PREDICTING CAR ACCIDENT SEVERITY

Julian Ariza

September 2020

Introduction

According to the World Health Organization, approximately 1.35 million people die each year as a result of road traffic crashes. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.

The 2030 Agenda for Sustainable Development has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2020. Saying that, the use of data science and artificial intelligence could have a huge impact in achieving this goal.

The addressing of this problem by data scientists teams has several benefits, starting by a continuous analysis of the important data to be collected on each crash, structuring the protocols of the different institutions handling these situations, creating reaction teams not only for first aid but also to restabilising the normality of the traffic, using real time applications to alert commuters and more. This benefits not only tackle the problem from a government perspective but also from the commuter point of view.

The goal of this project is to create a machine learning model to predict the severity of an accident as property damage only, or human injuries including death.

Data

The shared data is provided by the SDOT Traffic Management Division and Traffic Records Group from Seattle. It is a .csv file that contains 194.673 collisions registered from 2004 to 2019. Each collision provides 37 characteristics plus a severity classification given, from which we are going to obtain the features for the machine learning model.

Many of the characteristics have the purpose of collision identification for the government entities, these are values that are unique for each collision, because of this, no pattern applies. Other characteristics provide further description of numerically categorized characteristics, making them redundant. Finally, other characteristics provide details that will not be available until someone is in place taking the report not being available at the moment of the prediction, therefore, we discard them.

The following characteristics are discarded according to the reasons given above after the first approach. Further details are given for each when needed.

1. OBJECTID: Is a unique identifier for each collision.
2. INCKEY: Is a unique identifier for each collision.
3. COLDETKEY: Is a unique secondary key for each collision.
4. REPORTNO: Is a unique report identifier for each collision.
5. STATUS: No information is given for the values.
6. INTKEY: Is a unique identifier for the intersection associated with the collision.
7. LOCATION: Is the unique street address for the location associated with the collision.
8. EXCEPTRSNCODE: No information is given for the values.
9. EXCEPTRSNDESC: No information is given for the values.
10. SEVERITYCODE.1: Is the same as 'SEVERITYCODE', therefore, redundant.
11. SEVERITYDESC: Is the description for the 'SEVERITYCODE', therefore, redundant.
12. COLLISIONTYPE: Is a description of the collision that explains how it happened. I assume this information will not be available until the report.
13. PERSONCOUNT: Is the specific number of all the people involved. I assume this information will not be available until the report.
14. VEHCOUNT: Is the specific number of all the vehicles involved. I assume this information will not be available until the report.
15. INCDATE: Is the date of the collision. Redundant with INCDTTM with less information.
16. SDOT_COLCODE: Is redundant to SDOT_COLDESC.
17. SDOTCOLNUM: Is a unique identifier for each collision given by the SDOT.
18. ST_COLCODE: Is redundant to ST_COLDESC.
19. SEGLANEKEY: Is a unique identifier for the lane segment where the collision occurred. More than 75% of the information is unknown.
20. CROSSWALKKEY: Is a unique identifier for the crosswalk where the collision occurred. More than 75% of the information is unknown.

The following columns will have further analysis to build the machine learning model. Some of them will not be available until the report, however, we will try to infer patterns from these. Further details are given for each when needed.

1. SEVERITYCODE: Is the target of the machine learning model. 'Property Damage Only Collision' is classified as '1' and 'Injury Collision' is classified as '2'.
2. X: Coordinate longitude for the collision.
3. Y: Coordinate latitude for the collision.
4. ADDRTYPE: Classifies the address of the collision as 'Alley', 'Block' or 'Intersection'.
5. PEDCOUNT: Is the number of the pedestrian involved in the collision.
6. PEDCYLCOUNT: Is the number of people on bicycle involved in the collision.
7. INCDTTM: Is the date and time of the collision. Registered with the format MM/DD/YYY HH:MM: SS AM or PM.
8. JUNCTIONTYPE: Is the category of the junction where the collision took place.
9. SDOT_COLDESC: Is the description given to the SDOT_COLCODE. Provides details of the collision.
10. INATTENTIONIND: Indicates if the collision was due to inattention.
11. UNDERINFL: Indicates whether or not a driver involved was under the influence of alcohol or drugs.
12. WEATHER: Is a description of the weather conditions during the time of the collision.
13. ROADCOND: Is a description of the road condition during the collision.
14. LIGHTCOND: Is a description of the light condition during the collision.

15. PEDROWNOTGRNT: Indicates whether or not the collision involves not granting the pedestrian right of way.
16. SPEEDING: Indicates whether or not the speed was a factor on the collision.
17. ST_COLDESC: Is the description given to the ST_COLCODE. Provides details of the collision.
18. HITPARKEDCAR: Indicates whether or not the collision involved hitting a parked car.

Methodology

I loaded the .csv data and looked the general description of it. With the purpose of predicting the severity on mind, we separate the information that we are going to use in the model as mentioned in the previous section. With the remaining information, we notice that the columns have several unique values in many of them and the data is skewed, being more common to have 'Property damage only'.

First of all, we perform data cleaning. I check the existence of NaN values in every column and the following changes are made:

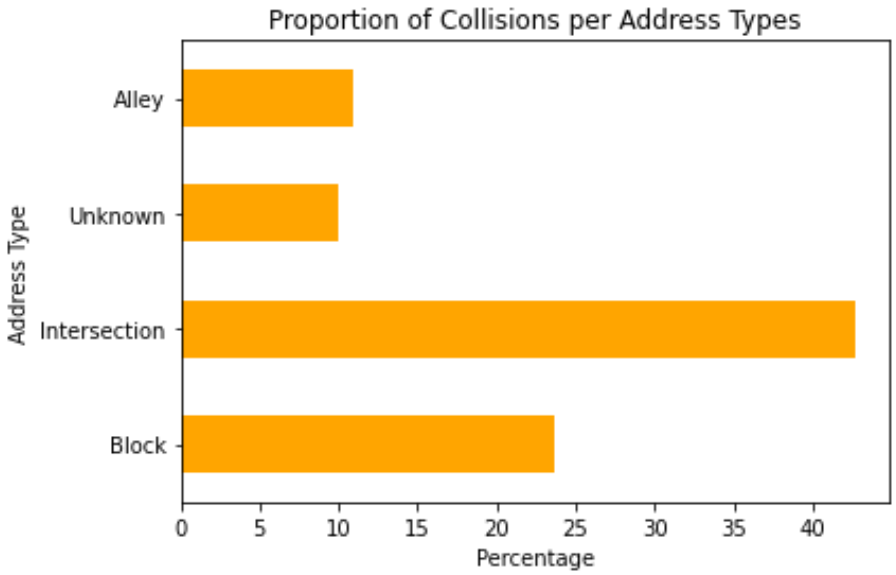
1. SEVERITYCODE '1' is changed to '0' and '2' is changed to '1'.
2. A total of 5334 NaN values on the 'X' are changed to the mean.
3. A total of 5334 NaN values on the 'Y' are changed to the mean.
4. A total of 1926 NaN values on ADDRTYPE are changed to the new value 'Unknown'.
5. A total of 6329 NaN values on JUNCTIONTYPE are changed to the existing value 'Unknown'.
6. A total of 4884 NaN values on UNDERINFL are changed to the existing value '0', 'N' values are changed to '0', 'Y' values are changed to '1' and data type is changed to integer.
7. A total of 5081 NaN values on WEATHER are changed to the existing value 'Unknown'.
8. A total of 5012 NaN values on ROADCOND are changed to the existing value 'Unknown'.
9. A total of 5170 NaN values on LIGHTCOND are changed to the existing value 'Unknown'.
10. A total of 9333 NaN values on SPEEDING are changed to the new value '0'.
11. A total of 4667 NaN values on PEDROWNOTGRNT are changed to the new value '0'.
12. 'N' values are changed to '0' and 'Y' values are changed to '1' on HITPARKEDCAR.
13. A total of 4904 NaN values on ST_COLDESC are changed to the existing value 'Not Stated'.
14. From the column INCDTTM we created the new columns HOURDAY, DAYWEEK and MONTH, this information could give us patterns on the behaviour related to the commuting activity. The column INCDTTM is then dropped.

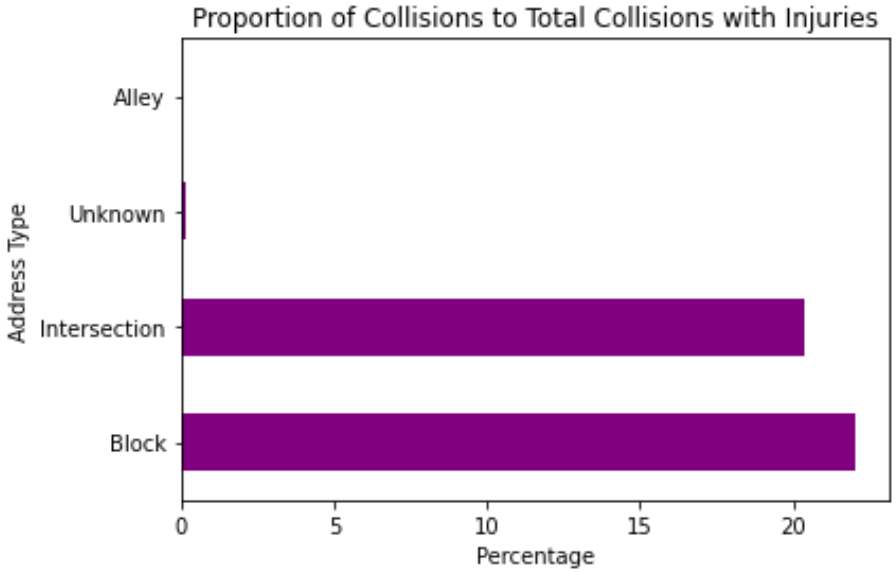
After having homogenous data, we proceeded with a frequency analysis, the goal of this analysis is to find the most common cases for having collisions with injuries, for this, we created a new data frame with the columns:

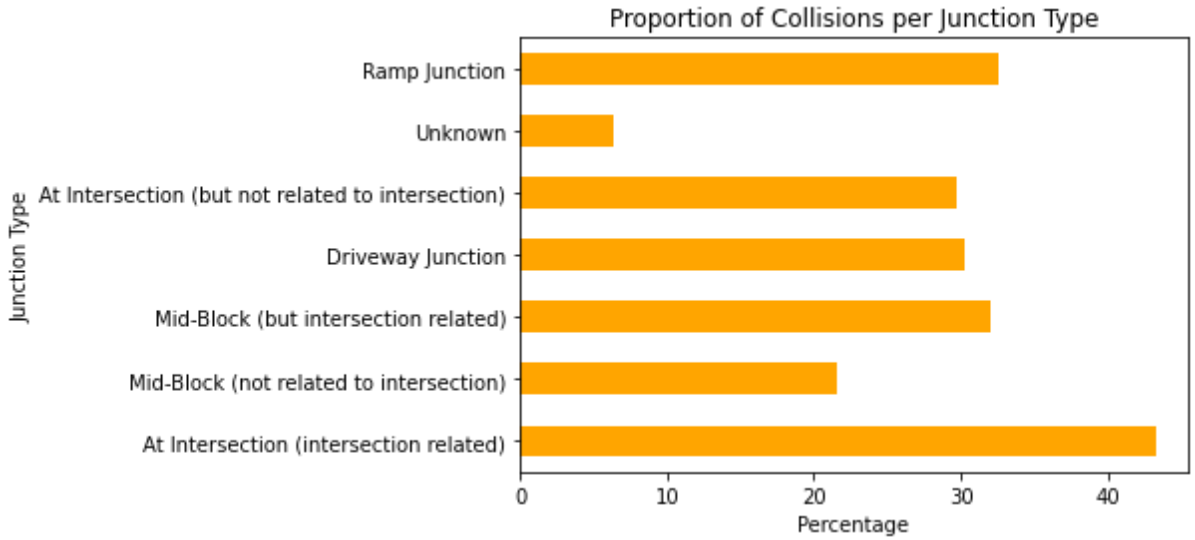
1. Column where we placed the original columns.
2. Value where we placed every single possible value of each column.
3. FreqDmg where we placed the count of the value domain with severity code '0'.
4. FreqInj where we placed the count of the value domain with severity code '1'.
5. PerInj where we placed the percentage of FreqInj respect to the value domain. FreqDmg plus FreqInj equals the hundred percent.
6. PerInjT where we placed the percentage of FreqInj respect to the total of severity code '1'. This is 58188.
7. PerT where we placed the percentage of FreqInj respect to the total of collisions. This is 194673.

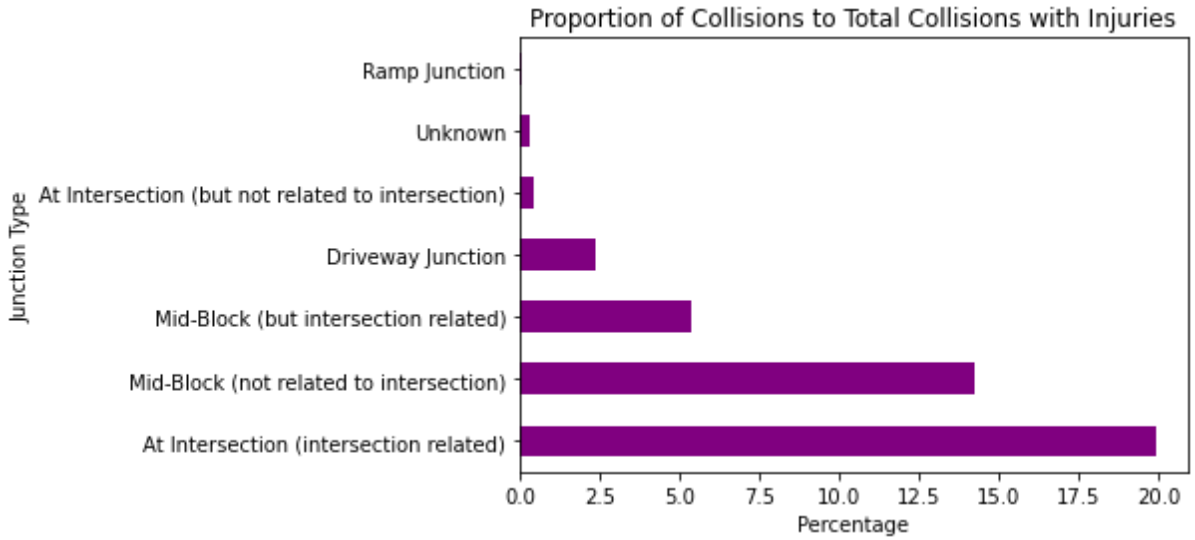
This new data frame was sorted in descending order and we got several important conclusions:

- The most common ST_COLDESC and SDOT_COLDESC values, are values with detailed information of the collision that is obtained once the report is done. Thus, we can discard these columns assuming that this information will not be available at the moment of the prediction.
- 43% of value 'Intersection' in the column ADDRTYPE are collisions with injuries, this represents 14% of the total of collisions. This is evident also in the JUNCTIONTYPE column, where the most common value is 'At intersection (intersection related)' for collisions with injuries.

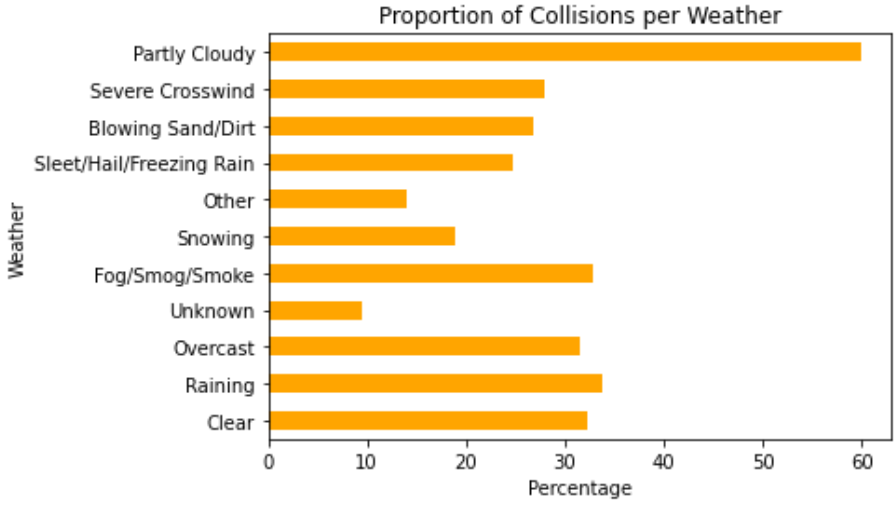


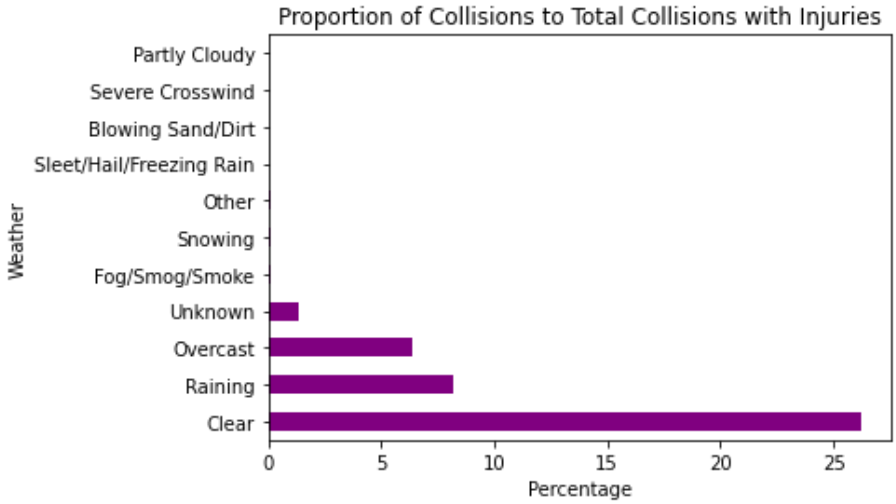




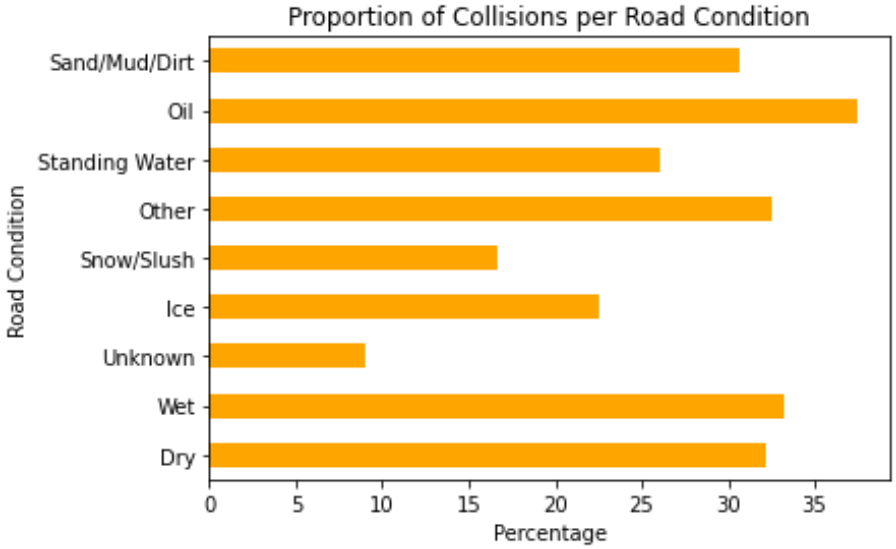


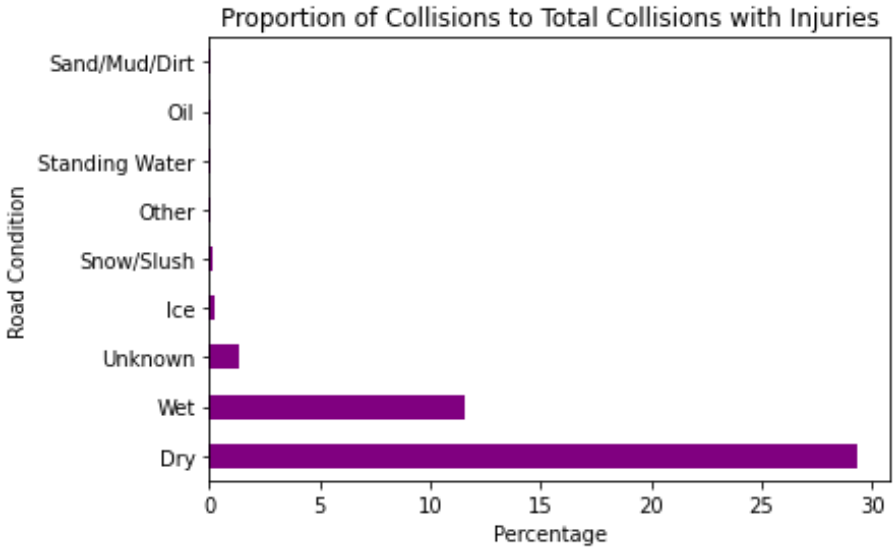
- WEATHER ‘Clear’, ‘Raining’ and ‘Overcast’ are more than 95% of the collisions with injuries and more than 85% of the total collisions, making other weather conditions rare. Each of these three have around 30% chance to be collisions with injuries, this is the highest probability among the weather that is supported with vast data.



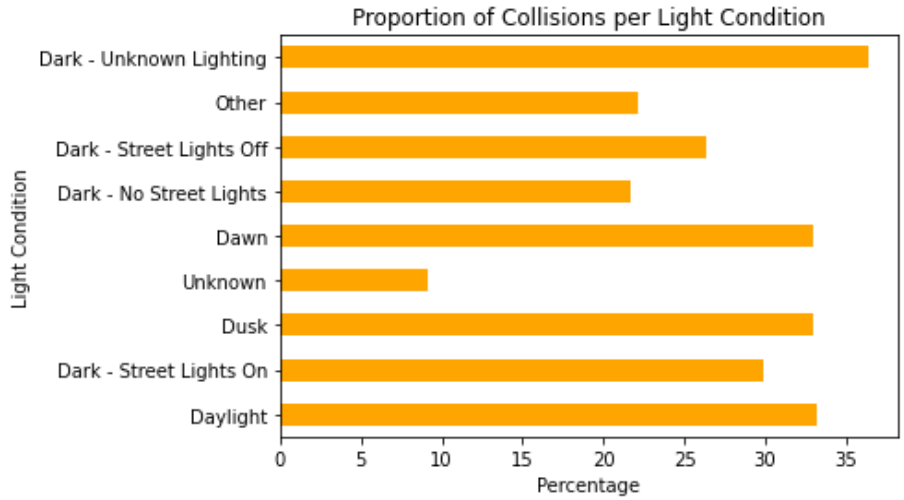


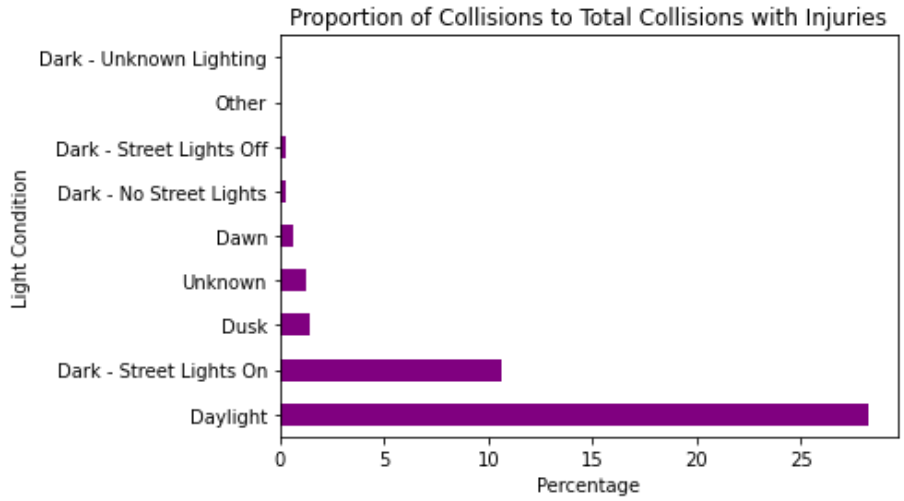
- ROADCOND is congruent with WEATHER conclusion. ‘Dry’ and ‘Wet’ are the most common values with more than 95% of total collision with injuries.





- For LIGHTCOND, normal conditions as ‘Daylight’ and ‘Dark-Street Lights On’, take over the 90% of total collisions with injuries.

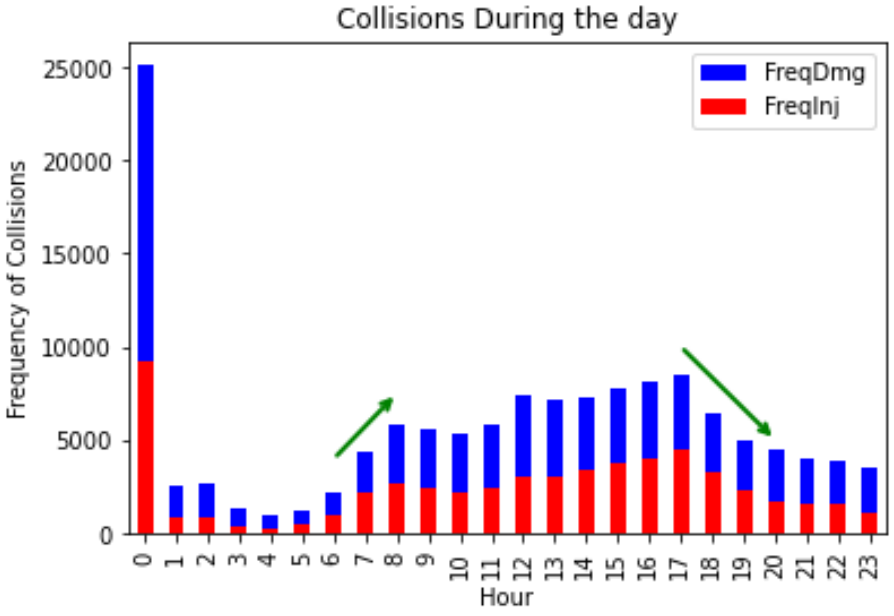


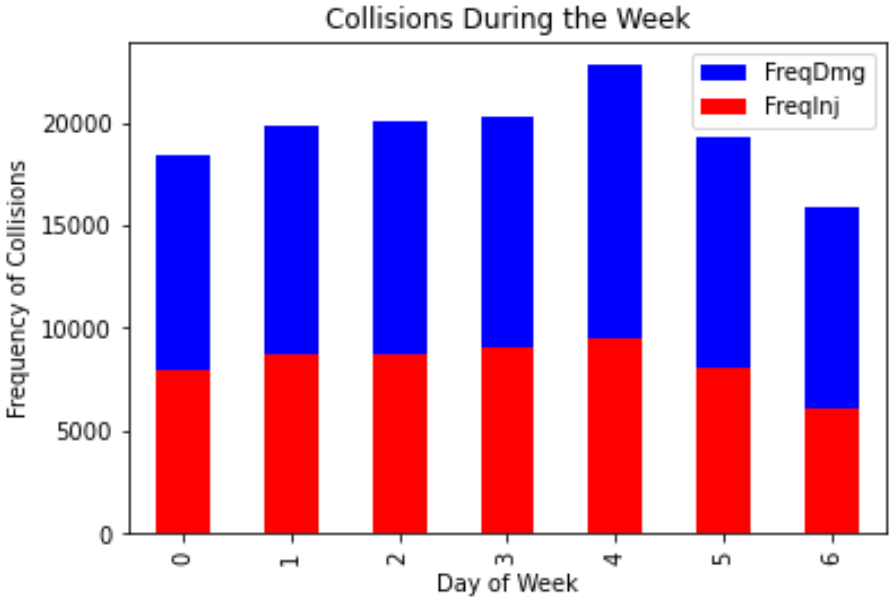


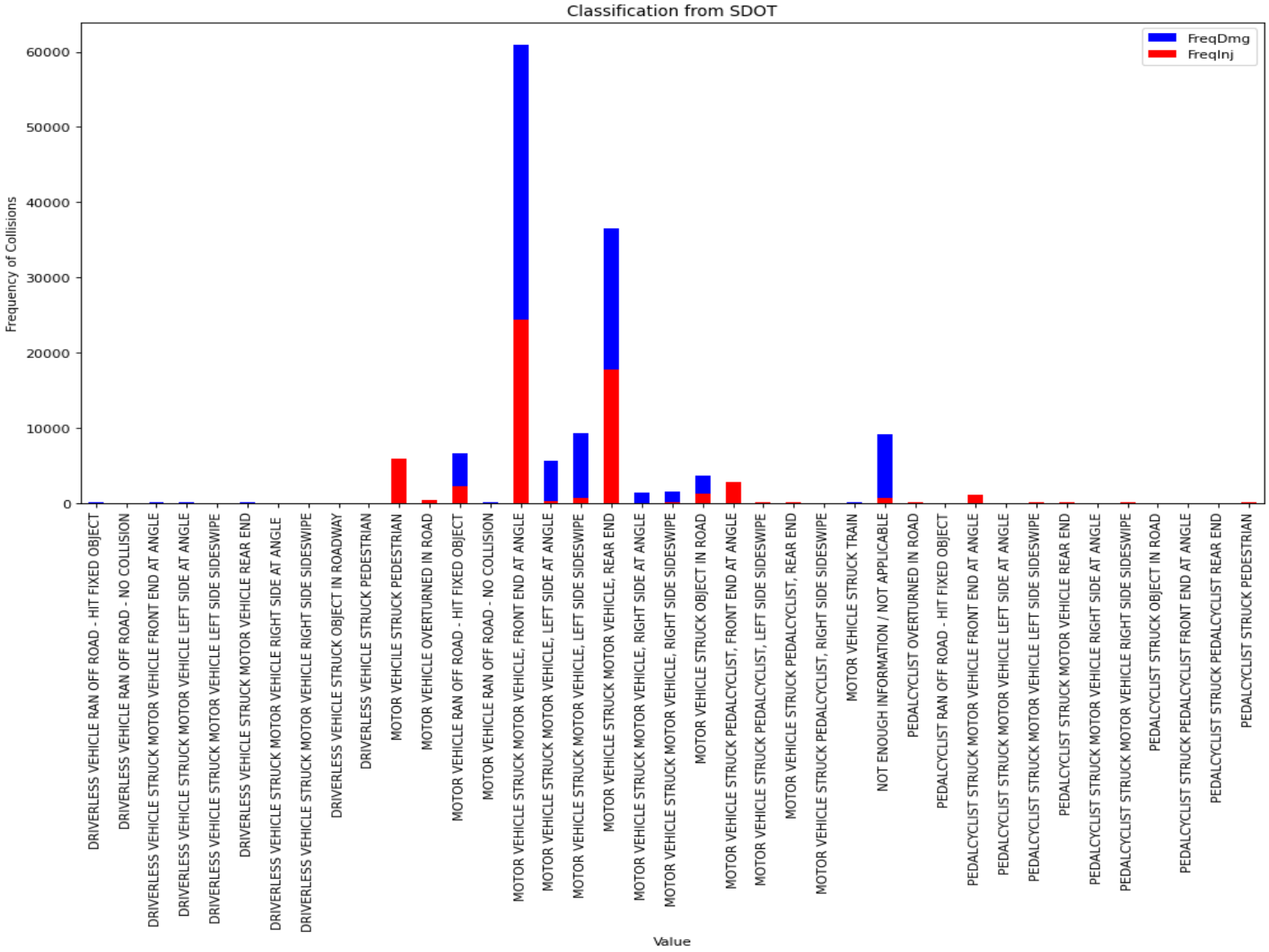
- More than 80% of the collisions for PEDCOUNT and PEDCYLCOUNT are collisions with injuries. Pedestrians and cyclist are the population with more risk of injuries when involved in a collision. This is relevant also in PEDROWNOTGRNT.

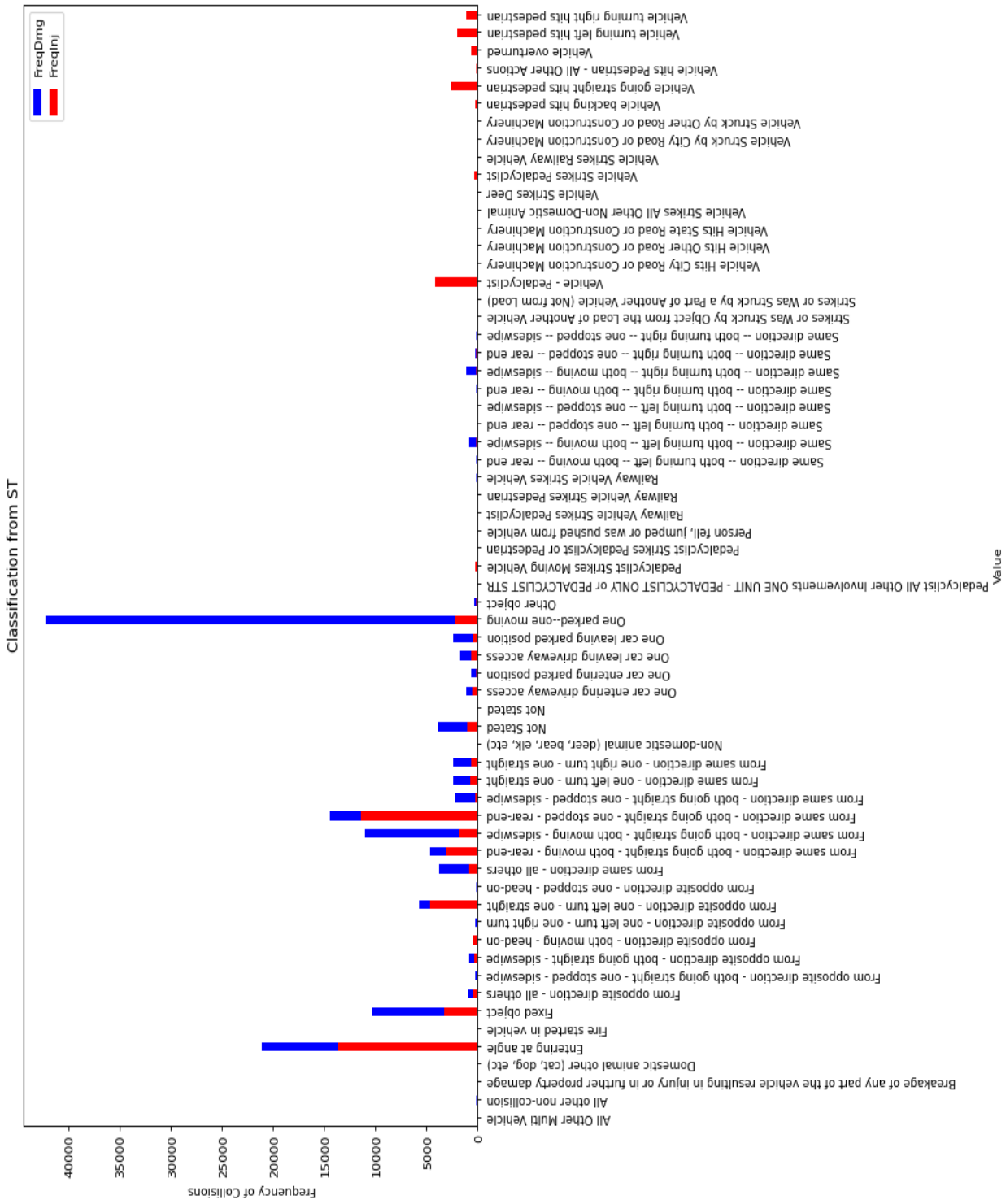
After the numerical comparison, histograms where created for the quantitative columns. The following conclusion were obtained.

- Collisions with injuries increase during business days, reaching a peak on Friday and a significant reduction on weekends.
- The most accidents occur the hour after midnight. However, there are important peak collision hours, these are the hours before and after business hours. In other words, peak collisions hours are those when people commute from their homes to their jobs and from their jobs to their homes.
- From SDOT_COLDESC and ST_COLDESC is noticeable that the cases where a pedestrian or a pedal cyclist is involved, the risk of presenting injuries increases.
- From ST_COLDESC there is a peak when a parked vehicle is involved for collisions with only property damage.


Collisions During the day
FreqDmg
FreqInj
Frequency of Collisions
25000
20000
15000
10000
5000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour

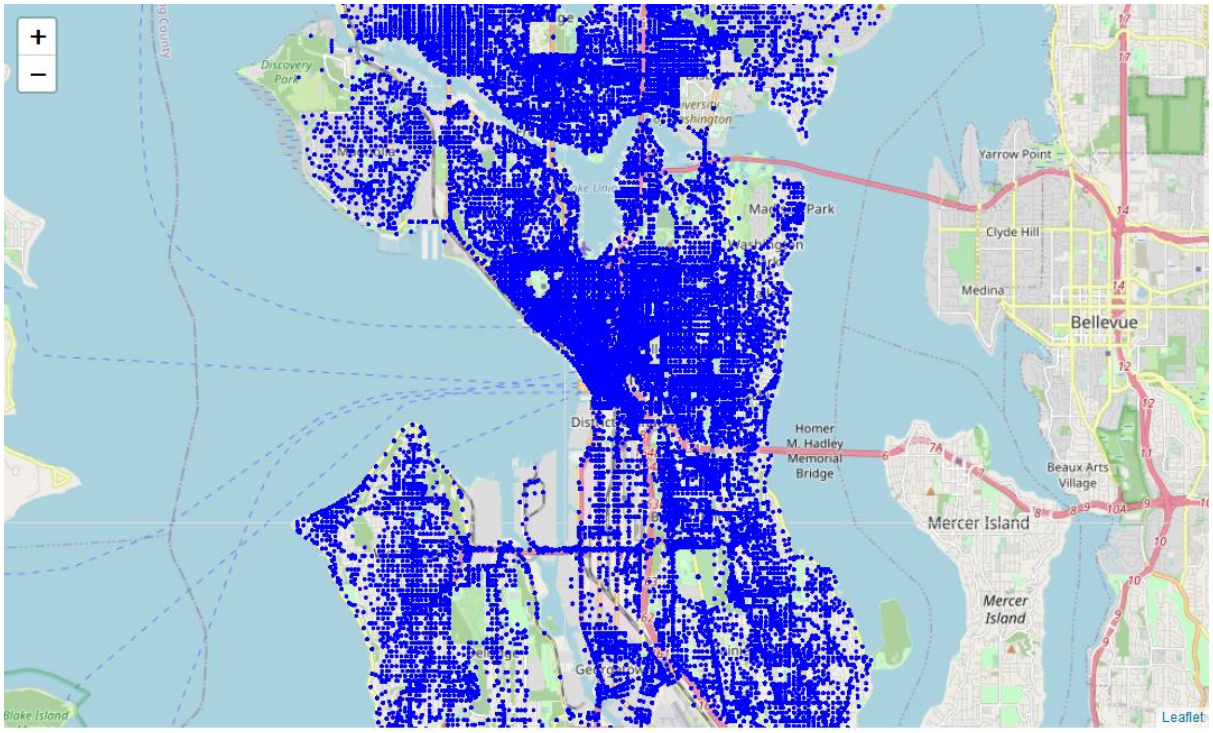

Collisions During the Week
FreqDmg
FreqInj
Frequency of Collisions
20000
15000
10000
5000
0
0 1 2 3 4 5 6
Day of Week


Classification from SDOT
FreqDmg
FreqInj
Frequency of Collisions
0
10000
20000
30000
40000
50000
60000
DRIVERLESS VEHICLE RAN OFF ROAD - HIT FIXED OBJECT
DRIVERLESS VEHICLE RAN OFF ROAD - NO COLLISION
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE FRONT END AT ANGLE
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE LEFT SIDE AT ANGLE
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE LEFT SIDE SIDESWIPE
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE REAR END
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE RIGHT SIDE AT ANGLE
DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE RIGHT SIDE SIDESWIPE
DRIVERLESS VEHICLE STRUCK OBJECT IN ROADWAY
DRIVERLESS VEHICLE STRUCK PEDESTRIAN
MOTOR VEHCILE STRUCK PEDESTRIAN
MOTOR VEHICLE OVERTURNED IN ROAD
MOTOR VEHICLE RAN OFF ROAD - HIT FIXED OBJECT
MOTOR VEHICLE RAN OFF ROAD - NO COLLISION
MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END AT ANGLE
MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE AT ANGLE
MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE SIDESWIPE
MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END
MOTOR VEHICLE STRUCK MOTOR VEHICLE, RIGHT SIDE AT ANGLE
MOTOR VEHICLE STRUCK MOTOR VEHICLE, RIGHT SIDE SIDESWIPE
MOTOR VEHICLE STRUCK OBJECT IN ROAD
MOTOR VEHICLE STRUCK PEDALCYCLIST, FRONT END AT ANGLE
MOTOR VEHICLE STRUCK PEDALCYCLIST, LEFT SIDE SIDESWIPE
MOTOR VEHICLE STRUCK PEDALCYCLIST, REAR END
MOTOR VEHICLE STRUCK PEDALCYCLIST, RIGHT SIDE SIDESWIPE
MOTOR VEHICLE STRUCK TRAIN
NOT ENOUGH INFORMATION / NOT APPLICABLE
PEDALCYCLIST OVERTURNED IN ROAD
PEDALCYCLIST RAN OFF ROAD - HIT FIXED OBJECT
PEDALCYCLIST STRUCK MOTOR VEHICLE FRONT END AT ANGLE
PEDALCYCLIST STRUCK MOTOR VEHICLE LEFT SIDE AT ANGLE
PEDALCYCLIST STRUCK MOTOR VEHICLE LEFT SIDE SIDESWIPE
PEDALCYCLIST STRUCK MOTOR VEHICLE REAR END
PEDALCYCLIST STRUCK MOTOR VEHICLE RIGHT SIDE AT ANGLE
PEDALCYCLIST STRUCK MOTOR VEHICLE RIGHT SIDE SIDESWIPE
PEDALCYCLIST STRUCK OBJECT IN ROAD
PEDALCYCLIST STRUCK PEDALCYCLIST FRONT END AT ANGLE
PEDALCYCLIST STRUCK PEDALCYCLIST REAR END
PEDALCYCLIST STRUCK PEDESTRIAN
Value

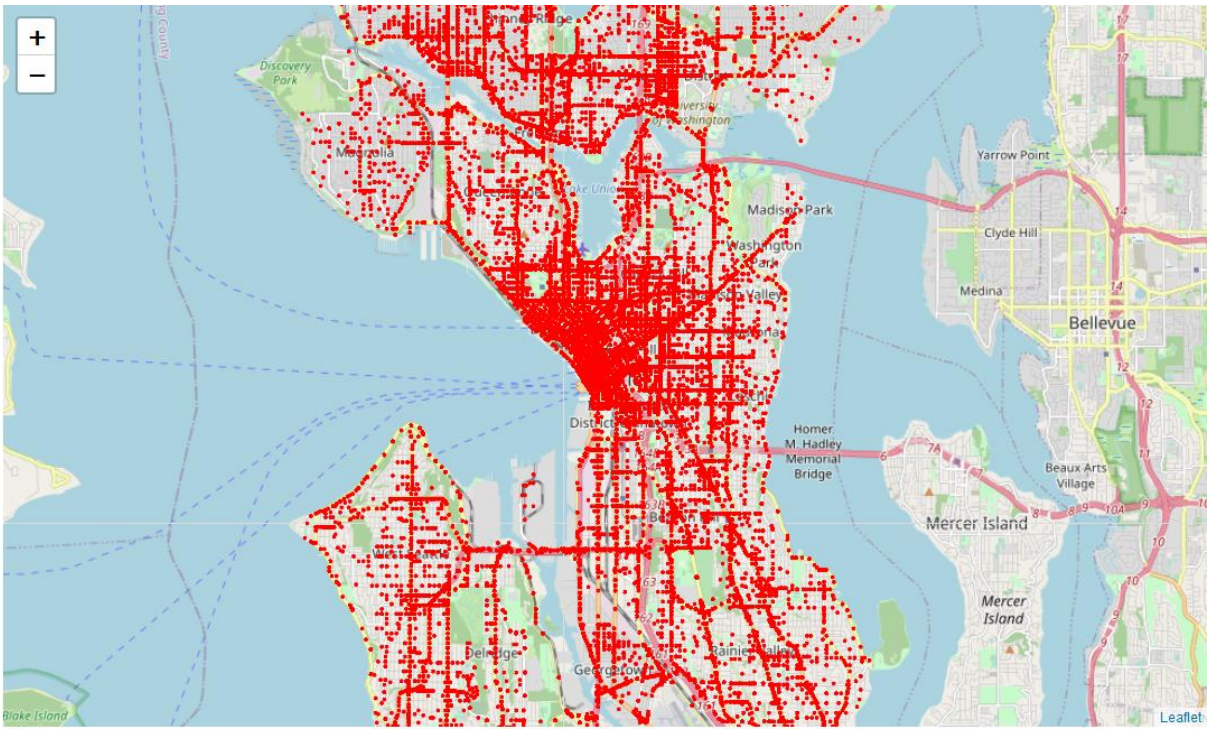


For the columns ‘INATTENTIONIND’, ‘UNDERINFL’, ‘SPEEDING’, ‘HITPARKEDCAR’, ‘PEDCOUNT’ and ‘PEDCYLCOUNT’ we performed a different approach. This information may have patterns on their locations, having more density on areas like avenues, commercial areas, parking lots and so on. Even though this is information that will not be obtained until the report, we could redefine this characteristic as whether or not the location of the collision is close to a place where those conditions are registered in previous collisions. For example, an avenue may have high density on accidents where speeding was a factor, thus, an accident in this avenue has high probability of

having speeding as an accident. On the machine learning model, this data must be dropped on the test data and estimated with the training data only, otherwise, the evaluation of the model is will not reflect the real performance of the algorithm to estimate these values.



Seattle map with Collisions with property damage only. The collisions are evenly dispersed.



Seattle map with Collision with Injuries. There are clear regions and avenues.

From the maps obtained for each column we can conclude:

- These factors, less HITPARKEDCAR, have recognizable zones when plotted in the map. I dropped HITPARKEDCAR because of it even distribution.

Finally, we perform the following changes to the data before being fed to the modelling tools, all of them supported by the data analysis made previously and its conclusions:

1. The columns to be fed will be 'X', 'Y', 'ADDRTYPE', 'JUNCTIONTYPE', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SPEEDING', 'HOURDAY', 'DAYWEEK', 'MONTH', 'PEDCOUNT' and 'PEDCYLCOUNT'.
2. On WEATHER, values 'Overcast' and 'Fog/Smog/Smoke' are joined for their similar conditions of low visibility and similar probability of being a collision with injured.
3. On WEATHER, values 'Snowing' and 'Sleet/Hail/Freezing Rain' are joined for their similar conditions of low temperatures and similar probability of being a collision with injured.
4. On WEATHER, values 'Blowing Sand/Dirt' and 'Severe Crosswind' are joined for their similar conditions related to the wind and similar probability of being a collision with injured.
5. On WEATHER, values 'Unknown', 'Other' and 'Partly Cloudy' are joined for their similar probability of being a collision with injured.
6. On ROADCOND, values 'Ice' and 'Snow/Slush' are joined for their similar conditions related to low temperatures.
7. On ROADCOND, values 'Wet' and 'Standing Water' are joined for their similar conditions related to water and similar probability of being a collision with injured.
8. On ROADCOND, values 'Oil' and 'Sand/Mud/Dirt' are joined for their similar conditions related with spread objects or substances and similar probability of being a collision with injured.
9. On LIGHTCOND, values that contain 'Dark' are joined for their similar condition and probability of being a collision with injured.
10. On LIGHTCOND, values 'Dusk' and 'Dawn' are joined for their similar probability of being a collision with injured.
11. On LIGHTCOND, values 'Unknown' and 'Other' are joined.
12. 'PEDCOUNT' and 'PEDCYLCOUNT' are joined to a new column 'PEDNUM' with a value '1' when there is, at least, one pedestrian or one pedal cyclist involved on the collision. For this we sum up the columns and give a threshold with zero or greater than 0. Previous columns are dropped.
13. For 'ADDRTYPE', 'JUNCTIONTYPE', 'WEATHER', 'ROADCOND' and 'LIGHTCOND' dummies are created.
14. For 'X' and 'Y', they are normalized, the mean is subtracted, and they are divided by their standard deviation.
15. For 'HOURDAY', 'DAYWEEK', and 'MONTH' are normalized being divided by their maximum value.

For building the prediction model we split the data on two. Sixty percent of the data is used as training data and forty percent of the data is used for cross validation and testing. From the forty percent data we drop the columns 'INATTENTIONIND', 'PEDNUM', 'UNDERINFL', 'PEDROWNOTGRNT' and 'SPEEDING', these were calculated based on the proximity to collisions with these characteristics in the training data. As stated before, these characteristics are supposed to not be available until the final report of the accident but will be estimated according to a distance epsilon to a collision with this attribute.

This forty percent of data is then split in two groups with equal amount of collisions, one group is used as cross validation, for tuning the machine learning model parameters. The second group is treated as unseen data for final evaluation.

I used the four different machine learning algorithms, these are K Nearest Neighbours, Decision Tree Classifier, Support Vector Machine and Logistic Regression. For their evaluation it is used Jaccard Index, F1 Score and Log Loss when possible. Overall, all the algorithms had a similar performance having Jaccard Index around 25%, F1 Score for 'Property Damage Only' of around 80% and F1 Score for 'Injury Collision' of around 40% with a weighted average of around 65%. Specific values are shown on the table.

Algorithm	Jaccard	F1-Score	Log Loss
KNN	0.27	0.65	-
Decision Tree	0.27	0.65	-
SVM	0.18	0.68	-
Logistic Regression	0.19	0.69	0.53

Discussion

As seen in the data analysis, most accidents happen under the best conditions for driving, for example during daylight, with clear weather and dry road. This could be interpreted as overconfidence. With exception of the involvement of pedestrians and pedal cyclist, most conditions in the collisions had a probability of 30% of being one with injured people. This is reflected in the machine learning models results, reflecting this proportion.

The accidents could be related to activities surrounding the location of the accident as shown on the map's patterns, so, integrating information related to these topics from entities of commerce for example, could increase the performance of the model.

Conclusion

In this study, I analyse collision traffic information in order to predict the severity of the collisions. For this purpose, we inferred information that was considered relevant for the prediction but not necessarily available at the moment of the prediction, depending on the method implemented to gather this information. The prediction results reflected the probability found in the data; however, complexity can be added, and data be collected to improve the performance.