IBM Data Science Capstone Project
Predicting the car accident severity of Seattle

Daniel Souza Lima

November 7, 2020

1 Introduction

1.1 Background

Road traffic accidents are a great force of unexpected hurdles, from which stem consequences ranging from the economic nature, with property damage, to those irreversible ones, with human injuries and death. The road traffic accidents along its financial ecosystem of medical treatment, judicial costs, and property damage/lost have costed the United States´ economy around $871 billions only in 2010 [1]. Therefore, it is of great value the comprehension of which conditions drive such misfortune happenstances.

1.2 Problem

Over the past 17 years, the traffic accidents in the Seattle City Council Area have been recorded in their details, conditions, and final outcomes by the Seattle Department of Transport (SDOT). The central problem is to predict the final outcome based on some underlining conditions.

The data that may contribute to describing transit accidents might include its GPS coordinates, address, road condition, luminosity, day of the week, and hour of the day. This project aims to predict the severity of such accidents by counting on these features.

1.3 Interest

Naturally, the people of Seattle would be very interested in the accurate prediction of the road traffic accidents, for the sake of Public and Transportation Safeties. Policymakers and the own Seattle Police Department may also be inclined towards its public reality.

2 Data acquisition and cleaning

2.1 Data acquisition

The data of the SDOT on the traffic collisions is provided by Seattle GeoData and can be found here. Up to November, the data is composed of 222,389 rows. Each row is an accident. To extract the neighbourhoods of the collisions in a geojson map, I downloaded the map of Seattle from this GitHub Repository.

2.1.1 The target feature

The data has 40 columns, from which the SEVERITYCODE (SC), a classifying code for the severity of road traffic accident, is **the target feature**, which we must predict. SC may have the following attributes defined in the pdf "Attribute Information":

- 0: unknown,
- 1: property/vehicular damage,
- 2: minor injure,
- 2b: serious injury, and

[1] A 2015 study of the National Highway Traffic Safety Administration: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013

- 3: fatality.

The target features will be resumed between 0 (no injury: 0 and 1) and 1 (injury and/or fatalities: 2, 2b and 3) for the machine learning models. The target features will be predicted according to two groups:

- a binary classification: $\{0, 1\} \rightarrow \{0\}$, $\{2, 2b, 3\} \rightarrow \{1\}$, and
- a multi classification: $\{0, 1\} \rightarrow \{1\}$, $\{2, 2b\} \rightarrow \{2\}$, and $\{3\} \rightarrow \{3\}$.

In Appendix A is listed all the 40 columns (features).

2.2 Data cleaning

By using the Pandas `df.isnull().sum(axis=0)` and grouping the features by the SEVERITYCODE with the code `dataframe.groupby(["SEVERITYCODE"])[feature]`, we might inquire how the features .
These are the results:

- 7,478 rows with missing (X,Y) coordinate;
- most of the Nan 26,641 values for the WEATHER entry are concentrated on SC 0: for accidents with SC equal to 0, for the WEATHER entry, there are 1 entry for "Clear", 1 for "Rain", and 21,654 missing values;
- similiarly, most of the Nan 26,451 values for the COLLISIONTYPE are concentrated on 21,654 SC=0 cases;
- the same happening for the ROADCOND (road condition) feature: 24,654 from the 26,560 missing values;
- same fenomena for ST_COLCODE (9,413/9,413), LIGHTCOND (21,654/26,730), and UNDERINFL (21,655/26,431);
- all NaN are actually "N" entries for INATTENTIONIND, PEDROWNOTGRNT and SPEEDING, which are Y/N answers, because sometimes they are answered only when Y;
- 3,714 entries for the ADDRTYPE are missing, and
- 4,593 entries for the LOCATION are missing.

It will be categorically deleted all the rows with missing: (X,Y) coordinates, ADDRTYPE, and LOCATION.
The rows with SC equal to 0 with a missing feature will have it substituted by an average accordingly to its statistical distribution. However, the rows with SC greater than 0 and with one of above features missing will be simply dropped.
All others non-predictive and redudant features like INCDATE, OBJECTID, INCKEY, and STATUS will be dropped. Nonetheless, the LOCATION feature will also be dropped.
Unlike the feature INCDATE, the feature INCDTTM indicates the date and the time date and will be kept.

2.3 Date conversion

From the column INCDTTM, the date, the hour, the day of the week (`day_name`), month, and the year are extrated.

2.4 Getting the neighbourhoods

The data set provided by the SDOT contains only the longitude and latitudes coordinates of the transit accidents, not their neighbourhoods. For such, the extraction of the 175 possible neighbourhoods was performed from a geojson of Seattle [2]. The extracted neighbourhoods become a new column in the data frame called NEIGHBOURHOODS.

[2] https://raw.githubusercontent.com/seattleio/seattle-boundaries-data/master/data/neighborhoods.geojson

2.5 The choropleth map of the neighbourhoods

After identifying the neighbourhoods, we can visualize the neighborhoods scaled by their accident counts.

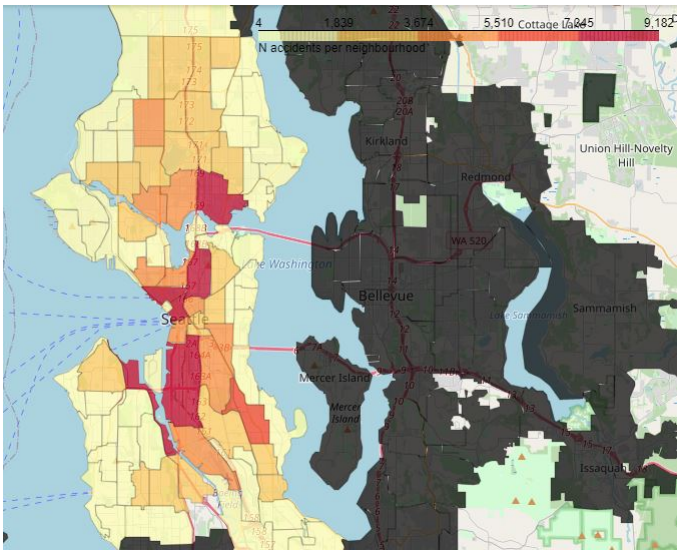


Figure 1: The neighbourhoods by their accidents counts. Belltwon has the most accidents.

After grouping each neighbourhood by the criterium of separating the accidents with non-serious injuries from those with serious injuries and fatalities:

- '0', '1' and '2': 0, and
- '2' and '3': 1.

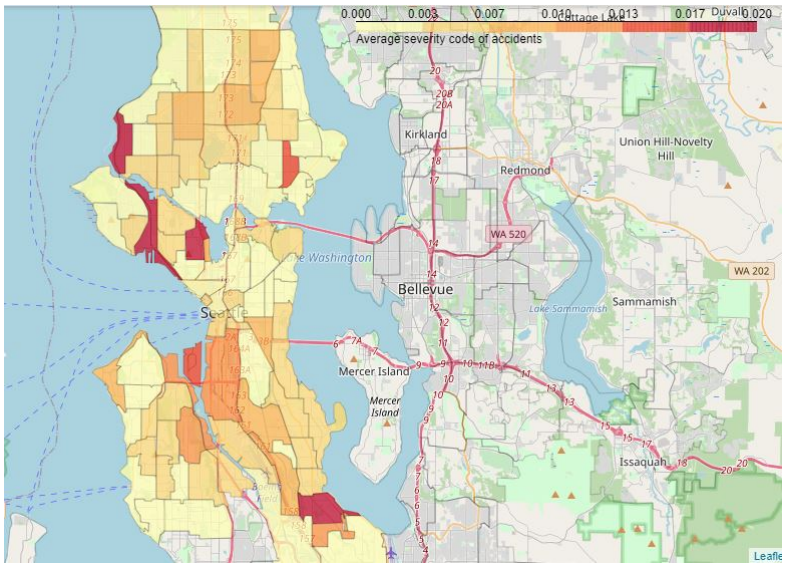
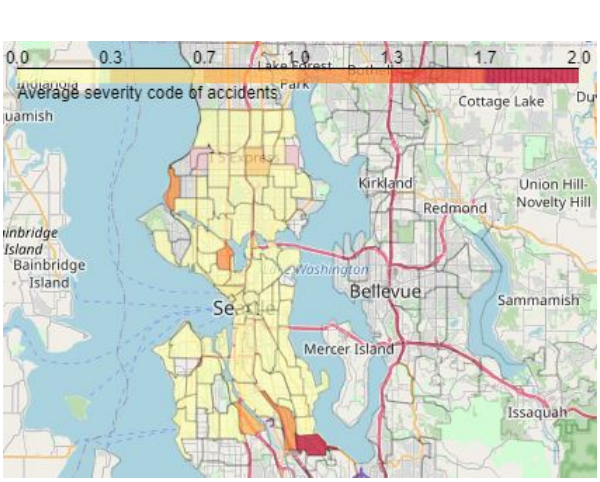
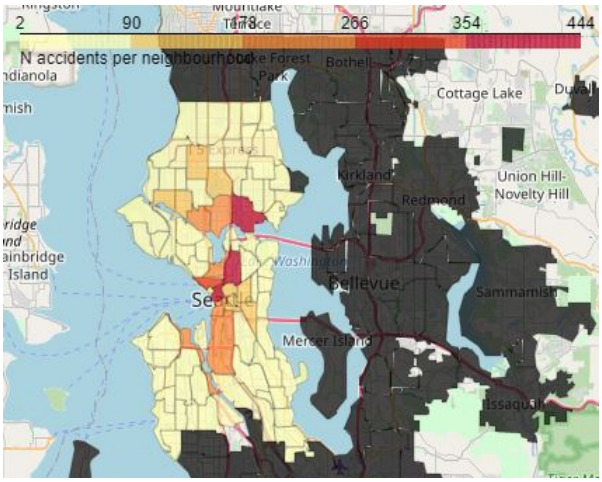


Figure 2: The neighbourhoods by their accident severity. East Queen Anne has the highest severity average.

Then, about the accidents involving cyclists:



(a) The mean severity accident involving cyclists.



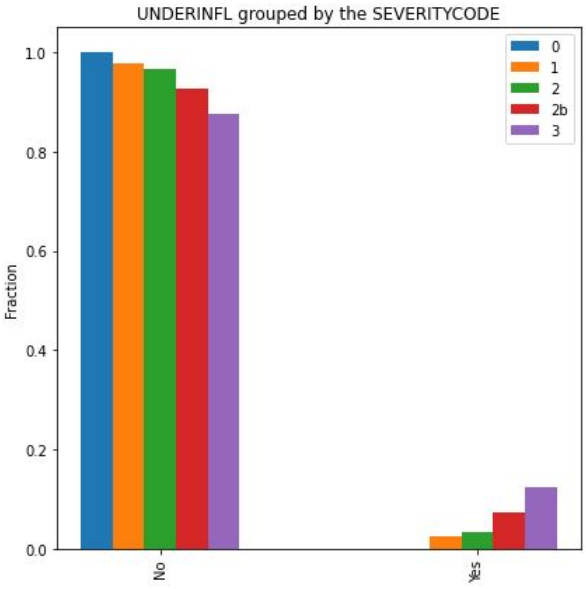
(b) The accidents count involving cyclists.

Figure 3: The choropleth map distribution by the neighbourhoods of the 5,944 accidents involving cyclists. Rainier Beach has the highest mean severity accident, while University District has the highest count.

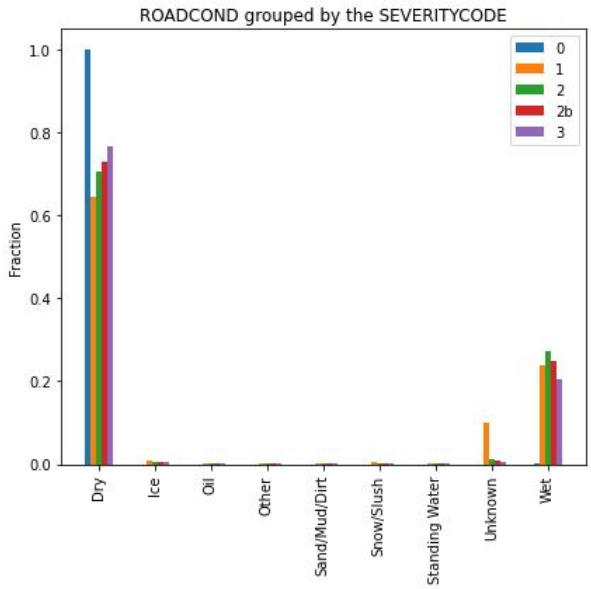
3 Methodology

3.1 Data exploratory, visualization and pre-processing: feature selection/extraction

The objective is to perceive how a determined feature could influence the target feature (the severity of the accident). With this in mind, after grouping the feature by the command `data_frame.groupby(["SEVERITYCODE"])(feature).val` we could make a histogram of each feature. For example, the feature UNDERINFL:



(a) UNDERINFL is the condition in which the driver is intoxicated with some drug.

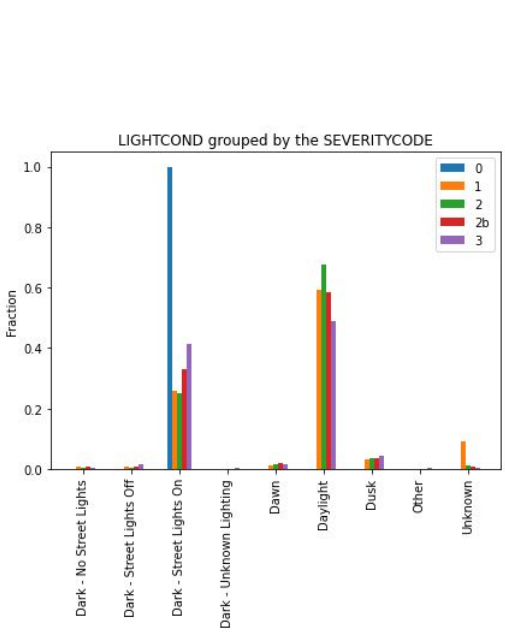


(b) ROADCOND is the condition of the road when the accident happened.

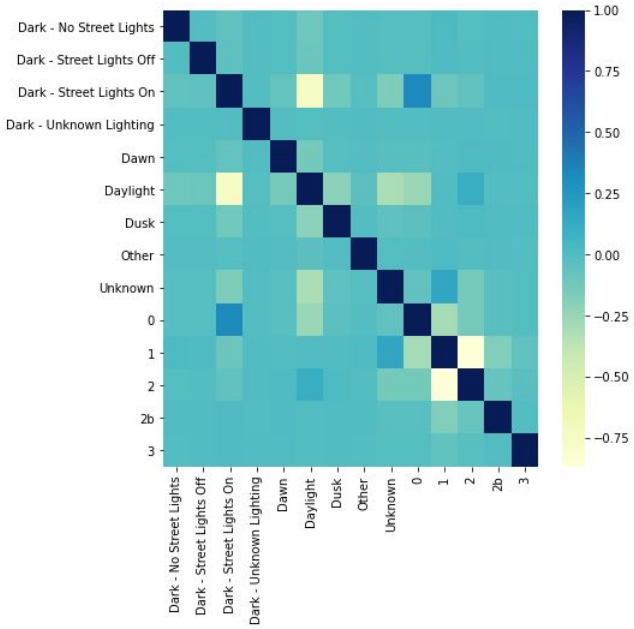
Figure 4: Histograms generated by grouping the data around the severity codes.

In accordance with the common sense, the histogram tells us that frequency of the more serious accidents rises as the intoxication grows.

It happens that we might also see a “threshold” kind of behaviour for the SC´s accordingly to the features. For example, the light conditions:



(a) Histogram of LIGHTCOND grouped by the severity codes



(b) The heatmap of the correlation matrix

Figure 5: The light conditions grouped by the severity codes. Observe that most of SC=0 accidents occur in just one condition, while the others SC´s occur might occur on others conditions.

The great majority of the SC=0 accidents does not happen on any other light condition besides `'Dark - Street Lights On'`. Thus, we could set a binary threshold for the LIGHTCOND variable:

```
dataframe.loc[(df["SEVERITYCODE"]=="0") & \
(df["LIGHTCOND"]=="Dark - Street Lights On"), "LIGHTCOND"] = 1

df.loc[(df["SEVERITYCODE"]=="0") & \
(df["LIGHTCOND"]!="Dark - Street Lights On"), "LIGHTCOND"] = 0
```

The LIGHTCOND column **will not** be required to be transformed by a `get_dummy()` into 9 binaries columns for each one of its possible categories. Such analysis can be justified by the Pearson Correlation Coeficient between 'Dark - Street Lights On' and 'SC = 0', which is 0.33 with a null p-value (which is the chance if the null hypothesis[3] were true). On other hand, e.g., the Pearson Correlation Coefficient for Daylight vs SC = 1 is 0.0037 with a P-value of 0.096.

The histograms and correlation heatmap for the others features in relation to the SEVERITYCODE are in Appendix B.

We could analyse the correlation between a feature and another non-target feature, like, the weather and road condition, or the weather and the month, and so on. Were this done, in principle, we could eliminated correlated, and thus redudant, features; therefore, causing the betterment of the machine learning accuracy.

For example, the Pearson Correlation Coefficient between the categorical feature of the WEATHER column 'Raining' and the feature 'Wet´ (ROADCOND) is 0.77 with a P-value much less than one: translation of the obviousness of roads getting wet by the rain.

However, any attempt of eliminating these correlated features will not be pursued.

Bearing in mind that the features influence differently the final outcome, the severity of the accident, we proceed on the methods of feature selection.

3.1.1 Method 1: selection by "hand-picking" the features

we visualize the histogram or table of grouped feature along the SC, calculate the mean frequency as a "threshold", and judge if such categorical feature should be valued as active or not.

That is, for example, the feature NEIGHBOURHOOD column has 175 different categorical features:

```
len(data_frame["NEIGHBOURHOOD"].unique().tolist())
>>> 175
```

[3]The null hypothesis here is: *the feature LIGHTCOND and SEVERITYCODE* ***are not*** *correlated.*

Were the frequency of each accident, no matter its severity, equal for each neighbourhood, it would be distributed with the fraction of 0.0057(=1/175) for each neighbourhood. Nonetheless, this is the result:

SC = 0

	NEIGHBOURHOOD	Fraction
0	University District	0.053582
1	Belltown	0.046127
2	Industrial District	0.039487
3	Broadway	0.035411
4	Central Business District	0.030518
5	Fremont	0.028771
6	Greenwood	0.028655
7	Wallingford	0.026092
8	First Hill	0.025743
9	South Lake Union	0.025160

SC = 1

	NEIGHBOURHOOD	Fraction
0	Belltown	0.048309
1	Industrial District	0.044811
2	University District	0.041159
3	Broadway	0.039359
4	Central Business District	0.038371
5	Columbia City	0.028402
6	Greenwood	0.024734
7	South Lake Union	0.024371
8	North Beacon Hill	0.023915
9	Fremont	0.023236

SC = 2

	NEIGHBOURHOOD	Fraction
0	Industrial District	0.046589
1	Belltown	0.042200
2	University District	0.038604
3	Central Business District	0.037758
4	Broadway	0.033016
5	Columbia City	0.030090
6	Greenwood	0.028768
7	South Lake Union	0.024238
8	North Beacon Hill	0.024202
9	Haller Lake	0.022810

SC = 2b

	NEIGHBOURHOOD	Fraction
0	Industrial District	0.057019
1	Belltown	0.041347
2	Broadway	0.036679
3	Central Business District	0.036679
4	Columbia City	0.030010
5	Greenwood	0.028343
6	University District	0.027342
7	Fremont	0.024341
8	South Lake Union	0.024341
9	Atlantic	0.021340

SC = 3

	NEIGHBOURHOOD	Fraction
0	Industrial District	0.096677
1	East Queen Anne	0.039275
2	Atlantic	0.036254
3	Columbia City	0.036254
4	Greenwood	0.036254
5	Belltown	0.033233
6	Rainier Beach	0.033233
7	Interbay	0.024169
8	Mid-Beacon Hill	0.024169
9	Pioneer Square	0.024169

Figure 6: Tables generated by grouping the feature NEIGHBOURHOOD by the SEVERITYCODE.

Totally against the presumption of equipartition, it is found that there are classes whose frequencies are way above the presumed average of 0.0057. In the case of SC = 3, it is even encountered that 'Industrial District' spikes alone at the top.

The idea of hand-picked feature selection is, then, to activate the categorical variables, whose frequencies are pronounced(dwarfed) in relation to a 'expected/presumed' average, into the binary entry 1(0). As a result, the type of the feature is not categorical anymore, but an integer number type now.

3.1.2 Method 2: selection by the Pearson Correlation Coefficient

Instead of simply hand-picking the values by their spikes in histograms, a more statistical reasoning may also be employed in the feature extraction.

Let $H(x - x_0)$ be the Heaviside function centered around x_0, let P_{xy} be Pearson Correlation Coefficient between a feature x and a severity code y. Then, we activated the column feature with 0 or 1 accordingly to $H(P_{xy} - 0)$:

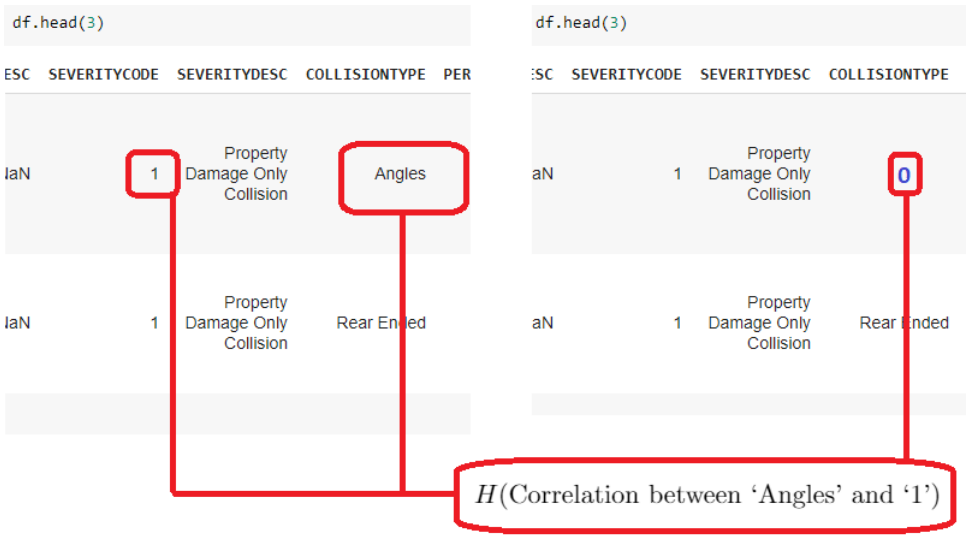


Figure 7: The Pearson Correlation Coefficient between the feature COLLISIONTYPE "Angles" and te SC "1" is equal to -0.194 with a P-value much less than 0. Then, as the coefficient is less than zero, the Heaviside function returns zero for the feature.

With this feature selection, the machine learning algorithms have an accuracy of 99% to distinguish SC '0' and anyone else (resumed in '1'). Nevertheless, it performs poorly when the target feature are the 5 possible severity codes outcomes.

3.1.3 Method 3: selection by the greatest χ^2

One more method that appeals to statistics is to choose the column features whose χ^2 are the greatest. The function from the `sklearn.feature_selection` package called `SelectKBest` selects the first k features with the greatest χ^2.

We can see that the features 5, 11, 14, and 7 (ST_COLCODE, WEATHER, PERSONCOUNT, and VEHCOUNT, respectively) have the greatest ^{2}.

Thus, we could assert that, in crescent order of influential power, that these are the most influential factors to an accident outcome:

1. the number of vehicles involved in the accident,
2. the weather conditions,
3. the number of persons envolved, and
4. mainly, the type of the collision.

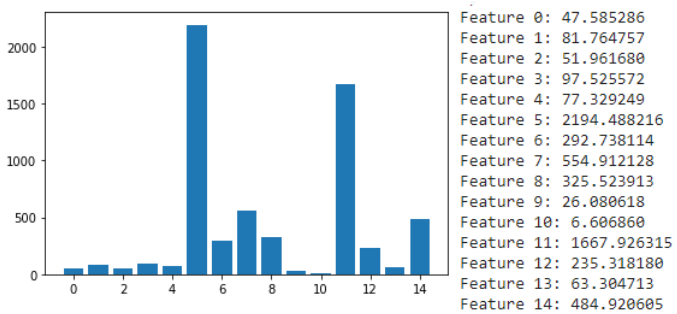


Figure 8: The χ^2 values of 15 features.

3.2 Balancing the dataset

The vast majority of the accidents has a low severity outcome, SC=1, i.e, does not involved people injured, only property damage. Consequently, any machine learning algorithm would then predict the severity of any accident to

be the most probable one, and with an great accuracy. For the machine learning to be capable to distinguish the cases, the dataset must have an equal number of different outcomes.

By typing the code line `data_frame["SEVERITYCODE"].value_counts()`, it gives this:

```
1 129499
2 56730
0 8585
2b 2999
3 331
Name:  SEVERITYCODE, dtype:  int64
```

Therefore, after shuffling the data frame and sampling it with an equal number of severity cases (331 cases for each one), from the almost 200,000 examples, only 1,655 random rows are picked.

3.3 Summary: the Final Selection

After the data cleaning and feature selection, the following columns are selected:

- X, Y, the longitude and latitude (except for the Method 3).
- the neighbourhood, NEIGHBOURHOOD,
- the light and road conditions, LIGHTCOND and ROADCOND, respectively,
- if the driver was inattentive, intoxicated or speeding, INATTENTIONIND, UNDERINFL and SPEEDING, respectively,
- PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT and VEHCOUNT, the number of persons, pedestrians, cyclists and vehicles involved, respectively,
- the type of the collision, given by ST_COLCODE,
- **COLLISIONTYPE column is droped, since it has the same information as ST_COLCODE, expect with less detail**,
- the day of the week, day_name, and
- the type of address, ADDRTYPE.

In short, it might be used up to a maximum of 17 columns. The method 3 utilizes up to 15 columns.

3.4 Predictive Modeling

Under the premise that the outcomes of the accidents might be grouped under the their physical, time and apparently happenstance conditions, the machine learning algorithms for classification kNN, Support Vector Machine (SVM), and Decision Tree (DT) will be used. Likewise, the Logistic Regression (LR) will be deployed as a regression model; nonetheless, this time, with the meaning of providing a probability of an accident to be of a certain severity, given the setup conditions.

4 Results

The results provided by just one Machine Learning algorithm are already plentiful. Therefore, only the plots of the best results are exhibited.

We recall that the target features to be predicted are separated as:

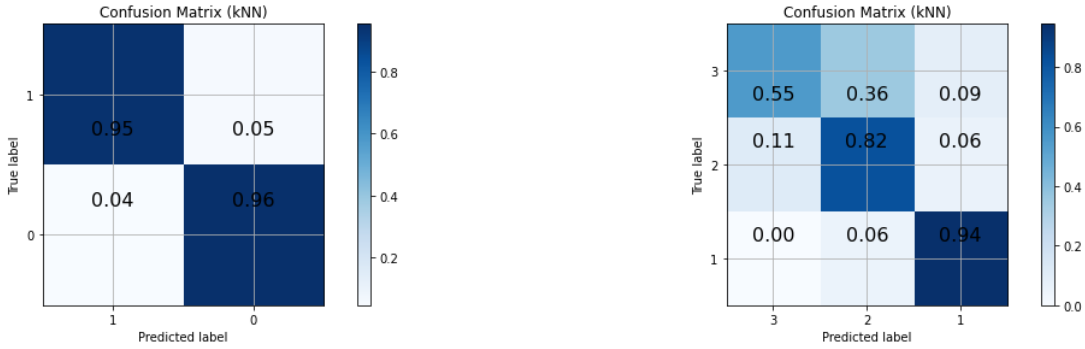
- binary classification: $\{0, 1\} \to \{0\}$, $\{2, 2\text{b}, 3\} \to \{1\}$, and
- multi classification: $\{0, 1\} \to \{1\}$, $\{2, 2\text{b}\} \to \{2\}$, and $\{3\} \to \{3\}$.

The resume of the results from the tree methods of data analysing are tabled bellow in the form Method 1/Method 2/Method 3.

Algorithm	Train set Accuracy(%)	Test set Accuracy(%)	Jaccard index	F1-score	R2-score
Binary kNN	86.10/97.50/83.99	86.10/**94.56**/80.66	0.86/0.95/0.81	0.86/0.95/0.80	0.44/0.78/0.04
Multiclass kNN	78.10/81.27/67.90	79.76/**83.08**/61.63	0.80/0.83/0.62	0.79/0.83/0.60	0.37/0.56/-0.31
Binary LR	85.80/95.24/77.11	85.26/**92.90**/78.25	0.85/0.93/0.78	0.85/0.93/0.78	0.36/0.71/-0.03
Multiclass LR	79.03/81.78/60.05	78.87/**80.64**/59.21	0.79/0.81/0.59	0.78/0.79/0.53	0.40/0.41/-1.00
Binary SVM	85.74/97.20/79.91	86.29/**95.43**/80.06	0.86/0.95/0.80	0.86/0.95/0.80	0.44/0.81/0.07
Multiclass SVM	78.43/83.57/64.12	79.89/**83.18**/62.24	0.80/0.83/0.62	0.79/0.82/0.59	0.43/0.56/-0.54
Binary DT	80.65/95.53/83.61	79.35/**94.56**/80.06	0.79/0.95/0.80	0.77/0.95/0.80	-0.32/0.78/0.11
Multiclass DT	75.33/78.14/67.82	70.65/**83.18**/61.93	0.71/0.83/0.62	0.69/0.82/0.62	0.06/0.52/-0.05

Table 1: The best Test set accuracy are in bold face.

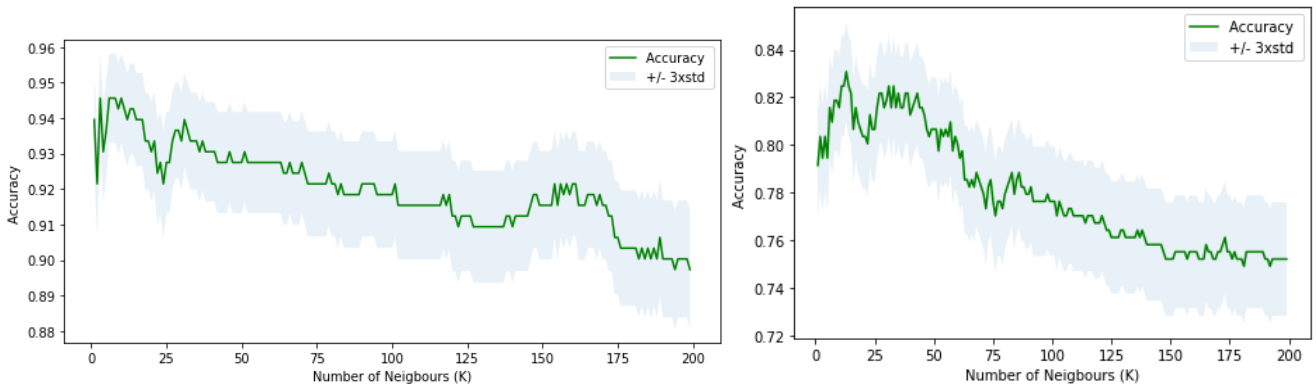
The results given after the Method 2 always prevail in both classifications, and the SVM and the Decision Tree prevail over the kNN just by 0.18% on the Test set accuracy. These are the normalised confusion matrices:



(a) Binary classification with kNN. The best k is 3.

(b) Multi classification with the kNN. The best k is 13.

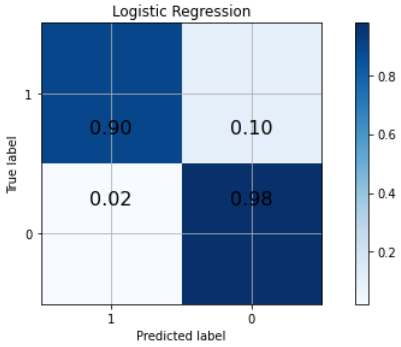
Figure 9: The confusion matrices of the classifications.



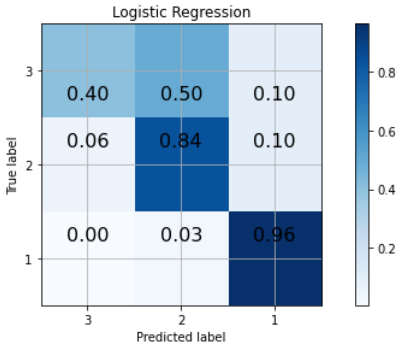
(a) Binary classification.

(b) Multi classification.

Figure 10: The Test set accuracy of kNN models such that $1 \leq k \leq 200$. The optimised accuracy of 94.56% and 83.08% corresponds to $k = 3$ and 13 for the binary and multi-classification problems, respectively.

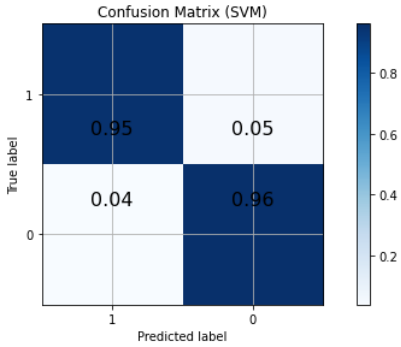


(a) Binary classification.

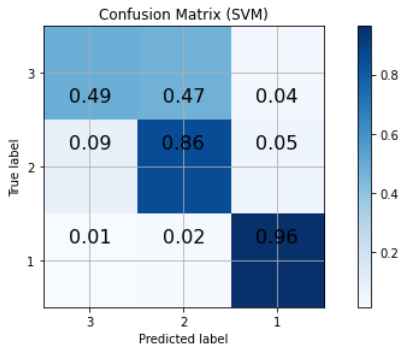


(b) Multi classification.

Figure 11: The confusion matrices of the Logistic Regression.

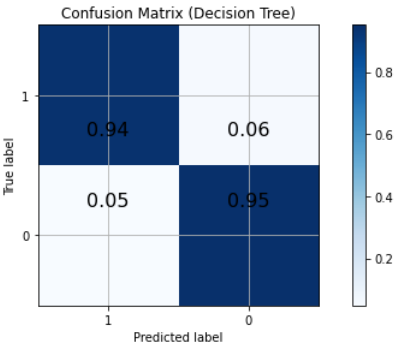


(a) Binary classification.

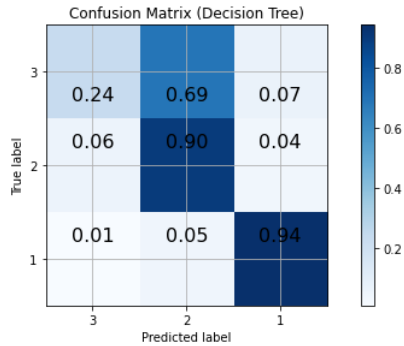


(b) Multi classification.

Figure 12: The confusion matrices of the Support Vector Machine.



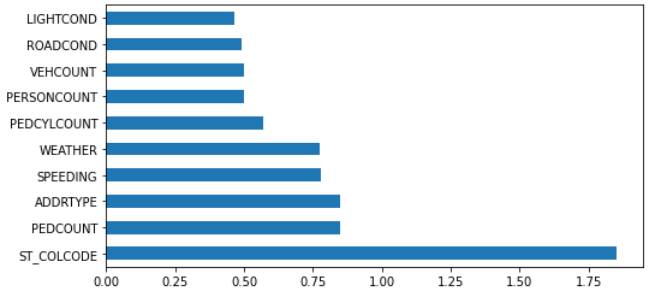
(a) Binary classification.



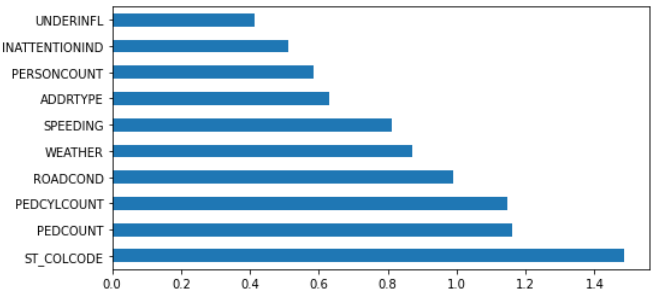
(b) Multi classification.

Figure 13: The confusion matrices of the Decision Tree.

The SVM has the advantage to rank the most influential factors for the target feature prediction:



(a) Binary classification.



(b) Multi classification.

Figure 14: The most relevant features given by the SVM classification algorithm for influencing the outcome of an accident.

In agreement with the Method 2 of data analysis, the greatest χ^2, the ST_COLCODE (the type of the collision) is the factor. The type of collision and the number of pedestrians may escalate the situation from 'only property damage' towards 'injury/severe injury/fatality'. Especially, the number of cyclists involved (PEDCYLCOUNT) distinguishes the fatality cases (SC ='3'), because the cyclists deaths represent almost 8.88% of the fatalities, and the SVM model captured this detail.

Against the intuition, probably because of the way by which the feature selections were made, the neighbourhoods in which an accident happened did not express any influence on the outcome.

5 Conclusions

In this study, it has been analysed how the severity of an transit accident might be predicted by some characteristics captured in general description. Primarily, it has been identified that the type of collision, number of pedestrians and cyclists determine mostly the severity of an accident.

Both regression model and classification models were built to predict how severe the outcome of an accident would be. These models can be helpful in the guide of public policies towards the public safety, and human life preservation. For example, by providing more compelling evidence for the speed limit near sidewalks and bike lanes.

6 Future directions

In these project, there have been some alternative strategies envisaged during the data analysis and the reach of the results.

The frequency of the accidents may not be homogeneous along the years: all the analysis made was based on the assumption that the number of accidents do not vary in a annual time period. Then, the data might divided in time periods in order to match a greater prediction accuracy.

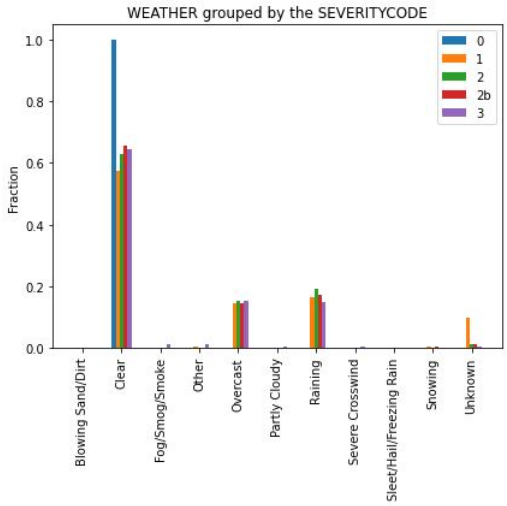
The specific collision types could be appointed: the most prominent feature for an outcome is the type of a collision. The feature extraction could take this in account by taking the most frequent types of collisions ('angles´, 'rear", etc), and transforming these in feature columns themselves. However, instead, by the methods here employed, the most frequent collisions is only discovered by analysis of a histogram.

A Appendix: the columns of the data frame

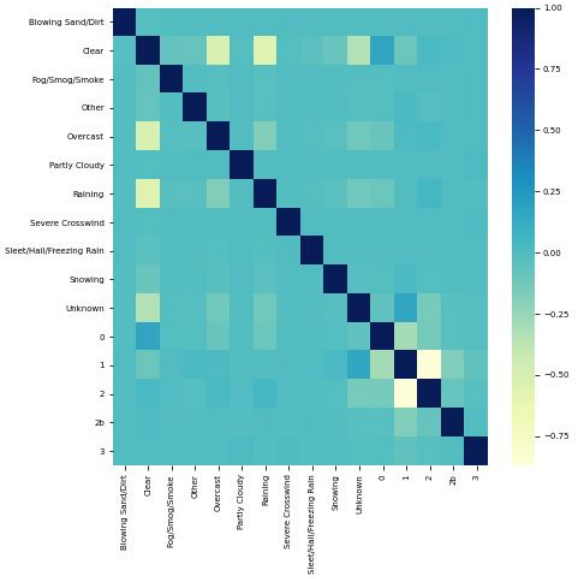
Column feature	Type	Description
X	float64	longitude(°)
Y	float64	latitude(°)
OBJECTID	int64	ESRI unique identifier
INCKEY	int64	A unique key for the incident
COLDETKEY	int64	Secondary key for the incident
REPORTNO	object	Number of the report
STATUS	object	
ADDRTYPE	object	Collision address type (alley, block or intersection)
INTKEY	float64	Key that corresponds to the intersection associated with a collision
LOCATION	object	Description of the general location of the collision
EXCEPTRSNCODE	object	
EXCEPTRSNDESC	object	
SEVERITYCODE	object	A code that corresponds to the severity of the collision
SEVERITYDESC	object	A detailed description of the severity of the collision
COLLISIONTYPE	object	Collision type
PERSONCOUNT	int64	The total number of people involved in the collision
PEDCOUNT	int64	The number of pedestrians involved in the collision
PEDCYLCOUNT	int64	The number of pedestrians involved in the collision
VEHCOUNT	int64	The number of vehicles involved in the collision
INJURIES	int64	The number of total injuries in the collision
SERIOUSINJURIES	int64	The number of serious injuries in the collision
FATALITIES	int64	The number of fatalities in the collision
INCDATE	object	The date of the incident
INCDTTM	object	The date and time of the incident
JUNCTIONTYPE	object	Category of junction at which collision took place
SDOT_COLCODE	float64	A code given to the collision by SDOT
SDOT_COLDESC	object	A description of the collision corresponding to the collision code
INATTENTIONIND	object	Whether or not collision was due to inattention (Y/N)
UNDERINFL	object	Whether or not a driver involved was intoxicated
WEATHER	object	Weather conditions
ROADCOND	object	The condition of the road during the collision
LIGHTCOND	object	The light conditions during the collision
PEDROWNOTGRNT	object	Whether or not the pedestrian right of way was not granted (Y/N)
SDOTCOLNUM	float64	A number given to the collision by SDOT
SPEEDING	object	Whether or not speeding was a factor in the collision (Y/N)
ST_COLCODE	object	A code provided by the state that describes the collision
ST_COLDESC	object	A description that corresponds to the state's coding designation
SEGLANEKEY	int64	A key for the lane segment in which the collision occurred
CROSSWALKKEY	int64	A key for the crosswalk at which the collision occurred
HITPARKEDCAR	object	Whether or not the collision involved hitting a parked car (Y/N)

Table 2: Table 1: description of the columns. Source: SDOT Traffic Management Division, Traffic Records Group

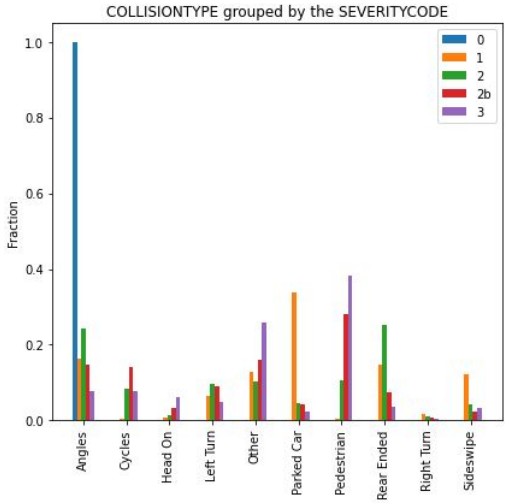
B Histograms and Correlation heatmaps



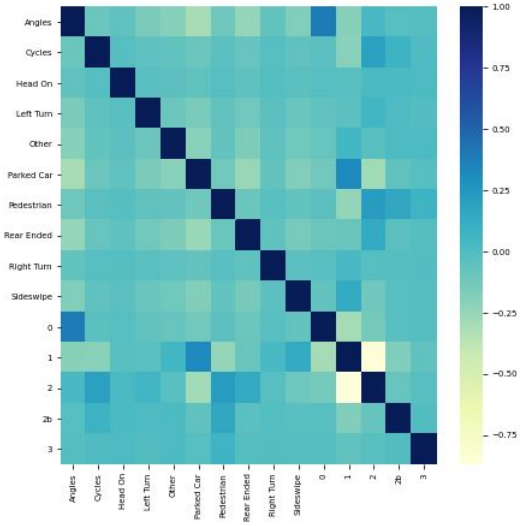
(a) Histogram of WEATHER gouped by the severity codes



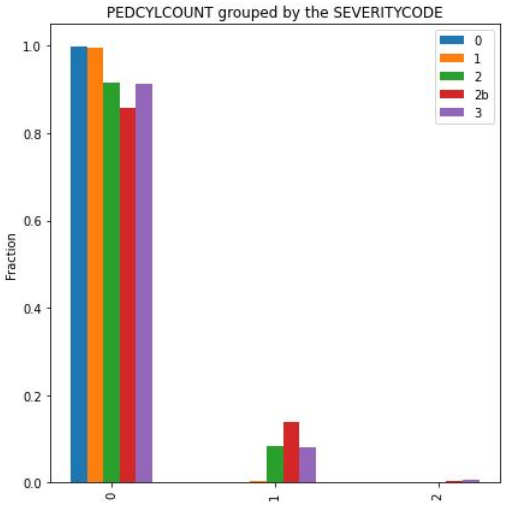
(b) The heatmap of the correlation matrix



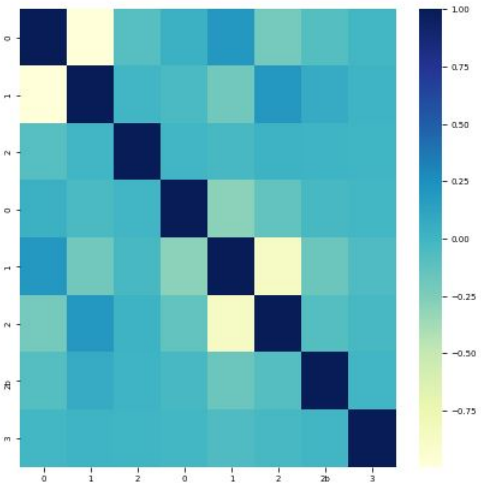
(a) Histogram of COLLISIONTYPE gouped by the severity codes



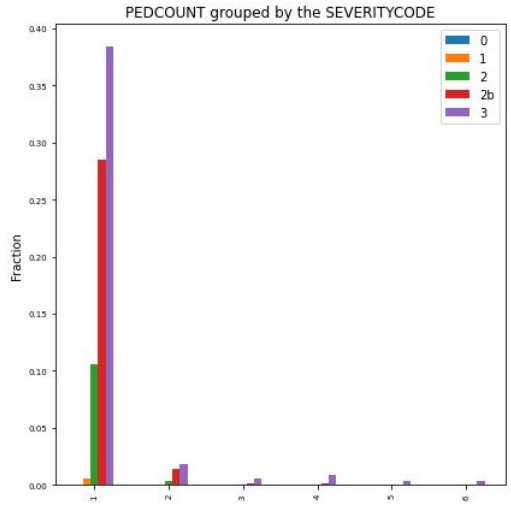
(b) The heatmap of the correlation matrix



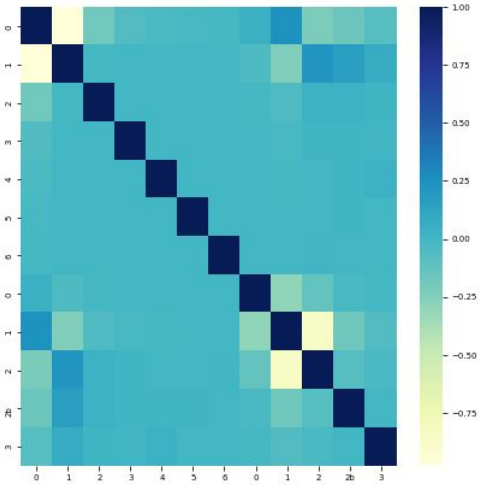
(a) Histogram of PEDCYLCOUNT gouped by the severity codes



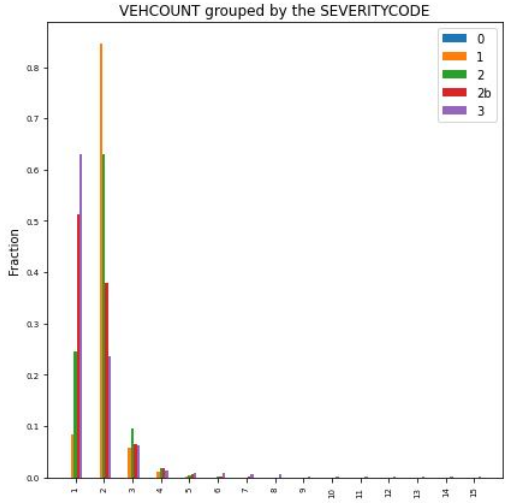
(b) The heatmap of the correlation matrix



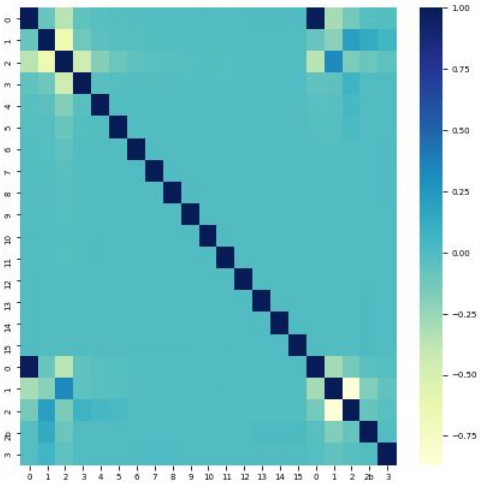
(a) Histogram of PEDCOUNT gouped by the severity codes



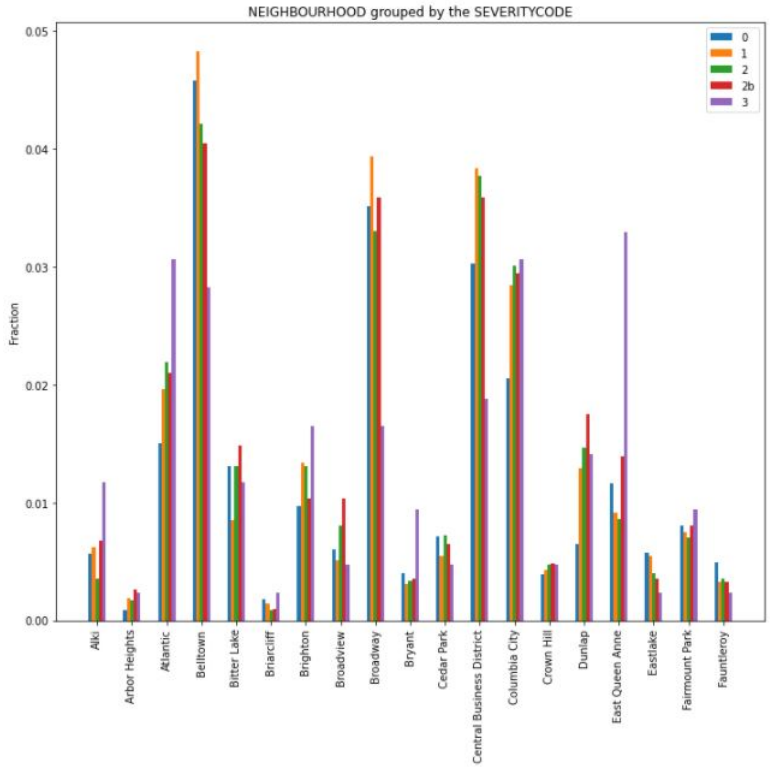
(b) The heatmap of the correlation matrix



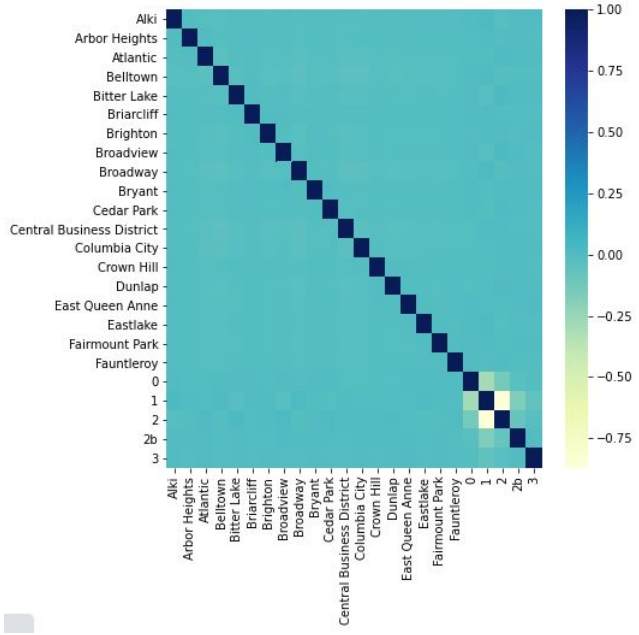
(a) Histogram of VEHCOUNT gouped by the severity codes



(b) The heatmap of the correlation matrix



(a) Histogram of NEIGHBOURHOOD gouped by the severity codes



(b) The heatmap of the correlation matrix

Figure 20: The case with NEIGHBOURHOOD is that most of the Pearson Coefficients are very close to zero, e.g., the correlation between 'East Queen Anne' and the severity codes (0,1,2,2b,3) are (0.0053, -0.0010, -0.0041, 0.0062, 0.0129) with P-values (0.017, 0.650, 0.070, 0.006, 10^{-8}). Although the correlations are around zero, the P-values are garantee the rejection of the null hypothesis.