1. Introduction

There has been an increase in traffic volume worldwide with more and more people having access to a car. Traffic accidents have become a significant cause of deaths and injuries and a major concern for public health. Recent reports of the World Health Organization (2018) show that car crashes represent 1.35 million deaths each year worldwide, and 2.4 million people seriously injured (requiring hospital admission). Road traffic crashes are a leading cause of deaths; this is the 1st cause of death among people aged 5 to 29 years old globally. Car crashes are the 8th leading cause of death across all age groups worldwide and are predicted to become the 7th leading cause by 2030. In the USA, 39,888 persons have died in a car accident in 2016. Reducing traffic accidents is a public safety challenge.

To reduce the number of accidents and improve safety on the road, the city of Seattle has provided a dataset containing the history of car accidents in the city over the last 16 years. The Seattle office of the mayor is interested in a model that will significantly reduce the number of accidents. Using the severity index, we aim at building a model that will warn drivers by sending real-time alerts when driving in a dangerous situation likely to lead to a severe accident. We will use the severity index of the dataset to differentiate the severe and minor accidents.

We will also use the dataset to extract insights about how to improve the security on the roads in Seattle. We can identify the dangerous areas of the city and provide recommendations to the local authorities of Seattle and police forces.

2. Data

This dataset has been collected by Seattle authorities. It describes the accidents that occurred in Seattle from 2004 to 2020. It contains 194,673 rows, each one corresponding to an accident. There are 40 columns describing the accident, these columns[1] can be grouped into the following categories:

- Identification numbers of the accident (OBJECTID, REPORTNO, INCKEY)
- Location of the accident (X, Y, LOCATION, INTKEY, JUNCTIONTYPE, SEGLANEKEY, CROSSWALKKEY)
- Description of the accident (INCDTTM, SEVERITYCODE, SEVERITYDESC, COLLISIONTYPE, SDOT_COLCODE, ST_COLCODE, HITPARKEDCAR)
- Description and number of victims (PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT)
- Consequences of the accident (INJURIES, SERIOUSINJURIES, FATALITIES)
- Weather conditions (WEATHER, ROADCOND, LIGHTCOND)
- Description of the driver's behavior (INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING)

The columns are either float, int or object type.

Given that we aim to give the drivers real time alerts, we will not use all the columns to train the model. Indeed, some information, such as the number of people involved in the accident or the collision type cannot be known before the accident happened and will be deleted to train the model. However, we will use these data to provide insights about the accidents that already happened. The severity column

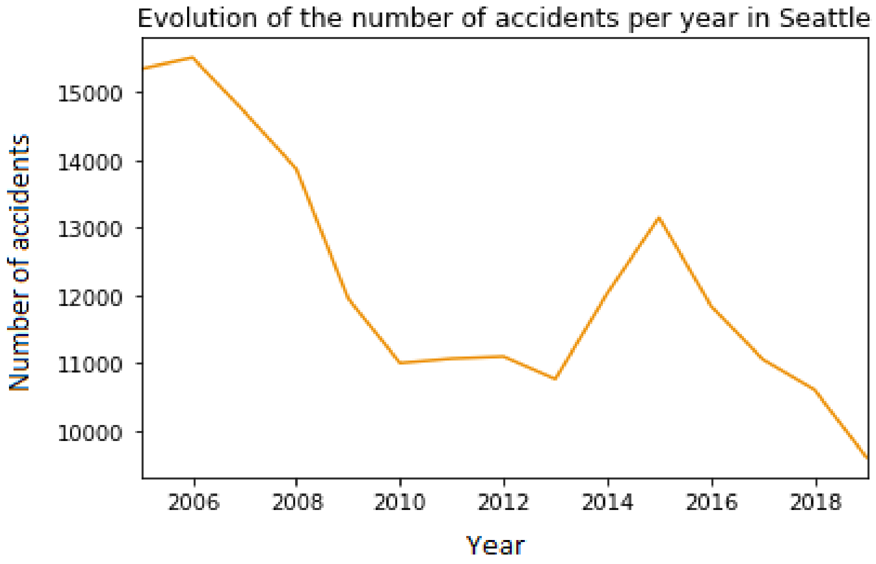
[1] Here is a non-exhaustive list of all the columns. You can check the notebook with the code to have a full description of the dataset.

contains 4 values : 1 (property damaged only collision), 2 (injury), 2.b (serious injury) and 3 (fatalities). To simplify the model and some of the graphs in the data visualization part, we have grouped the categories 2, 2.b and 3 together. Accidents will be categorized as property damaged only collision or severe.

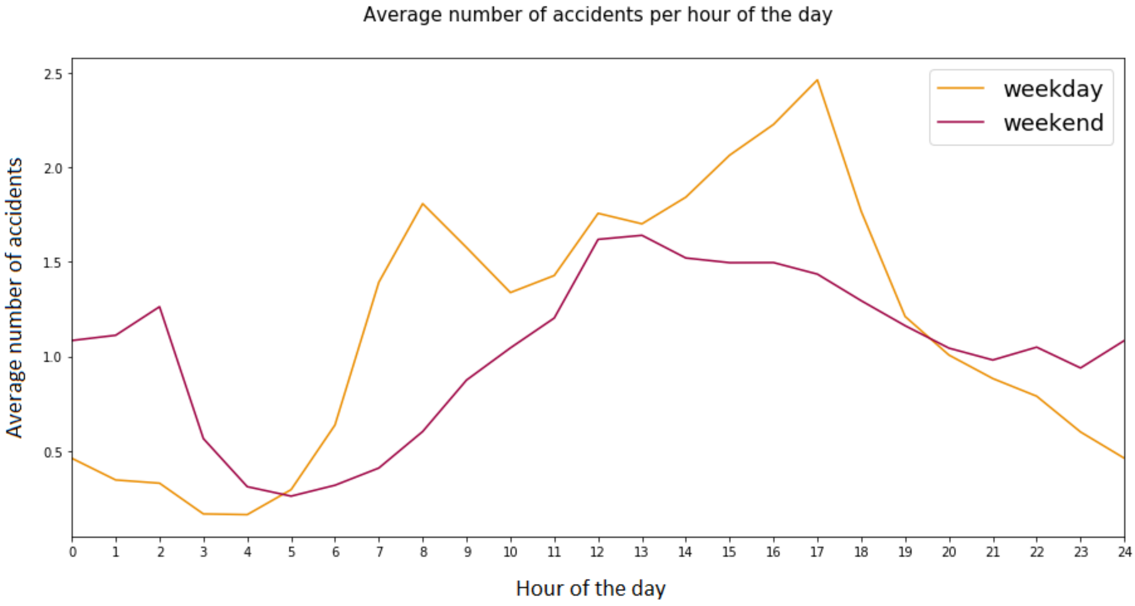
3. Methodology

3.1 Data visualization

We are first exploring the data to have some insights on the accidents happening in Seattle. We can see that the total number of accidents has decreased since 2004. This may be due to numerous road safety campaigns in the last decade.

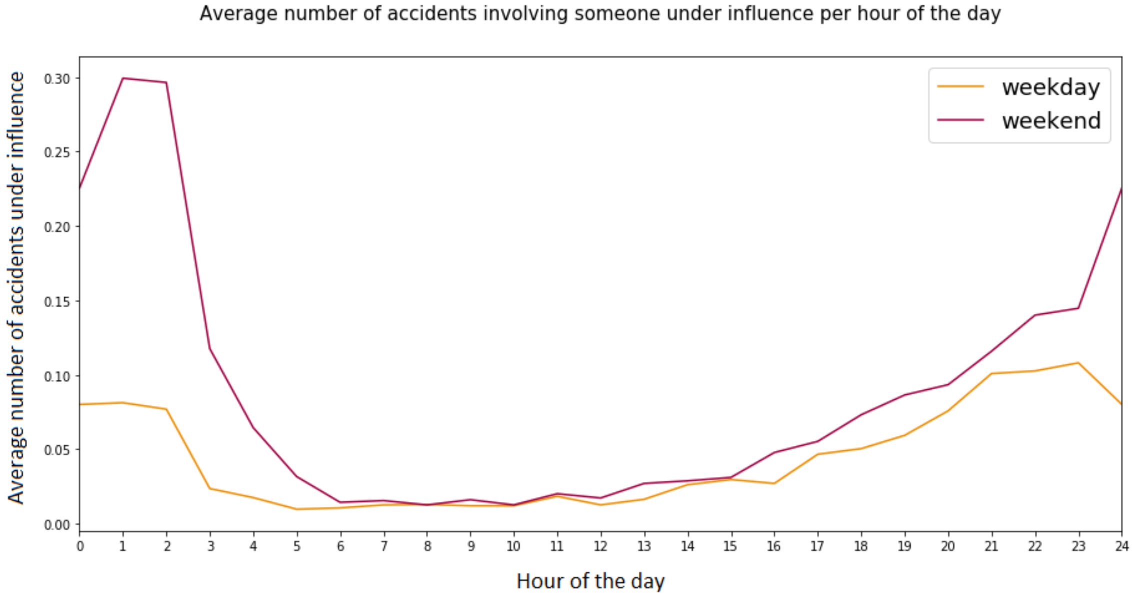


We would like to understand when accidents are happening. The graph below describes the distribution of the accidents during the day, both for weekdays and weekends.

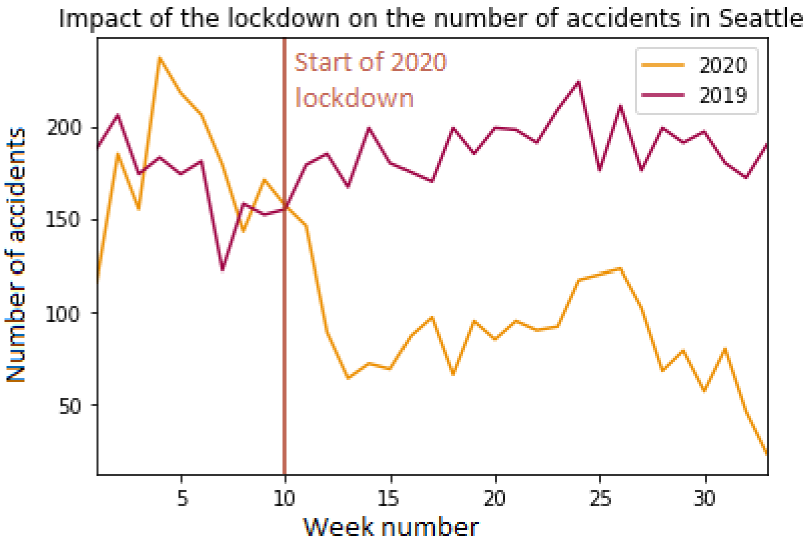


We notice a dramatic increase in the number of accidents at 7 am during the week. The number of accidents remain high until 7 pm. We can distinguish two peaks at 7.30 am and 5 pm. These peaks correspond to the rush hours when there are lot of drivers on the road.
During the weekend, we notice an increase in the number of accidents around 11 am with a peak at 1 pm. We also note a peak around 2 am during weekends, probably due to people going out.

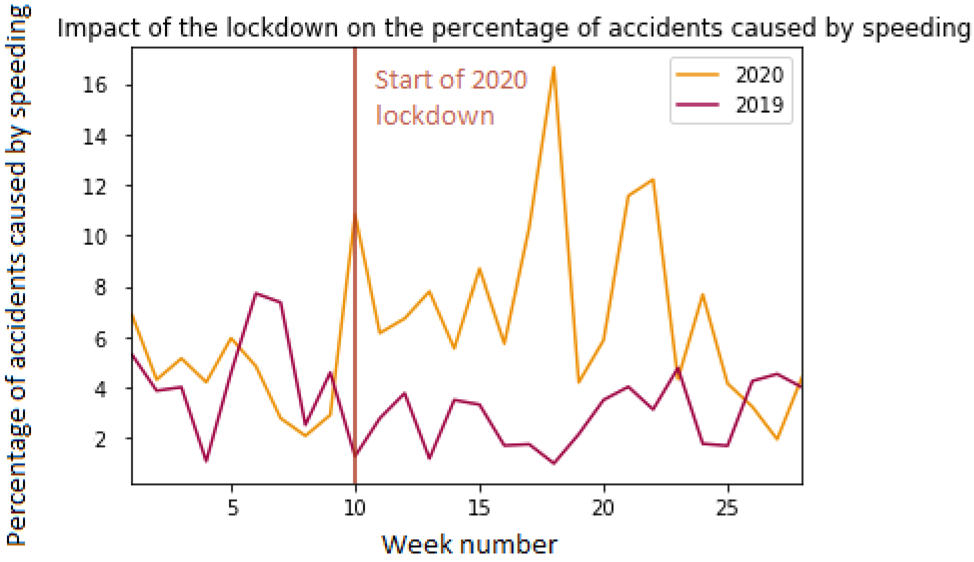
Concerning these accidents happening at night, we may find a reason in the following graph. Indeed, we can see that accidents caused by someone under influence happen mostly during the weekend with a peak around 2 am. Generally speaking, we see an increase in the number of accidents under influence after 5 pm every day of the week.


Average number of accidents involving someone under influence per hour of the day

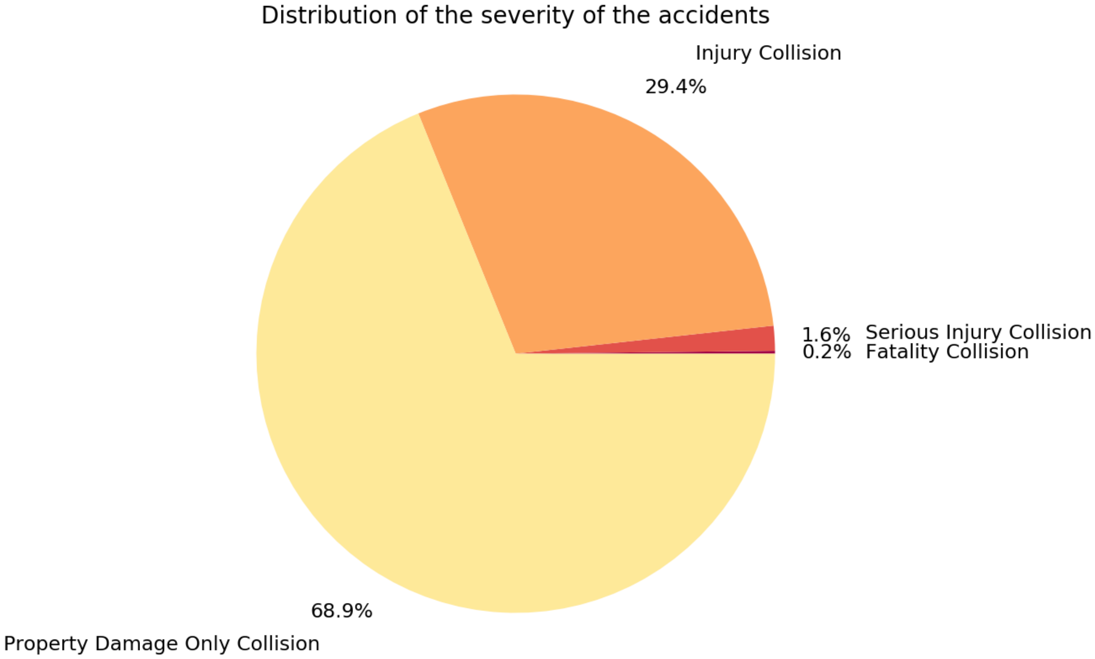
It is interesting to look at 2020 data to see if the Covid-19 and lockdown orders have had some impact on the number of accidents. Seattle authorities announced a stay-at-home order from March (week 10). The graph below compares the number of accidents in 2019 and 2020 (from January to August). We see that this number has considerably fallen since March 2020, creating a huge gap with the number of accidents within this same period in 2019.


Impact of the lockdown on the number of accidents in Seattle

If we look closer at the type of accidents happening in 2020, we can see that the number of accidents caused by speeding has increased since March and is much higher than 2019 during the same period. We can deduce that people drive much faster since there are less drivers on the road.

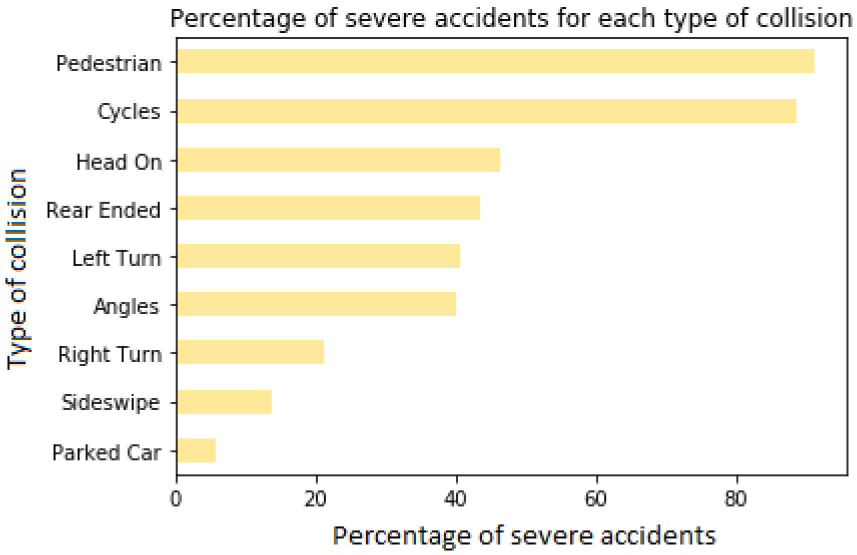


We can now look more specifically at the severity of the accidents since our model aims at predicting it. The following graph shows the distribution of the accidents' severity :

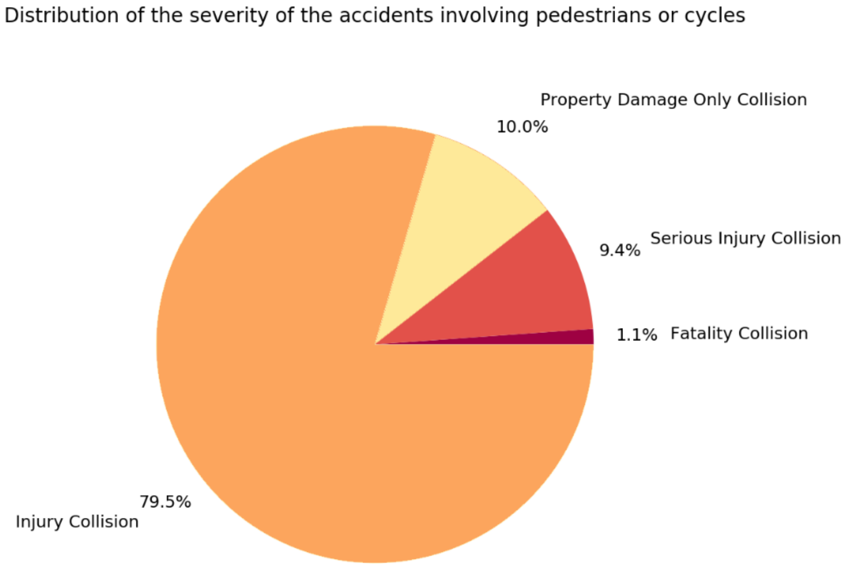


As we can see, most accidents only damage properties (68.9%) while around 29% of the accidents cause injuries and only 1.8% cause serious injuries or death. We can then look at the factors that may increase the severity of an accident.

First, we easily imagine that the type of collision greatly influences the severity of the accident. The graph below displays the percentage of severe accidents depending on the type of collision. We have only kept the most frequent types to display on the graph.



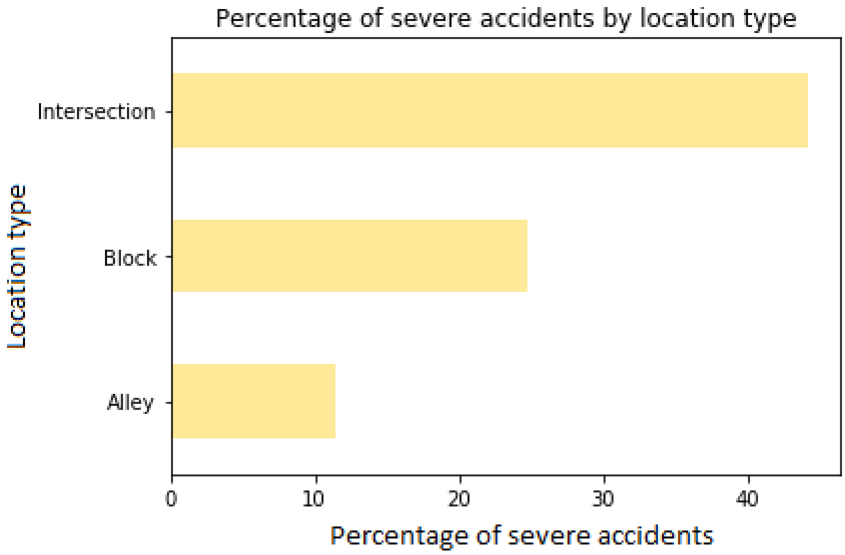
We can see that collisions involving pedestrians and cycles are the most severe, almost 90% caused at least an injury[2].



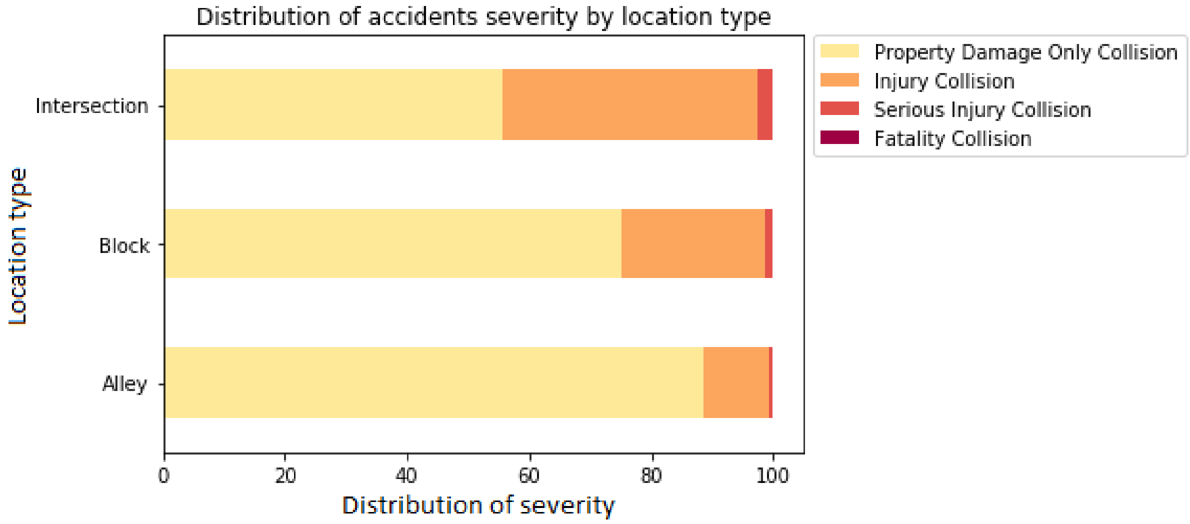
This pie chart gives more details on the collisions involving pedestrians and cycles. Almost 80% of the accidents caused injuries, 9.4% serious injuries and 1.1% were fatal. These numbers are much higher than in the first pie charts where 69% of accidents only caused properties damages.

Then, we presume that the location of an accident also influences its severity. To have a better understanding on which locations are the most dangerous, we visualize the percentage of severe accidents at different location types. Almost 50% of the accidents at an intersection are dangerous, 25% at a block and 11% in an alley.

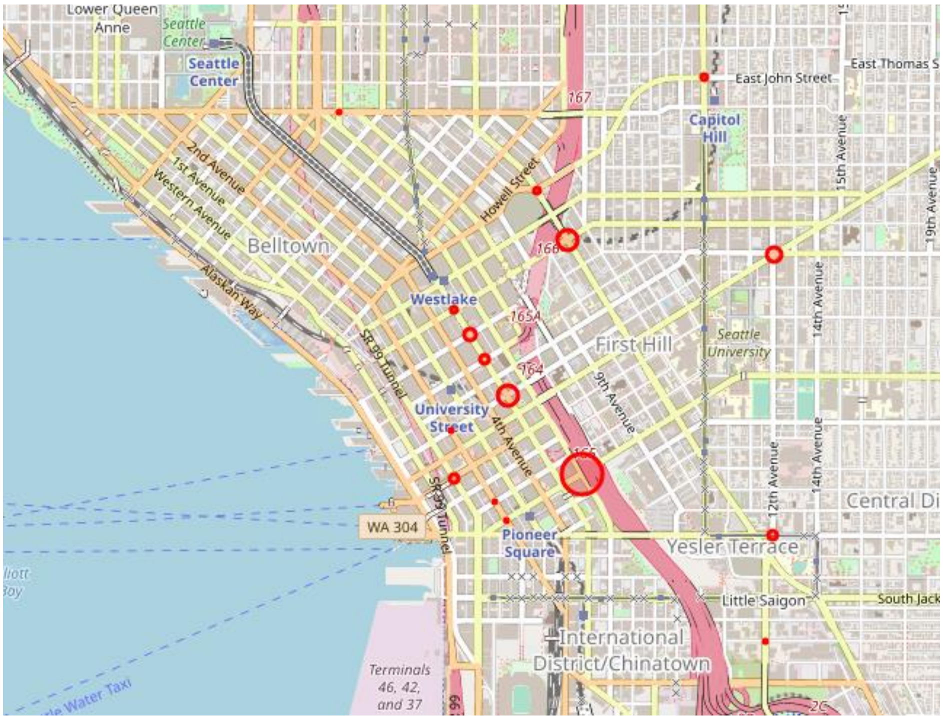
[2] Severe includes all the accidents where someone has been injured.



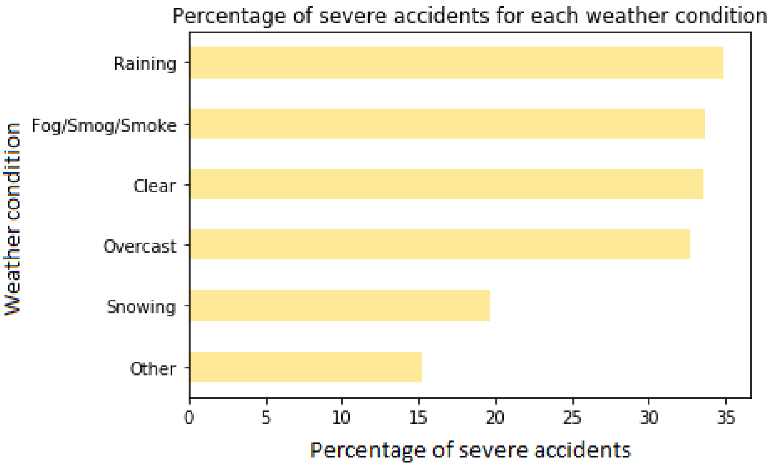
The next graph gives a detailed description of the distribution of the severity of the accidents by location type.



The map below shows the intersections that caused the most severe accidents. We notice a succession of dangerous intersections on the 5th avenue in the city center.



Finally, there is a last factor we can check : the weather. We expect the weather condition to influence the severity of an accident. The graph below shows the percentage of severe accidents depending on the weather conditions.



While we expected that bad weather influenced positively the severity of an accident, based on this dataset, we cannot conclude that there is a relationship between the severity and the weather. We would need the weather data of Seattle to have a better understanding of the relationship between the number of accidents and the weather conditions.

3.2 Data preprocessing

3.2.1 Outliers

We have not identified any outlier in this dataset. Indeed, the number of persons injured can be high, but not enough to be considered as an outlier.

3.2.2 Missing values

When performing the data exploration, we noticed that several columns had missing values. We have two solutions to deal with missing values: delete the rows with missing values or fill the missing values with a value of our choice.

First, we deleted the columns for which more than 20% of the values were missing. Then, we filled the missing values of the remaining columns with the mean of the values for the continuous features and the mode for the categorical ones.

The columns "SPEEDING" and "INATTENTION" only contained one value "Yes" and 96% of missing values. We considered that the missing values meant that the driver was not driving too fast or was not inattentive. We would check with the stakeholders who filled the dataset to confirm.

3.2.3 Feature engineering

Date and time

Since date and times types cannot be directly processed by Machine Learning algorithms, we have transformed them into numerical values or categories. In the data visualization section, we noticed that the accidents were not equally distributed among the hours of the day and the days of the week. Based on this observation, we have created two new features to replace the original date and time:

- WKDAY that shows if the accident happened during the week (Mon-Fri) or weekend (Sat-Sun)
- Moment_of_the_day that splits the hours of the day into 8 categories:
  - mid_night: the middle of the night, from midnight to 7 am
  - rush_am: the morning rush hours, from 7 am to 9 am
  - morning: the morning (during work hours), from 9 am to 12 pm
  - lunch: lunchtime, from 12 pm to 1 pm
  - afternoon: the afternoon (during work hours), from 1 pm to 5 pm
  - rush_pm: the afterwork rush hour, from 5 pm to 6 pm
  - evening: the time when people go out for drinks or diner, from 6 pm to 9 pm
  - night: from 9 pm to midnight

Location

The location column is the address of the accident, therefore there are thousands of different possible values. Since most of the Machine learning algorithms cannot handle categorical values, we will have to get the dummies of this feature if we want to use it to train the model. This would generate almost 4,000 columns to process, which is too many. Instead, we have decided to create two features that derive from the location column:

- LOCATION_number_accidents that indicates the number of accidents at the exact same location.
- Dangerous_locations that indicates if the locations had more than 5 severe accidents in the past.

3.3 Correlation analysis and features selection

3.3.1 Data leakage

As mentioned in the introduction, the objective of this model is to send real-time alerts to drivers on the road. Thus, there are several columns that we are not supposed to use. For example, we cannot know the accident ID or the number of pedestrians involved in the accident before the accident actually occur. These features would create data leakage, so we have deleted the following columns:

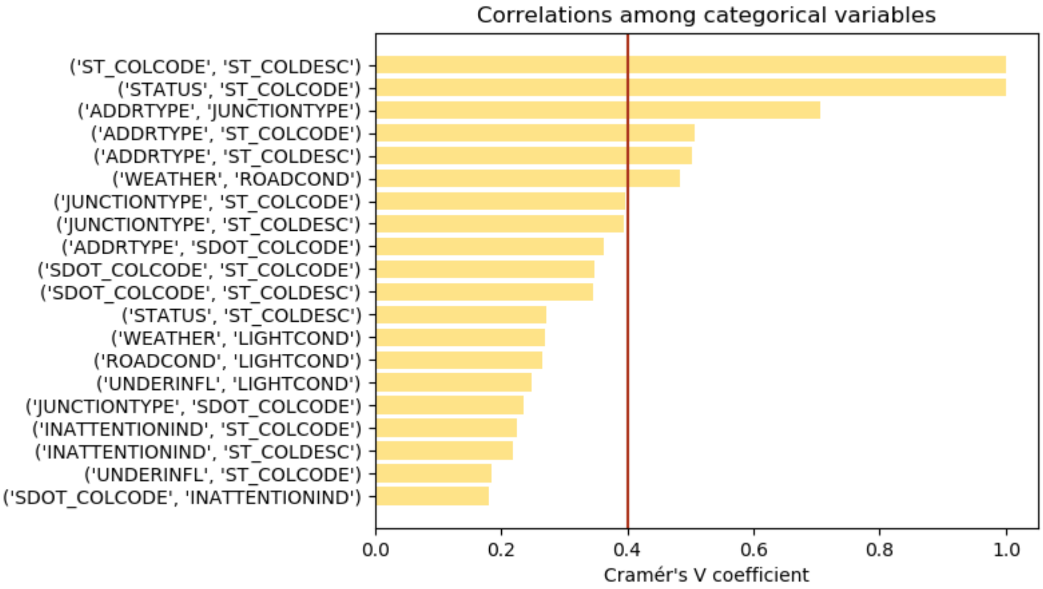
- OBJECTID, INCKEY, REPORTNO that indicate the ID numbers of the accident. Different authorities give an ID number to identify an accident.
- STATUS that gives the information that reports of an accident by different authorities match.
- SEVERITYDESC that gives the exact the same information than the target SEVERITYCODE.
- VEHCOUNT, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT that shows the number of victims.

3.3.2 Correlations analysis

We must delete highly correlated features because they can make the model unstable. We have used the Pearson's correlation coefficient to calculate the correlation among continuous features and the Cramér's V coefficient for the categorical ones. Indeed, we cannot compute the correlation between continuous and categorical features, so we use two different coefficients.

Correlation among features

The graph below displays the correlations among categorical variables.

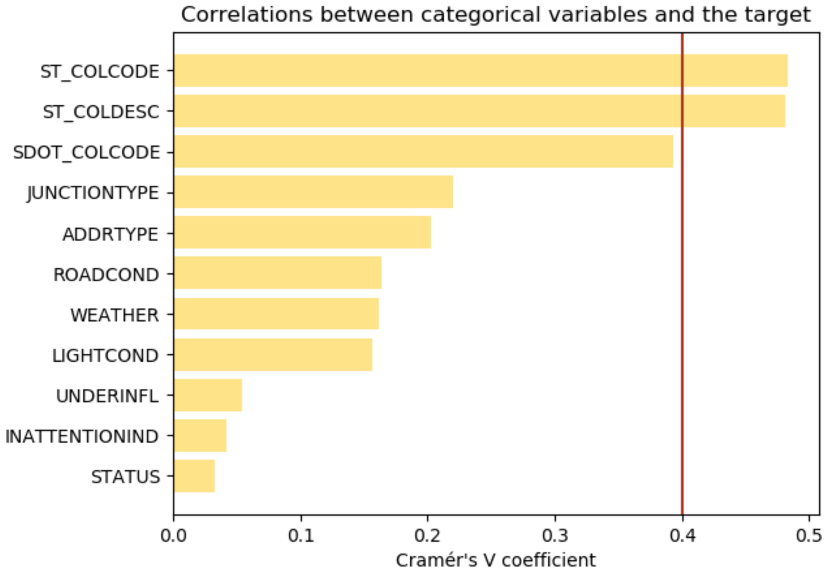


We can see that several pair of columns have a high correlation coefficient (above 0.4). For each correlated pair, we will remove the feature that has the smaller correlation with the target variable and therefore gives the less information on the target variable.

We repeated the same analysis with continuous variables using the Pearson's correlation.

Correlation with target

We have computed the correlation between each feature and the target variable to detect potential data leakage. Since binary variables are continuous and categorical, we can use Pearson's and Cramér's V coefficients to respectively compute correlations between continuous and categorical features with the target. The graph below displays the correlation of the categorical variables with the target.



On the graph above, we notice that some features have a high correlation with the target (for example, ST_COLCODE, ST_COLDESC) that can indicate a data leakage. By exploring these variables, we discovered they give information that are not available before the accident happen (such as the type of collision or number of victims) and so, cannot be used to send real-time alerts. We have removed these columns.

Encoding the categorical variables

Since all machine learning models cannot directly handle categorical features, we need to encode the categorical values into numerical values. We use the pandas function get_dummies on all the categorical features.

Below an example before getting the dummies:

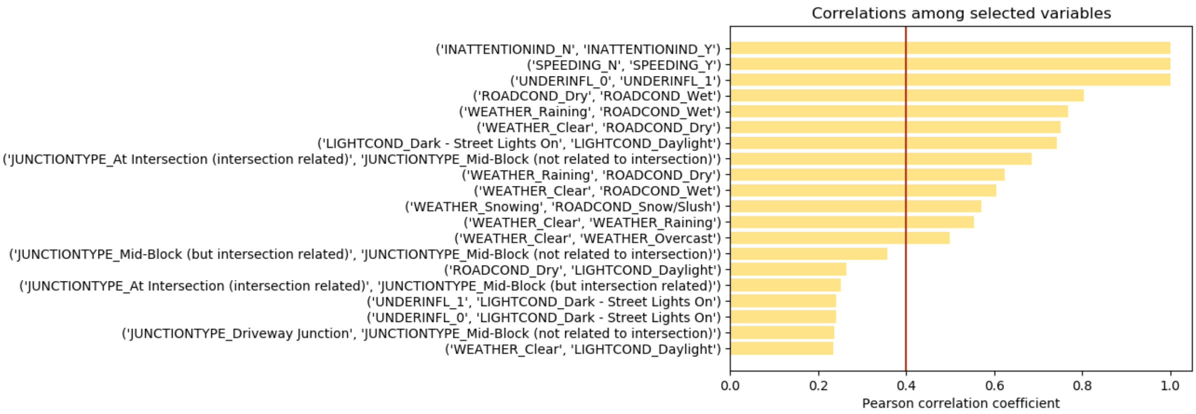
Accident ID	ADDRTYPE
43819849	Alley
7362849	Intersection
1284937	Block

After getting the dummies:

Accident ID	Intersection	Block	Alley
43819849	0	0	1
7362849	1	0	0
1284937	0	1	0

Final correlation check

Since we generated new features, we are doing a last pass of correlation analysis to detect potential new correlated features. We use the Pearson’s coefficient since all the features are now continuous. Some of the categorical variables will be entirely correlated now they are continuous. We will have to delete one of the variables for the correlated pairs. For example, “Underinfluence Yes” and “Underinfluence No” are obviously 100% correlated. The graph below displays the variables with the highest correlation (in absolute values).



3.4 Machine Learning

3.4.1 Choice of evaluation metric

Since the target variable is imbalanced, we chose the F1-score as evaluation metric. We computed the dummies of the categorical variables because the machine learning model only process continuous values.

3.4.2 Model Selection

We then performed a Gridsearch over the most used classifiers; decision tree, random forest, logistic regression and xgboost to select the best one. We tested each model with an extensive set of parameters to find the best combination of parameters. We also tested different methods of re-sampling (over-sampling and undersampling). The result table below summarizes the results of the Gridsearch :

Model	Decision tree	Random forest	Logistic Regression	XGBoost
F1-score	0.645	0.634	**0.665**	0.653

4. Results

4.1 Results

As mentioned above, we obtained an F1-score of 66% with the Logistic regression, combined with an undersampling ratio of 0.75. This score is obtained using a 5-fold cross validation.

4.2 Error analysis

Below the confusion matrix that summarizes the errors made by the model for each category (severe and not severe).

	Predicted not severe	Predicted severe
Actually not severe (0)	**27342**	6960
Actually severe (1)	9576	**6003**

The green cells correspond to the correct predictions. 6,960 is the number of accidents the model predicted severe while they were not and 9,576 represents the number of accidents the model predicted not severe while they were.

We could collect more information to improve the model : details on the driver (fatigue, distraction) and his behavior on the road (break of traffic rules). Predicting the severity remains a complicated task, severe accidents also happen when all the conditions are clear.

5. Discussions

The data visualization section provided insights that allow us to make some recommendations to local authorities. First, we noticed that the number of accidents had decreased since 2004. This is surely due to more intense road safety campaigns and that proves that the city should continue its efforts to sensitize the population to road safety. We noted that most accidents happened during rush hours during the week and in the afternoon during the weekends. We also discovered a peak of accidents the weekends during the night. We saw a peak in the number of accidents around 2 am that can be explained by the increasing number of accidents under influence during the weekend. The city could increase police controls at certain locations after 7pm and during the weekend.

The graph displaying the Covid-19 stay-at-home orders showed that the number of accidents has dramatically decreased since the beginning of the lockdown. This decrease is due to the reduction of traffic. The local authorities could consider crating incentives for companies that implement remote working for their employees. The city would encourage companies to let their employees work from home one and reduce traffic in the city.

As said before, road safety campaigns seem to have an impact on the reduction of the number of accidents. However, the graph displaying the number of accidents due to speeding during the Covid-19 lockdown showed that people are driving fast when they can. Road safety campaigns should emphasize the necessity to respect speed limit on the road. Local authorities could also increase the number of speed cameras in the city. Then, road safety campaigns must raise awareness on the growing number of people biking in the city. Accidents that involve pedestrians and cycles are often severe (causing at least an injury). While people are more and more riding bikes around the cities, it is necessary to build safe bike lanes. Seattle has already developed the construction of bike lanes but should increase its number to allow bikes to ride in the entire city. It is also important that authorities accentuate the necessity to use these bike lanes instead of the road when possible, or to have a helmet. Moreover, drivers should be more aware of the number of bikes in the city. It is important that bikes and cars learn to share the road.

Finally, we noticed that accidents happening at intersections are serious (50% of them) and identified the most dangerous intersections. The city can find solutions to reduce the number of accidents happening at intersections. One solution could be to increase the delay between red and green lights. This could allow drivers or pedestrians that cross the road late to cross safely before the other cars start. It could be interesting for the city to study what cause accidents at intersections.

6. Conclusion

Road traffic crashes are a leading cause of deaths and it is a public safety challenge to reduce the number of accidents. We explored the data and have been able to give some recommendations to improve the safety on the roads. It is necessary to find solutions to allow bikes to safely share the road with cars. Some dangerous intersections have been identified and the city can deploy solutions at these locations. Then, we recommended to find incentives that would allow employees to work from home more often.

We built a model that predicts the severity of an accident in order to send real-time alerts to drivers. Accidents are difficult to predict. They can happen anytime and do not need specific conditions. We would need more detailed information to improve our model. It would be interesting to have details on the drivers' behavior and the daily weather conditions in Seattle.