Marcelo Camera Oliveira
november - 2020

A DATA ANALYSIS OF TRAFFIC ACCIDENTS IN SEATTLE, WA

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

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.

OBJECTIVE

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.

STAKEHOLDERS

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.

THE DATA

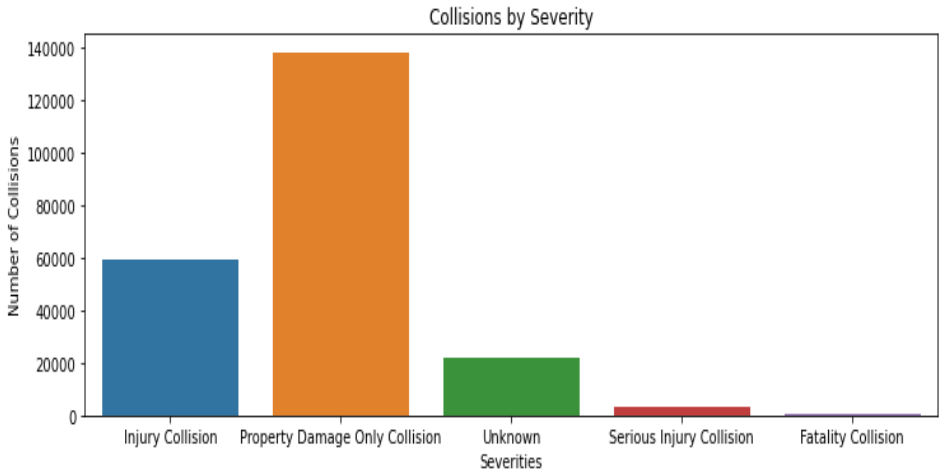
Data of the City of Seattle Open Data Portal (https://data.seattle.gov/).

Contains several records of traffic accidents in the city of Seattle, USA, from 2004 to the present.

THE DATA ATTRIBUTES

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.

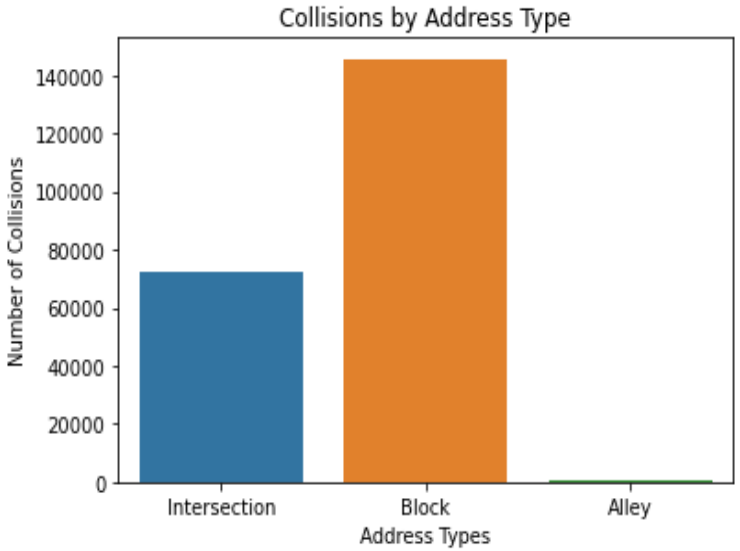
EXPLORATORY DATA ANALYSIS



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

- 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.

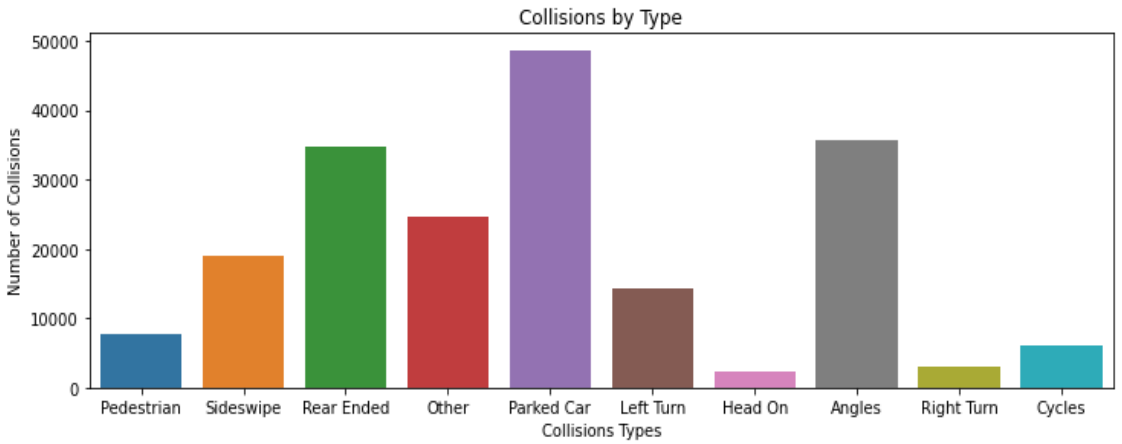
EXPLORATORY DATA ANALYSIS



	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.

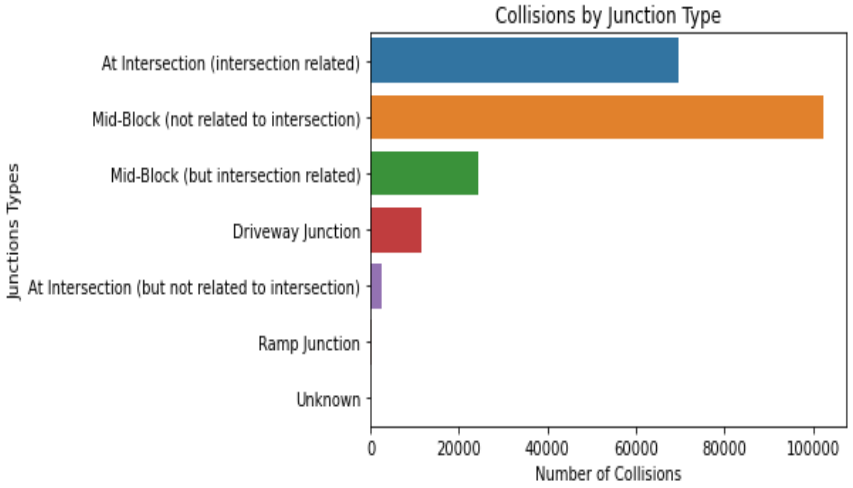
EXPLORATORY DATA ANALYSIS



	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

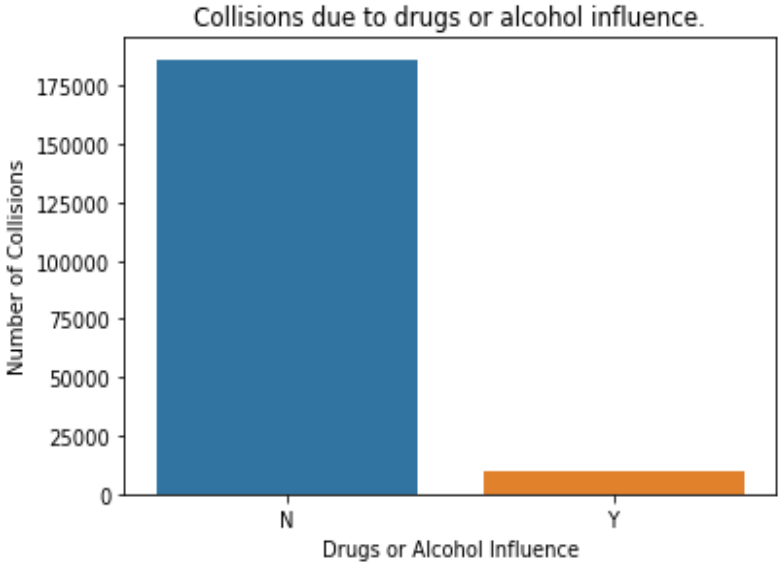
EXPLORATORY DATA ANALYSIS



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.

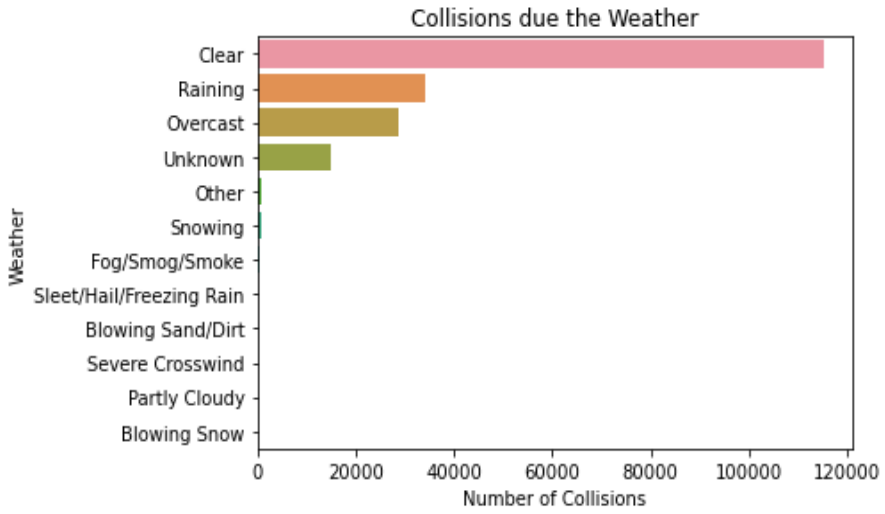
EXPLORATORY DATA ANALYSIS



UNDERINFL	
N	186372
Y	9629

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

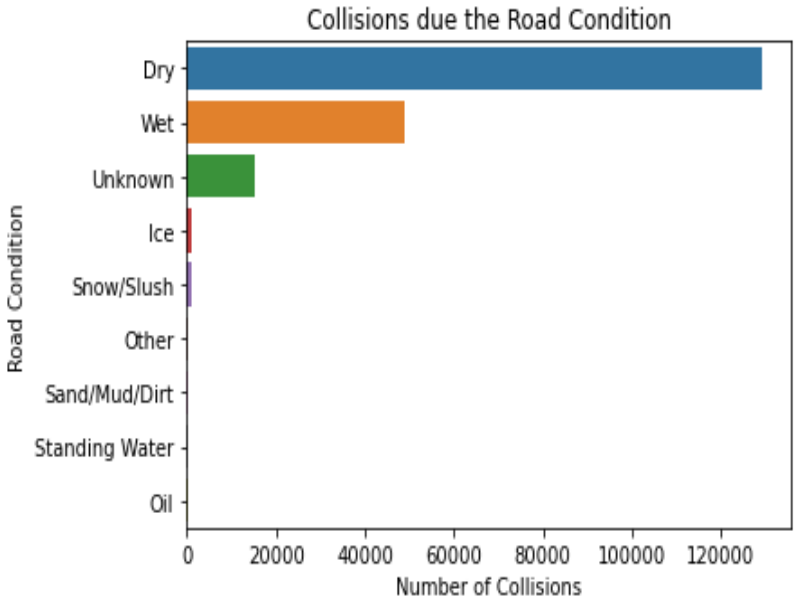
EXPLORATORY DATA ANALYSIS



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

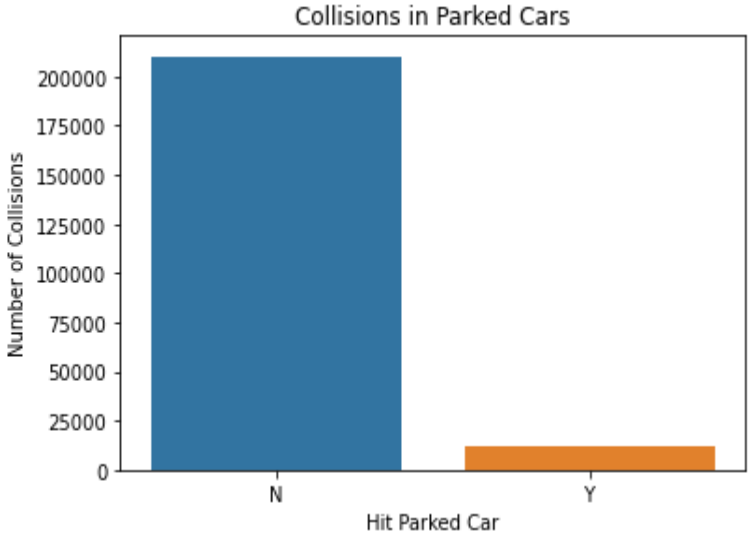
EXPLORATORY DATA ANALYSIS



	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.

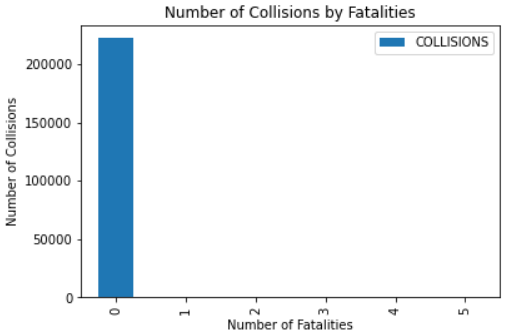
EXPLORATORY DATA ANALYSIS



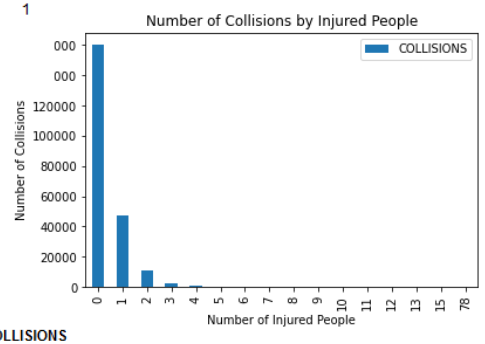
HITPARKEDCAR	
N	210285
Y	12295

- The great majority of accidents does not involve parked cars

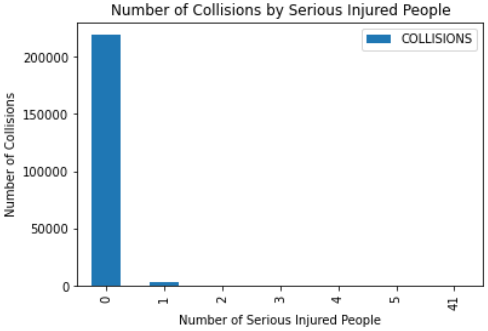
EXPLORATORY DATA ANALYSIS



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



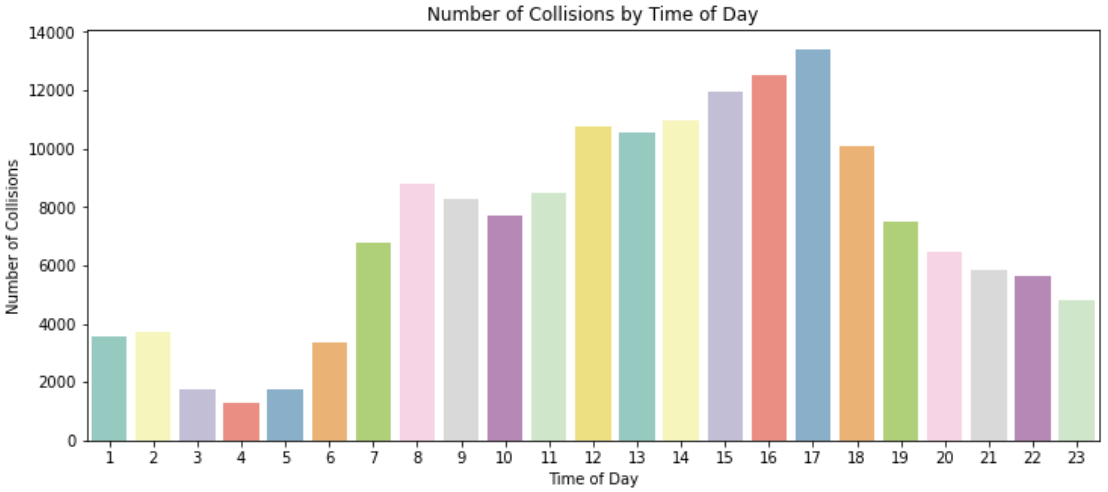
INJURIES	COLLISIONS
0	160273
1	47554
2	10737
3	2740
4	823
5	275
6	100
7	40
8	12
9	10
10	6
11	5
12	1
13	2
15	1
78	1

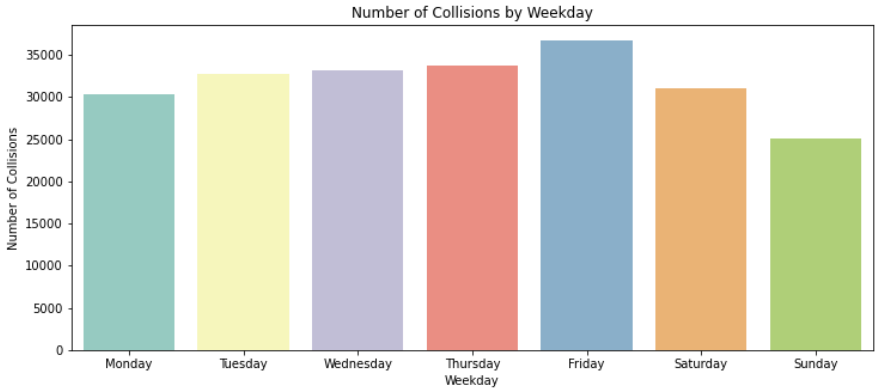


SERIOUS_INJURIES	COLLISIONS
0	219447
1	2964
2	134
3	23
4	6
5	5
41	1

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

EXPLORATORY DATA ANALYSIS





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

- 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.

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.

RESULTS AND DISCUSSION

- 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

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