Peer-graded Assignment: Capstone Project - Car accident severity

By Ildiko Balazs-Papp

Table of contents

Introduction

In today's modern world everyone can be connected to the Internet and with this we are able to generate a huge bulk of information. Our phones can track where we are shopping, what movies, music we like or how we used to go to work.
Despite this it could easily happen that we are on a certain road where a not-expected accident happens so we will be late.
In this project I will try to analyze the circumstances of car accidents based on the Seattle car accident collection (that is the data set shared in the Coursera Applied Data Science Capstone Project) and set up a model to predict the severity of an accident.
With this model warnings will be given the weather and the road conditions about the possibility of getting into a car accident. Hopefully this will lead to a possible mitigation of driving risks.

Methodology

Data analysis and preparation

As mentioned above the data used in this project will be the collection of accidents happened in Seattle. The earliest data point is from 2004 and the latest is from 2020. Attributes covered by the data set are listed in the first table of "*High level data analysis*". The accidents are described by a "severity code" that refers to either to an "injury collision" or a "property damage only collision." Please note that this is not realistic in a 6-year time period for a city like Seattle to have no severe or fatal car accidents. Other attributes cover location, intersection type, vehicle and people count, date and time, and environmental conditions.

In this section a data analysis is conducted as of the following:

- First, high level analysis is conducted, where the shape, size and column data are examined along with their correlation to each other.
- Second the data is corrected, missing data is dealt with so data visualization can happen for further analysis in the "*Data visualization*" section, so the meaningful columns can be chosen to build a custom model for prediction of the severity of the accidents.

High level data analysis

First of all, the car accident data set was downloaded from here. The fields in the dataset are described as the following by the Seattle Department of Traffic as stated in Table 1.

Table 1 Data description given by SDOT of the original source

Attribute	Data type, length	Description
OBJECTID	ObjectID	ESRI unique identifier
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: - Alley, - Block, - Intersection
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: - 3: fatality, - 2b—serious injury, - 2—injury - 1: prop damage, - 0: unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary in the end of this file.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.

CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

To examine the data first I imported the necessary libraries, such as numpy and pandas, and then the csv file will be read into a dataframe. The first thing I checked in the loaded dataframe were the data types, this is shown in Table 2.

Table 2 Data types in the loaded dataframe

Column name	Type
SEVERITYCODE	int64
X	float64
Y	float64
OBJECTID	int64
INCKEY	int64
COLDETKEY	int64
REPORTNO	Object
STATUS	Object
ADDRTYPE	Object
INTKEY	float64
LOCATION	Object
EXCEPTRSNCODE	Object
EXCEPTRSNDESC	Object
SEVERITYCODE.1	int64
SEVERITYDESC	Object
COLLISIONTYPE	Object
PERSONCOUNT	int64
PEDCOUNT	int64
PEDCYLCOUNT	int64
VEHCOUNT	int64
INCDATE	Object
INCDTTM	Object
JUNCTIONTYPE	Object
SDOT_COLCODE	int64
SDOT_COLDESC	Object
INATTENTIONIND	Object
UNDERINFL	Object
WEATHER	Object
ROADCOND	Object
LIGHTCOND	Object
PEDROWNOTGRNT	Object
SDOTCOLNUM	float64

SPEEDING	Object
ST_COLCODE	Object
ST_COLDESC	Object
SEGLANEKEY	int64
CROSSWALKKEY	int64
HITPARKEDCAR	Object

This shows that many of the probably necessary columns for the model development such as weather and road conditions are not numerical so they have to be transformed.

As a next step I checked the correlation between the data set columns so it would give a nice starting point for the exploratory data analysis conducted in Data visualization. Please note, that Table 3 only shows the top correlations.

Table 3 Top correlations between the columns

Column1	Column2	Correlation coefficient
SEVERITYCODE	SEVERITYCODE.1	1
SDOTCOLNUM	SDOTCOLNUM	1
SEVERITYCODE	SEVERITYCODE	1
INCKEY	COLDETKEY	0.999996
COLDETKEY	INCKEY	0.999996
SDOTCOLNUM	INCKEY	0.990571
COLDETKEY	SDOTCOLNUM	0.990571
INCKEY	SDOTCOLNUM	0.990571
OBJECTID	SDOTCOLNUM	0.969276
SDOTCOLNUM	OBJECTID	0.969276
OBJECTID	INCKEY	0.946383
INCKEY	OBJECTID	0.946383
OBJECTID	COLDETKEY	0.945837
COLDETKEY	OBJECTID	0.945837
CROSSWALKKEY	PEDCOUNT	0.565326
PEDCOUNT	CROSSWALKKEY	0.565326
PEDCYLCOUNT	SEGLANEKEY	0.453657
SEGLANEKEY	PEDCYLCOUNT	0.453657
PEDCYLCOUNT	SDOT_COLCODE	0.382521
SDOT_COLCODE	PEDCYLCOUNT	0.382521
PERSONCOUNT	VEHCOUNT	0.380523
VEHCOUNT	PERSONCOUNT	0.380523
SDOT_COLCODE	VEHCOUNT	0.365814
VEHCOUNT	SDOT_COLCODE	0.365814
PEDCOUNT	VEHCOUNT	0.261285

VEHCOUNT	PEDCOUNT	0.261285
SDOT_COLCODE	PEDCOUNT	0.260393
PEDCOUNT	SDOT_COLCODE	0.260393
PEDCYLCOUNT	VEHCOUNT	0.253773
VEHCOUNT	PEDCYLCOUNT	0.253773
SEVERITYCODE.1	PEDCOUNT	0.246338
SEVERITYCODE	PEDCOUNT	0.246338
PEDCOUNT	SEVERITYCODE	0.246338

While checking the above table description and the correlation coefficient, some conclusions can be made about the data:

- SEVERITYCODE and SEVERITYCODE.1 columns probably contain the same data, so SEVERITYCODE.1 can and will be dropped.
- INCDATE and INCDTTM contain the date and the date-time of the accident, therefore INCDATE is redundant so it will be dropped.
- SDOT_COLCODE depends on the severity and PERSONCOUNT of the accident.
- LOCATION and X,Y contain the address and the coordinates of the accident, therefore LOCATION is redundant so it will be dropped.

Identifying and dealing with missing values

In the next step I've checked which columns contain missing data and found that the following columns contain missing values:

- X
- Y
- ADDRTYPE
- INTKEY
- EXCEPTRSNCODE
- EXCEPTRSNDESC
- COLLISIONTYPE
- JUNCTIONTYPE
- INATTENTIONIND
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- PEDROWNOTGRNT
- SDOTCOLNUM
- SPEEDING
- ST_COLCODE
- ST_COLDESC

- SEGLANEKEY

Before dealing with the missing values the following train of though was conducted:

- As car accidents not necessarily happen in intersection, missing values in the INTKEY and JUNCTIONTYPE are fine.
- For the location related columns (X,Y) the empty fields will be filled with the mean of the other values.
- Missing data in categorical columns (ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER,UNDERINFL, ROADCOND, LIGHTCOND, ST_COLCODE) will be filled with the most frequent data in the given column.
- As for EXCEPTRSNCODE and EXCEPTRSNDESC columns no description is provided anyway, these columns will be dropped.

Correct data format

As a next step I checked the value counts in each column. I found that unfortunately the UNDERINFL column not only just contained nan-s but it's value set is the following: ['N', '0', nan, '1', 'Y']. Therefore, in this column first 'N' and 'Y' are converted to 0 and 1 respectively and then the most frequent of these two values will be substituted for the nan values.

Data visualization – explanatory analysis

As an explanatory analysis I created some figures and tables that can shed light on the nature of the data.

Cross-tab counts and representations for number of pedestrians, cyclists, and people involved in accidents and the severity of the accident

First the distribution of the severities of the accidents were examined regarding the people count involved in the accident. This is shown in Table 4 and Figure 1.

Table 4 Numerical show of the distribution

SEVERITYCODE	1	2
PERSONCOUNT		
0	3782	1762
1	9858	3296
2	86420	27811
3	22092	13461
4	8365	6295
5	3615	2969
6	1345	1357
7	494	637
8	249	284
9	87	129
10	54	74
11	23	33
12	13	20
13	9	12
14	12	7
15	4	7
16	3	5
17	3	8
18	5	1
19	3	2
20	5	1
21	2	0
22	2	2
23	1	1
24	1	1
25	5	1
26	4	0
27	2	1
28	2	1
29	2	1
30	1	1

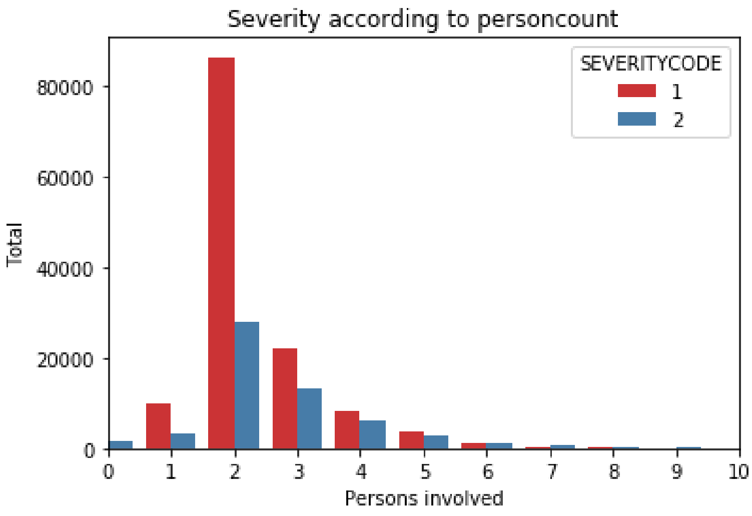


Figure 1 Severity according to person count

I found that the percentage of SEV 1 (= property damage) incidents is 70.11 %, while the percentage of SEV 2 (= injurie collision) incidents is 29.89 %. So low ratio of injuries indicates which is also shown later in the histogram of collision types, that these accidents usually happen including another car not a pedestrian or cyclist and a car.

To check that theory I've created the following 2 figures and tables (Figure 2 Figure 3 Table 5 Table 6) which show that either 0 or just 1 pedestrians or cyclists were involved in the majority of the accidents.

Number of pedestrians involved in the accident	Percentage of total number of accidents
0	96.44
1	3.43
2	0.12
3	0.01
4	0.0
5	0.0
6	0.0

Table 5 Distribution of accidents according to involved pedestrians

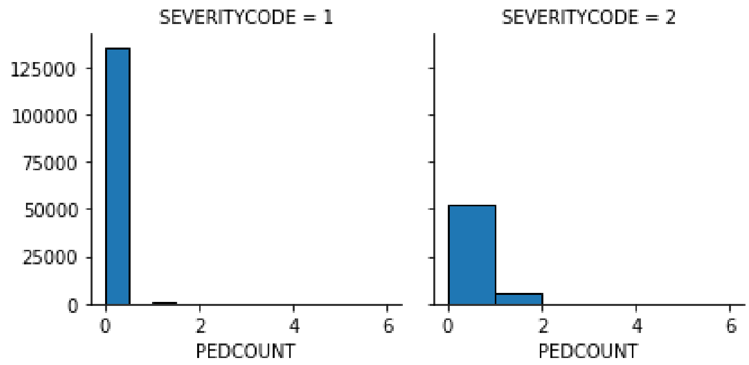


Figure 2 Pedestrian count in accidents

Table 6 Distribution of accidents according to involved cyclists

Number of cyclists involved in the accident	Percentage of total number of accidents
0	97.18
1	2.79
2	0.02
3	0.0

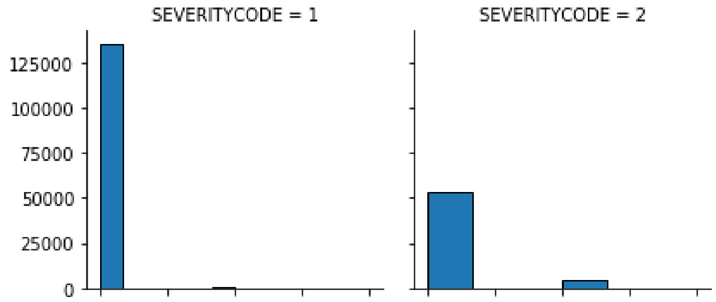


Figure 3 Cyclist count in accidents

4	0.0

Let's see what conclusions can be drawn from the first sight of these figures:

- The majority (70 %) of the accidents fell into the category represented by severity code 1, which is property damage.
- More than 99.8 % of the accidents involved less then or equal to 10 people.
- In the 96 % of the accidents no pedestrians were involved in the accident.
- In the 97 % of the accidents no cyclists were involved in the accident.

Analysis and representations for weather, light and road conditions and the severity of the accident

The next stop was to check the impact of the road, light and weather conditions on the severity of the accidents. These are shown in the following figures (Figure 4 Figure 5 Figure 6). Sev 1 accidents are signed with color red while Sev 2 accidents with blue.

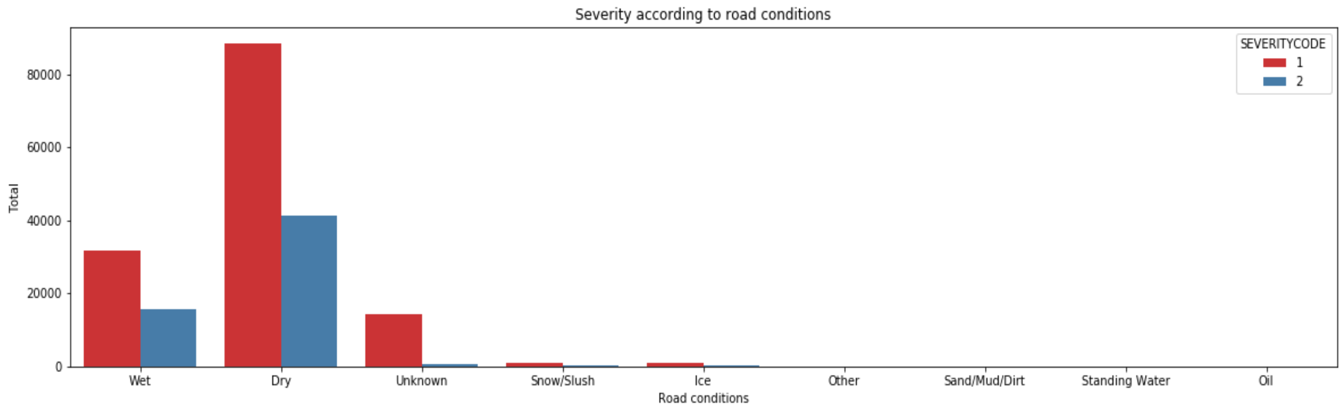
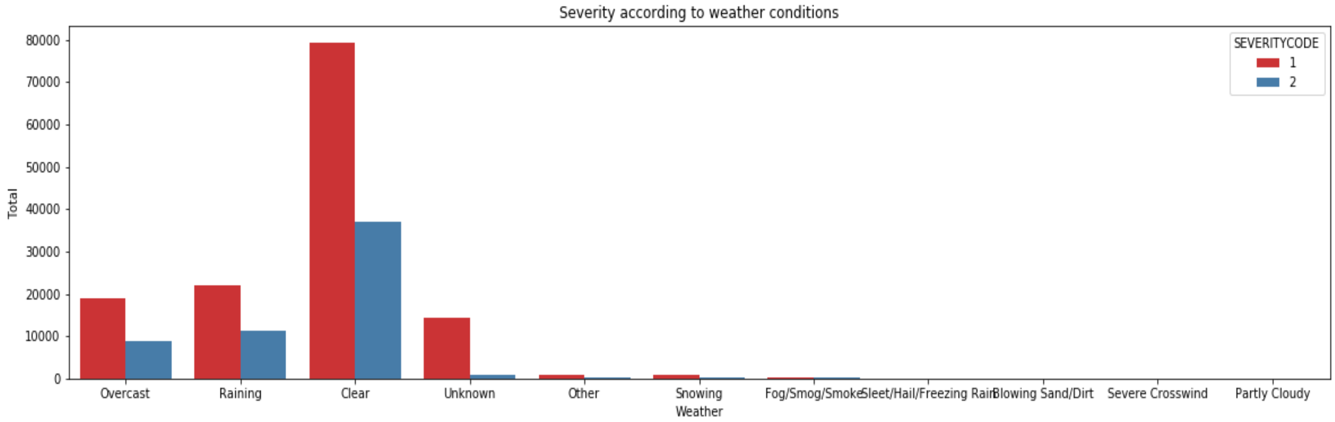


Figure 4 Distribution of accidents according to road conditions



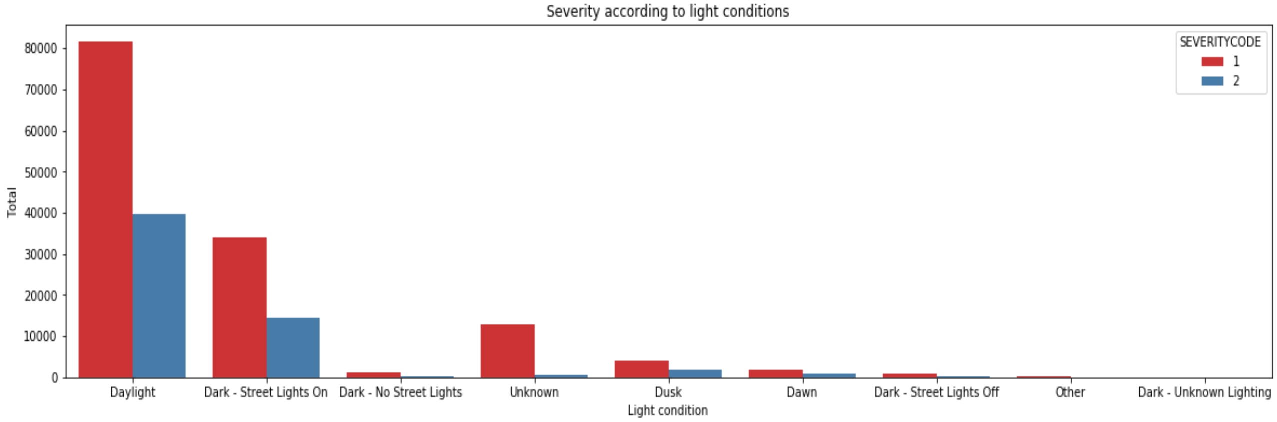


Figure 6 Distribution of accidents according to light conditions

From the first sight of these figures the following conclusion can be drawn. Most of both severity 1 and 2 accidents happened in clear weather conditions with dry roads in daylight. This again indicates that the given data set is not realistic.

Figure 5 Distribution of accidents according to weather conditions

Distribution of the accidents in time and type

In this section I checked the distribution of the accidents in time and in types. The month of the year was extracted from all incident date and the number of accidents is shown in Figure 7.

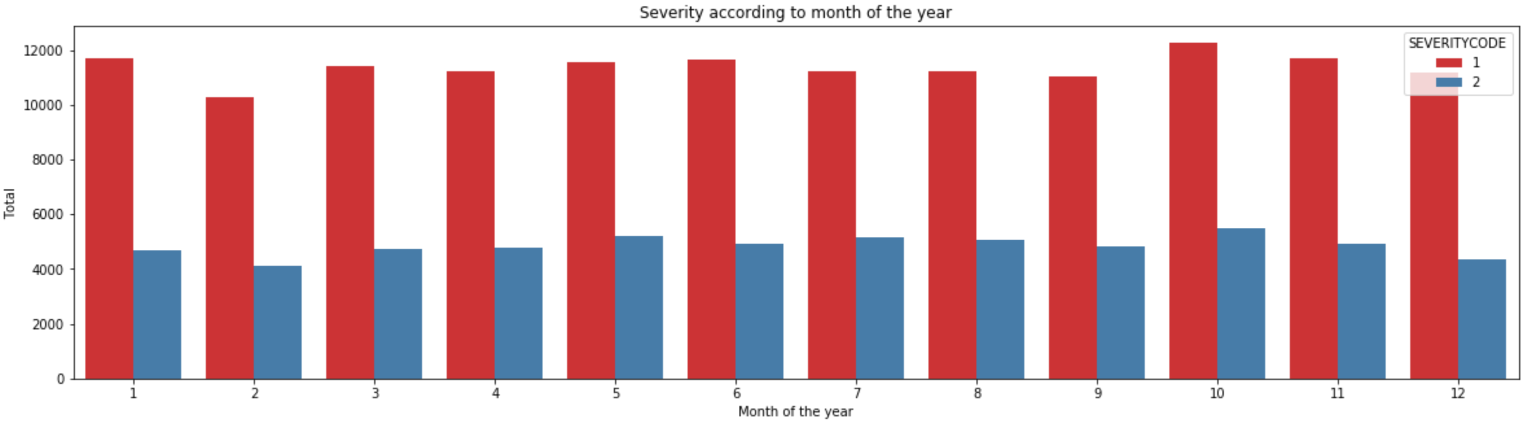


Figure 7 Distribution of accidents throughout the year

The figure suggests that the distribution of accidents is approximately constant in in regards with the month of the year.

To shed light on why this could happen I also checked the number of each accident type in the data set, and plotted a pie chart from them shown in Figure 8.

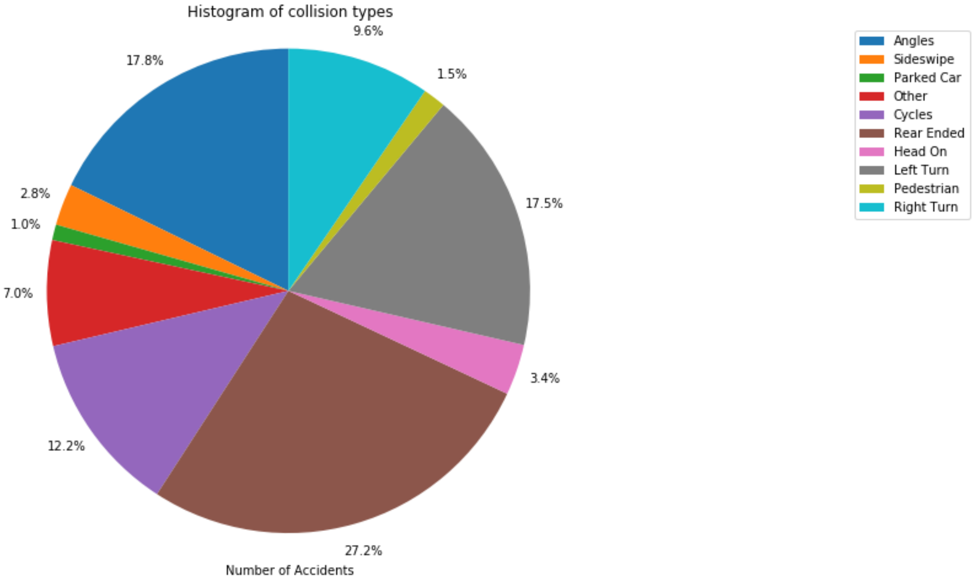


Figure 8 Pie chart of accident types

The chart shows that the most likely accidents of all are rear ending each other (this usually happens in cities with traffic jams). This could explain the high number of property damage incidents and that this is approximately constant during the year.

Spatial distribution of accidents

The aim of Figure 9 is to demonstrate how the accidents are distributed in Seattle to support the above theory about the traffic jams as the reason of accidents. Here the first 1000 data points were used to create the figure.

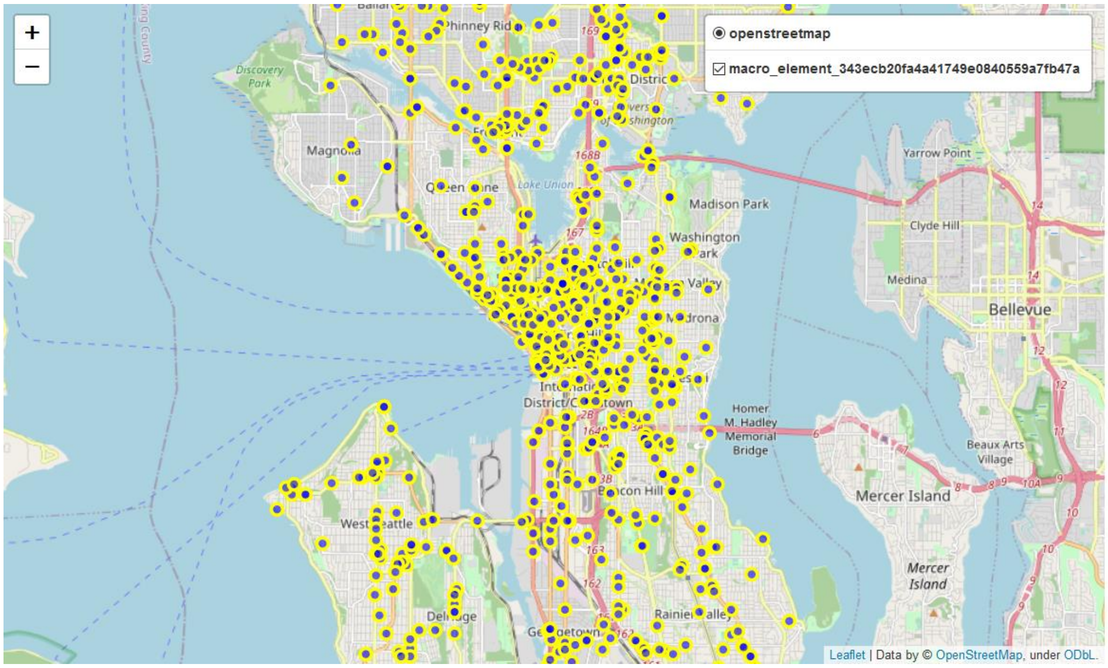


Figure 9 Map of Seattle with accidents

In Figure 10 the typical traffic on a Monday morning is shown in the most heavily effected part of Seattle.

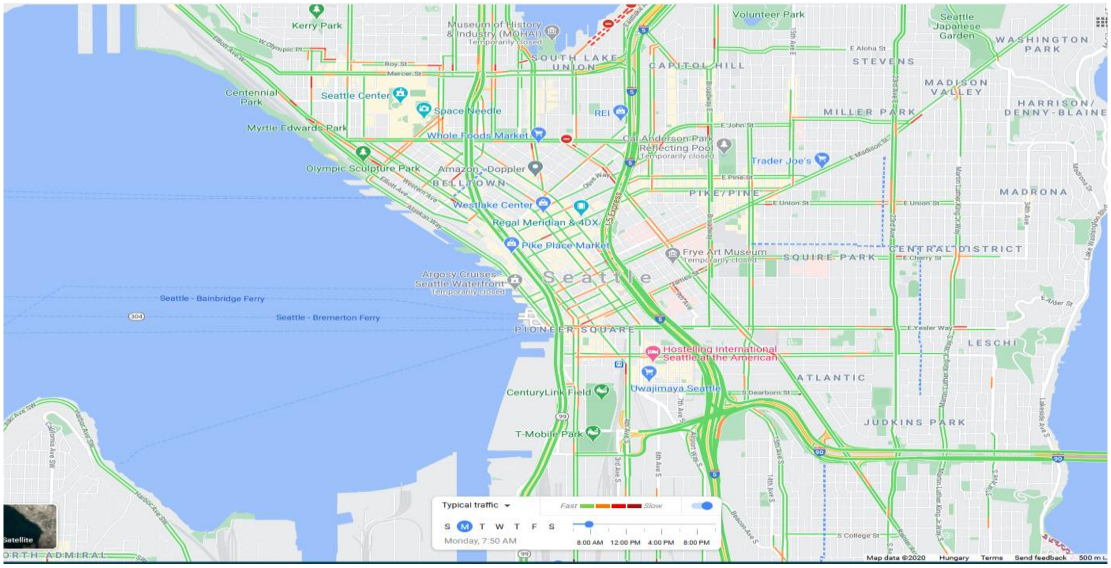


Figure 10 Typical traffic in a Monday morning according to Google Maps

It could be seen that this heavily effected area has many traffic jams and closures.

Building a classification model

When building my prediction model, I decided to use the following columns based on the above figures:

ADDRTYPE, JUNCTIONTYPE, INATTENTIONIND, WEATHER, ROADCOND, LIGHTCOND, PEDROWNOTGRNT, SPEEDING, HITPARKEDCAR, X, Y, OBJECTID, INCKEY, COLDETKEY, INTKEY, SEVERITYCODE.1, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, SDOT_COLCODE, UNDERINFL, SDOTCOLNUM, SEGLANEKEY, CROSSWALKKEY, month.

Unfortunately, not all the columns in our data set is numerical therefore columns that will be the base of the K Nearest Neighbor classification model needed to be converted into numerical.

After that I normalized the data with sklearn StandardScaler and created a train set of data as the 80 % of the data available for me, leaving 20 % to test it. After that I built a K Nearest Neighbor model (KNN model) with k=6. I chose this model as I think that accidents among the same circumstances will be similar in the outcome as well.

Model evaluation

I evaluated my model with determining the followings for it:

- KNN model accuracy for test set: 0.988
- KNN model f1 score for test set: 0.988
- KNN model Jaccard score for test set: 0.984

From these it seems that the model is working correctly however with more real life data it should be evaluated again.

Results

In this notebook I analyzed the Seattle car accident data set and as I found it to be a multidimensional and labeled. After some data preparation and cleaning I created some figures so the data could be better understood. From the first visualizations of the data set I was able to conclude the followings:

- The majority (70 %) of the accidents fell into the category represented by severity code 1, which is property damage.
- More than 99.8 % of the accidents involved less then or equal to 10 people.
- In the 96 % of the accidents no pedestrians were involved in the accident.
- In the 97 % of the accidents no cyclists were involved in the accident.
- The majority of both severity 1 and 2 accidents happened in clear weather conditions with dry roads in daylight.
- The distribution of accidents is approximately constant in in regards with the month of the year.
- Most likely accidents of all are rear ending each other (this usually happens in cities with traffic jams). This could explain the high number of property damage incidents and that this is approximately constant during the year.

Discussion

These findings not necessarily mirror reality as here eg. most of the accidents happened in clear weather conditions with dry roads in daylight. Besides that, no fatal accidents were recorded in the data set. This is probably due to previous manipulation of the dataset which I am not aware of.
I also created, optimized and evaluated a classification based on the K Nearest Neighbors method. This algorithm classifies the accidents based on their similarity to other cases and to be able to predict the severity of a car accident this is exactly what we need as probably accidents among the same circumstances will be similar in the outcome as well.

Conclusion

As stated above the reality of the original data set is questionable therefore the model trained on it could be unrealistic. That is why I think that going forward a collection of more real life data could be advisable, so the model could be further developed and specified.