PREDICTING SEVERITY OF CAR ACCIDENT IN THE CITY OF SEATTLE

1. Business problem

In 2015 the Seattle Department of Transportation (SDOT) launched the "Vision Zero" program which aims to end traffic deaths and serious injuries by 2030. The foundation of the project is "traffic collisions aren't accidents - they're preventable through smarter street design, targeted enforcement, partnerships, and thoughtful public engagement".

The 2015 Vision Zero Action Plan states that traffic collisions are a leading cause of death for Seattle residents age 5-24. Older adults are also disproportionately affected, and as its population ages, this trend could grow. In 2013, there were 10,310 police-reported collisions in Seattle. 155 people were seriously injured and 23 were killed.

This report aims to identify how internal and external conditions relate to car collisions and to calculate the probability of getting property damage or get injured in a car accident. This information might be used in street signs to alert drivers and strengthen an attentive driving.

The targets of the project are the Seattle Department of Transportation and other public or private organization working on decrease de number of traffic accidents in Seattle, Washington.

2. Data acquisition and cleaning

2.1. Data source

The “City of Seattle Open Data” portal provides under the Transportation category a variety of data sets which include traffic flow counts, collisions, intersections, marked crosswalks, one-way Streets and others.

For this project we will be working with the Collision dataset. The dataset consists of 40 columns with 221266 number of observations from 2004 to present. There is no duplicate rows and the percentage of missing cells is 15.8%. Across the columns we have different data types: categorical, numerical and Boolean.

Source of dataset: https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d

Detailed description of attributes:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

2.2. Data cleaning

According to the dataset summary provided by SDOT some of the attributes represent a unique key for the collision. For the purpose of this project this kind of information it is of no use. Because of that, the following columns will be deleted from the data set:

Attribute	Description
OBJECTID	ESRI unique identifier
INCKEY	A unique key for the incident
COLDETKEY	Secondary key for the incident
INTKEY	Key that corresponds to the intersection associated with a collision
SEGLANEKEY	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
SDOTCOLNUM	A number given to the collision by SDOT.

The columns with a description of another attribute add no information to the analysis. Because of that, the following columns will be deleted from the data set:

Attribute	Description
LOCATION	Description of the general location of the collision
SEVERITYDESC	A detailed description of the severity of the collision
SDOT_COLDESC	A description of the collision corresponding to the collision code.
ST_COLDESC	A description that corresponds to the state's coding designation.

There are also some columns which are not included in the dataset summary or has no description in it. Also, its values will no represent additional important information for this project. Because of that, the following columns will be deleted from the data set:

Attribute	Description
REPORTNO	A combination of numbers and letters.
EXCEPTRSNCODE	Unique value: NEI
EXCEPTRSNDESC	Not Enough Information
STATUS	Values: Matched - Unmatched

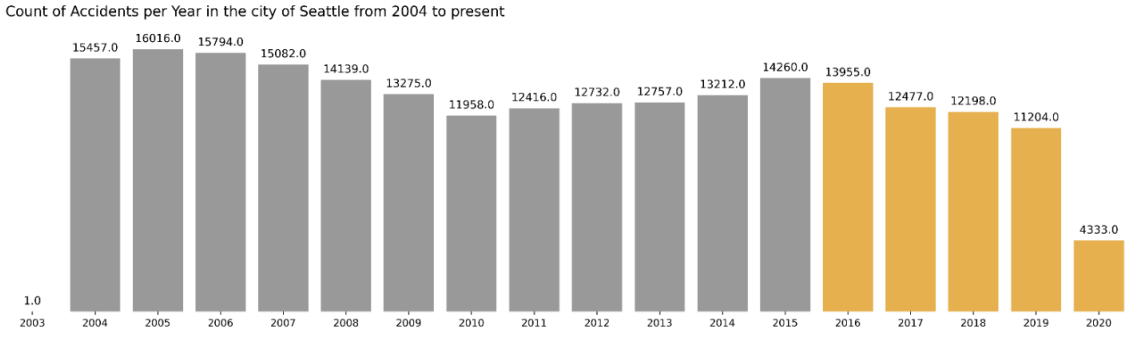
The SDOT_COLCODE columns gives a particular code for every type of the collision. According to the State Collision Code Dictionary there are 84 different types of collision. To simplify the analysis instead of using this column we will be use the COLLISIONTYPE attribute which is a reduced classification of collisions.

To analyse the evolution of collision across the years, months, days and hours we will generate the following columns from the INCDATE and INCDTTM attribute which has values about the date and time of the incident: - YEAR, -MONTH, -DAY, -HOUR. The INCDTTM has no time values for some records because of that, first we need to convert to missing value those entrances with no time information and then create the HOUR feature.

Once generated those new features, we will drop the INCDATE and INCDTTM columns.

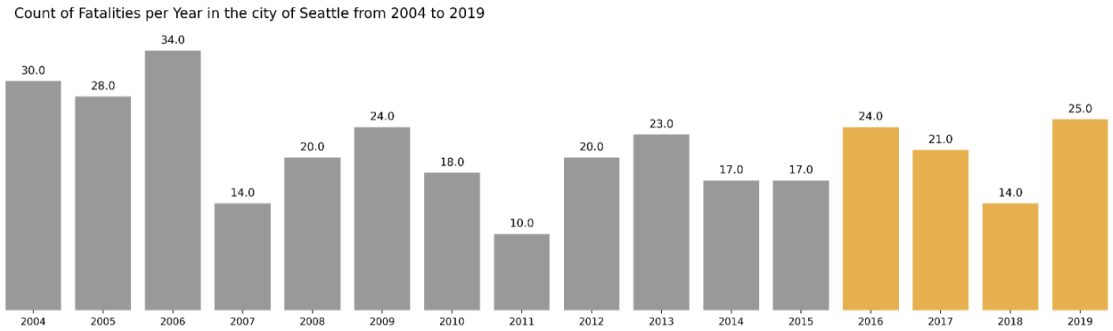
3. Methodology: Exploratory Data Analysis

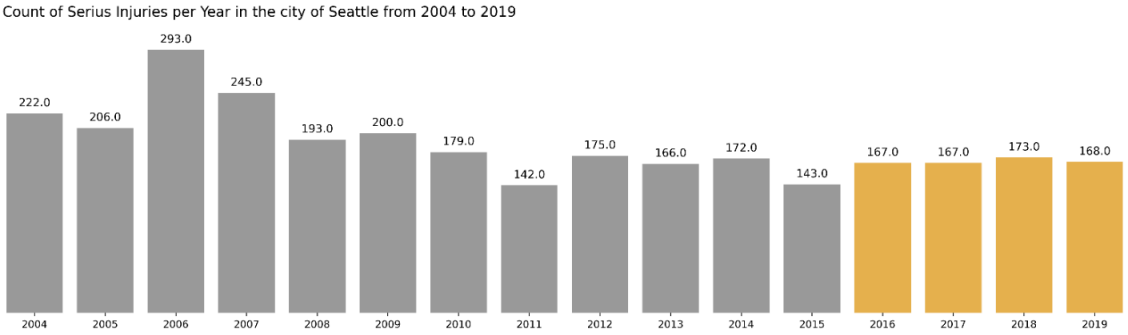
For a better understanding of the problem let us explore the evolution of collision across the years.



First, there is a low value in 2003. Probably this is a wrong entry so we should delete it. From 2016 we can see a reduction tendency of the global number of collisions. The low number from 2020 is probably related to COVID19 and some lockdown and circulation restriction. Because that is an extremely rare situation, we will drop values from 2020 too.

The interest of the Vision Zero project is to reduce the number of death and serious injuries, so let us analyse its evolution across the years.

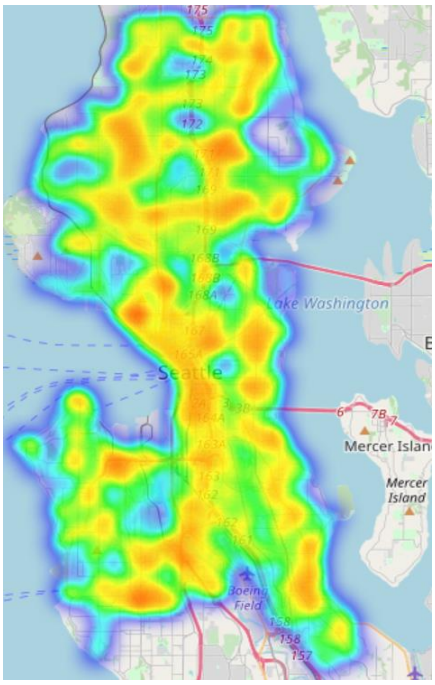




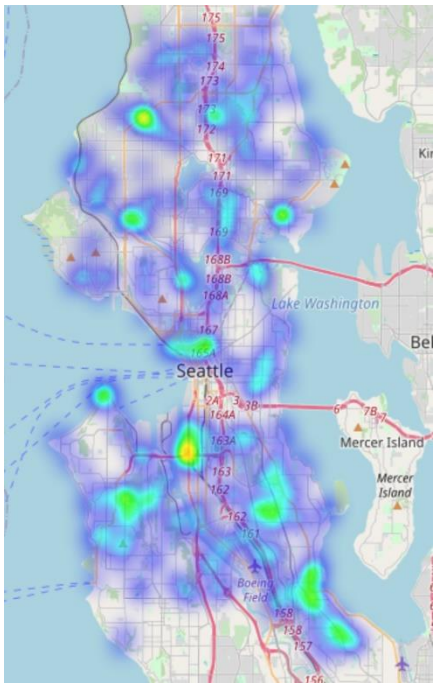
Despite there is a reduction in the number of collisions since 2016, the number of fatalities and cases with serious injuries have no change for the same period. On average, 21 families lost someone who loved because of the fatality of the accident and 675 people got injured.

Let us look to the geographical distribution of those collisions which end in a fatality or serious injury.

SERIOUS INJURY – MAP



FATALITIES - MAP



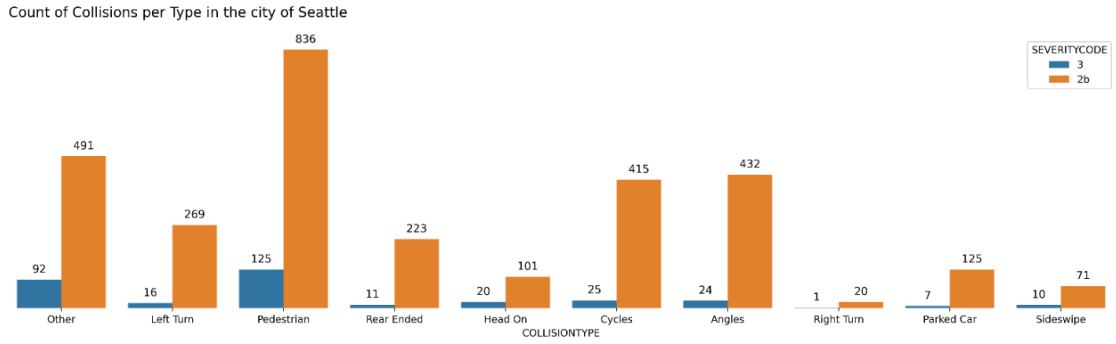
Those type of collisions take place across the city of Seattle. There are some locations with a higher number of occurrences. The following are the top five:

LOCATION	SERIOUS
AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	46
OLSON PL SW BETWEEN 3RD AVE SW AND SW CAMBRIDGE PL	13
RAINIER AVE S BETWEEN S MOUNT BAKER EB BV AND S HANFORD ST	8
BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB	7
AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST	7

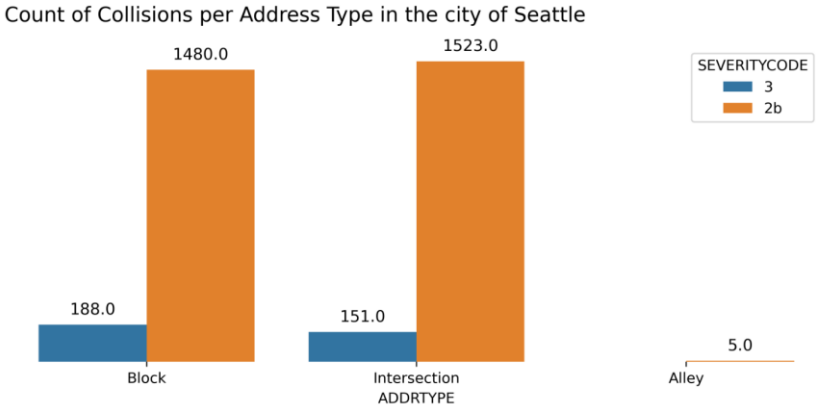
LOCATION	FATALITY
AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	5
ALASKAN WY VI SB BETWEEN ALASKAN WY VI SB EFR OFF RP AND S HOLGATE ST	4
ALASKAN WY VI NB BETWEEN S ROYAL BROUGHAM WAY ON RP AND SENECA ST OFF RP	4
ALASKAN WY VI SB BETWEEN COLUMBIA ST ON RP AND ALASKAN WY VI SB EFR OFF RP	4
RAINIER AVE S BETWEEN 57TH AVE S AND ITHACA PL S	4

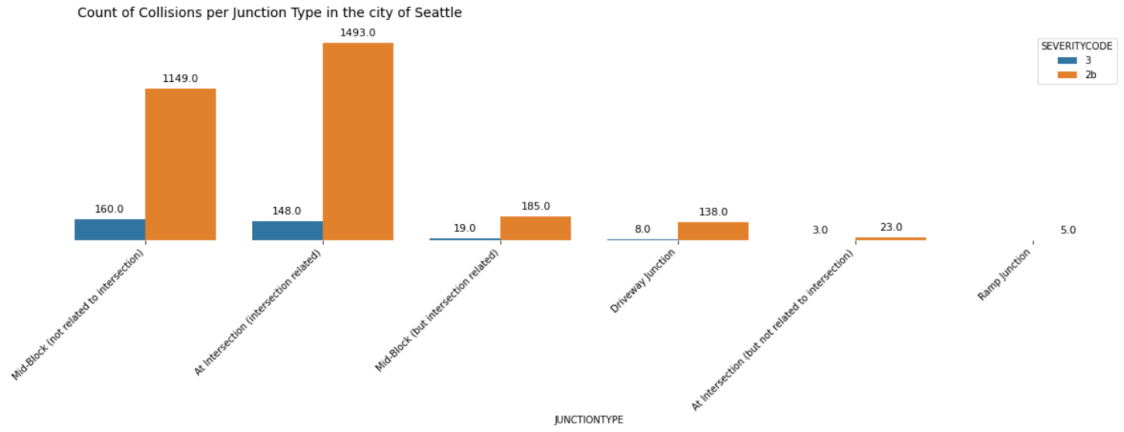
Because the aim of the report is to generate insights for the Vision Zero project or for another institution which works to decrease the number of fatalities and serious injuries, from now and on we will analyse those types of collisions.

Let us look what are the types of collisions.



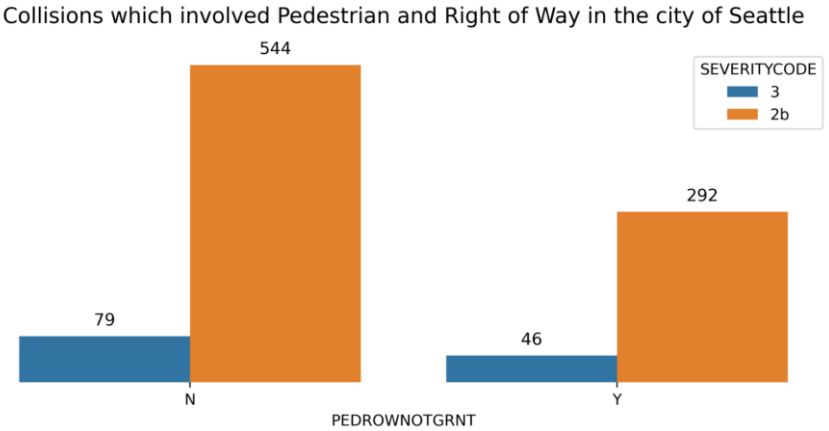
The first type of collision which cause a serious injured or fatality involve a pedestrian. Also, bicycles occupied the 3rd position. Let us see the collision address type and the category of junction at which collision took place.



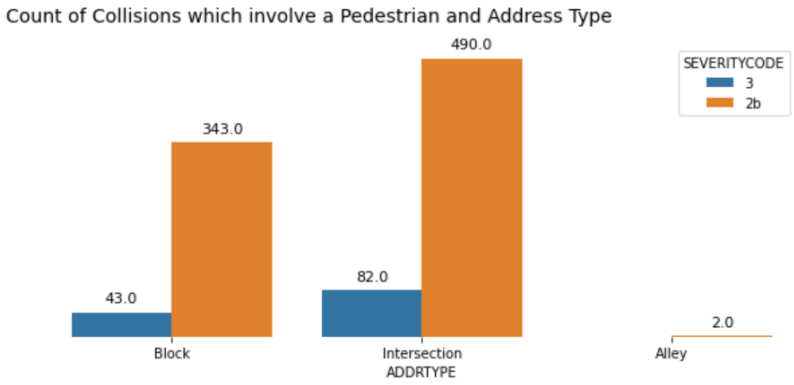


It would be interesting to analyse the number and position of traffic light across the city of Seattle and the existence of crosswalk and bicycle line. But this is not the scope of this project. It would be a theme for future analysis. Nevertheless, the PEDROWNOTGRNT feature could give us some insight about it.

This feature content a high number of missing data. An analysis of its values reveals that the only no null value is Y = yes. So, first we need to replace those missing values to N = no.



According to the Washington State Department of Transportation drivers and bicyclists must yield to pedestrians on sidewalks and in crosswalks. But every pedestrian crossing a roadway at any point other than within a marked crosswalk or within an unmarked crosswalk at an intersection shall yield the right of way to all vehicles upon the roadway. From the graph we can say that the most cases occurs when the pedestrian right of way is not granted. We need a deeper analysis to understand if the collision is due to inattention of the pedestrian when crossing the street.
Let us analyse again the distribution of collision per address type but in this case only for collisions which involve a pedestrian.

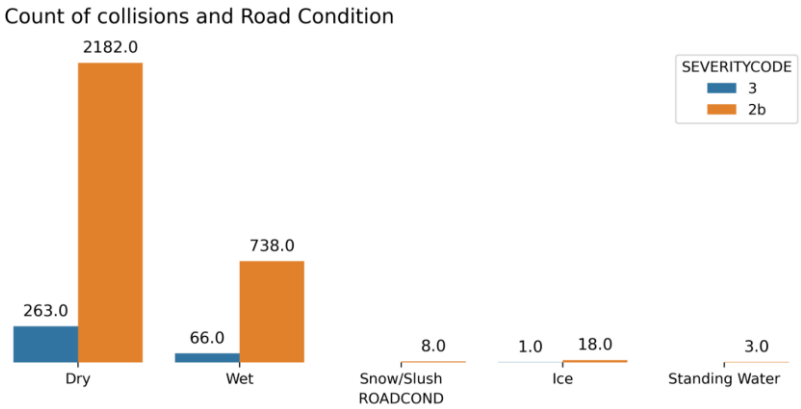


Most cases occur at intersection. Again it would be valuable to analyse the existence of crosswalks and traffic lights at intersections.

Let us see now other external condition that may affect the driving like road condition, weather and level of light.

Road Condition

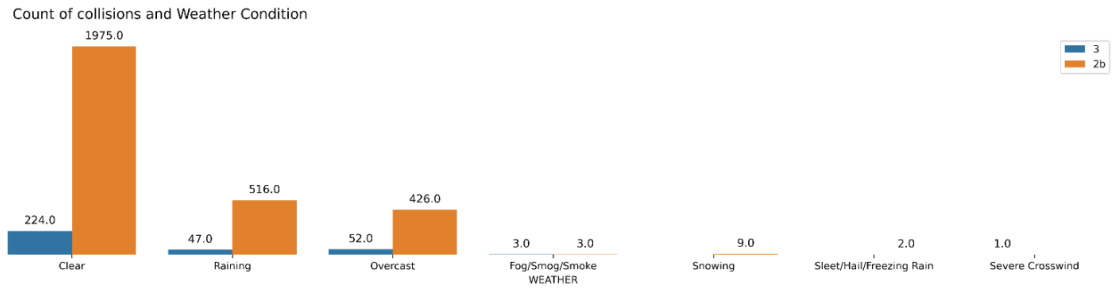
First, we need to replace the Other values and Unknown values to missing data because it does not add useful information.



Most of the time the road condition is dry, so it seems that this feature is not the main cause of collisions. For fatality cases, 20% of them occur under wet road.

Weather condition

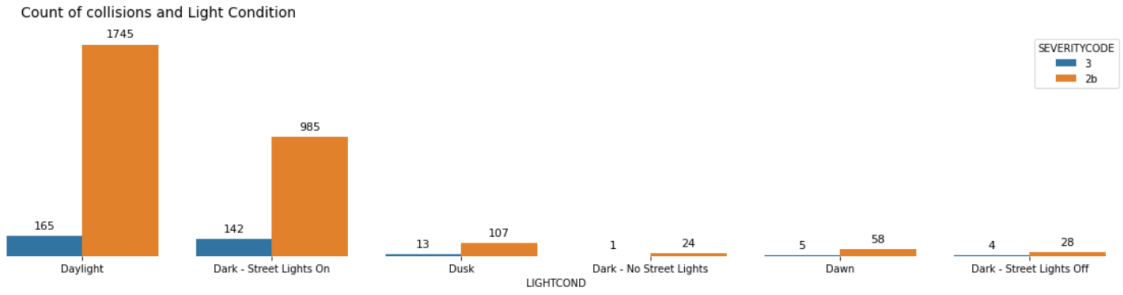
First, we need to replace the Other values and Unknown values to missing data because it does not add useful information.



In most cases the collisions occur under clear weather. It makes sense because we see this tendency under dry road. For fatality cases, 30% of them occur under raining or overcast condition.

Light Condition

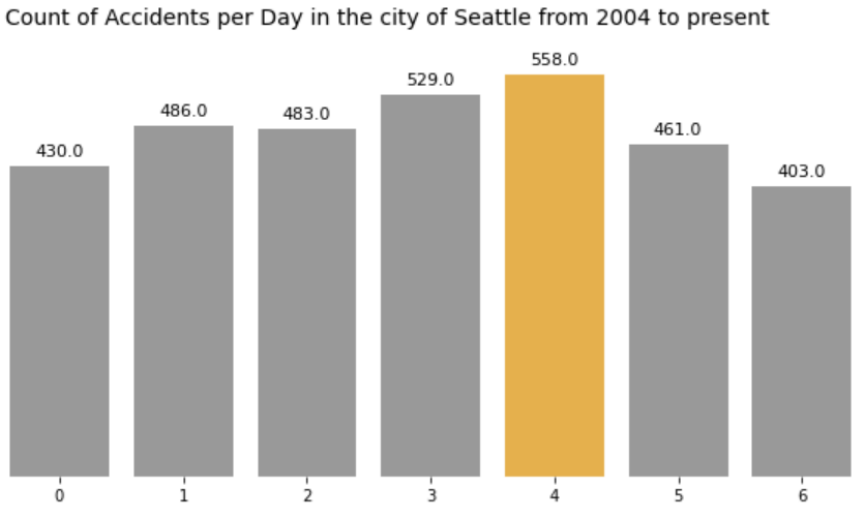
First, we need to replace the Other values and Unknown values to missing data because it does not add useful information.

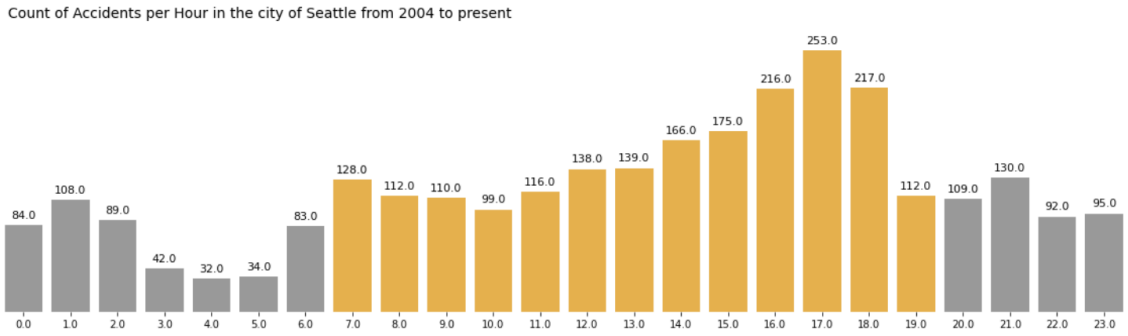


Most cases occur under daylight and at night with streets lights on.

Time

Let us see if it is a tendency across days, hours.



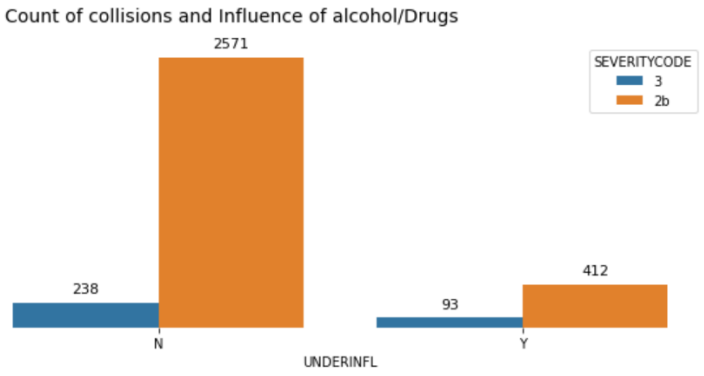


In most cases the collisions occur on Friday. On the other hand, the 75% of cases occur between 7 A.M and 19 A.M. It is related with what we saw in Light Condition chart.

Let us analyse some internal conditions which may be related with collisions.

Under Influence of alcohol or drugs

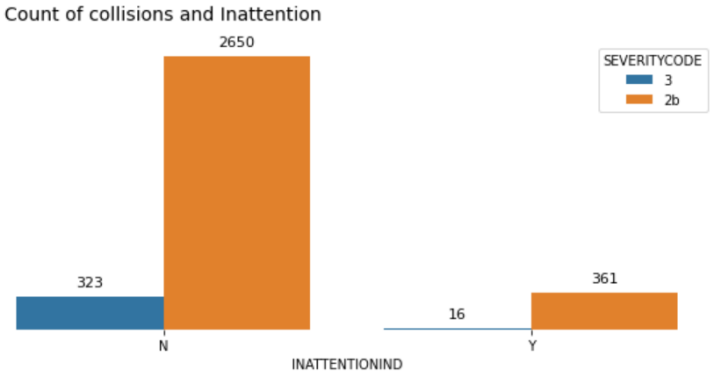
The values of this feature are Y and 1 for yes, N and 0 for no. Let us replace 1 to Y and 0 to N.



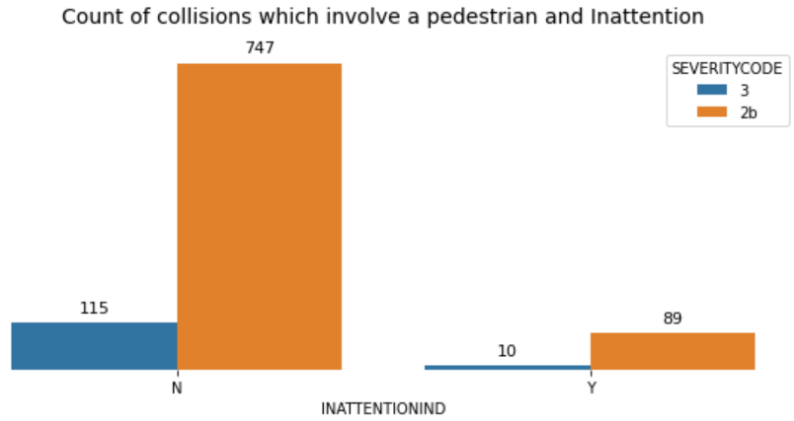
The 28% of cases which lead on fatality were related to a driver under influence of alcohol or drugs.

Inattention

This feature content a high number of missing data. An analysis of its values reveals that the only no null value is Y = yes. So, first we need to replace those missing values to N = no.

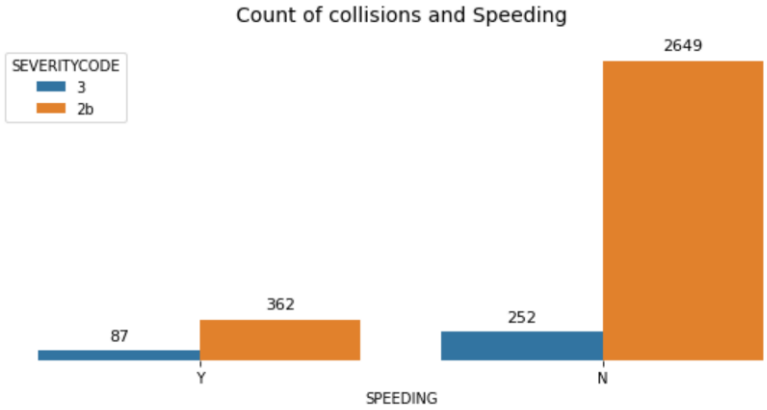


Most of the cases, the collision was not due to inattention. But let us take a better look for the cases where a pedestrian is involved.



We can see the same distribution of cases.

Speeding



The 30% of cases which lead on fatality were related to speeding.

4. Modelling

The second part of this project aims to develop a model which predict the severitycode and calculate its probability. Let us take a better look of the target variable on the entire data set.

SEVERITYCODE	%
0	9.8
1	62.20
2	26,5
2b	1.4
3	0.015

For SEVERITYCODE equals to 0 we have not a detailed description of the severity of the collision or collision type. Because of that we should drop every line of data with SEVERITYCODE == 0. We will do the same with NaN values.

On the other hand, we will focus on SEVERITYCODE equals to 2b and 3b. So, to simplify the analysis let us replace 1 and 2 categories to 0 and 2b, 3 to 1.

The new distribution looks like this:

SEVERITYCODE	%
0	98.2
1	1.7

This distribution of values represents an unbalanced dataset. In this situation, the predictive model using conventional machine learning algorithms could be biases and inaccurate. A widely adopted technique for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (SEVEIRYCODE == 0) (under-sampling) and / or adding more examples from the minority class (SEVERITYCODE == 1) (over-sampling). Here we will evaluate both methodologies.

Using simpler metrics like accuracy_score can be misleading. In a dataset with highly unbalanced classes, if the classifier always "predicts" the most common class without performing any analysis of the features, it will still have a high accuracy rate. Instead to evaluate the results, we will use a confusion matrix, which shows the correct and incorrect predictions for each class. In the first row, the first column indicates how many classes 0 were predicted correctly, and the second column, how many classes 0 were predicted as 1. In the second row, we note that all class 1 entries were erroneously predicted as class 0.

Feature Selection

After the analysis of collision cases under different internal and external conditions we will select the following features:

UNDERINFL: 28% of fatality cases occur under this condition.

SPEEDING: 30% of fatality cases are related to speeding.

WEATHER: For fatality cases, 30% of them occur under raining or overcast condition.

ROADCOND: For fatality cases, 20% of them occur under wet road.

LIGHTCOND: 40% of cases with SEVERITYCODE 2b and 3 occur at dark with lights on.

PEDROWNOTGRNT: 51% of cases with SEVERITYCODE 2b and 3 occur when pedestrian right of way were not granted.

ADDRTYPE: Most cases occur at intersection for pedestrian collision type.

JUNCTIONTYPE: most cases occur at midblock or at intersection.

COLLISIONTYPE: for SEVERITYCODE 2b and 3 the main collision cases are related to pedestrian, cyclist and collision on angles.

Modelling

Because of the lack of balance on the dataset we will evaluate three different method to counter its effect and then applied a logistic regression model. The following technics will be evaluated:

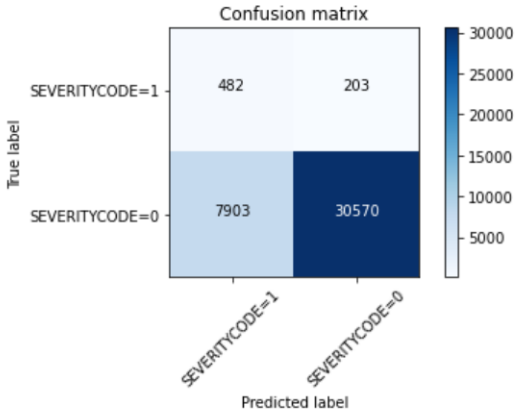
1. Upsampling with resample from sklearn
2. Downsampling with resample from sklearn
3. SMOTE

UPSAMPLING

In upsampling, for every observation in the majority class, we randomly select an observation from the minority class with replacement. The result is the same number of observations from the minority and majority classes.

After splitting the dataset into train and test set. We need to perform an upsampling of the minority class which in this case is SEVERITYCODE equals to 1.

Then we will fit our data into a Logistic Regression model. The confusion matrix generated for this case it is represented in the following chart.

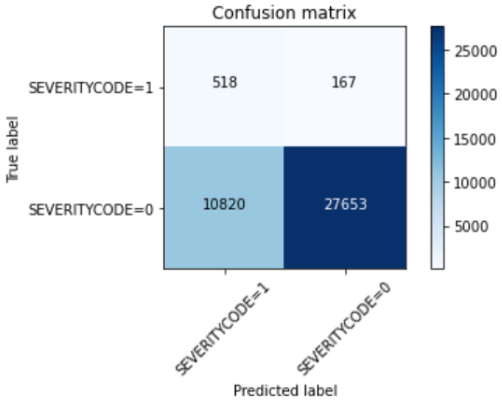


DOWNSAMPLING

Down-sampling involves randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm.

After splitting the dataset into train and test set. We need to perform an downsampling of the majority class which in this case is SEVERITYCODE equals to 0.

Then we will fit our data into a Logistic Regression model. The confusion matrix generated for this case it is represented in the following chart.

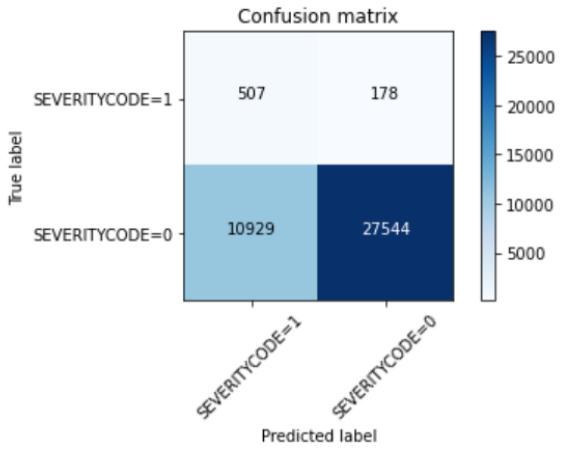


SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a type of over-sampling procedure that is used to correct the imbalances in the groups. This technique creates new data instances of the minority groups by copying existing minority instances and making small changes to them.

After splitting the dataset into train and test set. We need to perform an upsampling of the minority class which in this case is SEVERITYCODE equals to 1.

Then we will fit our data into a Logistic Regression model. The confusion matrix generated for this case it is represented in the following chart.



5. Results and discussion

For the last five years we can see a downward trend for global collisions in the city of Seattle. Probability it is related to the plan Vision Zero were the core of the project is the belief that death and injury on city streets is preventable. However, we cannot see the same tendency for fatality or serious injured cases.

In the first part of the project we analysed how different internal and external condition are related to collisions with SEVERITYCODE equals to 2b and 3. In most cases those collisions are related to pedestrian and cyclist. It makes sense because they have no kind of material protection but ¿why they are involved in accidents? For pedestrians, multiple times the right of way is not granted. To solve this puzzle, we need a deeper understanding of the city design. We need to answer questions like, ¿Are there enough crosswalks? / ¿Are there enough traffic lights?

Also, for fatalities cases we see collisions were the driver was speeding or under influence of alcohol or drugs. Probably we need to focus on speed limits and how the city monitors this and analyse the necessity of implement an alcohol control on Friday evening.

Regarding modelling selection, we generate three different models. According to the confusion matrix charts we can conclude that the best performance we obtain under the downsampling model. Nevertheless, we need to improve the model because for both SEVERITYCODES, approximately 25% of entrances were predicted wrong.

6. Conclusion

The purpose of the project was to generate valuable insights to help decrease the number of collisions which end into fatalities and severity injuries.

Most of the times the cause of them is bad human behaviour and the first step to eradicate them is to be aware of the consequences they might bring. Much of the information produced in the project may be useful for yard signs and for awareness campaigns. On the other hand, this report invites us to go deeper into the city design to improve the experience of the pedestrian and cyclists on the street and make their travel to home safer.

The modelling will help to predict situations were the probability of a collision is higher and act preventively.