TÉCNICO LISBOA

DS 2021

Data Science Project

Team nr: 7	**Student 1 :** Alexandre Antunes Rodrigues	**IST nr:** 89404
	Student 2 : Guilherme Jardim	**IST nr:** 92471
	Student 3 : Ricardo Monteiro de Santos Ferreira	**IST nr:** 87701

1 DATA PROFILING

1. Data Dimensionality

The **NYCC** dataset has 45669 instances and 21 attributes (*Fig.1*) and the **AQC** dataset has 169273 records and 32 variables (*Fig.2*), however we only used 20% of the observations because our computer machines don't have high processing capacity. Therefore we are not dealing with high dimensional data.
The relation between the observations and the features is promising to obtain good results on later analysis, and it prevents us from facing the curse of dimensionality problem.
NYCC dataset has four features in which the missing values surpass the 85% (PED_LOCATION, CONTRIBUTING_FACTOR_2, CONTRIBUTING_FACTOR_1, PED_ACTION) and therefore will probably not be very useful in the sense that they will not contribute to obtaining better results. To finish this section, we also observed that only four attributes are of numeric type. The remaining ones, 2 belong to date, and 14 belong to symbolic types. The fact that we have so many symbolic features can cause us difficulties when trying to convert them to ordinal ones later.
For the **AQC** dataset most of the variables are numeric (27), four are symbolic, one is the binary variable target "ALARM" and zero is date. About missing values after an analysis it is concluded that all chemical compounds in the air have missing values between 1500 and 1600 and the variable FIELD_1 has 3488 missing values.

2. Data Granularity

In order to study the granularity for the numerical features, we plotted histograms varying the number of bins for each variable.
For the **NYCC** dataset regarding the PERSON_AGE feature, the chart obtained only presents one bar with 10, 100 and 1000 bins. This might be due to the extreme negative and positive values that we address in the data distribution section. With this, we decided to do the same as we did in the data distribution and narrow the data to people whose age is between 0 and 120. We obtained a much clearer chart, and we can observe that throughout the three charts, the shape remains the same and therefore we can confirm a log normal distribution. With this we can also think about grouping the ages, going from 10 to 10 since we probably don't need such fine granularity and with this improve the results of the information discovery stage.
Regarding the CRASH_DATE feature, we decided to draw a time hierarchy that could help us to aggregate our date variables into a coarser one. We decided to group the year in weeks and the weeks in the days of the week.
Regarding the other symbolic variables, we applied data taxonomy in order to have the data organized with a specific criterion and to try to improve the efficiency of discovering information.
For **AQC** dataset the binary variable which was just one, the target variable, we just used 2 bins because it just has 2 possible values (SAFE and DANGEROUS) and for the numeric values we tried with 10, 100 and 1000 bins to get a global perspective about which one better fits all the variables.

3. Data Distribution

After analyzing the histogram of the CRASH_TIME feature of the NYCC dataset we can see that most accidents happen in the afternoon, more specifically from 2 pm to 6 pm. This might be useful information to use in the discovery information phase.

To analyze the data distribution of the numerical variables we used box plots to have an univariable analysis about mean, quartiles and outliers and a visual representation of the scale and we noticed that the PERSON_AGE feature of the **NYCC** dataset, had some outliers that will need to be handled later. We observed that the values ranged between -971 and 9999 and due to this, it was a little difficult to observe what type of distribution this variable followed. We found it acceptable to remove values less than 0 and greater than 120 and re-study their distribution. We observed that the data is close to a log-normal distribution and because of this standard deviation will probably be a better approach to identify and remove them. Also, depending on our approach towards the date features it is probably a good idea to treat them due to their different scales.

We also found that there were four ID features that, after evaluating the charts, we decided that would make sense to discard them later, since they do not seem to encode a meaningful feature for the output and will not contribute to obtaining better results.

Our target variable is a binary variable very unbalanced (ratio of 253:45416) where the INJURED value represents the majority and KILLED represents the minority. This can lead the algorithms to be biased towards the injured class and obtain poor results.

For the **AQC** dataset we analyzed that mean, min, max and std of CO, NO2, O3, PM2.5, PM10 and SO3 have lots of outliers above a specific value. The numeric variables with more outliers are FID, SO2_max and SO2_Std with more than 2800 outliers.

We also used a function to compare histograms with Normal, Exponential and LogNormal distribution. All the numeric variables have all three distributions. The normal distribution is right skewed.

4. Data Sparsity

We used scatter plots to study the correlation between the variables, but we were not very lucky in discovering interesting results because since we are dealing with a lot of qualitative features it makes it very hard to read this type of plots. However, we were able to see that there is a very good cover of the data throughout the dataset. Also, the **NYCC** target variable seems to not be correlated with any feature. Saying this, we decided to use a heatmap to study their correlation, using the Kendall measure, and we obtain some good correlations between some of the symbolic features as we can see in the appendix. However, since the Kendall correlation operates with ordinal, interval or ratio variables that happen to not be our variable type, the correlation values may not be very accurate although some make sense.
For the **AQC** dataset we analyzed that there are lots of redundant variables….

2 DATA PREPARATION

Before we begin, we must define the measure we are going to use to evaluate our models. We quickly realized accuracy was always very high, especially in set 1, because the data set is extremely unbalanced. A model that only predicts the negative class will still be evaluated with a high accuracy. We are not sure which are more prejudicial: false positives or false negatives. For this reason, we couldn't choose between precision and recall, and decided to use f1_score, which takes both into consideration. We will measure the quality of our models mostly using it, but we will always show accuracy and in specific situations we will also present recall and precision.

In terms of training techniques, since both datasets had a big number of records, train-test splitting was the right technique.

5. Missing Value Imputation

As we saw in the data profiling phase, the **NYCC** dataset has four features in which the missing values surpass the 85% (PED_LOCATION, CONTRIBUTING_FACTOR_2, CONTRIBUTING_FACTOR_1, PED_ACTION). Due to this, we decided to remove these features completely. Before proceeding with the imputation values, we decided to remove the samples with missing information beyond 50% for both sets. To the remaining features which had missing values, we imputed the mean value to missing values of numeric features and the most frequent value to symbolic ones. We found

this could be a good solution because the missing value percentage never exceeds 15% which is not a significant proportion. We also verified that choosing between mean, median and most frequent had very little impact on the performance of the models, as shown below.

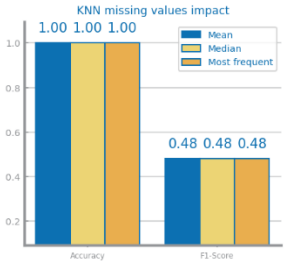


*Fig. 1a) **NYCC** KNN missing values impact*

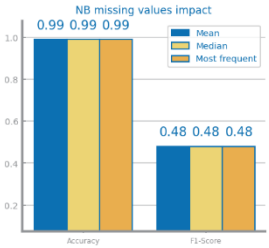


*Fig. 1b) **NYCC** NB missing values impact*

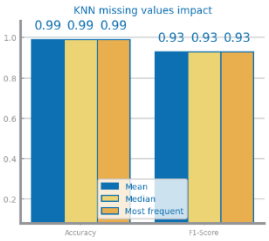


*Fig. 1c) **AQC** KNN missing values impact*

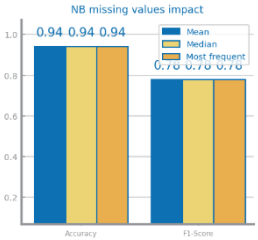


*Fig. 1d) **AQC** NB missing values impact*

6. Dummification and other transformations

Regarding the **NYCC** dataset, we started by dropping the ID features (VEHICLE_ID, PERSON_ID, UNIQUE_ID, COLLISION_ID) that we also talked about in the data profiling phase. After that, we decided that the minutes of the collision hour won't be of our interest, so we created a new feature (CRASH_HOUR) only with the hour of the day and dropped the CRASH_TIME. We think that the moment of day might be a good feature to describe the target. We did something similar to the CRASH_DATE, where we dropped the column and created a new one (CRASH_DAY_WEEK) with the day of the week. This is very important since we can work with data as an "Ordered factor" which will be helpful when moving for further analysis. Having the date variable months, quarters or semesters is better than have it in days. However, since we are dealing with data within less than a year, we might not have enough information to make conclusions about that period if we aggregate in those time periods. Therefore, we found a better option to aggregate the year in days of the week and with this we maximize the number of periods we have in the dataset.

In the next step, we dropped the samples in which the BODILY_INJURY variable had the value "Does not apply" since there were only 5 samples with that value, and this will cause a problem on making this feature ordinal.

In order to minimize the number of features that would need Dummification, we applied the taxonomies that we created in the study of the granularity to the variables and next, mapped it to values in order to turn the variables in ordinal ones. After this, we mapped all features that had the values "Unknown", "Unspecified" and "Does Not Apply" to missing values and then we proceeded as already explained in the Missing value imputation section.

Regarding the **AQC** dataset, we also dropped the ID's, and we decided to drop the date. We considered that the date of the measure does not influence the level of alarm. This dataset had just four more symbolic variables, Prov_EN, City_EN, ProvGb and CityGb. We dropped CityGb and ProvGb because they were redundant. Also, we felt city level was a level of granularity we didn't need since there were hundreds of different cities, so we dropped City_EN. We ordered the remaining provinces by decreasing population density, as this affects the province's pollution, and we transformed it into an ordinal variable.

7. Outliers Imputation

About the **NYCC** dataset for the treatment of the outliers, we decided to use Standard deviation since our feature was close to a normal distribution. We can observe in the picture below that we obtained a better boxplot when compared to before removing the outliers.

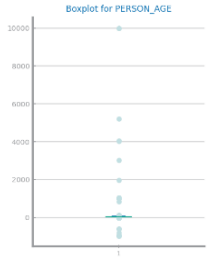


Fig. 2a) Boxplot before removing noise

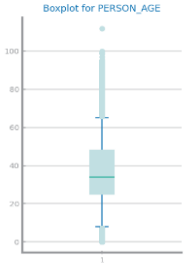


Fig. 2b) Boxplot after removing noise

*Fig. 2c) **NYCC** KNN outliers impact*

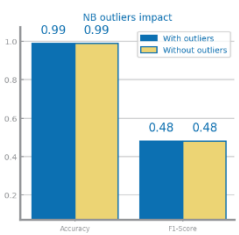


*Fig. 2d) **NYCC** NB outliers impact*

On the **AQC** dataset, we have thousands of outliers in the gas measures. Our first idea was to increase the n in the standard deviation calculation, meaning we would just delete the records with values 3 standard deviations away from the mean and 1.5 IQR away from the first and third quadrants. However, this softer method still eliminated a big number of records (20000) and, even worse, a big proportion of those were positive records (8000), which means these are interesting outliers relevant to the class instead of noise. Therefore, we tried eliminating the values where the boxplot was no longer continuous, and we tried setting the records higher than the 95th percentile to the 95th percentile value. The impact showed keeping the outliers was the best alternative.

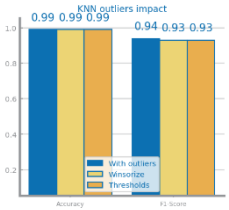


*Fig. 3a) **AQC** KNN outliers impact*

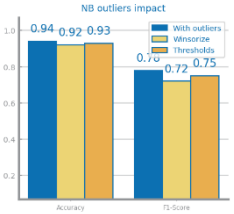


*Fig. 3b) **NYCC** NB outliers impact*

8. Scaling

In terms of scaling, we had to decide between the zscore and minmax methods. We measured their impact in KNN and Naïve Bayes. On the **NYCC** dataset, zscore showed an improvement on the KNN results, so it was the chosen method. As for the **AQC** dataset, no improvement was observed on the KNN, and the Naïve Bayes performed the same regardless of the scaling. We can explain this phenomenon by noting that most of the features of this dataset are measurements in the same unit. Therefore, their scale is already similar and further away values are supposed to have more impact. Nevertheless, we still chose zscore because it deals well with outliers.

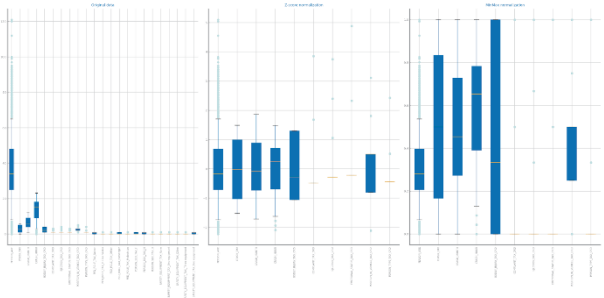


*Fig. 4a) **NYCC** scaling*

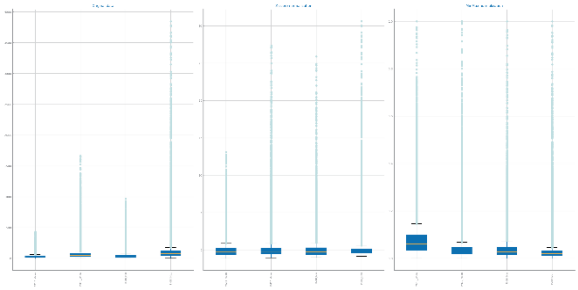


*Fig. 4b) **AQC** scaling example*

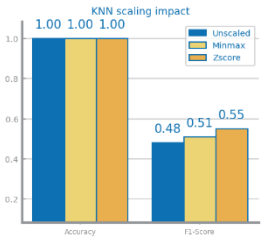


*Fig. 5a) **NYCC** KNN scaling impact*

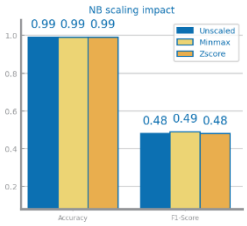


*Fig. 5b) **NYCC** NB scaling impact*

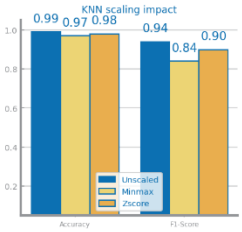


*Fig. 5c) **NYCC** NB scaling impact*

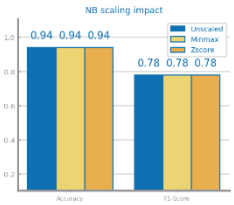


*Fig. 5d) **NYCC** NB scaling impact*

9. Balancing

Regarding the balancing, we have three main techniques: oversampling, undersampling and SMOTE. Even though the first set is highly unbalanced, balancing techniques made the predictions worse, not only in terms of accuracy, but in terms of f1-score. We were expecting the recall to improve and were surprised that wasn't the case. Similarly, on the set 2, predictions also worsened. In this case, we believe it's because we already have enough positive records for the recall to be high, and an equal proportion caused more false positives and consequently caused a decline of the precision, worsening the f1-score. Due to these results, we decided to move forward without balancing in either set.

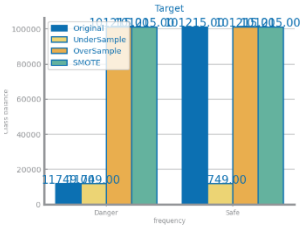


*Fig. 6a) **AQC** balancing*

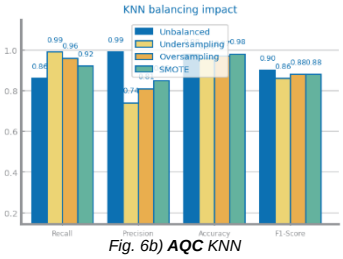


*Fig. 6b) **AQC** KNN balancing impact*

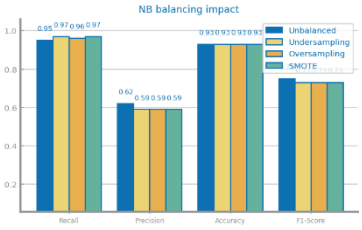


*Fig. 6c) **AQC** NB balancing impact*

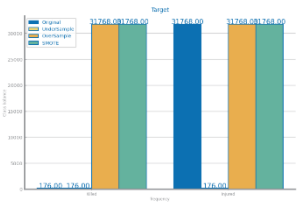


*Fig. 7a) **NYCC** balancing*

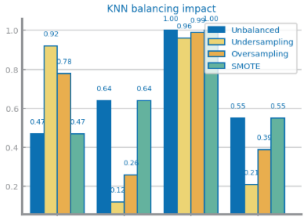


*Fig. 7b) **NYCC** KNN balancing impact*

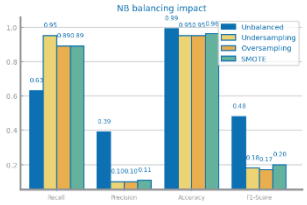


*Fig. 7c) **NYCC** NB balancing impact*

2.1 Feature Engineering

10. Feature Selection

To remove all the redundant variables, we studied the correlation between them and defined a threshold at 0,90. The groups of variables highly correlated were removed, leaving only one variable and we should not lose any information. The variables removed in the **AQC** dataset were ['CO_Std', 'PM2.5_Min', 'PM2.5_Max', 'PM10_Std', 'SO2_Std'].

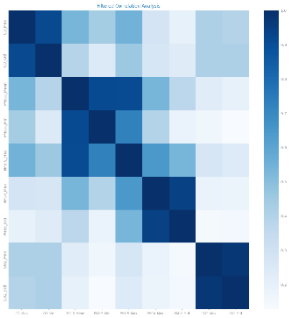


*Fig. 8a) **AQC** correlation between variables*

*Fig. 8a) **NYCC** correlation between variables*

Afterwards, we need to remove the irrelevant variables. We used another unsupervised method consisting of studying the features variance and discarding the ones below a threshold. If variable has little variance, it also adds very little information about the class. In the **AQC** dataset, there was a very clear interval between variables with the lowest variance and the rest.

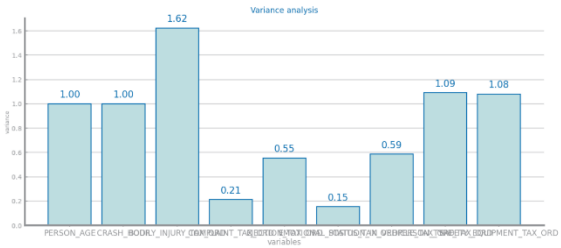


*Fig. 8c) **NYCC** variance analysis*

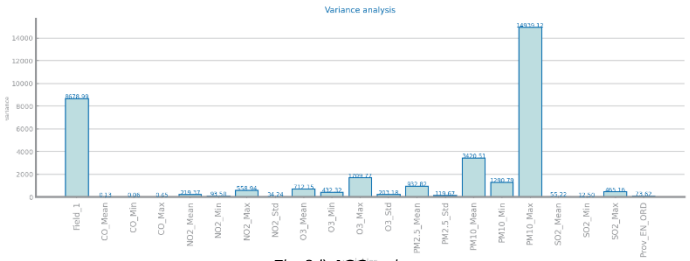


*Fig. 8d) **AQC** variance analysis*

Finally, we have the supervised technique, which consists of evaluating the information each variable gives about the class. There are many options to measure this. The two we considered were the chi-square and the mutual information, which works the same way as the entropy. Since the chi-square is similar to the correlation which we are already using above, we decided to use mutual information. On the **NYCC** dataset, the redundant and irrelevant variables intersected, and consequently we opted to remove the 6 more irrelevant variables, according to the measure explained above

(['PERSON_AGE', 'PERSON_SEX_TAX_U', 'CRASH_HOUR', 'CRASH_DAY_WEEK', 'PED_ROLE_TAX_Other', 'PED_ROLE_TAX_In-Line Skater']). On the **AQC** dataset, we discovered a very interesting phenomenon. We kept reducing the variables and ran KNN and NB models until the results stopped improving. We discovered that a very little number of variables was needed to predict the class, as shown below. We ended up only keeping 4 variables ([PM2.5_Mean, PM10_Mean, PM10_Min, PM10_Max]). The feature selection impact will be shown in each classification model and therefore won't be shown here to avoid redundancy.

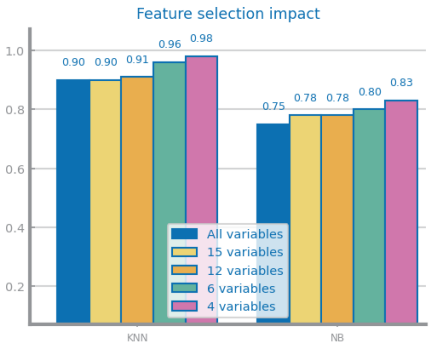
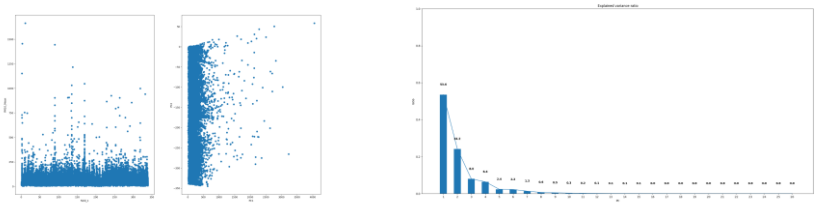


*Fig. 8e) **AQC** feature selection impact on KNN and NB*

11. Feature Extraction

Shall contain all relevant information and charts respecting to feature extraction, in particular PCA. The different choices and their impact on the modeling results shall be presented and explained.

set2



For both datasets the target variable was deleted because we are in an unsupervised learning field.
About the **AQC** dataset we choose two variables with high variance (FIELD_1 and PM10_MEAN). The variance ratio is used to measure how dispersed or clustered the set of events are. With the "Explained Variance Ratio" plot we noticed that the blue bars show the percentage variance explained by each principal component. The first principal component explains 53.6% of the variance of the data set, the second one explains 24.2%, the third one 8.0% and the last 23 explains above 8%. We can reduce the number of variables from 26 to 1 which means that 1 variable is still 53.6% of the information within the original dataset. As we have a ration between 0 and 1 we can say that the distribution could be binomial.

3 CLASSIFICATION

Summarizing, both our sets' missing values were replaced with the mean for numeric variables and most frequent for symbolic variables. On the **NYCC** dataset, the age noise was removed and on the **AQC** dataset the outliers were kept.

We used zscore scaling and no balancing for both sets. In terms of feature selection, on the **NYCC** dataset we kept the best 12 variables and on the **AQC** dataset only 4.

3.1 Naïve Bayes

In the Naïve Bayes model, we weren't able to run the Multinomial regression because zscore generates negative values, not accepted by it.

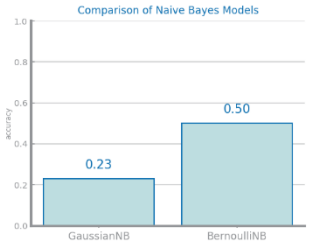


*Fig. 9a) **NYCC** without feature selection*

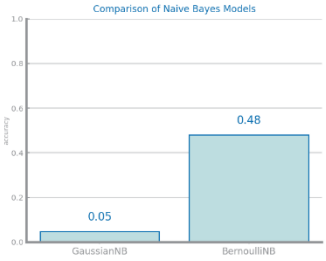


*Fig. 9b) **NYCC** with feature selection*

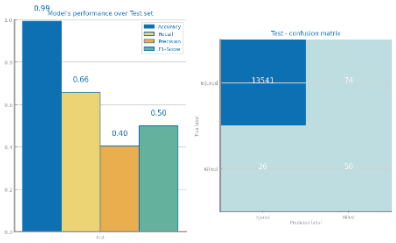


*Fig. 9c) **NYCC** best NB model*

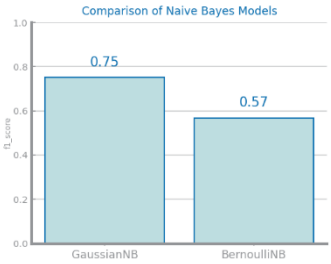


*Fig. 9a) **AQC** without feature selection*

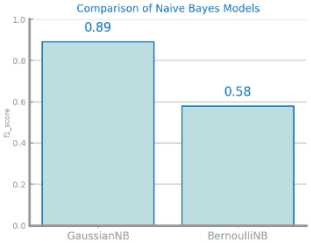


*Fig. 9b) **AQC** with feature selection*

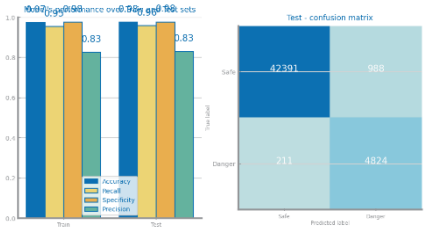
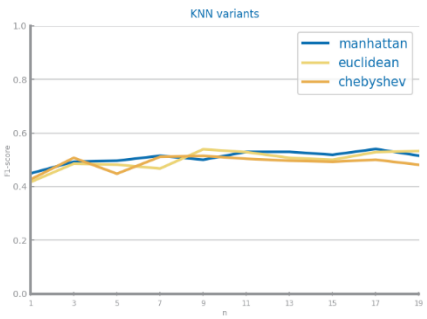


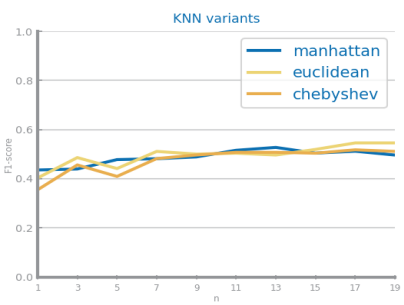
*Fig. 10c) **AQC** best NB model*

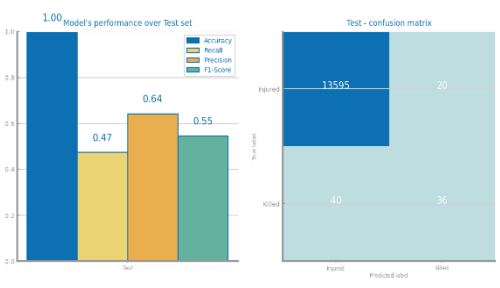
3.2 KNN

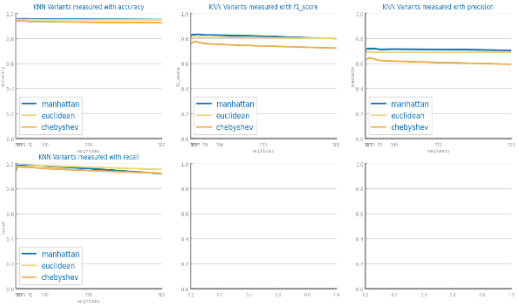
We used three different metrics for the KNN, Manhattan, Euclidean and Chebyshe. We didn't use Jaccard since it always predicted the negative class and therefore had a score of 0 with our chosen measure. For **AQC**, we can also see the impact of the feature selection. Since KNN attributes the same weight to every variable, removing the ones who didn't give us information about the class allowed our model to focus on the right variables and there was a big increase in the score.

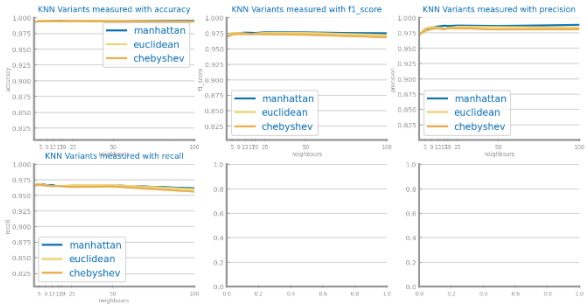
We can also conclude that the KNN model wasn't affected by overfitting. This model doesn't learn any parameters and therefore is not subject to overfitting.

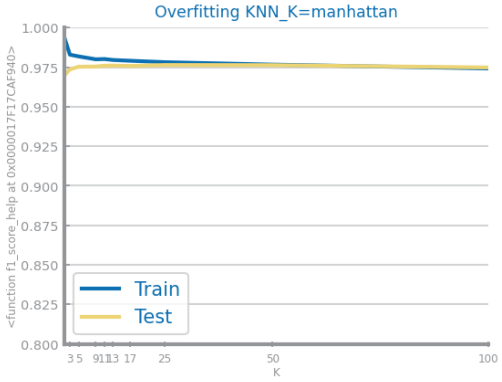


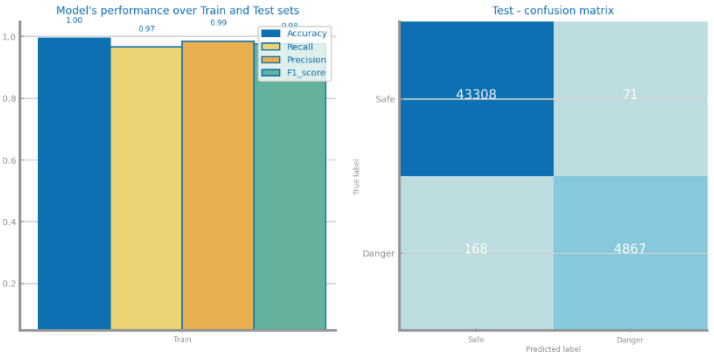








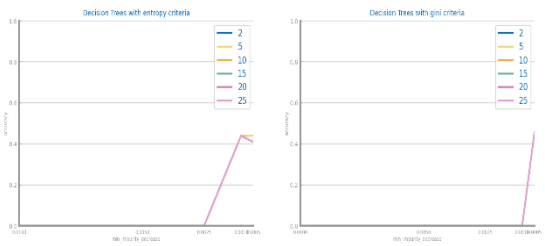


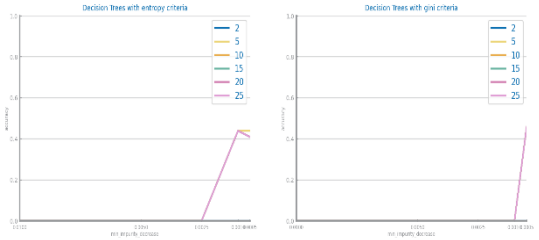


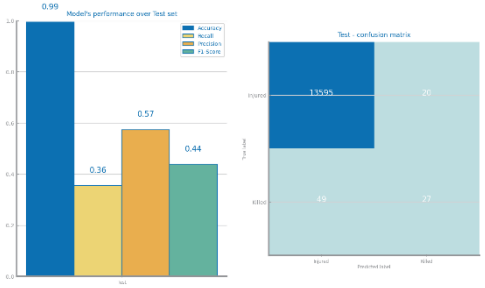
3.3 Decision Trees

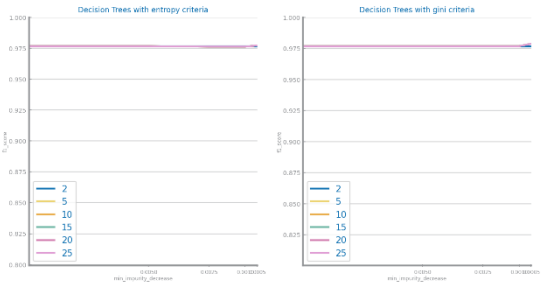
A decision tree has nodes that represent variables and in order to find the optimum split of the features we used gini and entropy criteria. Decision trees are based on the thresholds that maximize the impurity decrease of each node.

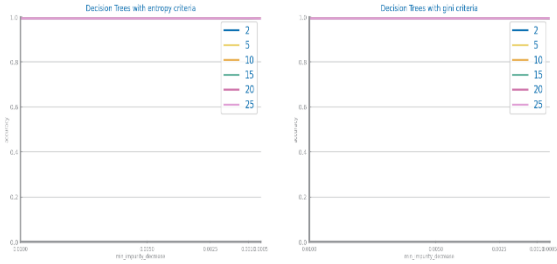
For both datasets we carried out a study without and with Feature Selection.

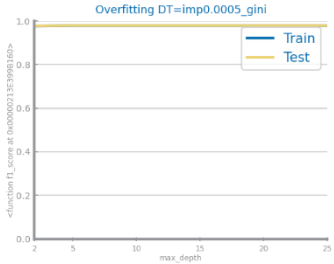








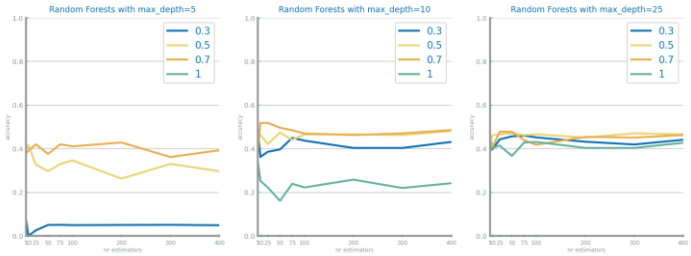




3.4 Random Forests

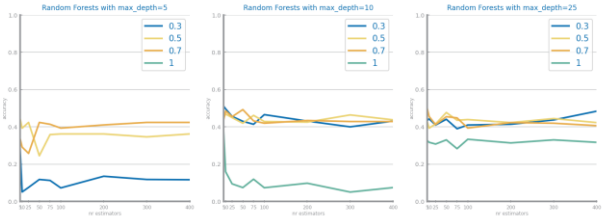
For both datasets we carried out a study with and without Feature Selection. First, we got the parametrizations of the best model to be used later in the same model with Feature Selection.

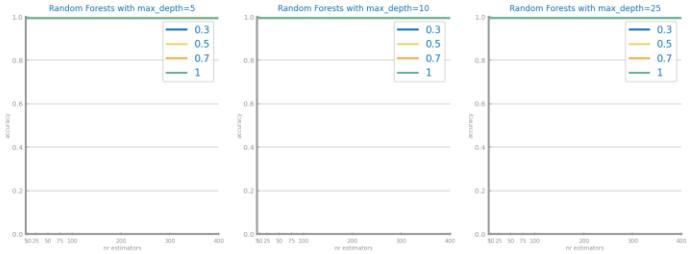
sem FS ->0.513888888888889

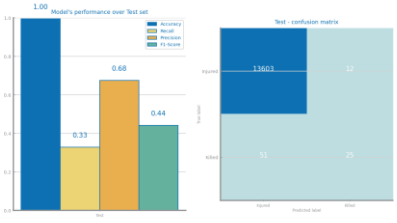


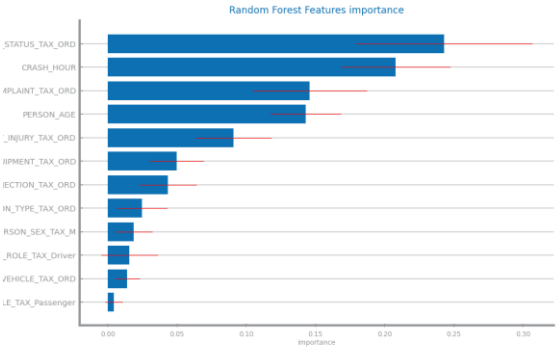
---

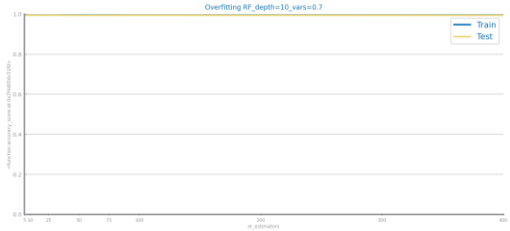
com FS -> 0.5333333333333333









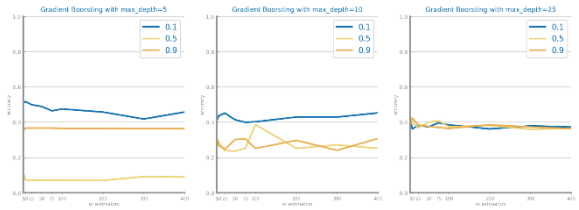


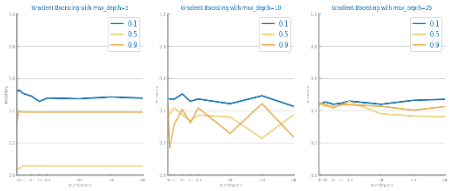
Best results achieved with entropy criteria, depth=5 and min_impurity_decrease=0.00 ==> accuracy=0.99

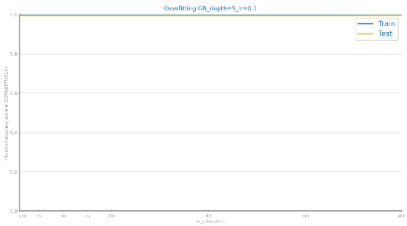
3.5 Gradient Boosting

For both datasets we carried out a study with and without Feature Selection. First, we got the parametrizations of the best model to be used later in the same model with Feature Selection.

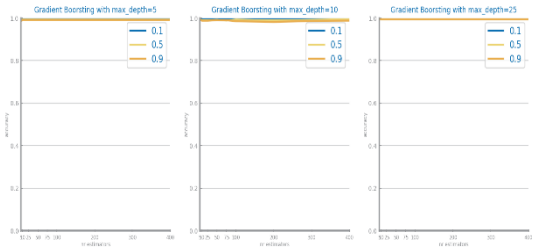
Set 1

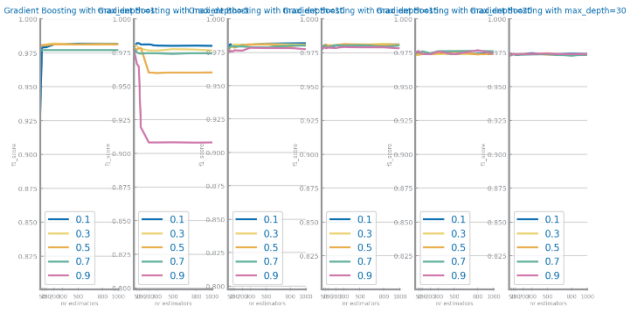


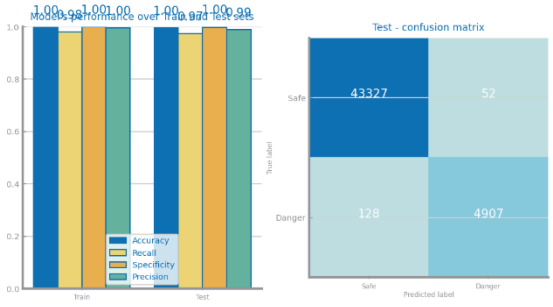


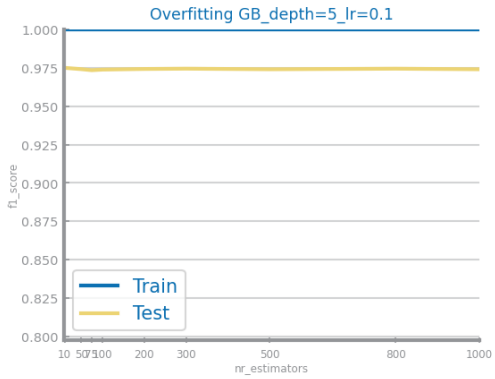


Set 2





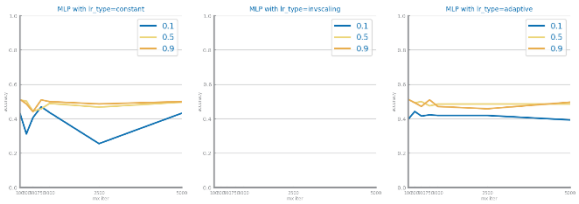


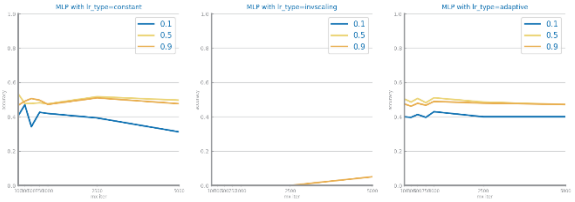


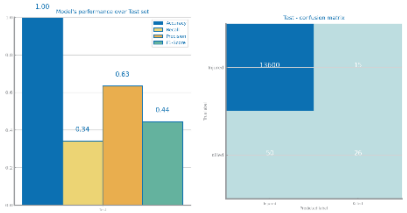
3.6 Multi-Layer Perceptrons

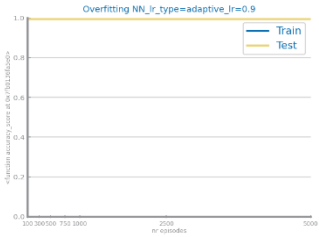
For both datasets we carried out a study with and without Feature Selection. First, we got the parametrizations of the best model to be used later in the same model with Feature Selection.

Set 1

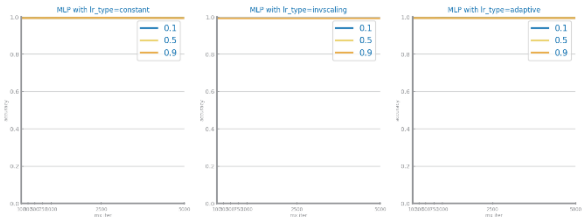


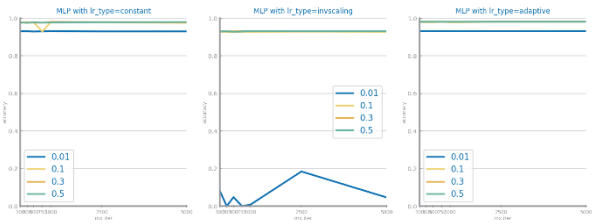


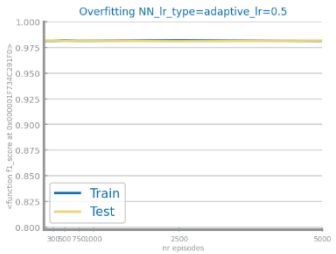


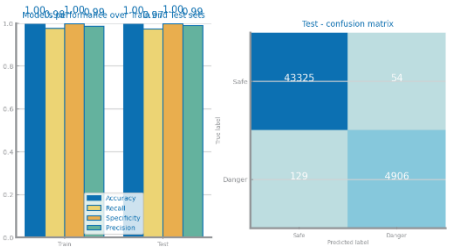


Set 2









4 CLUSTERING

For clustering we decided to choose the features with the highest variance. On set NYCC we choose the variables PERSON_AGE and BODILY_INJURY and for AQC we choose Field_1 and PM10_Max. For both data sets, we observe a decrease of MSE with the increase of clusters regarding the KMeans. We obtained the best results using PCA and feature selection. Using the elbow rule, regarding the fig. 1a) and 2a) we choose 3 and 5 clusters, respectively, for both data sets, which were also the number of clusters with the highest silhouette coefficient. For EM we also obtained better results using PCA and feature Selection and we obtained the best results when using 3 clusters, for both datasets. The same happened with Hierarchical and, regarding the cluster and the PCA and feature selection. The best result was for an average link with cityblock and k=3 for NYCC dataset and we obtained the same best result for an average link with cityblock and chebyshev. Jaccard metric performed bad in general maybe because it works better for binary, and we have none in both datasets. Regarding DBSCAN, we considered that we obtained poor results. We achieved the best result with cosine eps=0.08 and k=2 for NYCC data set and

chebyshev eps=169.94 and k=8. In general, the clustering algorithms improve with PCA and feature selection applied.

Fig. 1a) ***NYCC*** Measures KMeans

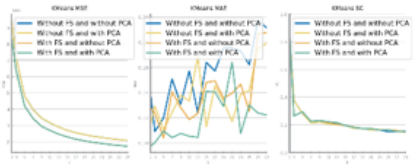


Fig. 2a) ***AQC*** Measures KMeans

Fig. 1b) ***NYCC*** Measures EM

Fig. 2b) ***AQC*** Measures EM

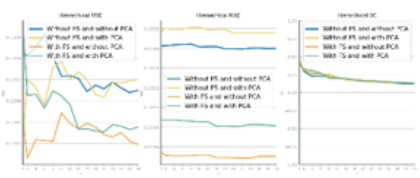


Fig. 1c) ***NYCC*** Measures Hierarchical

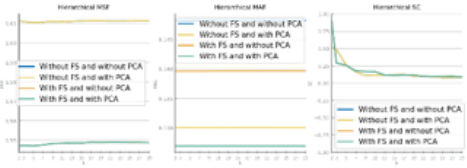


Fig. 2c) ***AQC*** Measures Hierarchical

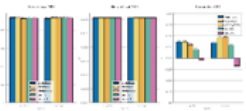


Fig. 1d) ***NYCC*** Measures Hierarchical – Without FS and without PCA

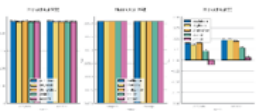


Fig. 1e) ***NYCC*** Measures Hierarchical – Without FS and with PCA

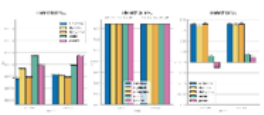


Fig. 2d) ***AQC*** Measures Hierarchical – Without FS and without PCA

Fig. 2e) ***AQC*** Measures Hierarchical – Without FS and with PCA

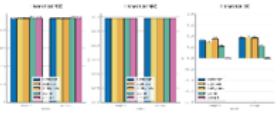


Fig. 1f) ***NYCC*** Measures Hierarchical – With FS and without PCA

Fig. 1g) ***NYCC*** Measures Hierarchical – With FS and without PCA

Fig. 2f) ***AQC*** Measures Hierarchical – With FS and without PCA

Fig. 2g) ***AQC*** Measures Hierarchical – With FS and without PCA

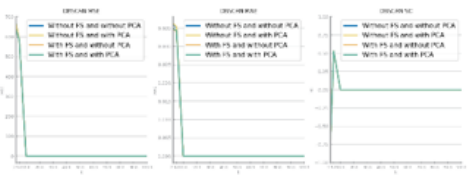


Fig. 1h) ***NYCC*** Measures DBSCAN

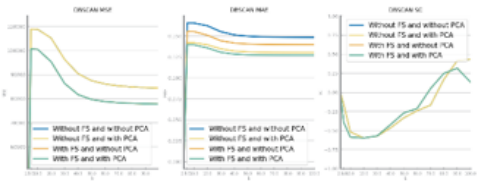


Fig. 2h) ***AQC*** Measures DBSCAN

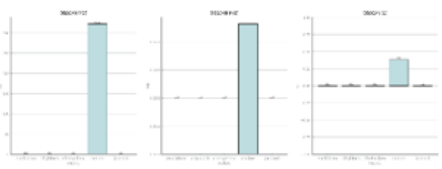


Fig. 1i) ***NYCC*** Measures DBSCAN

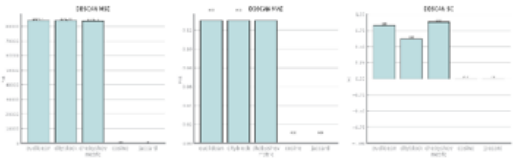


Fig. 2i) ***AQC*** Measures DBSCAN

5 ASSOCIATION RULES

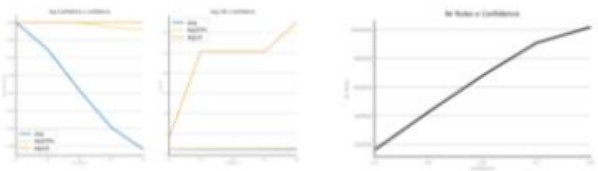


Fig. Quality evaluation per support **NYCC** dataset

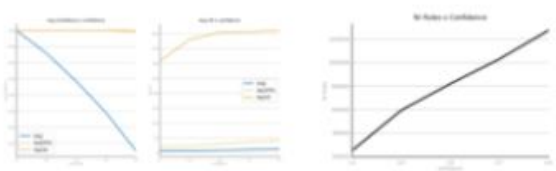


Fig. Quality evaluation per support **AQC** dataset

About evaluation per confidence for the **NYCC** dataset confidence and average confidence are inversely proportional when we talk about average rules and average confidence is constant when the confidence increases when we talk about top 25% of the rules and top 10 rules. Average lift and confidence are directly proportional, constant and directly proportional again. The number of rules and confidence are directly proportional. For the **AQC** dataset the results are similar except about Average lift and confidence with a less accentuated growth in relation to the top 10 rules.

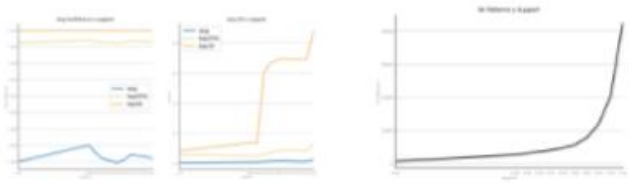


Fig. Quality evaluation per confidence **NYCC** dataset

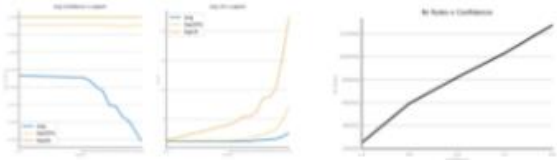


Fig. Quality evaluation per confidence **AQC** dataset

About evaluation per support for the **NYCC** dataset the average confidence and support are constant when we talk about the top 25% of the rules and top 10 rules and have highs and lows in relation to the average. About average lift and support there is an increase of both in relation to top 10 rules. For the **AQC** dataset the results are similar except about Average confidence and confidence with a sharp drop in relation to the average.

6 CRITICAL ANALYSIS

In terms of the **NYCC** dataset, we concluded we had a too low number of positive records. There was no balancing technique able to reduce the number of false negatives, without increasing a lot the false positives. This made it impossible to learn good models. Although every model will have a very high accuracy score, their ability to actually identify the positive records is always bad. Therefore, no models would be useful in a real situation. Regarding the **AQC** dataset, we found that we need very few variables to predict the class variable. It was almost a direct correlation and therefore our models were close to perfect.

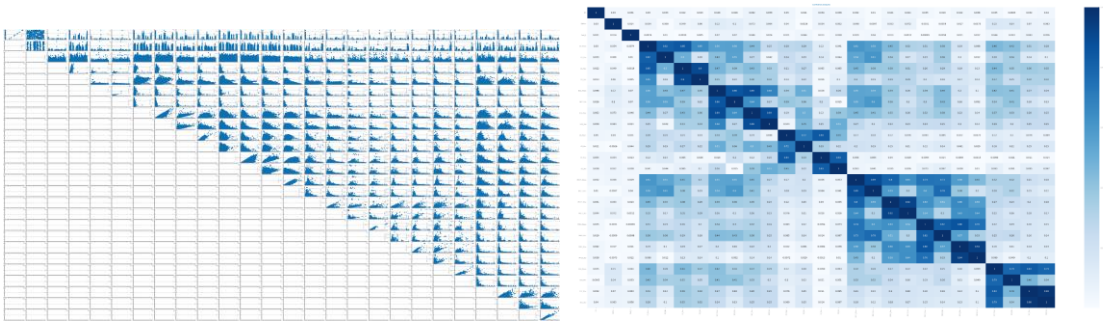
2. APPENDIX (OPTIONAL)

To be used only for presenting less relevant charts for data profiling, and following the same order as before. No text will be considered, only the charts.

1.

Set2

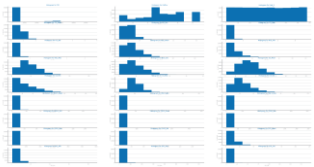
2. Correlation



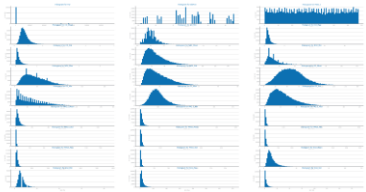
3.

4. Granularity

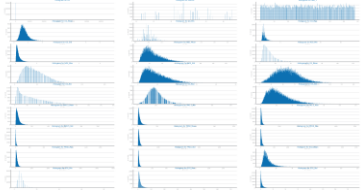
10bins



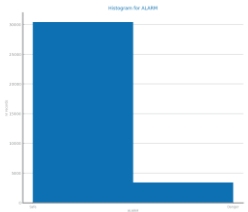
100bins



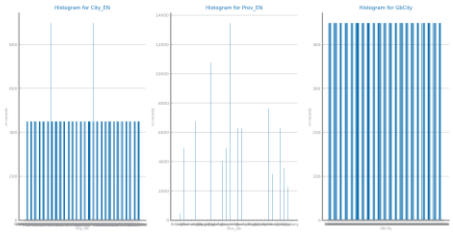
1000bins



Target Variable



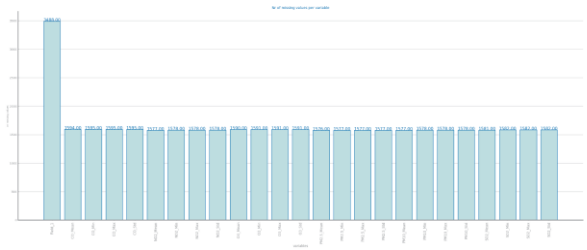
Symbolic variables

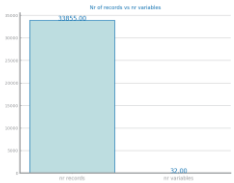


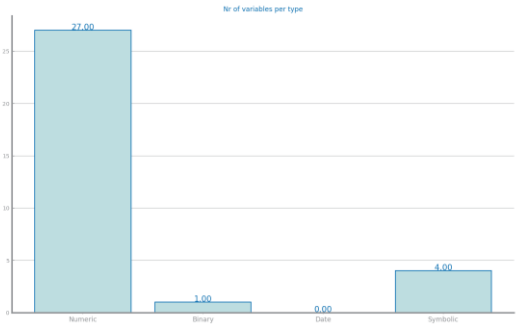
5.

6. Dimensionality

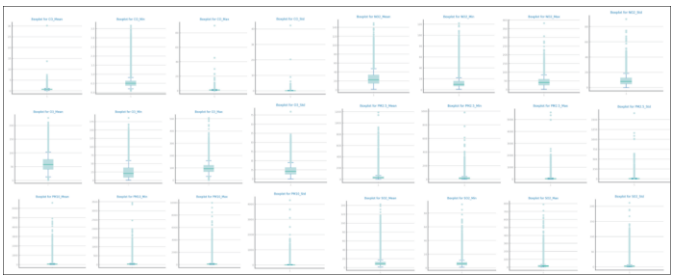
1.

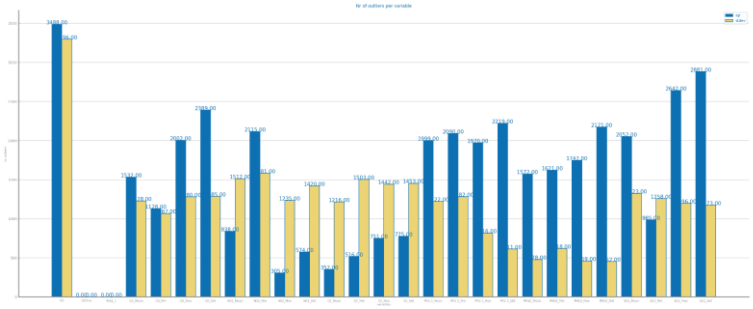


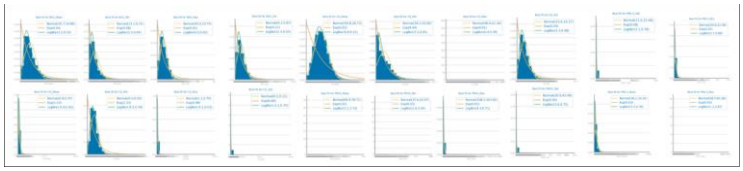




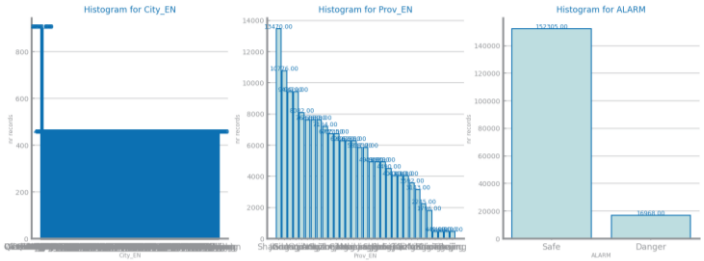
7. Data distribution

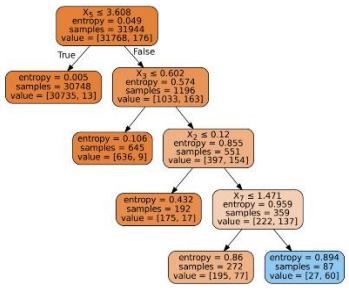


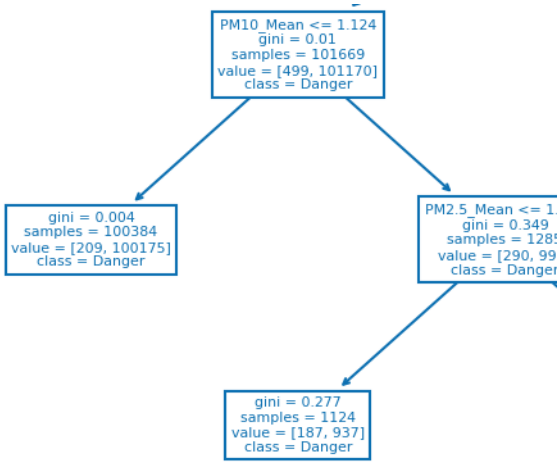


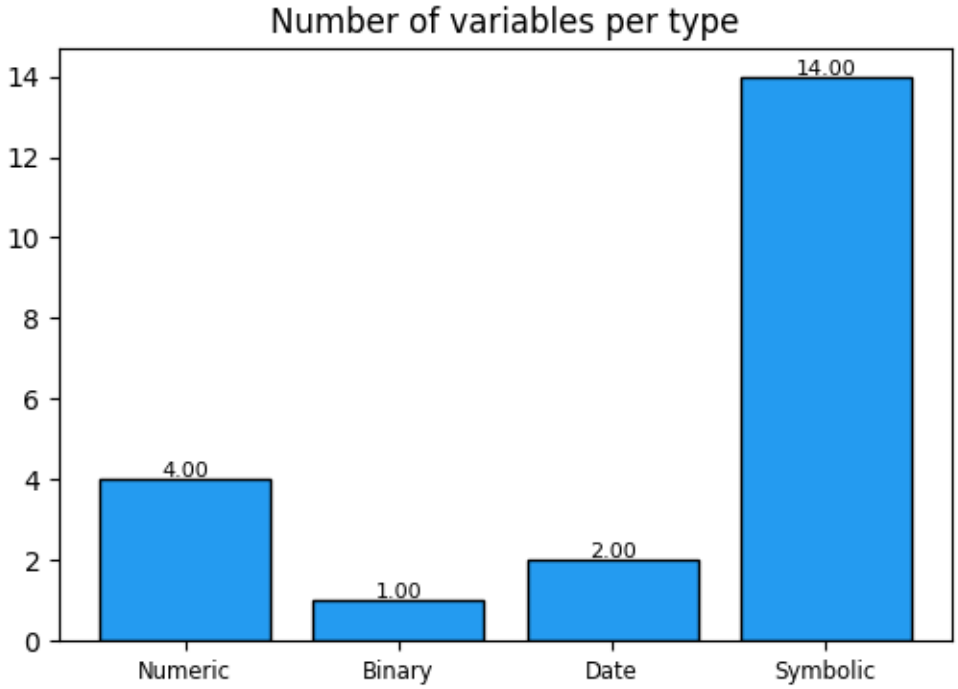


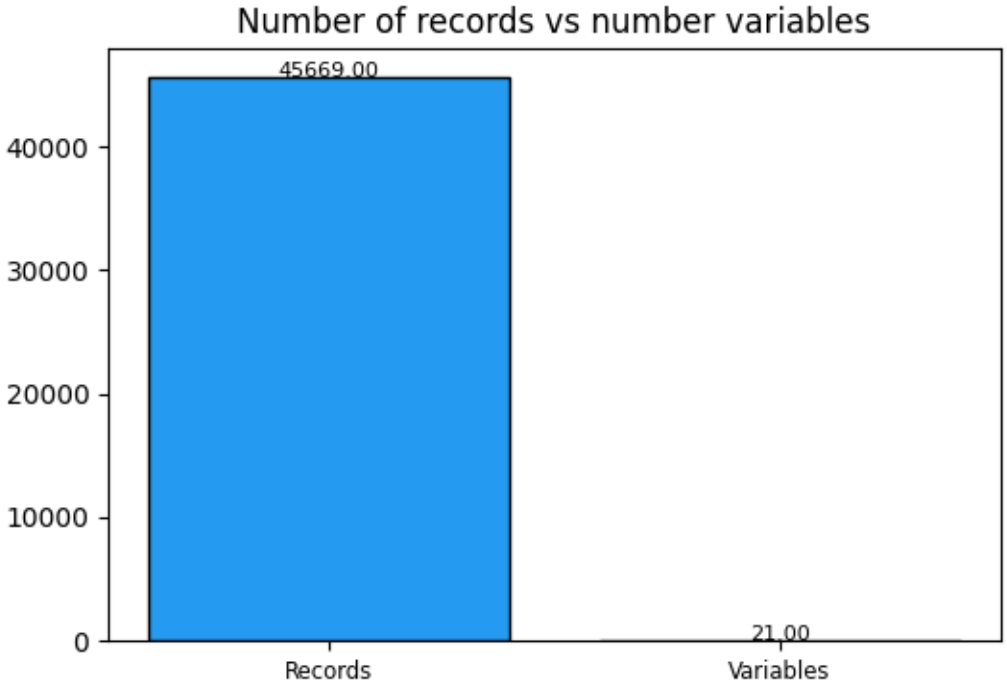
Symbolic variables

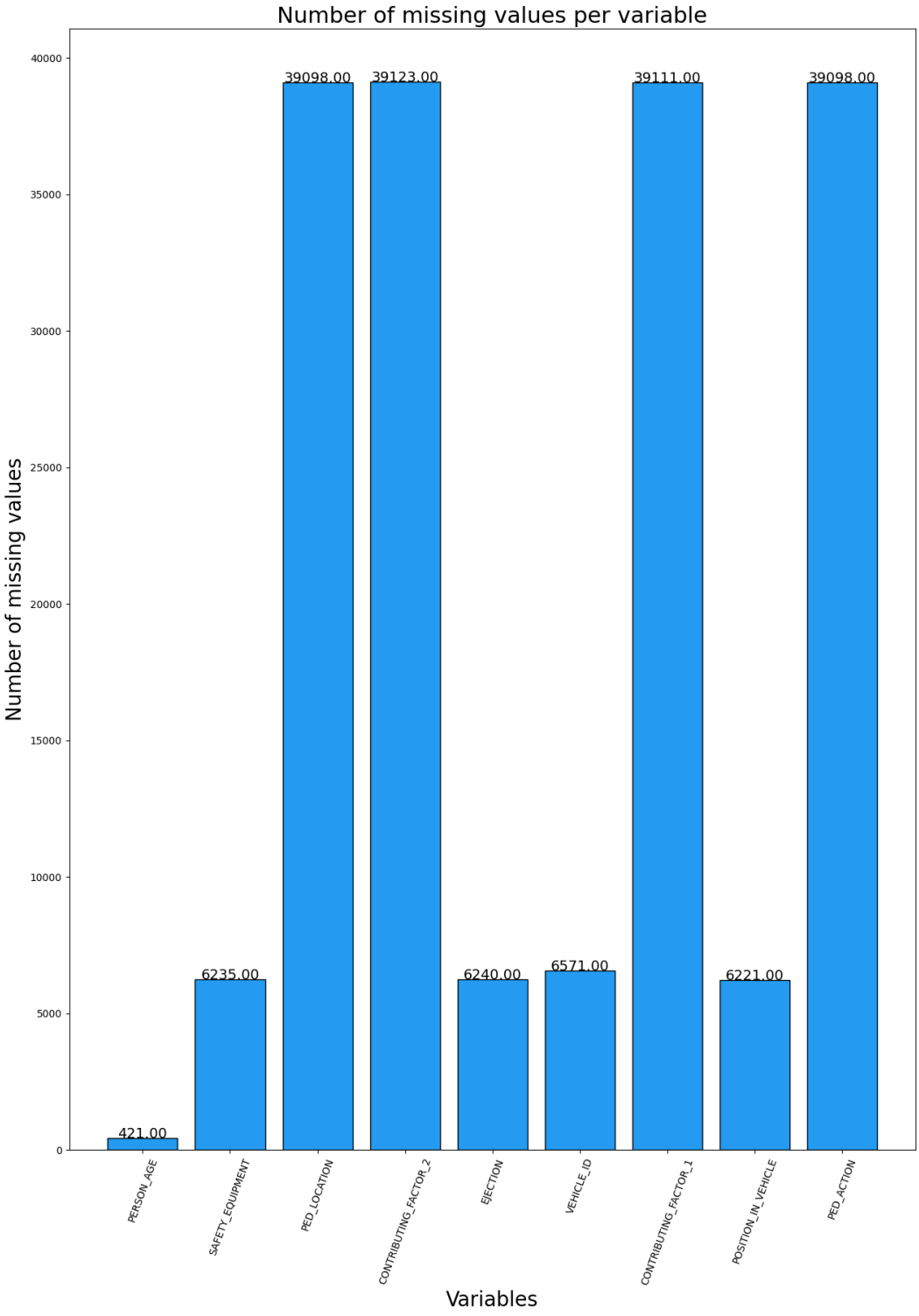


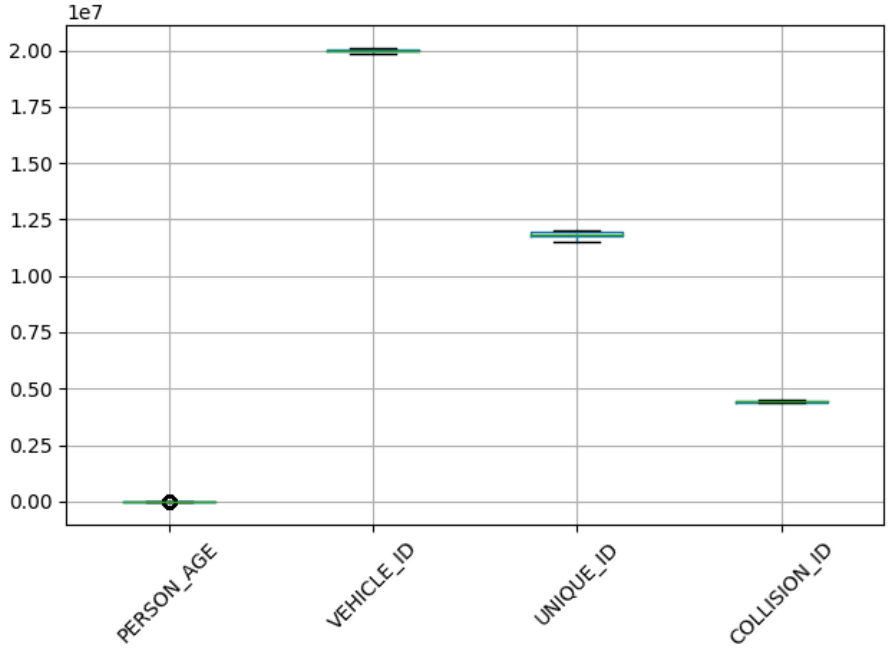


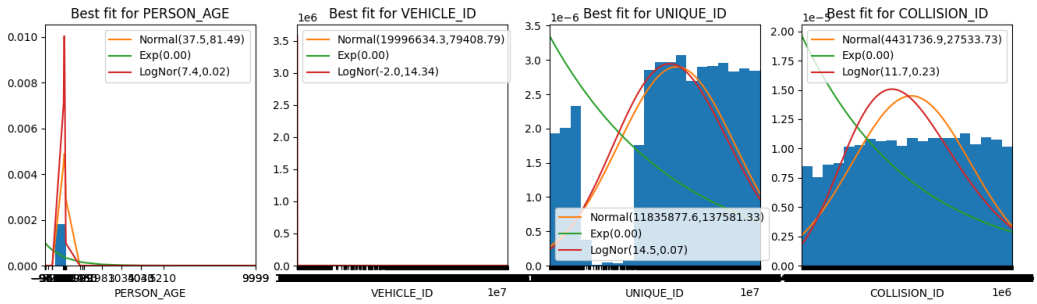


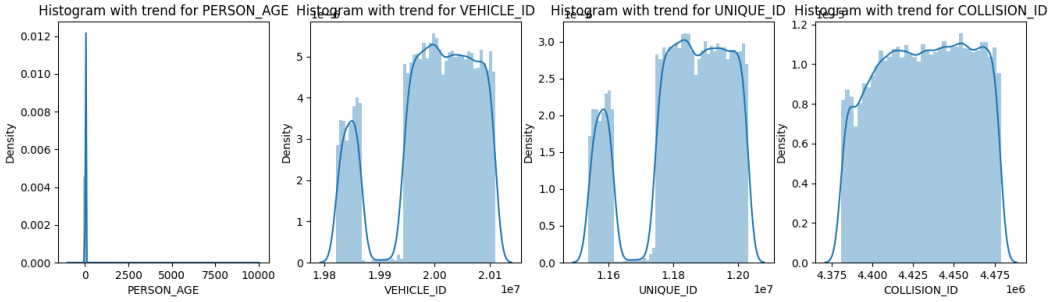

Number of variables per type
14
12
10
8
6
4
2
0
4.00
1.00
2.00
14.00
Numeric
Binary
Date
Symbolic

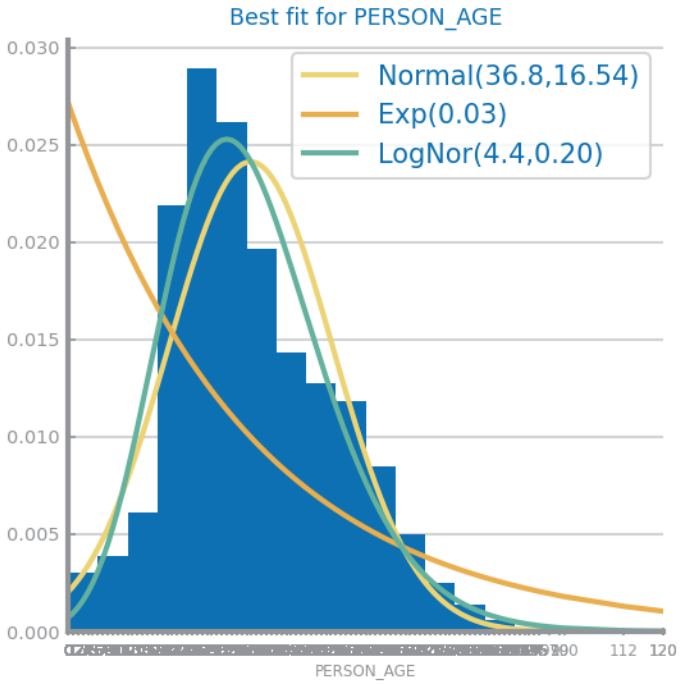

Number of records vs number variables
40000
30000
20000
10000
0
45669.00
21.00
Records
Variables

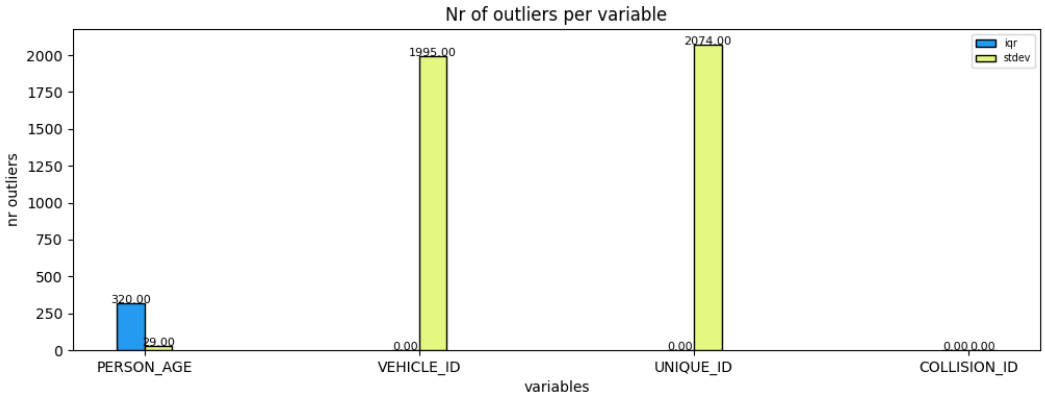

Number of missing values per variable
Number of missing values
Variables
0
5000
10000
15000
20000
25000
30000
35000
40000
421.00
6235.00
39098.00
39123.00
6240.00
6571.00
39111.00
6221.00
39098.00
PERSON_AGE
SAFETY_EQUIPMENT
PED_LOCATION
CONTRIBUTING_FACTOR_2
EJECTION
VEHICLE_ID
CONTRIBUTING_FACTOR_1
POSITION_IN_VEHICLE
PED_ACTION

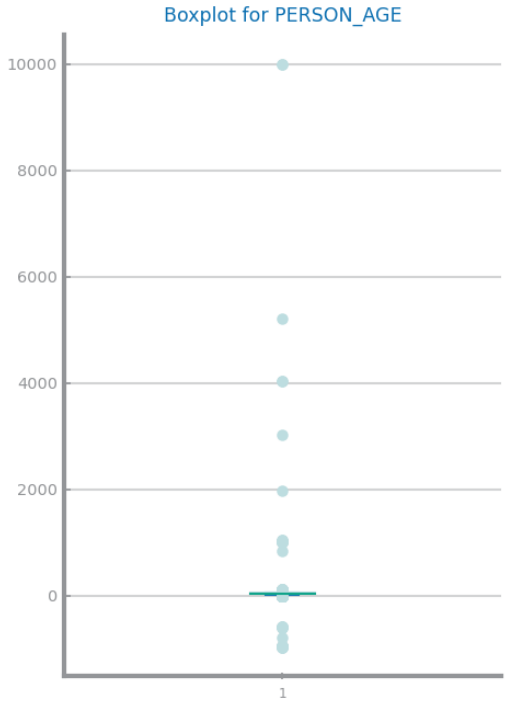

1e7
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
PERSON_AGE
VEHICLE_ID
UNIQUE_ID
COLLISION_ID

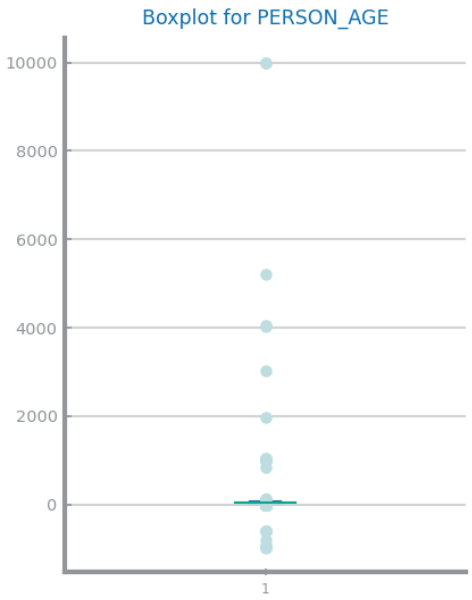

Best fit for PERSON_AGE
Normal(37.5,81.49)
Exp(0.00)
LogNor(7.4,0.02)
PERSON_AGE
Best fit for VEHICLE_ID
Normal(19996634.3,79408.79)
Exp(0.00)
LogNor(-2.0,14.34)
VEHICLE_ID
1e7
Best fit for UNIQUE_ID
Normal(11835877.6,137581.33)
Exp(0.00)
LogNor(14.5,0.07)
UNIQUE_ID
1e7
Best fit for COLLISION_ID
Normal(4431736.9,27533.73)
Exp(0.00)
LogNor(11.7,0.23)
COLLISION_ID
1e6

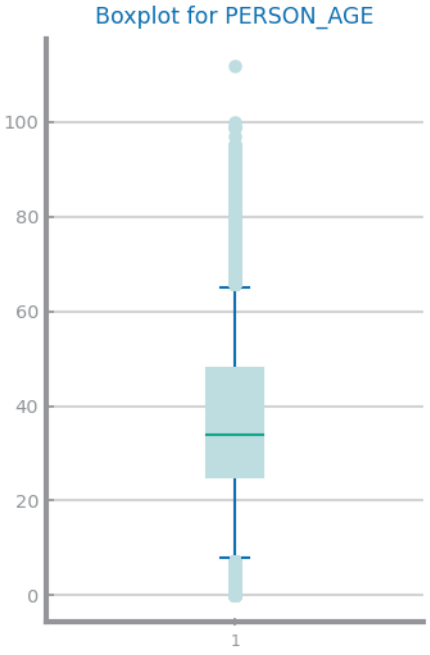

Histogram with trend for PERSON_AGE
Density
PERSON_AGE
Histogram with trend for VEHICLE_ID
Density
VEHICLE_ID
1e7
Histogram with trend for UNIQUE_ID
Density
UNIQUE_ID
1e7
Histogram with trend for COLLISION_ID
Density
COLLISION_ID
1e6

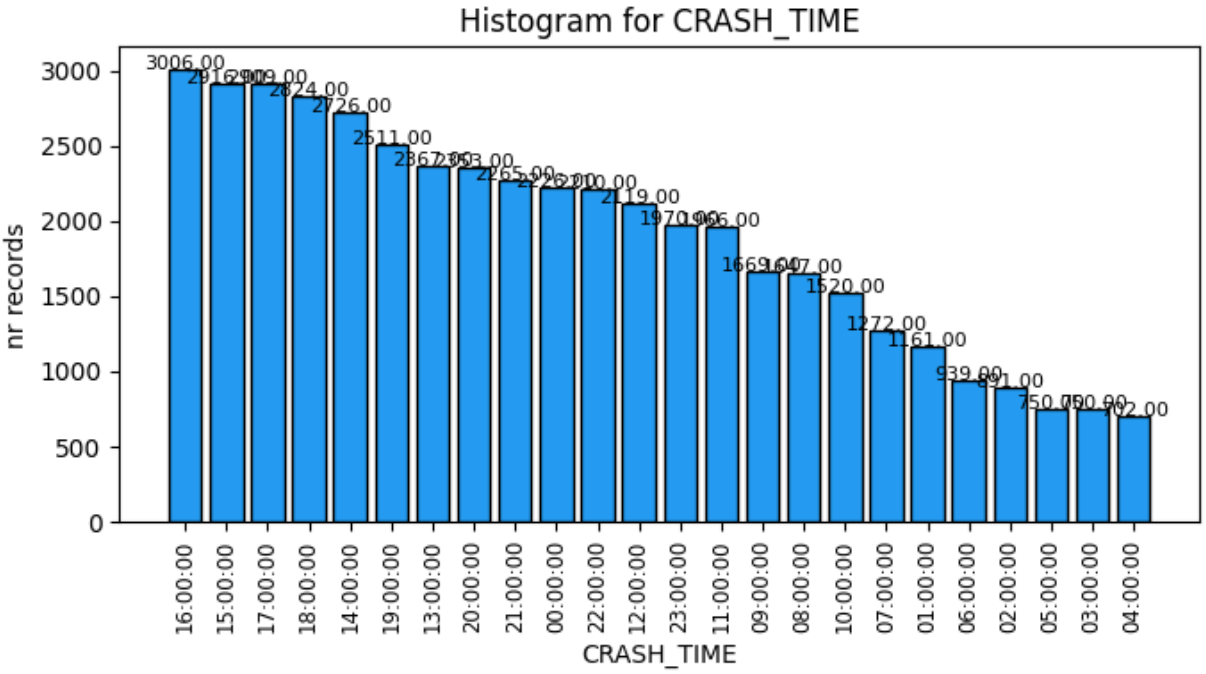

Best fit for PERSON_AGE
Normal(36.8,16.54)
Exp(0.03)
LogNor(4.4,0.20)
0.030
0.025
0.020
0.015
0.010
0.005
0.000
112
120
PERSON_AGE

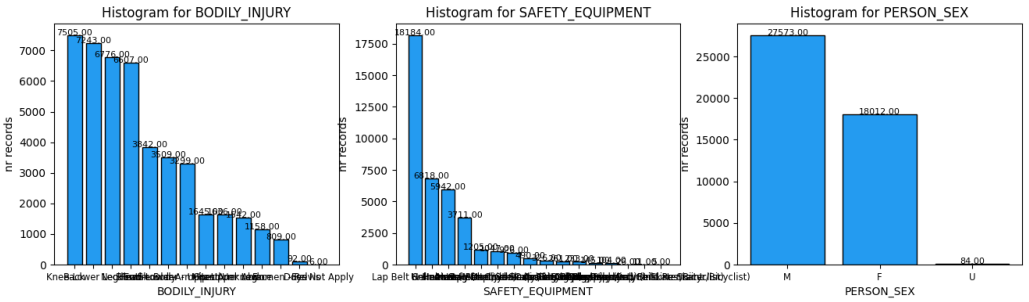

Nr of outliers per variable
nr outliers
variables
iqr
stdev
PERSON_AGE
VEHICLE_ID
UNIQUE_ID
COLLISION_ID
320.00
29.00
1995.00
0.00
2074.00
0.00
0.00
0.00

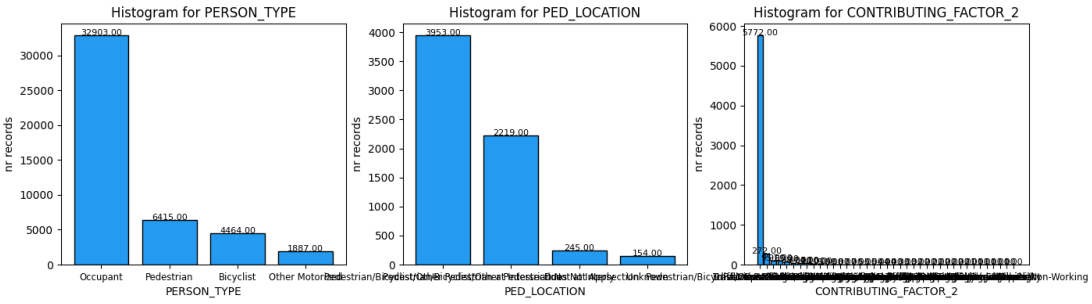

Boxplot for PERSON_AGE
10000
8000
6000
4000
2000
0
1

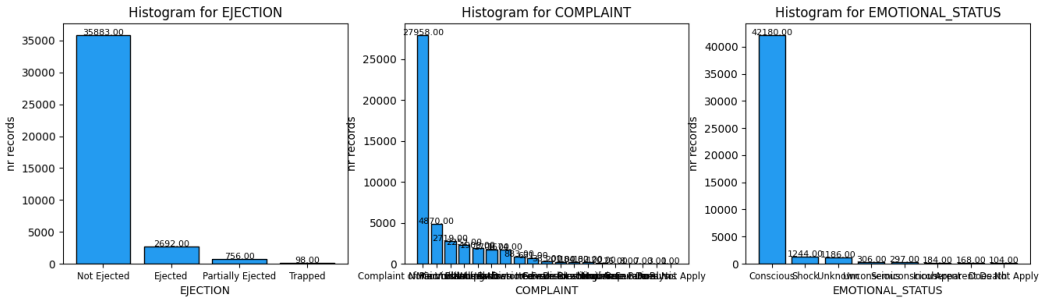

Boxplot for PERSON_AGE
10000
8000
6000
4000
2000
0
1

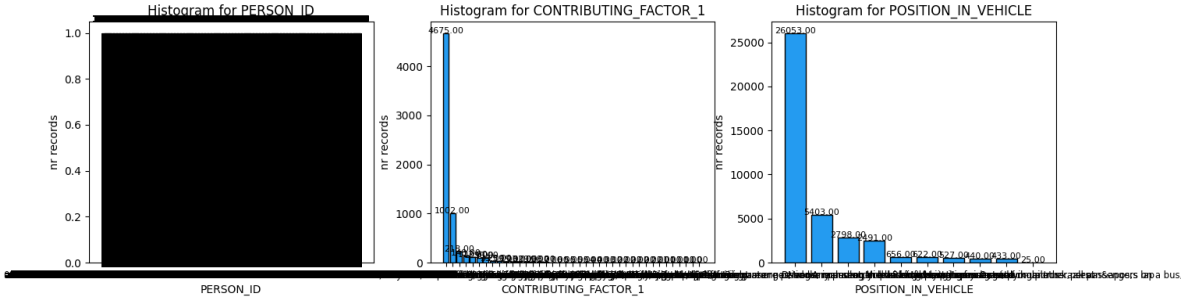

Boxplot for PERSON_AGE
100
80
60
40
20
0
1

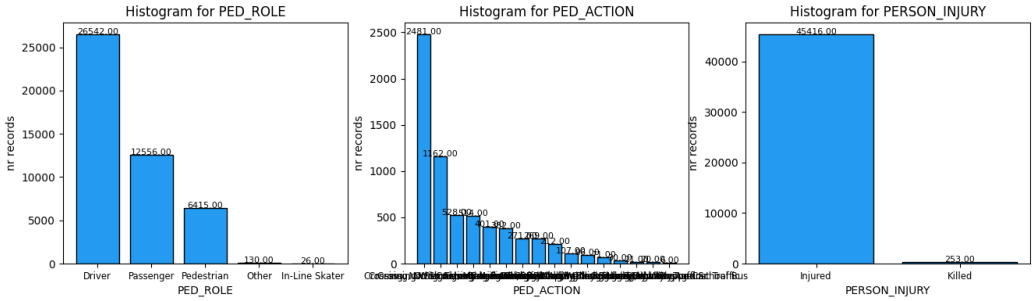

Histogram for CRASH_TIME
nr records
CRASH_TIME
16:00:00
15:00:00
17:00:00
18:00:00
14:00:00
19:00:00
13:00:00
20:00:00
21:00:00
00:00:00
22:00:00
12:00:00
23:00:00
11:00:00
09:00:00
08:00:00
10:00:00
07:00:00
01:00:00
06:00:00
02:00:00
05:00:00
03:00:00
04:00:00

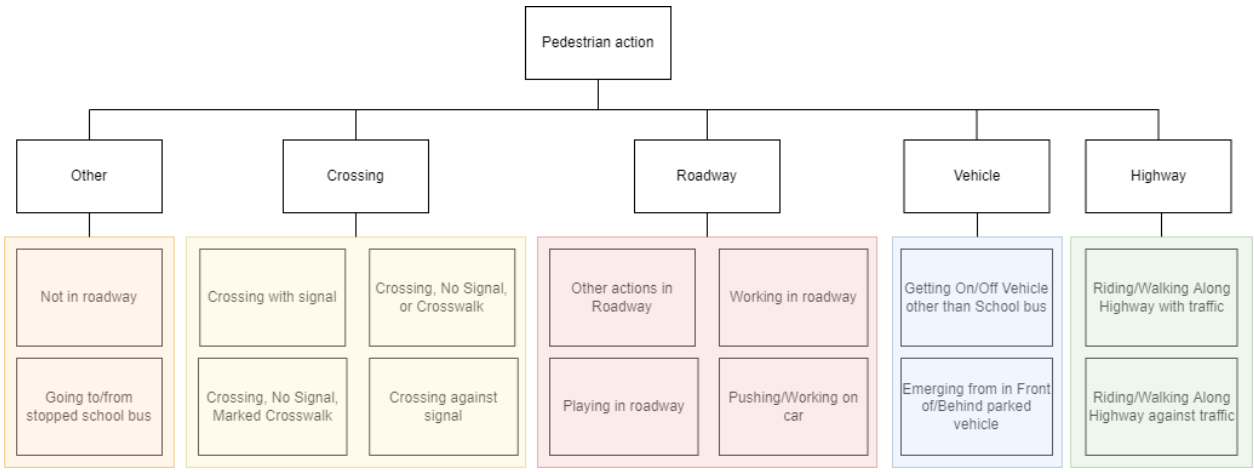

Histogram for BODILY_INJURY
nr records
BODILY_INJURY
Histogram for SAFETY_EQUIPMENT
nr records
SAFETY_EQUIPMENT
Histogram for PERSON_SEX
nr records
M
F
U
PERSON_SEX

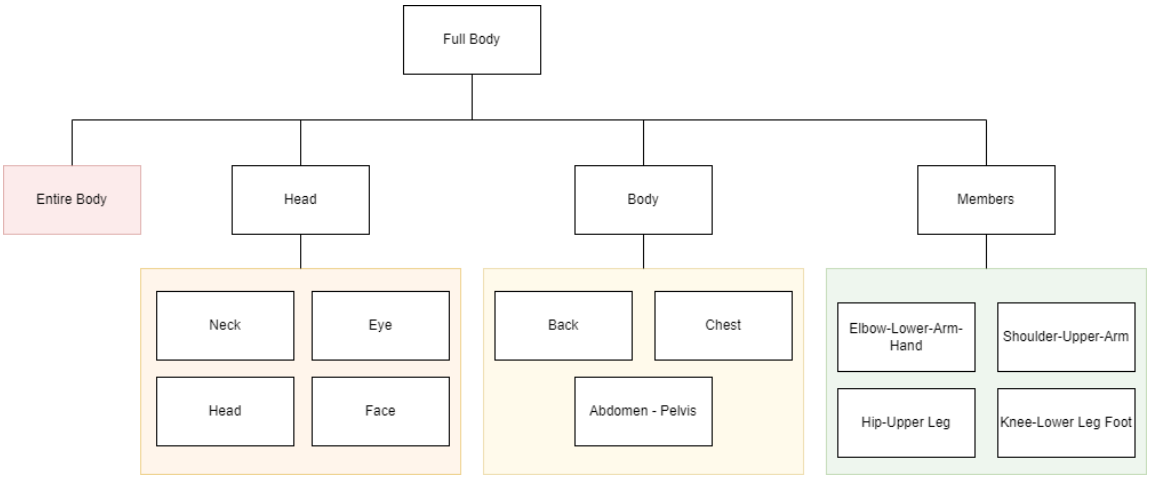

Histogram for PERSON_TYPE
nr records
Occupant
Pedestrian
Bicyclist
Other Motorized
PERSON_TYPE
Histogram for PED_LOCATION
nr records
PED_LOCATION
Histogram for CONTRIBUTING_FACTOR_2
nr records
CONTRIBUTING_FACTOR_2

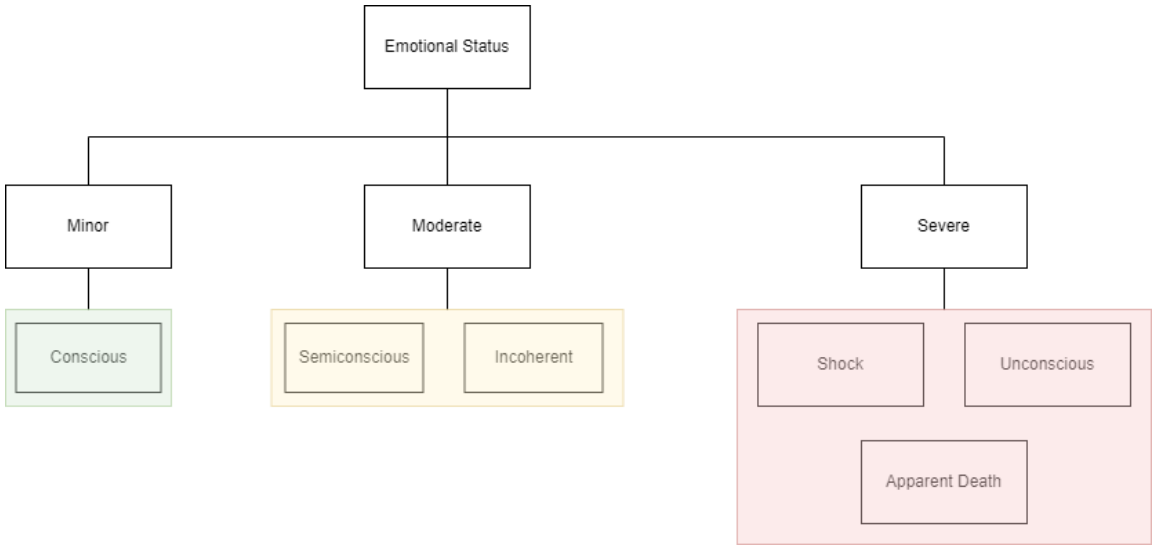

Histogram for EJECTION
nr records
EJECTION
Not Ejected
Ejected
Partially Ejected
Trapped
35883.00
2692.00
756.00
98.00
Histogram for COMPLAINT
COMPLAINT
27958.00
4870.00
2719.00
Histogram for EMOTIONAL_STATUS
EMOTIONAL_STATUS
42180.00
1244.00
1186.00
206.00
297.00
184.00
168.00
104.00

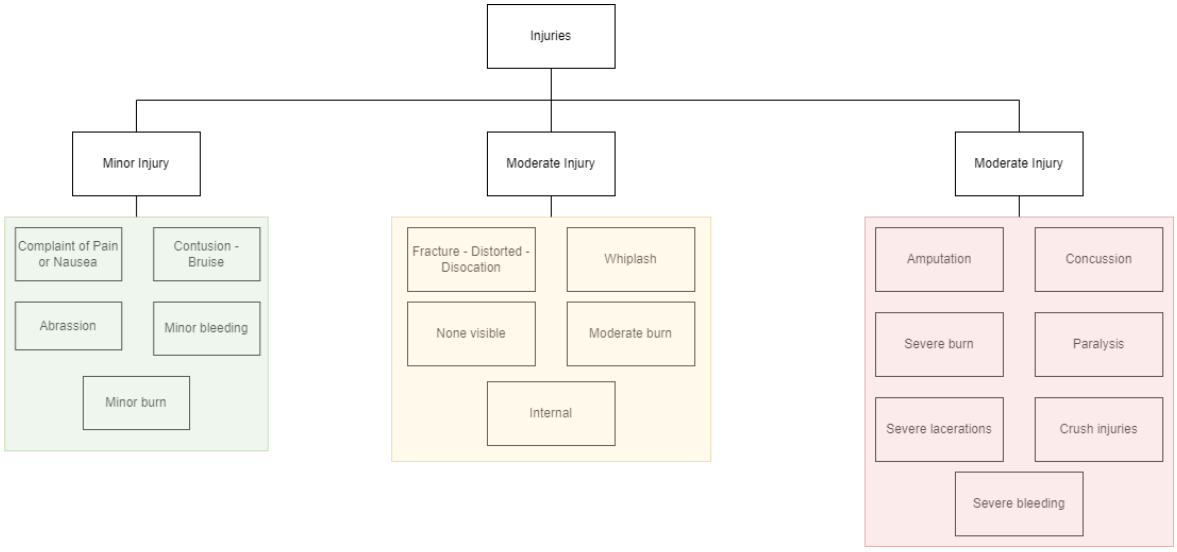

Histogram for PERSON_ID
nr records
PERSON_ID
Histogram for CONTRIBUTING_FACTOR_1
CONTRIBUTING_FACTOR_1
4675.00
1062.00
Histogram for POSITION_IN_VEHICLE
POSITION_IN_VEHICLE
26053.00
6403.00
2798.00
2491.00
656.00
522.00
527.00
340.00
323.00
25.00


Histogram for PED_ROLE
nr records
PED_ROLE
Driver
Passenger
Pedestrian
Other
In-Line Skater
26542.00
12556.00
6415.00
130.00
26.00
Histogram for PED_ACTION
PED_ACTION
2481.00
1162.00
Histogram for PERSON_INJURY
PERSON_INJURY
Injured
Killed
45416.00
253.00

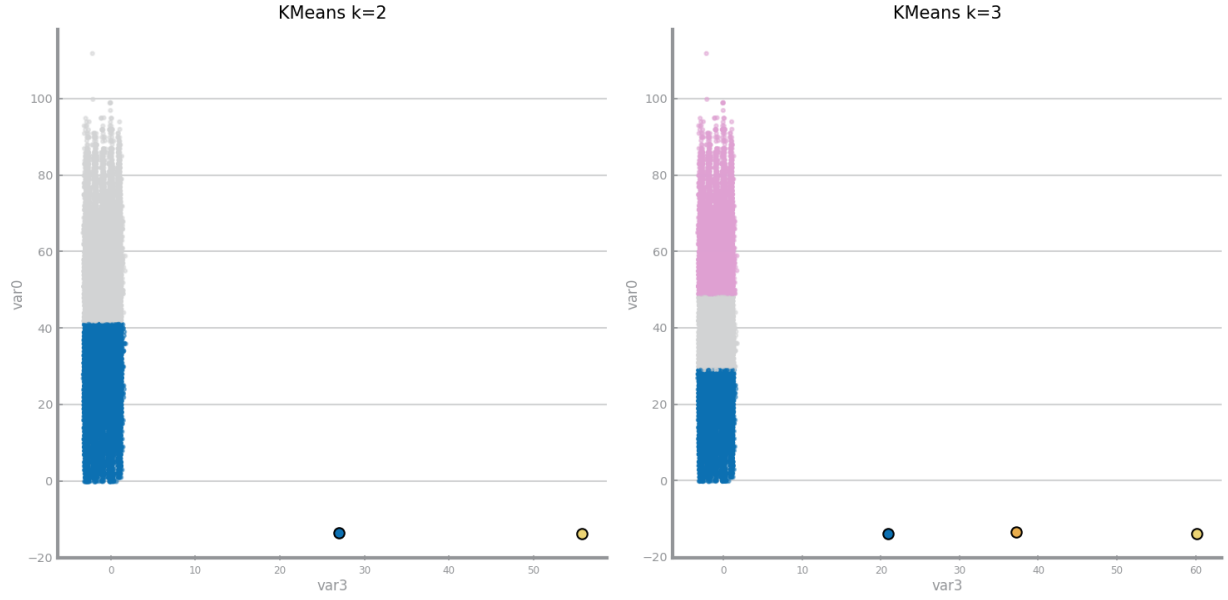

Pedestrian action
Other
Crossing
Roadway
Vehicle
Highway
Not in roadway
Going to/from stopped school bus
Crossing with signal
Crossing, No Signal, or Crosswalk
Crossing, No Signal, Marked Crosswalk
Crossing against signal
Other actions in Roadway
Working in roadway
Playing in roadway
Pushing/Working on car
Getting On/Off Vehicle other than School bus
Emerging from in Front of/Behind parked vehicle
Riding/Walking Along Highway with traffic
Riding/Walking Along Highway against traffic


Full Body
Entire Body
Head
Body
Members
Neck
Eye
Head
Face
Back
Chest
Abdomen - Pelvis
Elbow-Lower-Arm-Hand
Shoulder-Upper-Arm
Hip-Upper Leg
Knee-Lower Leg Foot

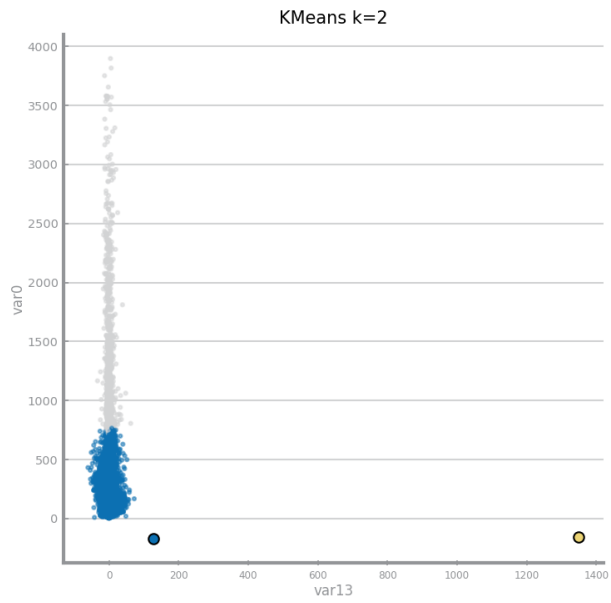

Emotional Status
Minor
Moderate
Severe
Conscious
Semiconscious
Incoherent
Shock
Unconscious
Apparent Death

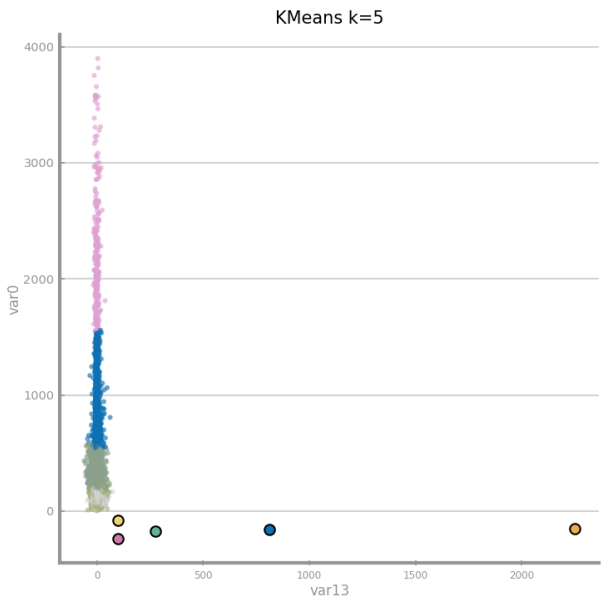

Injuries
Minor Injury
Moderate Injury
Moderate Injury
Complaint of Pain or Nausea
Contusion - Bruise
Abrassion
Minor bleeding
Minor burn
Fracture - Distorted - Disocation
Whiplash
None visible
Moderate burn
Internal
Amputation
Concussion
Severe burn
Paralysis
Severe lacerations
Crush injuries
Severe bleeding

NYCC

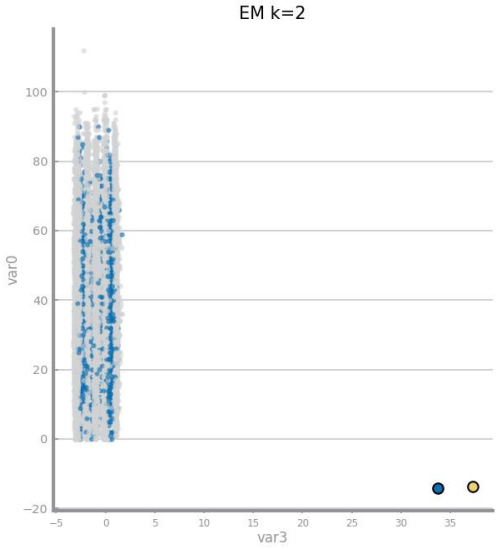

KMeans k=2
KMeans k=3
var0
var3
100
80
60
40
20
0
−20
0
10
20
30
40
50
60

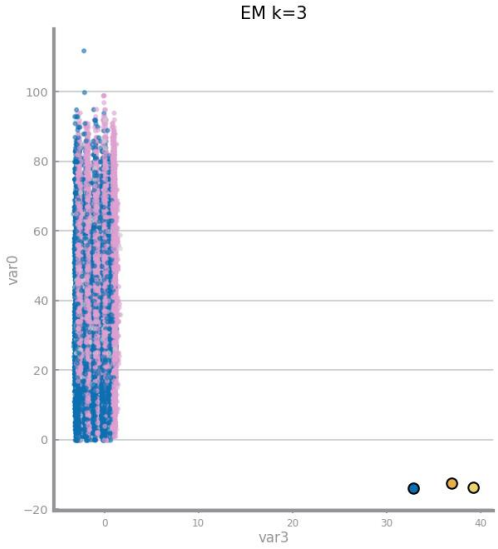
AQC


KMeans k=2
var0
var13
4000
3500
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400

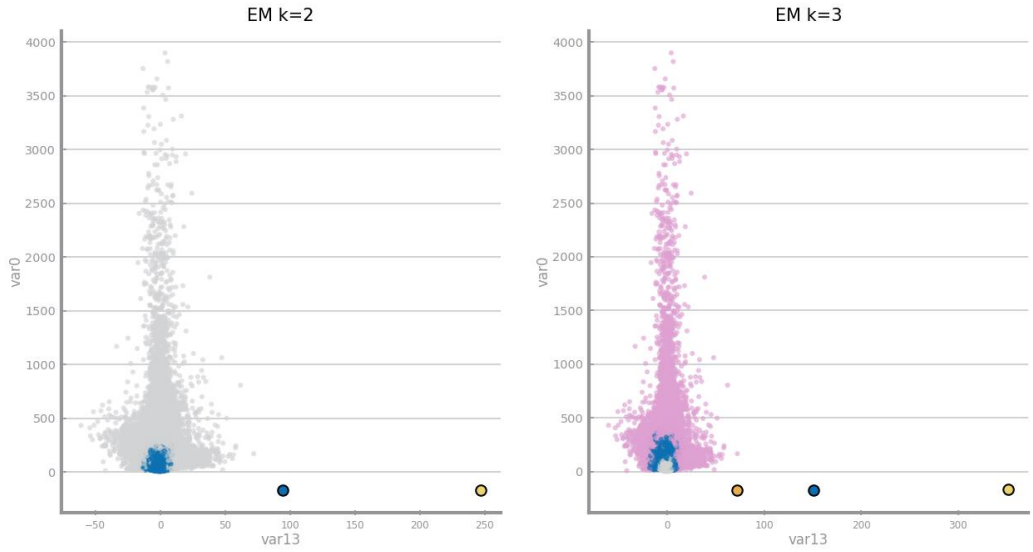

KMeans k=5
var0
var13
4000
3000
2000
1000
0
0
500
1000
1500
2000

NYC

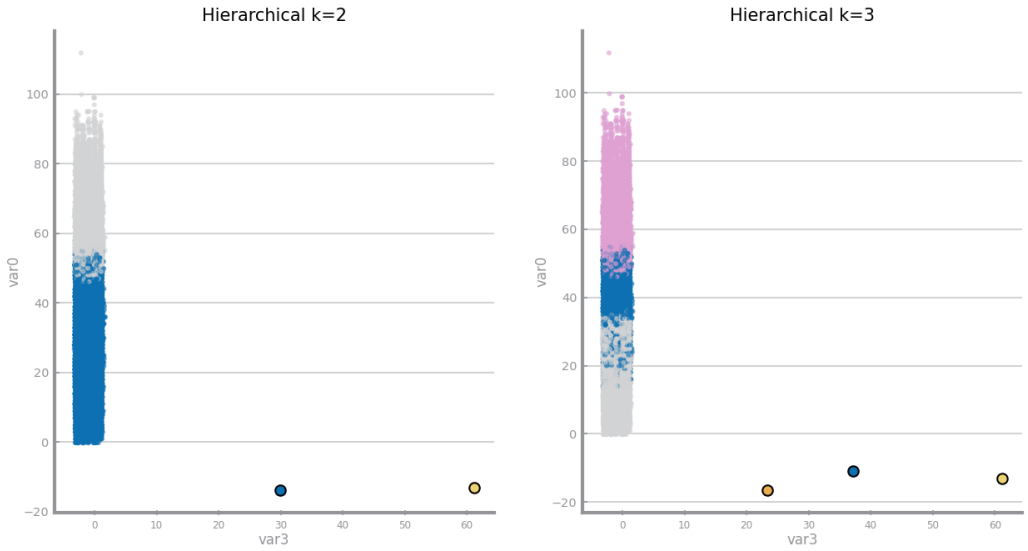

EM k=2
var0
var3
100
80
60
40
20
0
−20
−5
0
5
10
15
20
25
30
35


EM k=3
var0
var3
100
80
60
40
20
0
−20
0
10
20
30
40

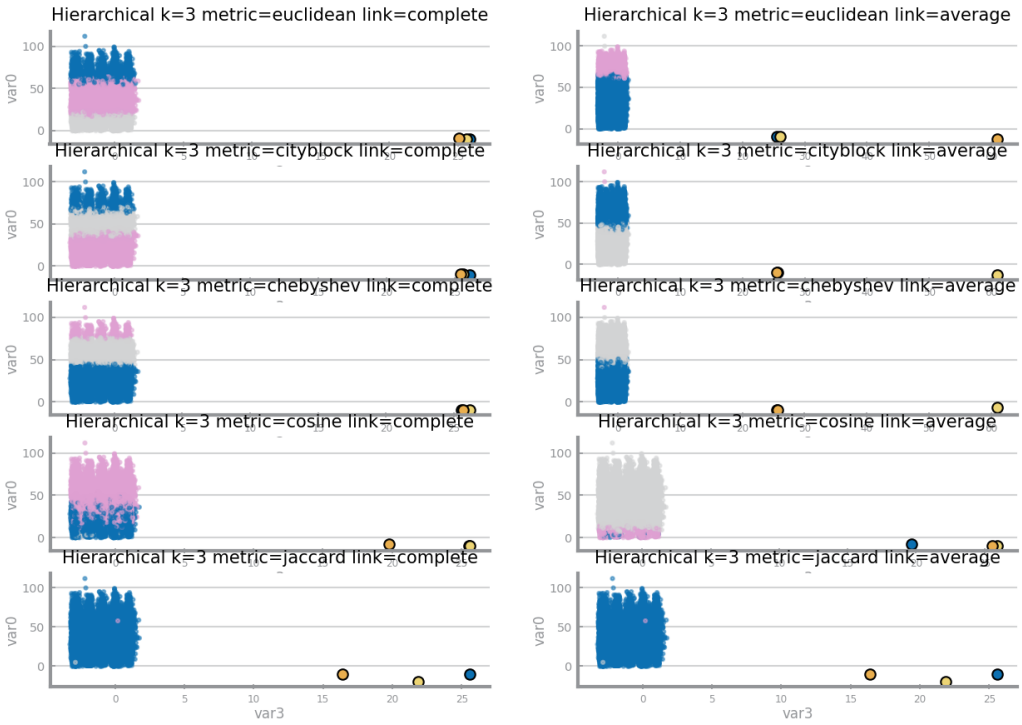
AQC


EM k=2
EM k=3
var0
var13
4000
3500
3000
2500
2000
1500
1000
500
0
−50
0
50
100
150
200
250
300

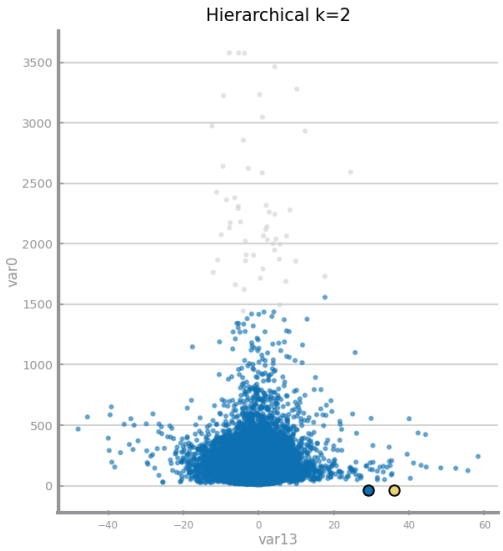
NYC

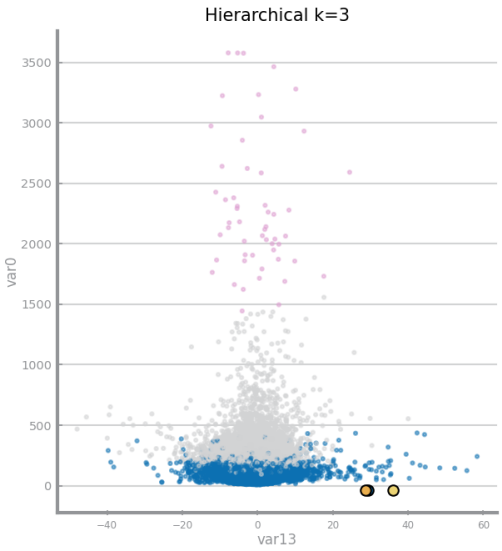

Hierarchical k=2
Hierarchical k=3
var0
var3
100
80
60
40
20
0
−20
0
10
20
30
40
50
60

NYC

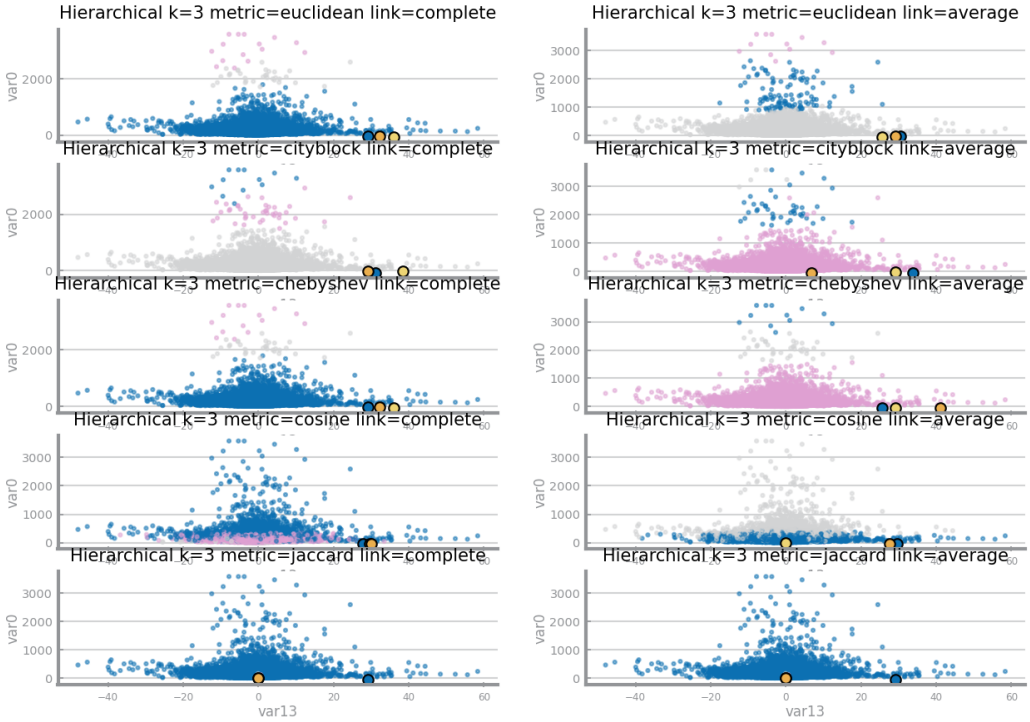

Hierarchical k=3 metric=euclidean link=complete
Hierarchical k=3 metric=euclidean link=average
Hierarchical k=3 metric=cityblock link=complete
Hierarchical k=3 metric=cityblock link=average
Hierarchical k=3 metric=chebyshev link=complete
Hierarchical k=3 metric=chebyshev link=average
Hierarchical k=3 metric=cosine link=complete
Hierarchical k=3 metric=cosine link=average
Hierarchical k=3 metric=jaccard link=complete
Hierarchical k=3 metric=jaccard link=average
var0
var3

AQC

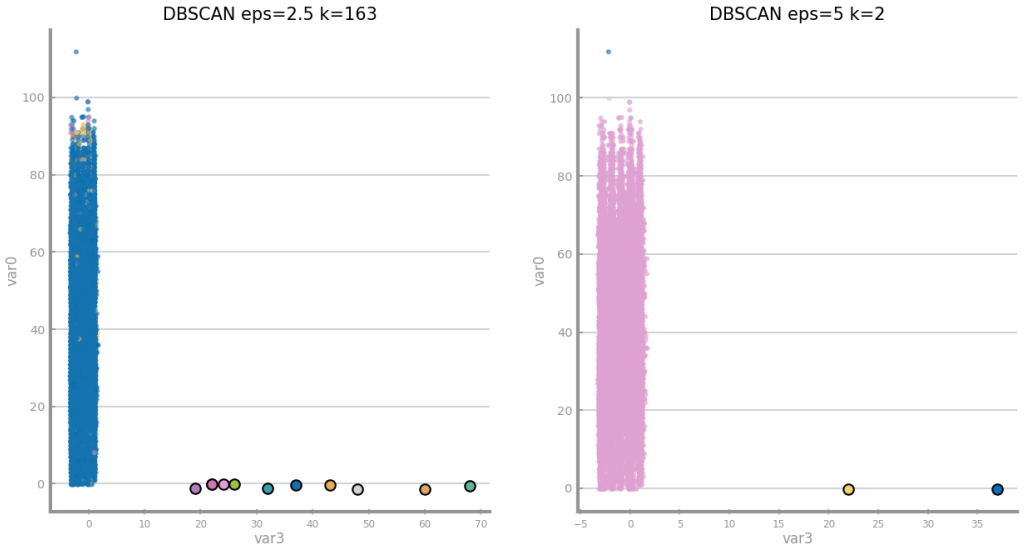

Hierarchical k=2
var0
var13
3500
3000
2500
2000
1500
1000
500
0
−40
−20
0
20
40
60

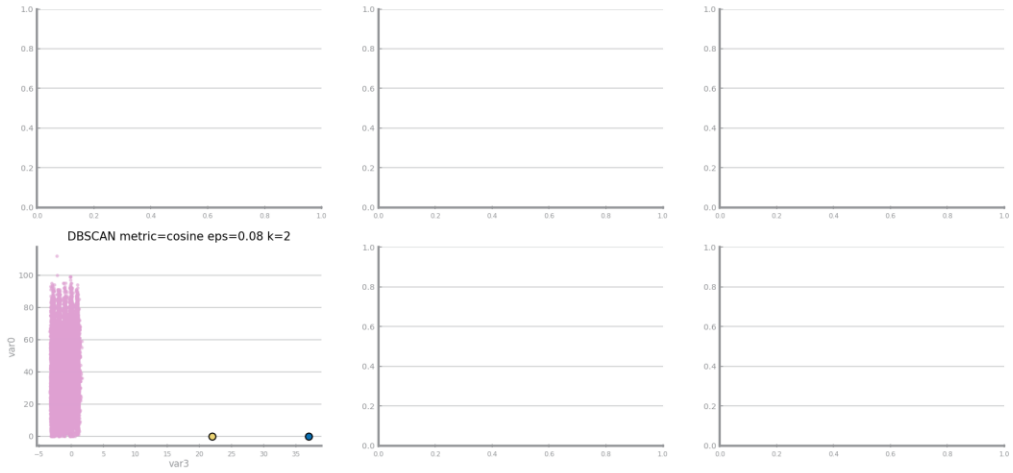

Hierarchical k=3
var0
var13
3500
3000
2500
2000
1500
1000
500
0
−40
−20
0
20
40
60

AQC

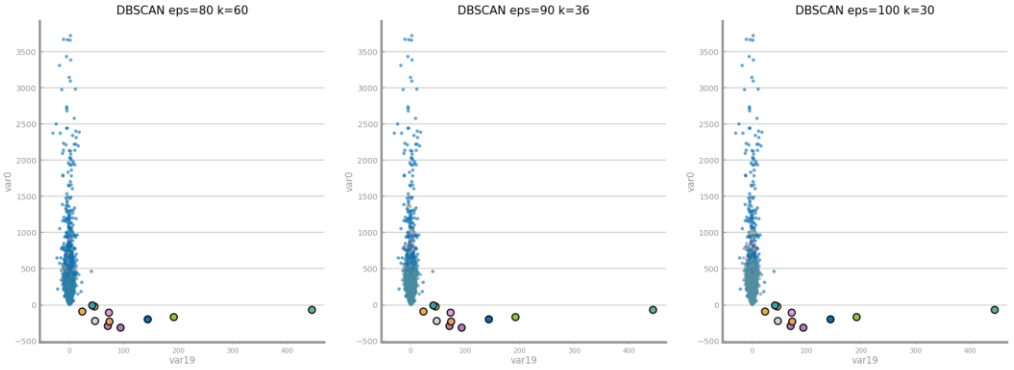

Hierarchical k=3 metric=euclidean link=complete
Hierarchical k=3 metric=euclidean link=average
Hierarchical k=3 metric=cityblock link=complete
Hierarchical k=3 metric=cityblock link=average
Hierarchical k=3 metric=chebyshev link=complete
Hierarchical k=3 metric=chebyshev link=average
Hierarchical k=3 metric=cosine link=complete
Hierarchical k=3 metric=cosine link=average
Hierarchical k=3 metric=jaccard link=complete
Hierarchical k=3 metric=jaccard link=average
var0
var13

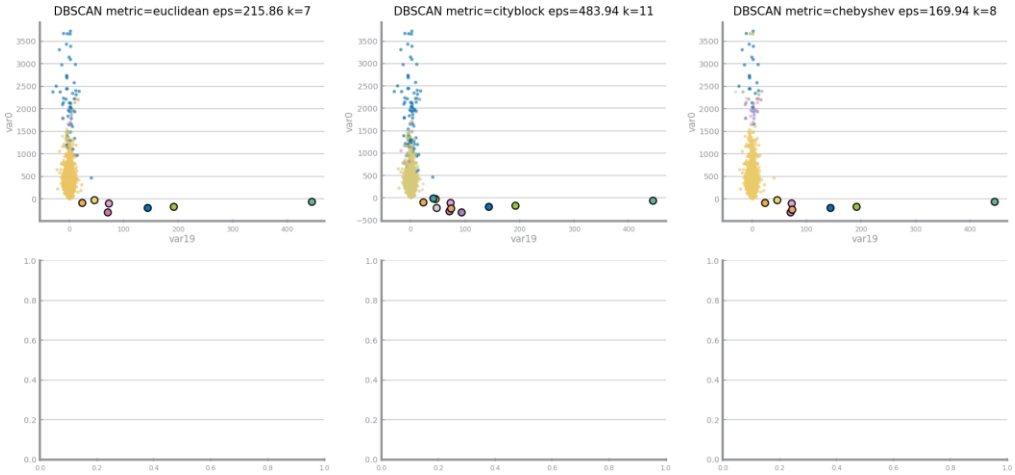
NYC


DBSCAN eps=2.5 k=163
DBSCAN eps=5 k=2
var0
var3


DBSCAN metric=cosine eps=0.08 k=2
var0
var3

AQC


DBSCAN eps=80 k=60
DBSCAN eps=90 k=36
DBSCAN eps=100 k=30
var0
var19


DBSCAN metric=euclidean eps=215.86 k=7
DBSCAN metric=cityblock eps=483.94 k=11
DBSCAN metric=chebyshev eps=169.94 k=8
var0
var19