A DATA ANALYSIS OF TRAFFIC ACCIDENTS IN SEATTLE, WA

Applied Data Science Capstone for the IBM Data Science Professional Certificate by IBM/Coursera

Marcelo Camera Oliveira

november - 2020

TABLE OF CONTENTS

1. Business Understanding

The increasing number of cars on the roads brings with it a worrying reality: the increase of the accident rate. Such accidents cause enormous consequences, the most important of which, of a human nature, is the loss of life. Other effects appear in the ride of this reality, one of which is the financial loss, due to the long traffic jams and roadblocks, which have a negative impact on the logistics of goods.

In parallel, it is necessary to contextualize the importance of Information and Communication Technology (ICT) in solving urban problems. Given the worldwide technological advancement, cities are becoming increasingly “smart”. The terminology "Smart Cities" basically goes back to the concept of a city that makes use of ICT, through various physical devices connected to the IoT (Internet of Things) network, in order to optimize its operations and services, in addition to connecting citizens.

Faced with such a problem, it is suggested the development of a mathematical model that, in view of the knowledge of initial situations, a risk classification should be carried out. From this perspective, given the driver's entry into a certain road, he is given knowledge of the degree of risk to which he will be subject and, thus, the necessary preventive measures can be taken, such as: changing his route, reducing the car speed, or increasing your attention.

Thus, having the knowledge of historical accident data in a given location and the creation of a mathematical model based on this data, a city could use this tool to implement data capture devices (rain, traffic, object recognition on the roads) with the purpose of issuing alerts to the drivers (electronic boards, mobile phone applications, among others).

1.1. The Stakehoders

This project aims to inform to the Seattle city drivers (audience) about the conditions of the road he will be joining. In this way, he will be able to analyze the situation in order to take the necessary preventive measures.

2. The Data

For this project it was used the data of the City of Seattle Open Data Portal (https://data.seattle.gov/). This data contains several records of traffic accidents in the city of Seattle, USA, from 2004 to the present.

2.1. Dataset Download

The dataset was downloaded from this URL:

https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions

2.2. Data Sumary

Title	Collisions — All Years
Abstract	All collisions provided by SPD and recorded by Traffic Records.
Description	This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
Update Frequency	Weekly.
Keyword(s)	SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
Contact Organization	SDOT Traffic Management Division, Traffic Records Group.
Contact Person	SDOT GIS Analyst
Contact Email	DOT_IT_GIS@seattle.gov

2.3. Attribute Information

The complete attribute information for this data source can be found via this link:

https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

Analyzing the available data dictionary, with focus on the problem, it was clear which features we will necessary. These were the chosen ones:

ATTRIBUTE	DESCRIPTION
OBJECTID	Unique record identifier;
X	Longitude - Geographic coordinate;
Y	Latitude - Geographic coordinate;
ADDRTYPE	Collision address type: Alley, Block or Intersection;
LOCATION	Description of the general location of the collision;
SEVERITYCODE	Target variable. Severity of the collision (3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown);
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type;
INJURIES	The number of total injuries in the collision;

SERIOUSINJURIES	The number of serious injuries in the collision;
FATALITIES	The number of fatalities in the collision;
INCDTTM	The date and time of the incident;
JUNCTIONTYPE	Category of junction at which collision took place;
INATTENTIONIND	Whether or not collision was due to inattention;
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol;
WEATHER	A description of the weather conditions during the time of the collision;
ROADCOND	The condition of the road during the collision;
LIGHTCOND	The light conditions during the collision;
SPEEDING	Whether or not speeding was a factor in the collision;
HITPARKEDCAR	Whether or not the collision involved hitting a parked car.

The target variable will be the "SEVERITYCODE", which indicates the severity of the accident.

3. Methodology

For this report, it was used the Cross-Industry Standard Process for Data Mining (CRISP-DM) which consist of the following steps:

1. **Business Understanding:** The initial phase is to understand the project's objective from the business or application perspective.

2. **Data understanding:** In this phase, the dataset is downloaded and filtered by the attributes (columns) that we will use.

3. **Data Preparation:** The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

4. **Modeling:** In this phase, various algorithms and methods will be tested to build the model including supervised machine learning techniques.

5. **Evaluation:** Before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the application's objectives are achieved. Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others.

4. Analysis

4.1. Data Acquisition

The first step was to download and to load the data into a dataframe, with the attributes already chosen.

4.2. Data Understanding

	X	Y	OBJECTID	ADDRTYPE	LOCATION	SEVERITYCODE	SEVERITYDESC	COLLISIONTYPE	INJURIES	SERIOUSINJURIES	FATALITIES	INCDTTM	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	HITPARKEDCAR
0	-122.315658	47.675815	1	Intersection	12TH AVE NE AND NE 65TH ST	2	Injury Collision	Pedestrian	1	0	0	2020-03-09 11:34:00	At Intersection (intersection related)	NaN	N	Clear	Dry	Daylight	NaN	N
1	-122.316780	47.608643	2	Block	12TH AVE BETWEEN E CHERRY ST AND E COLUMBIA ST	1	Property Damage Only Collision	Sideswipe	0	0	0	2013-03-27 14:02:00	Mid-Block (not related to intersection)	NaN	N	Raining	Wet	Daylight	NaN	N
2	-122.344569	47.694547	3	Block	AURORA AVE N BETWEEN N 90TH ST AND N 91ST ST	2	Injury Collision	Rear Ended	1	0	0	2013-03-29 14:47:00	Mid-Block (not related to intersection)	NaN	N	Clear	Dry	Daylight	NaN	N
3	-122.365999	47.691729	4	Block	8TH AVE NW BETWEEN NW 86TH ST AND NW 87TH ST	0	Unknown	NaN	0	0	0	2019-08-10 00:00:00	Mid-Block (not related to intersection)	NaN	NaN	NaN	NaN	NaN	NaN	Y
4	NaN	NaN	5	Block	ALASKAN WY VI NB BETWEEN S ROYAL BROUGHAM WAY ...	1	Property Damage Only Collision	Other	0	0	0	2004-12-23 00:20:00	Mid-Block (not related to intersection)	NaN	1	Clear	Dry	Dark - Street Lights On	NaN	N

The dataset has presented 222,581 records and 20 columns (target variable and attributes).

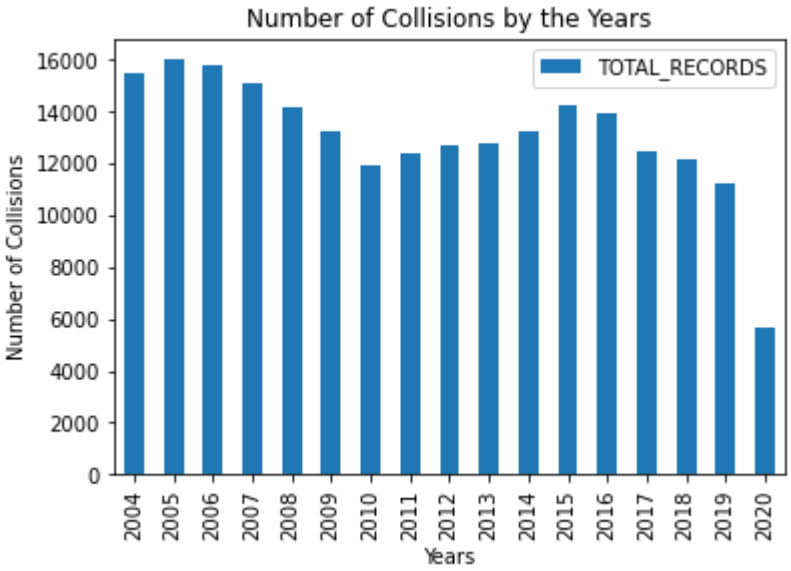
4.3. Preliminary Data Preparation

The attributes "INATTENTIONIND" and "SPEEDING" attracted our attention. There were just a few registers of them: 9,982 registers for the "SPEEDING" attribute and 30,195 for the "INATTENTIONIND" atrribute. This amounts to of 4.5% and 13.5% of the data, respectively. Because of this few quantity of records, we choose to delete them.

The values of the "UNDERINFL" attribute had some problem. All the zero's values were transformed to "N" and all the one's values to "Y".

Were added columns for years and months values, to expand our analysis.

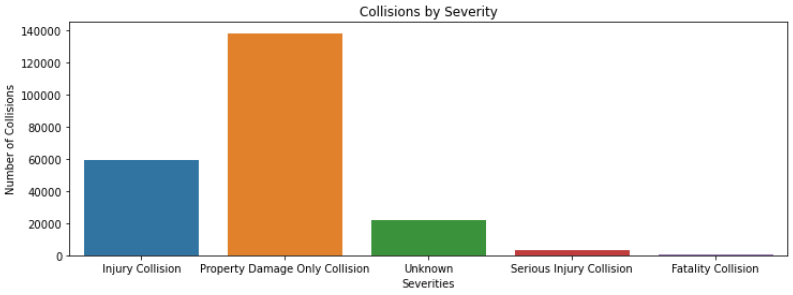
Were identified that there was only one record for all the 2003 year. So, this information was deleted. The graph below shows the number of accidents over the years 2004 to 2020.



The first record was on 2004-01-01, and the last was on 2020-11-04.

4.4. Exploratory Data Analysis

In this section some analysis were made from the dataset.



	SEVERITYDESC
Property Damage Only Collision	138252
Injury Collision	59058
Unknown	21781
Serious Injury Collision	3131
Fatality Collision	358

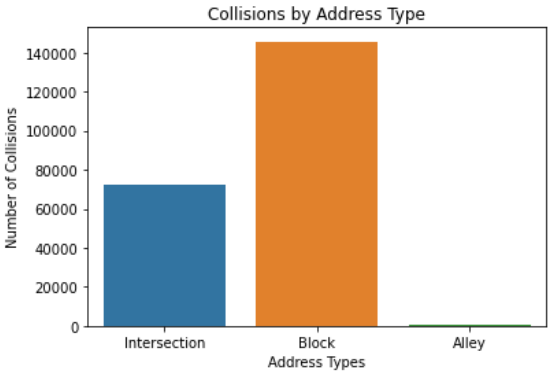
The attribute SEVERITYDESC is our target variable. It will be important when we are modeling our machine learning algorithm. He tells us how serious an accident was.

The attribute SEVERITYDESC is the same as SEVERITYCODE. The difference is that the SEVERITYCODE contains the records in numerical values.

The great majority of recorded accidents are damage to property. Second, with more than twice as many records as first place, there are injuries accidents.

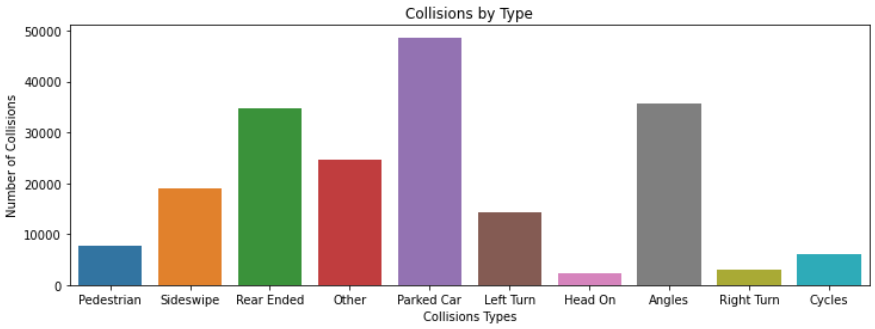
There are a significant number of accidents with unidentified severity. We will have to decide what to do with these records during the data preparation stage to create the machine learning model.

Finally, there are accidents with few records, which are those with serious severity and fatal accidents.



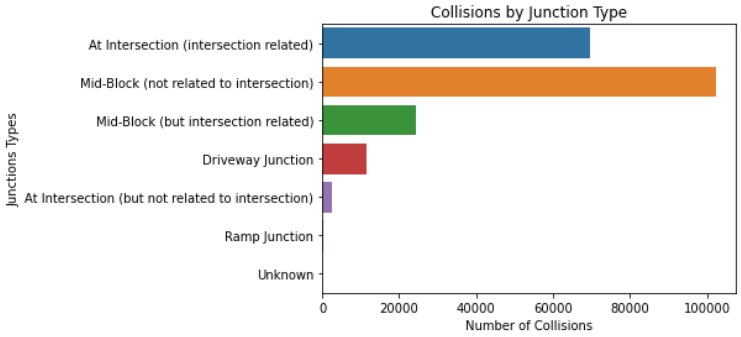
ADDRTYPE	
Block	145679
Intersection	72299
Alley	879

There are twice as many accidents at the Blocks as at the intersections. There are just a few accident records at Alleys.



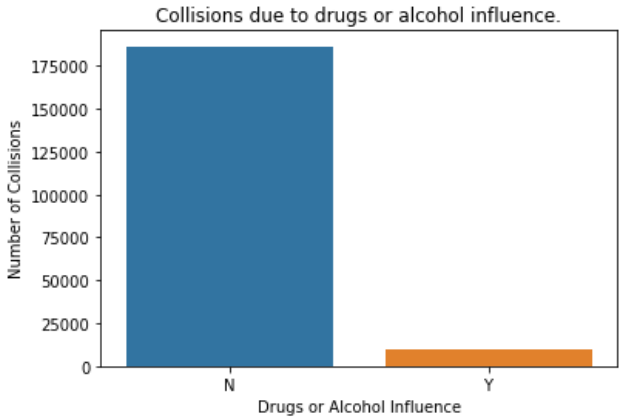
COLLISIONTYPE	
Parked Car	48673
Angles	35759
Rear Ended	34825
Other	24699
Sideswipe	18962
Left Turn	14173
Pedestrian	7698
Cycles	5958
Right Turn	3029
Head On	2205

Collisions with parked cars are the most common. Secondly, accidents at the rear end and at angles are tied. Approximately 25,000 records have not been categorized. They are responsible for the fourth most common type of collision. It will discuss later what to do with this missing data.



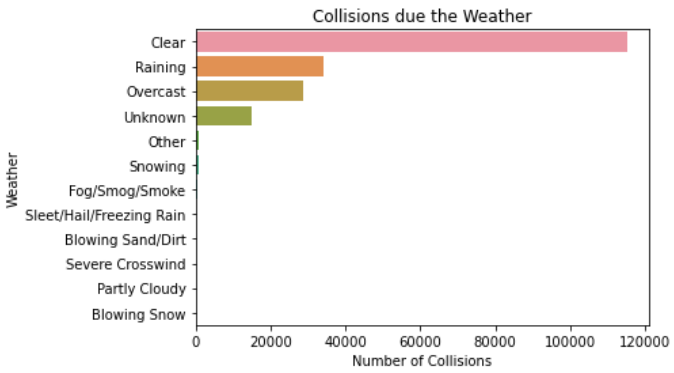
JUNCTIONTYPE	
Mid-Block (not related to intersection)	102327
At Intersection (intersection related)	69588
Mid-Block (but intersection related)	24453
Driveway Junction	11499
At Intersection (but not related to intersection)	2501
Ramp Junction	190
Unknown	21

The analysis of this attribute seems to confirm the previous analysis on accidents by type of address. Accidents at intersections and blocks are the most evident.



UNDERINFL	
N	186372
Y	9629

The influence of drugs or alcohol beverages is not the main cause of accidents.

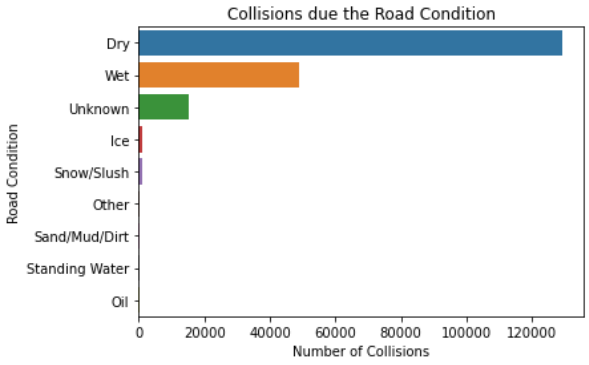


WEATHER	
Clear	115305
Raining	34096
Overcast	28640
Unknown	15131
Snowing	919
Other	867
Fog/Smog/Smoke	623
Sleet/Hail/Freezing Rain	116
Blowing Sand/Dirt	56
Severe Crosswind	26
Partly Cloudy	10
Blowing Snow	1

The great majority of accidents occur in clean weather.

The second factor is in rainy times, but it is still at least three times less frequent than in clean times.

The other types of weather have just a few records.

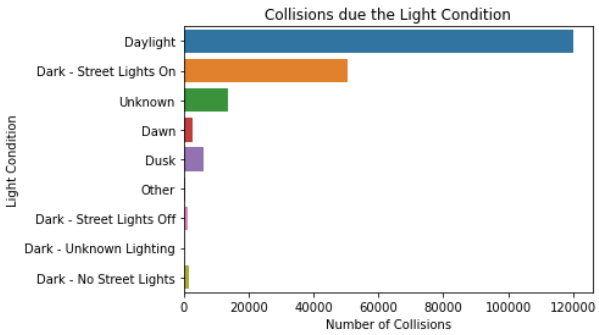


ROADCOND	
Dry	129251
Wet	48828
Unknown	15149
Ice	1233
Snow/Slush	1014
Other	136
Standing Water	119
Sand/Mud/Dirt	77

The great majority of accidents occur in dry roads.

The second factor is in wet roads, but it is still at least twice less frequent than in dry roads.

The other types of road conditions have just a few records.

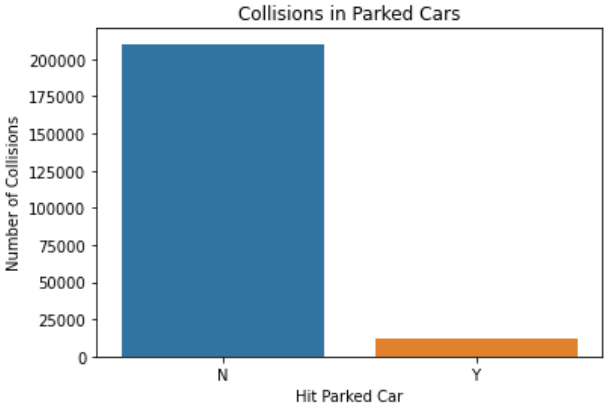


	LIGHTCOND
Daylight	120015
Dark - Street Lights On	50321
Unknown	13539
Dusk	6113
Dawn	2616
Dark - No Street Lights	1581
Dark - Street Lights Off	1240
Other	245
Dark - Unknown Lighting	31

The great majority of accidents occur on daylight.

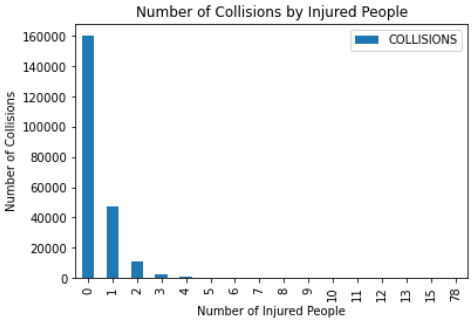
The second factor is on dark, with streetlights, but it is still at least twice less frequent than on daylights.

The other types of light conditions have just a few records.



	HITPARKEDCAR
N	210285
Y	12295

The great majority of accidents does not involve parked cars.

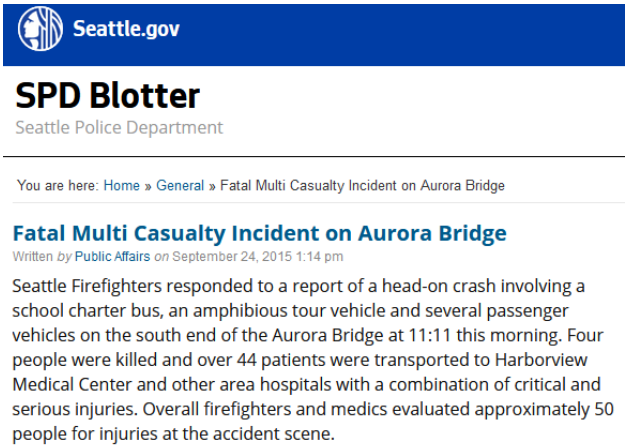


INJURIES	COLLISIONS
0	160273
1	47554
2	10737
3	2740
4	823
5	275
6	100
7	40
8	12
9	10
10	6
11	5
12	1
13	2
15	1
78	1

The great majority of accidents had no injuries.

At first, it seems that we have an outlier. There is a record of 78 wounds. Maybe it's an error in the accident record?

After researching on the Internet about the address and date of the accident, we found that there was really a major accident involving a school bus, an amphibious tour vehicle and several passenger vehicles!



Seattle.gov

SPD Blotter

Seattle Police Department

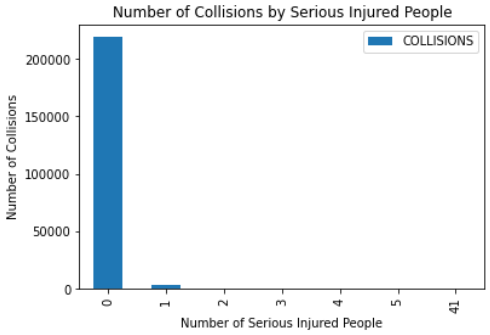
You are here: Home » General » Fatal Multi Casualty Incident on Aurora Bridge

Fatal Multi Casualty Incident on Aurora Bridge

Written *by* Public Affairs *on* September 24, 2015 1:14 pm

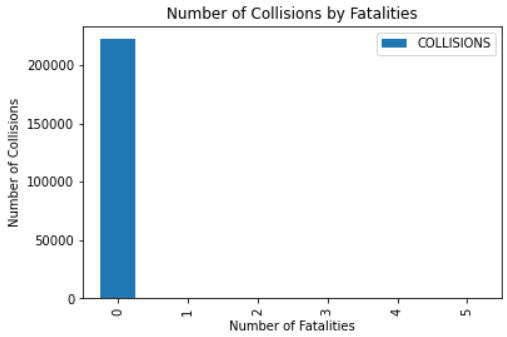
Seattle Firefighters responded to a report of a head-on crash involving a school charter bus, an amphibious tour vehicle and several passenger vehicles on the south end of the Aurora Bridge at 11:11 this morning. Four people were killed and over 44 patients were transported to Harborview Medical Center and other area hospitals with a combination of critical and serious injuries. Overall firefighters and medics evaluated approximately 50 people for injuries at the accident scene.

Fatal Multi Casualty Incident on Aurora Bridge (https://spdblotter.seattle.gov/2015/09/24/fatal-multi-casualty-incident-on-aurora-bridge/)



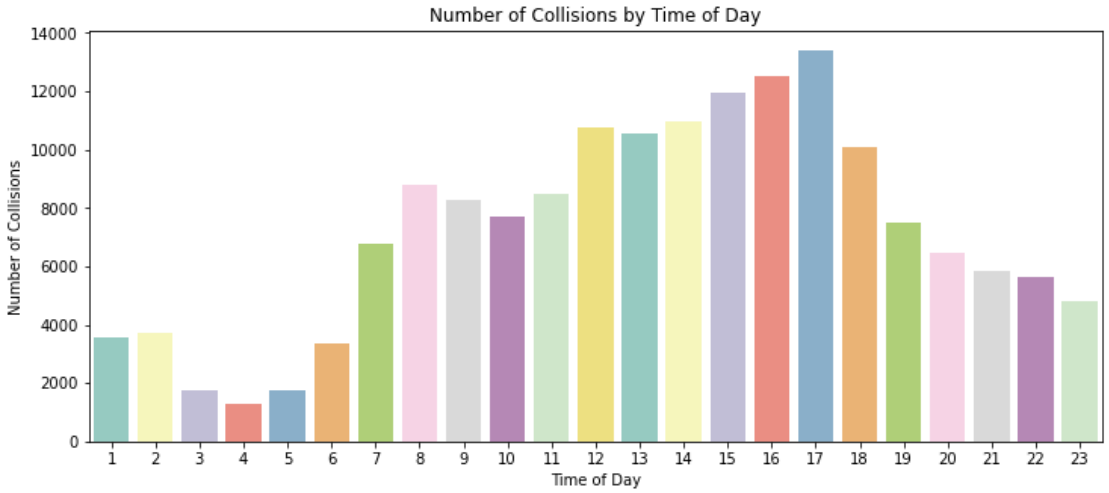
SERIOUS_INJURIES	COLLISIONS
0	219447
1	2964
2	134
3	23
4	6
5	5
41	1

The great majority of accidents had no serious injuries. The record with 41 serious injuries is that what we had reported before.



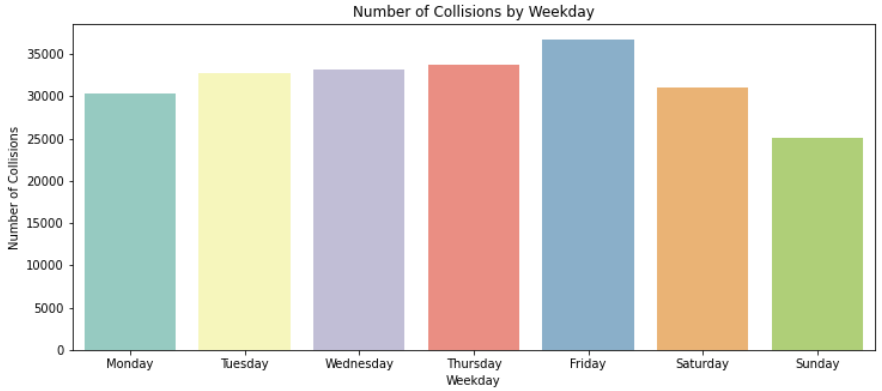
FATALITIES	COLLISIONS
0	222222
1	341
2	13
3	2
4	1
5	1

The great majority of accidents had no fatalities.



Less accidents happen during the night between 1 am and 5 am.

At 6 am, the number of accidents starts to increase, peaking at 5 pm, when it starts to decrease.



WEEKDAY	COLLISIONS
Friday	36710
Monday	30266
Saturday	31039
Sunday	25034
Thursday	33740
Tuesday	32689
Wednesday	33102

The number of accidents per weekday is close one each other. here seems to be a small increase on Fridays. Sunday is the day with the least amount of accidents reported.

	COLLISIONS
BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB AND AURORA AVE N	298
N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N	297
BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB	291
AURORA AVE N BETWEEN N 117TH PL AND N 125TH ST	283
6TH AVE AND JAMES ST	277
AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST	273
RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLELLAN ST	260
ALASKAN WY VI NB BETWEEN S ROYAL BROUGHAM WAY ON RP AND SENECA ST OFF RP	256
ALASKAN WY VI SB BETWEEN COLUMBIA ST ON RP AND ALASKAN WY VI SB EFR OFF RP	230
WEST SEATTLE BR EB BETWEEN ALASKAN WY VI NB ON RP AND DELRIDGE-W SEATTLE BR EB ON RP	225

These are the 10 places that most accidents occurred.

4.5. Feature Selection

As the focus of this project is to create a mathematical model in which the input variables can be obtained through electronic mechanisms (devices), we will reduce the number of attributes. We have identified the following attributes as possible for use:

ATTRIBUTE	DESCRIPTION
ADDRTYPE	Collision address type: Alley, Block or Intersection;
SEVERITYCODE	Target variable. Severity of the collision (3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown);
COLLISIONTYPE	Collision type;
JUNCTIONTYPE	Category of junction at which collision took place;
WEATHER	A description of the weather conditions during the time of the collision;
ROADCOND	The condition of the road during the collision;
LIGHTCOND	The light conditions during the collision;

4.6. Data Cleaning

The dataset has 222580 records and 7 columns (target variable and attributes). It was found some missing values.

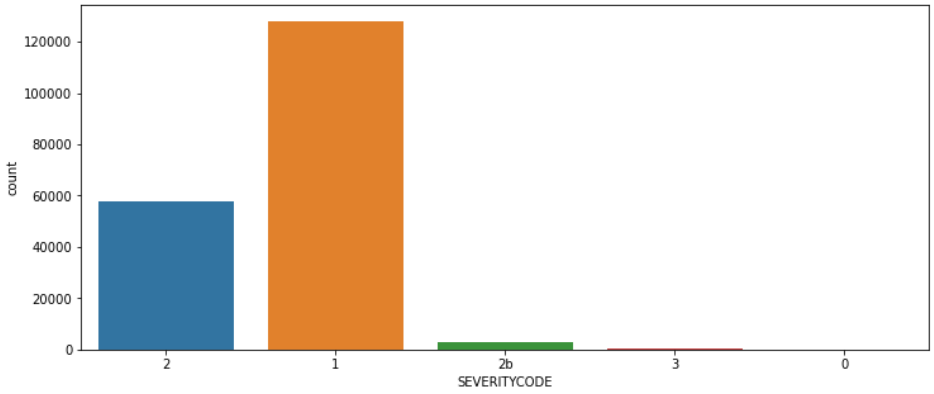
ADDRTYPE	3723
SEVERITYCODE	1
COLLISIONTYPE	26599
JUNCTIONTYPE	12001
WEATHER	26790
ROADCOND	26709
LIGHTCOND	26879

It was decided to drop all the records that has missing values.

4.7. Data Balancing

The target variable is the attribute "SEVERITYCODE", which indicates the severity of the accident. The dataset has these types of severity:

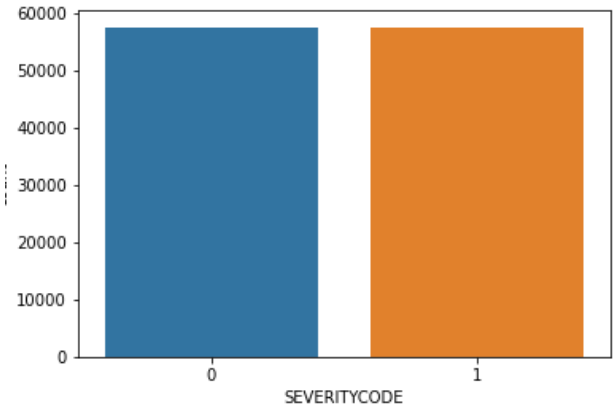
CODE	SEVERITY
1	Property Damage Only Collision
2	Injury Collision
2b	serious injury Collision
3	Fatality Collision
0	Unknown



	SEVERITYCODE
1	128027
2	57592
2b	3050
3	342
0	2

As we can see, there are twice as much "Property Damage Only Collision" data as "Injury Collision" data. The other types of severities won't help us because of too few records. Because of this we had to remove they.

This situation could cause a bias in the data results. So, we equalize this data. This was the result:




```
1    57592
0    57592
```

4.8. Data Binning

In order to improve the effectiveness and accuracy of predictive models, we have grouped the common values of the attributes.

4.9. Data Split

In this section the data was splitted in main blocks: data training and data test. It was splitted in a 70 (training) / 30 (test) ratio.

The data training was splitted again in other two blocks: data training and data evaluation.

```
X data train has shape: (64502, 6)
Y data train has shape: (64502,)
X data validation  has shape: (16126, 6)
Y data validation  has shape: (16126,)
```

4.10.Modelling

Three different algorithms were modeled in order to verify which one will be the most suitable for solving the proposed problem. Are they:

- Decision Tree
- K Nearest Neighbors
- Logistic Regression

4.10.1. Decision Tree Algorithm

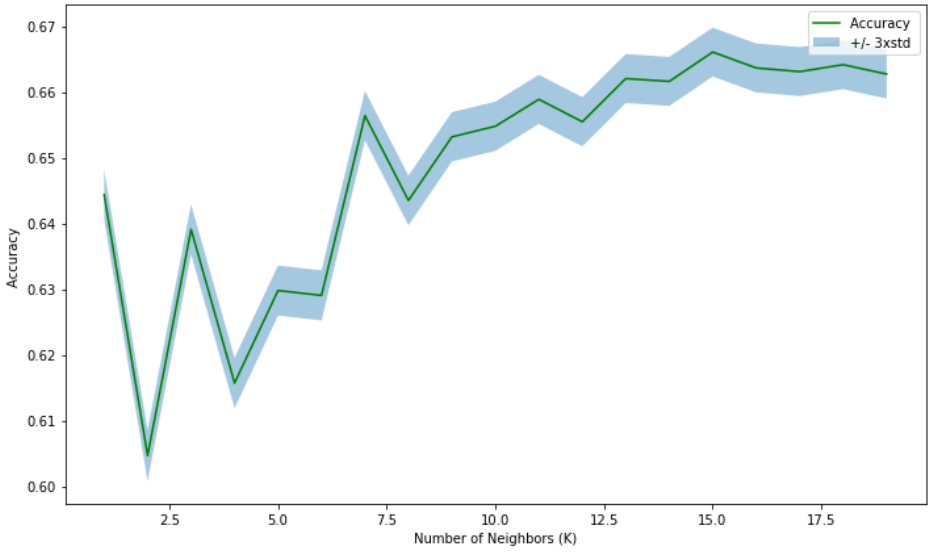
These were the configuration and accuracy result for the Decision Three Algorithm:

```
tree_model = DecisionTreeClassifier(criterion='entropy', max_depth = 4)
tree_model.fit(X_train, y_train)
predTree = tree_model.predict(X_val)
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_val, predTree))
```

```
DecisionTrees's Accuracy:  0.6944685600892968
```

4.10.2. K Nearest Neighbors Algorithm

A test was performed to define which parameter would be the most suitable for this algorithm and the result was as follows:




```
The best accuracy was with 0.6660672206374798 with k= 15
```

4.10.3. Logistic Regression Algorithm

These were the configuration and accuracy result for the Logistic Regression Algorithm:

```
LR_model = LogisticRegression(C=0.01, solver='sag', max_iter=1000)
LR_model.fit(X_train, y_train)
predLR = LR_model.predict(X_val)
print("Logistic Regression's Accuracy: ", metrics.accuracy_score(y_val,predLR))
```

```
Logistic Regression's Accuracy:  0.5999007813468932
```

4.11. Model Evaluation

The same algorithms configurations were used on the data test, that was created before (section 4.9 Data Split). There are the results:

4.11.1. Decision Tree Evaluation

```
Tree model Accuracy Score 0.6971003588378285
Tree model Jaccard Score:  0.5659187989880977
Tree model F1 Score:  0.6944754571044064
```

4.11.2. K Nearest Neighbors Evaluation

```
K Nearest Neighbors model Accuracy Score 0.6893158930431763
K Nearest Neighbors model Jaccard Score:  0.5487938135664453
K Nearest Neighbors model F1 Score:  0.6879382478557854
```

4.11.3. Logistic Regression Evaluation

```
Logistic Regression model Accuracy Score 0.6009376085195046
Logistic Regression model Jaccard Score:  0.38912022680960395
Logistic Regression model F1 Score:  0.5974902504703452
Logistic Regression mode Log loss  0.6678049101020683
```

5. Results and Discussion

The great majority of recorded accidents are damage to property, at the blocks.

The influence of drugs or alcohol beverages is not the main cause of accidents.

The most of accidents occur in clean weather, on daylight and doesn't involve parked cars.

The great majority of accidents had no injuries, serious injuries or fatalities.

Less accidents happen during the night between 1 am and 5 am. At 6 am, the number of accidents starts to increase, peaking at 5 pm, when it starts to decrease.

The best performed machine learning algorithm was the Decision Tree. It obtained the best accuracy and also performed better after validation with the test data.

Algorithm	Accuracy	Jaccard	F1-score	LogLoss
Decision Tree	0.697100	0.565919	0.694475	
K Nearest Neighbors	0.689316	0.548794	0.687938	
Logistic Regression	0.600938	0.389120	0.597490	0.667805

6. Conclusion

The accuracy of the chosen machine learning model (Decision Tree) shows us that it can be used to solve the proposed problem, however, its result can be considered only satisfactory.

Despite the large amount of data studied, many of the records had problems with their classification and had to be removed. The imbalance of the data was also a considerable factor in the loss of data during the analysis.

I believe that with a greater amount of data, the proposed mathematical model will have its effectiveness improved considerably.