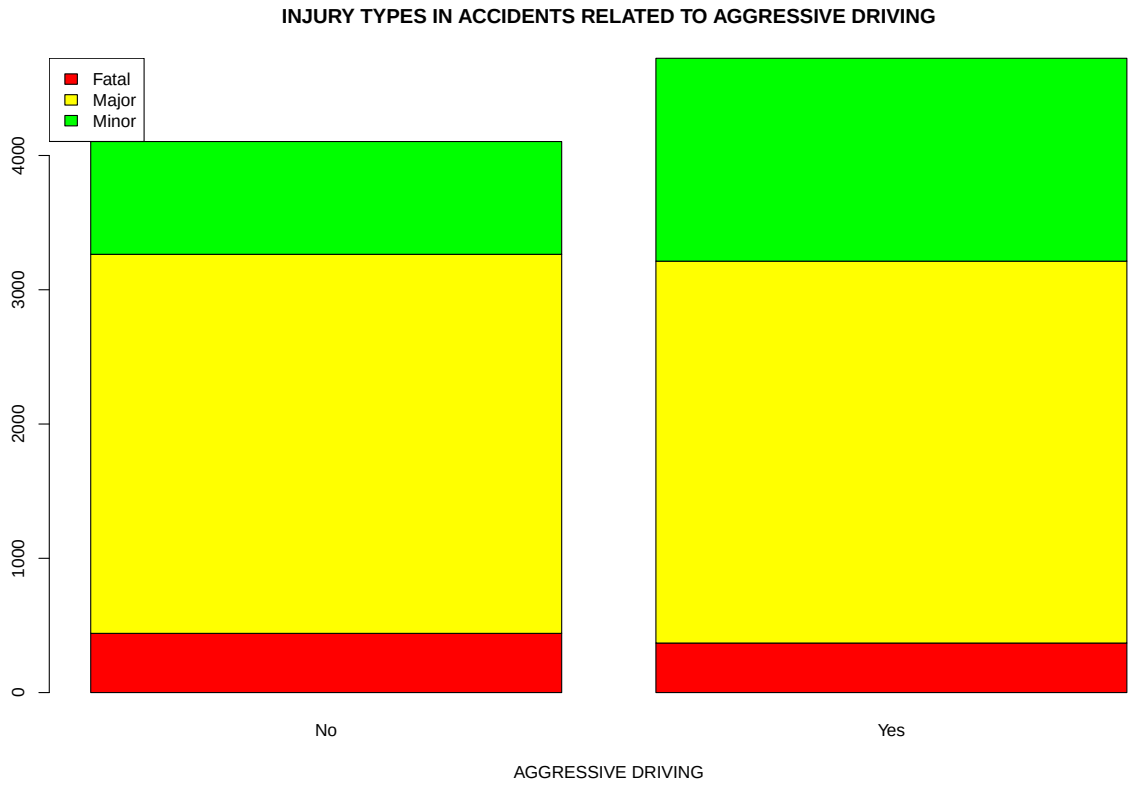


Here we see that in alcohol related incidents, 8.4% result in fatal injuries, 51.4% result in major injuries and 40.2% result in minor injuries whereas in non-alcohol related incidents these numbers stand at 7.5%, 63.9%, and 28.3%. Again, we see that the rate of fatal incidents is lower in non-alcohol related incidents, however those with serious injuries are higher.

In our dataset, most incidents are those of aggressive driving. It is always dangerous to indulge in road rage and this has often been identified as the leading cause of accidents. Let us quickly do a similar analysis of aggressive driving incidents before we move on to discussing these results.



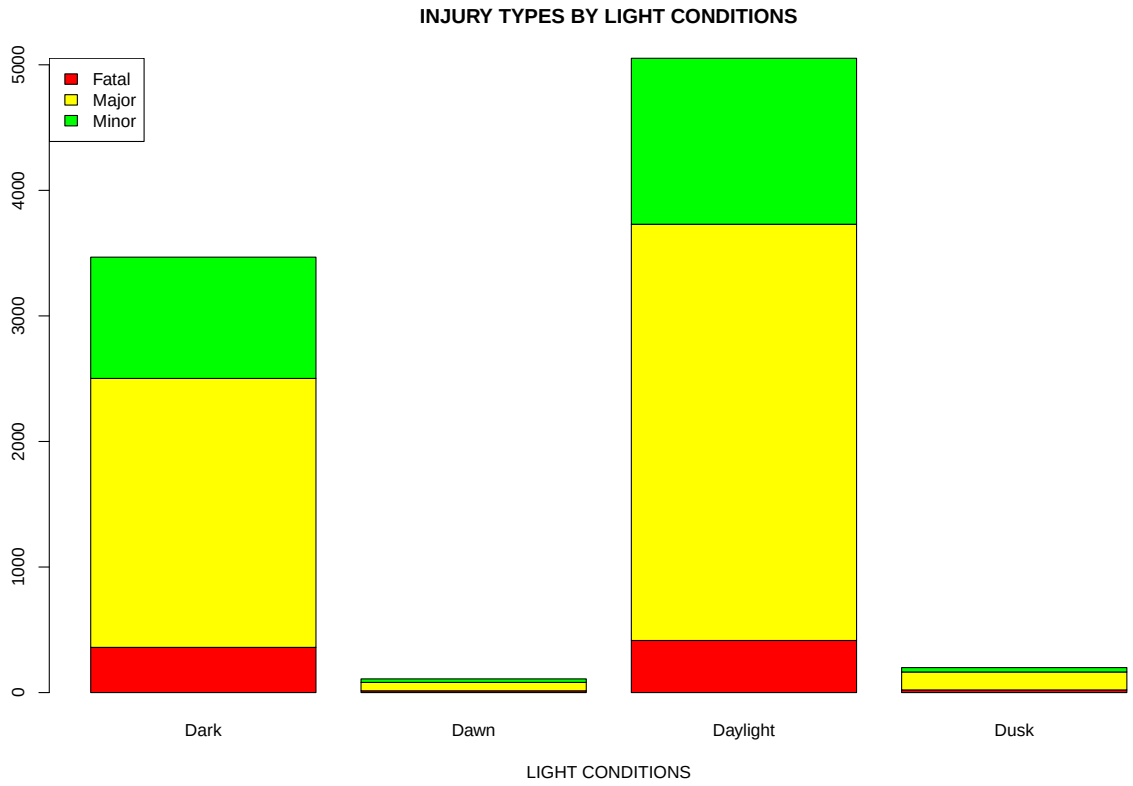
It is clear from this plot that approximately 50% of all accidents involving killed or seriously injured persons are due to aggressive driving irrespective of the hour of day. Here we see that in aggressive driving related incidents, 10.7% result in fatal injuries, 68.7% result in major injuries and 20.46% result in minor injuries whereas in non-aggressive driving related incidents these numbers stand at 7.8%, 60.2%, and 31.9%.



We can see that in non-aggressive driving cases the chances of a fatal or serious accident are lower than that in the case of aggressive driving.

4 Discussion

In our analysis, in figure 11, we saw that most accidents recorded were due to aggressive driving. Most common examples of this are jumping red lights, not following stop signs, failing to signal turns or lane changes. At high speeds on highways, failing to signal a lane change is a serious error. At highway speeds any car accident has the potential to be fatal. These errors are seen much more than accidents involving alcohol or driving under influence. This may be due as getting caught while driving under influence has a much higher perceived cost in terms of legal and employment implications. However, while driving on highways a little bit of carelessness does not have the same perceived cost. This leads to more incidents of aggressive driving. We observe that the average fatality rate for speeding is 11.54% compared to 10.7% in cases of aggressive driving and 8.4% in alcohol-related incidents. In the above analysis we saw that the rates of accidents resulting in major injuries seemed to be higher in non-speeding related incidents than speeding related incidents, it seemed to be higher in non-alcohol related incidents that alcohol related incidents. This was unintuitive. However, on taking a closer look this may be due to some inclusion of aggressive driving incidents in "non-speeding related" and "non-alcohol related" incidents. Another method to correct this is discussed in the weakness in terms of a model which may be better suited. This could also be due to some confounding error with underlying variables. For example, it may be the case that the real determining factor in fatality rates may be visibility or road conditions which we did not explore which may come to light in a multinomial logistic regression. However, on analyzing other variables and their potential relationship with fatality rates, I did not suspect anything. One example of light conditions is shown below. Under "Dark", "Dusk", "Dawn" and "Daylight", there seemed to be the same proportion of injuries of each type. This was also the case for other variables not included here.



5 Weaknesses and Next Steps

An analysis is only as strong as the data provided is. This data on motor vehicle collisions is collected by the Toronto Police Service. Such data is often taken from police reports of accidents which are sometimes incomplete. There were many missing entries which were omitted. One would assume these are of the more horrific/fatal accidents in which collecting the data is harder. This may have impacted our analysis and could explain some of the unintuitive results we saw. Additionally, when collecting such data on traffic accidents, the police officer often must hear contradicting viewpoints of parties involved in the accident before making a decision. This introduces a subjective element in this process and introduces any biases, cognitive or otherwise, that the police officer or the parties involved may have. Moreover, a better model of ordinal or nominal logistic regression with the severity of injury as the dependent variable may have been more apt, but it is a more sophisticated statistical method which comes with its own complexities. In further studies and research, it could prove useful to explore this avenue.

6 Appendix

Extract of the questions from Gebru et al. (2021)

Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. • The data was collected by the City of Toronto, Toronto Police Services. Each accident involving a fatality or serious injury needs to be recorded and the data is collected by the police officer on duty.

2. Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)? • The dataset was created by the Toronto Police Service on behalf of the City of Toronto.
3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. • We took the dataset from the Open Data Toronto website where the final report is published. No funding was needed on our part.
4. Any other comments? • No, no other comments.

Composition

1. What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. • The instances are of traffic collisions. They include details of the rough neighborhood in which they occurred. It also contains information about what the caused the accident, who was involved, the driving conditions at the time and any type of rule infringements like driving under influence/ speeding/ jumping a red light etc.
2. How many instances are there in total (of each type, if appropriate)? • In the initial dataset there were a total of 16,860 instances. After cleaning we had about 8,828 instances.
3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable). • Our dataset is not a random sample. It is the initial data but cleaned for errors such as columns where there were no entries. We also removed certain characteristics which were not necessary for our analysis. Most of these entries are recorded by the police and are reliable, except the entries about the exact location of the accident, which are randomized to the nearest node for privacy purposes.
4. What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description. • Each instance contains information about what the caused the accident, who was involved, the driving conditions at the time and any type of rule infringements like driving under influence/ speeding/ jumping a red light etc.
5. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text. • The exact location of the accident, which are randomized to the nearest node for privacy purposes.
6. Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. • Relationships between individual instances is made explicit by graphing. The graphs highlight trends and differences between individual instances.
7. Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. • No.
8. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. • There are no errors, sources of noise, or redundancies in the dataset.
9. Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate. • This dataset is archived by the City of Toronto, so since it is a government institution, there are guarantees that it will exist over time. • There are official archival versions of the complete dataset. • There are no licenses or fees that would restrict someone's access to any of the external resources.

10. Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description. • Since the data is anonymous, it is not considered confidential.
11. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. • The data is not offensive, insulting, threatening, and does not cause anxiety.
12. Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. • This dataset does not identify sub-populations such as people with different religions, education history, residence, and age.
13. Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how. • It is impossible to identify individuals directly or indirectly from the dataset because all the information that was collected has no name attached to it and measures to randomize locations of accidents have been taken. The car number plates are not included in the data either.
14. Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

• The dataset does not contain data that might be considered offensive in any way. 15. Any other comments?
• No other comments.

Collection process

1. How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. • The data was collected by the City of Toronto, Toronto Police Services. Each accident involving a fatality or serious injury needs to be recorded and the data is collected by the police officer on duty. .
2. What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? • Scraping and parsing with R.
3. If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)? • This data set is not a sample from a larger population.
4. Who was involved in the data collection process (for example, students, crowd workers, contractors) and how were they compensated (for example, how much were crowd workers paid)? • The data is collected by the on-duty police officers who are on the payroll for the city of Toronto.
5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. • The data is collected from 2006 to 2020. It is associated with the time frame that instances were created.
6. Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. • No.
7. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)? • The data was collected by the City of Toronto, Toronto Police Services. Each accident involving a fatality or serious injury needs to be recorded and the data is collected by the police officer on duty.

8. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

• Yes. Those involved in an accident are present at the time of data collection. The data is collected with their permission, or a lawyer can be seeked. 9. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. • Yes, the city can use this data. 10. Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. • Yes. We will be using the data to identify causes and signs that worsen accidents. This can help us put more safety measures in place. 11. Any other comments? • No.

Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. • The data is cleaned, and some labels are changed. We dropped columns which we did not require.
2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. • Yes, the saw data is available on opentoronto.ca under Motor Vehicle Collisions Involving Killed or Seriously Injured Persons.
3. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point. • R was used.
4. Any other comments? • No.

Uses

1. Has the dataset been used for any tasks already? If so, please provide a description. • No.
2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. • No.
3. What (other) tasks could the dataset be used for? • To track and improve road safety.
4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms? • The data set is only used for STA304 Final Paper, which is this task.
5. Are there tasks for which the dataset should not be used? If so, please provide a description. • No.
6. Any other comments? • No other comments.

Distribution

1. Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. • No.
2. How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? • The dataset will not be distributed.
3. When will the dataset be distributed? • The dataset will not be distributed.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. • No.
5. Have any third parties-imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. • No.
6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. • No.
7. Any other comments? • No.

Maintenance

1. Who will be supporting/hosting/maintaining the dataset? • The dataset will not be maintained.
2. How can the owner/curator/manager of the dataset be contacted (for example, email address)? • N/A
3. Is there an erratum? If so, please provide a link or other access point. • No.
4. Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)? • The dataset will not be updated.
5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. • N/A
6. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers. • N/A
7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. • N/A
8. Any other comments? • No. ("Impaired Driving: Get the Facts. Centers for Disease Control and Prevention." n.d.) (Wickham et al. 2021)("Canadian Motor Vehicle Traffic Collision Statistics: 2020. Transport Canada." n.d.)(Xie 2014)("The Leading Causes of Car Accidents." n.d.)

References.

"Canadian Motor Vehicle Traffic Collision Statistics: 2020. Transport Canada." n.d. *Transport Canada.* https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2020.

"Impaired Driving: Get the Facts. Centers for Disease Control and Prevention." n.d. *Centers for Disease Control and Prevention.* https://www.cdc.gov/transportationsafety/impaired_driving/impaired-drv_factsheet.html.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

"The Leading Causes of Car Accidents." n.d. *Waterdown Collision.* https://waterdowncollision.com/safe-driving/leading-causes-of-car-accidents/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.