Report 3: Automobiles data set

Unsupervised learning: Clustering and density estimation

Mathias Husted Torp, s133547
Damian Kowalczyk, s166071

May 2nd, 2017

1 Clustering

In order to perform clustering of automobiles based on horsepower, 4 intervals are used to define classes (labels):

- 0-60 -> class(0)
- 60-120 -> class(1)
- 120-160 -> class(2)
- 160+ -> class(3)

1.1 Gaussian Mixture Model

Due to predefined number of 4 classes, the GMM cross validation, was not able to improve the clustering performance beyond K=4, regardless of configuration.

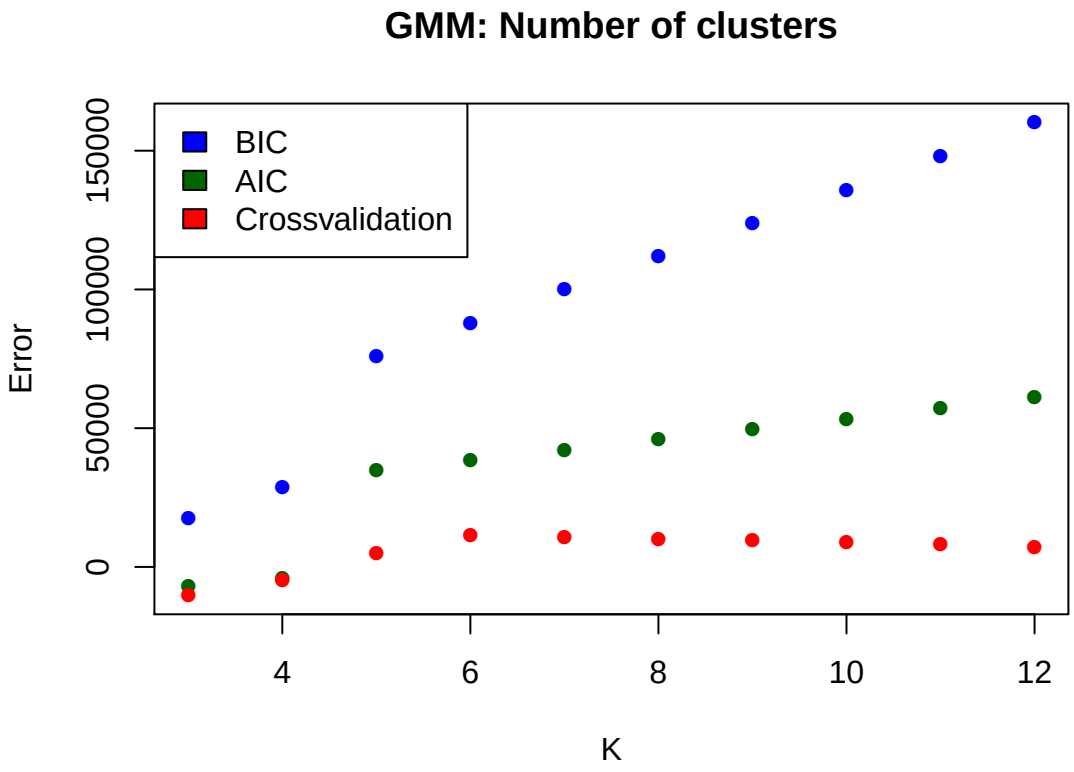

Figure 1: GMM: Selecting K. Here, the error is plotted as a function of the number of clusters

The lowest attained error is 47.8 %, and the clusters are visualized in Figure 2.

2 out of 4 predicted cluster centers are of low confidence.

1.2 Hierarchical clustering

Hierarchical clustering comes as an attractive alternative to finding a single agreed-upon value of K.

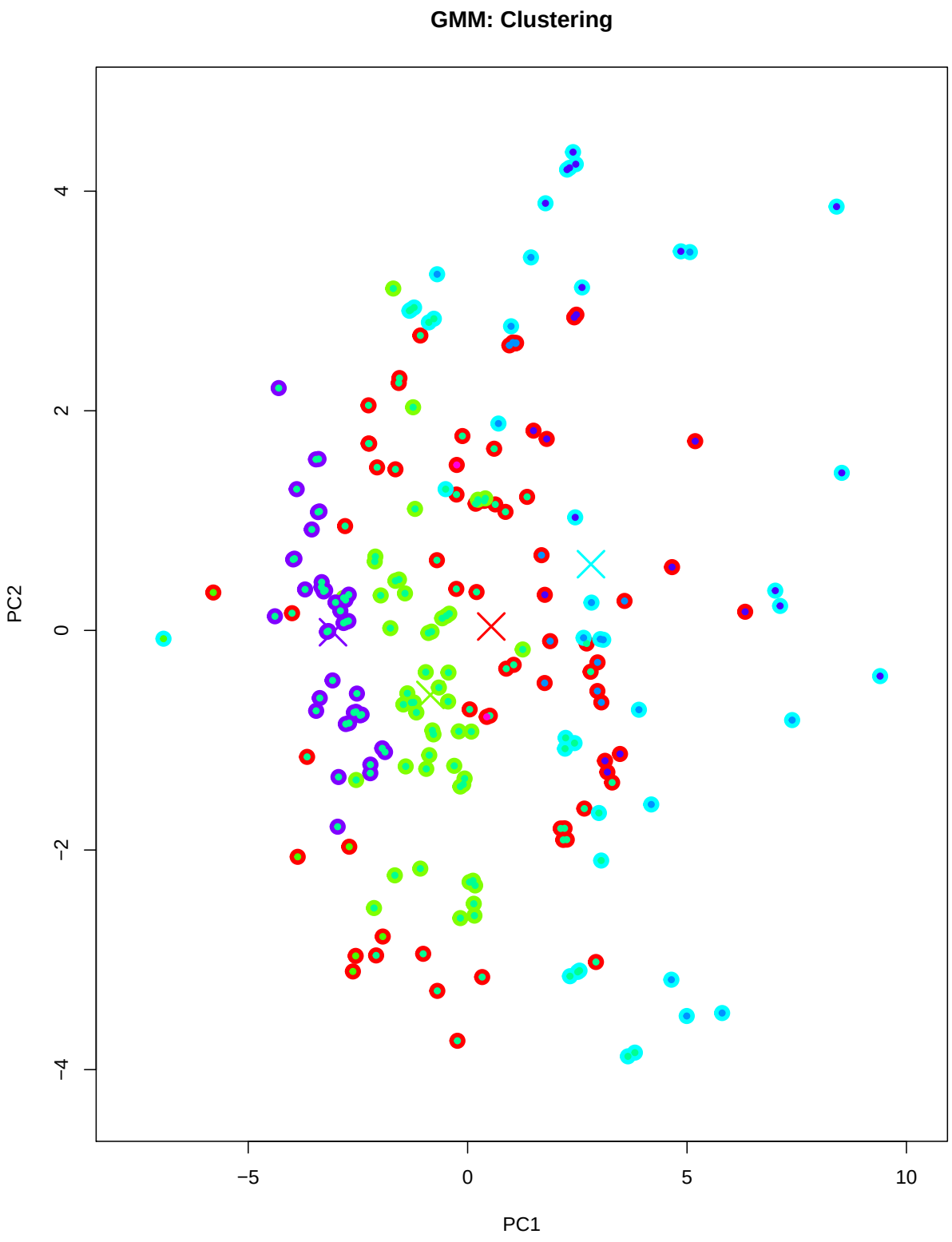

Figure 2: A visualization of the clusters predicted by GMM for K = 4.

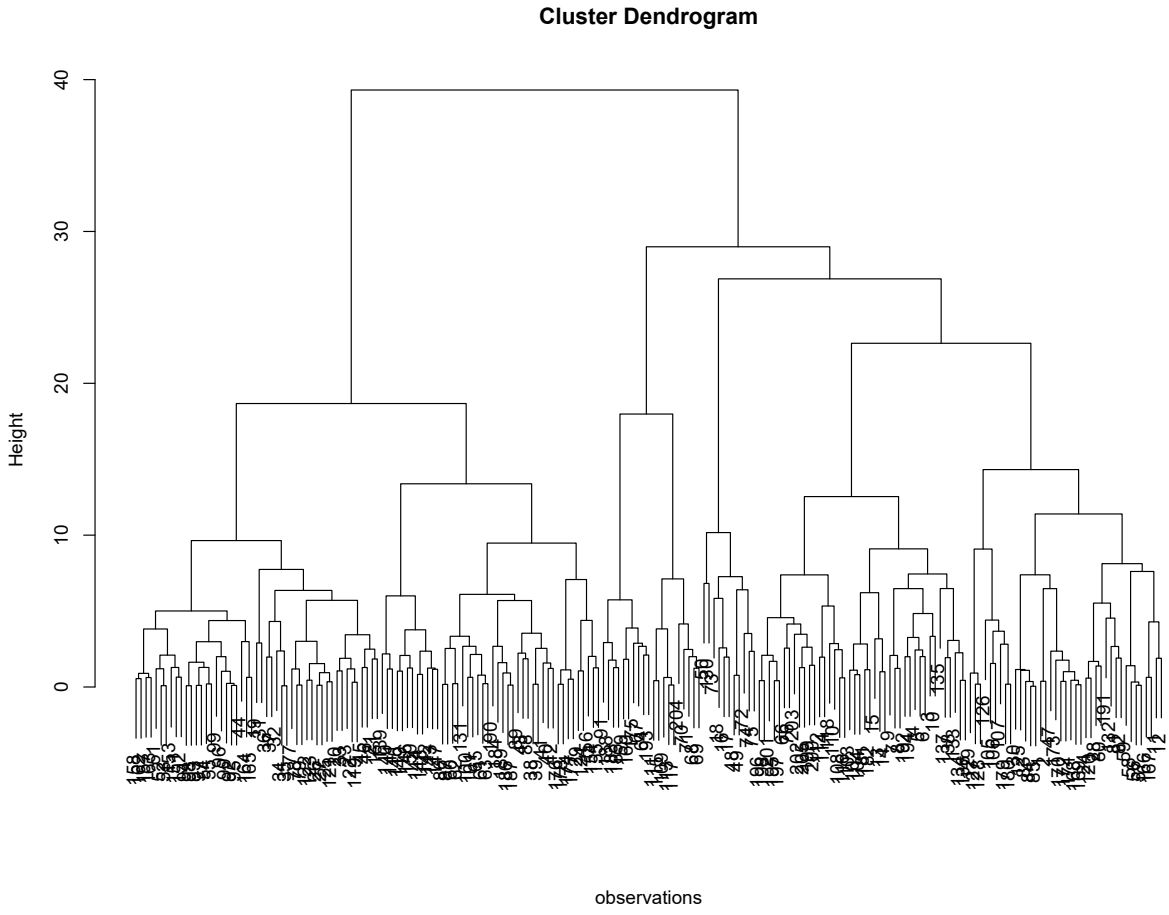

Figure 3: Dendrogram of the possible clusters predicted by hierarchical clustering.

The dendrogram in Figure 3 arranges the automobile dataset in a nested sequence of partitions organized as a hierarchy. The bottom of the hierarchy corresponds to the finest partition (each automobile is a unique cluster). The top level hierarchy corresponds to the coarsest possible partition corresponding to putting every automobile in the same cluster.

For this report, 4 clusters are selected as with GMM-clustering. The results from this cut-off can be seen in Figure 4.

Hierarchical clustering dimensionality reduction leads to better isolation of clusters, and better predictive performance (36% error rate vs 47.8% for GMM) as calculated by mclust::classError

2 Outlier detection

It is possible to identify possible outliers through the density distribution of the data. In practice, the density of each data point is calculated, and the points with the lowest densities have the highest probability of being outliers.

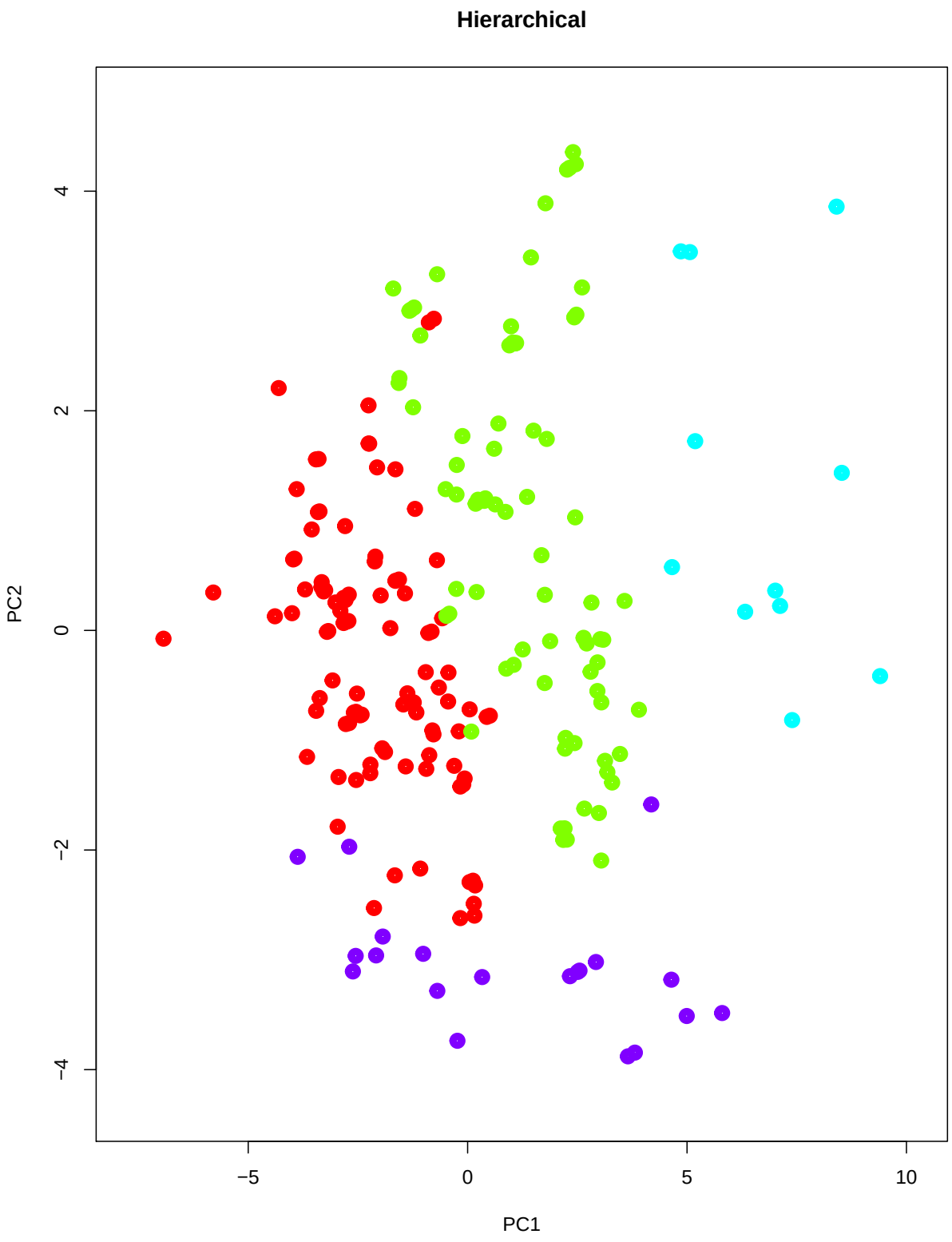

Figure 4: A visualization of the clusters predicted by hierarchical clustering for a cutoff at 4 clusters.

Table 1: The eight data points with the lowest density as predicted by the Gaussian Kernel Density Estimator. The presented points correspond to the eight left most bars in Figure 5

	symboling	make	body-style	bore	horsepower	highway-mpg	price
50	0	jaguar	sedan	3.54	262	17	36000
130	1	porsche	hatchback	3.94	288	28	N/A
135	3	saab	hatchback	2.54	110	28	15040
73	3	mercedes-benz	convertible	3.46	155	18	35056
191	3	volkswagen	hatchback	3.19	90	29	9980
3	1	alfa-romero	hatchback	2.68	154	26	16500
126	3	porsche	hatchback	3.94	143	27	22018
204	-1	volvo	sedan	3.01	106	27	22470

2.1 Gaussian Kernel density

The first method used in this report is the Gaussian Kernel Density estimator, which estimates the Gaussian Kernel Density using a specified width. The optimal width for describing the data set is found using cross-validation. The results are shown in Figure 5.

Figure 5: Outlier detection using Gaussian Kernel Density Estimator. The log(density) is used, because there is a very large difference in densities between the points.

Based on estimated densities, some possible outliers are: 50, 130, 135, 73. These correspond to: The top Jaguar, the top Porsche, a pretty standard Saab with a perhaps mistyped bore, and a very risky (symboling 3) convertible Mercedes Benz. Further details about the outliers are in Table 1.

Table 2: The eight data points with the lowest density as predicted using the KNN Density Estimator. The presented points correspond to the eight left most bars in Figure 6

	symboling	make	body-style	bore	horsepower	highway-mpg	price
50	0	jaguar	sedan	3.54	262	17	36000
74	0	mercedes-benz	sedan	3.80	184	16	40960
75	1	mercedes-benz	hardtop	3.80	184	16	45400
130	1	porsche	hatchback	3.94	288	28	N/A
135	3	saab	hatchback	2.54	110	28	15040
49	0	jaguar	sedan	3.63	176	19	35550
73	3	mercedes-benz	convertible	3.46	155	18	35056
48	0	jaguar	sedan	3.63	176	19	32250

2.2 KNN density

The second method used for predicting outliers in this report is the K nearest neighbor density. Using this method, a free parameter, K, indicating the number of neigbors used to estimate the density must be set. Here, K = 30 is chosen, corresponding to about 14.6 % of the data set. The results can be seen in Figure 6.

Figure 6: Outlier detection using KNN Density Estimator.

Based on estimated densities, some possible outliers are: 50, 74, 75, 130. A table of the top eight outliers is seen in Table 2.

2.3 KNN average relative density

The third method used to identify possible outliers is the KNN average relative density. This method is better suited for identifying outliers, when cluster densities differ a lot. Again, K is set to 30. The results can be seen in Figure 7.

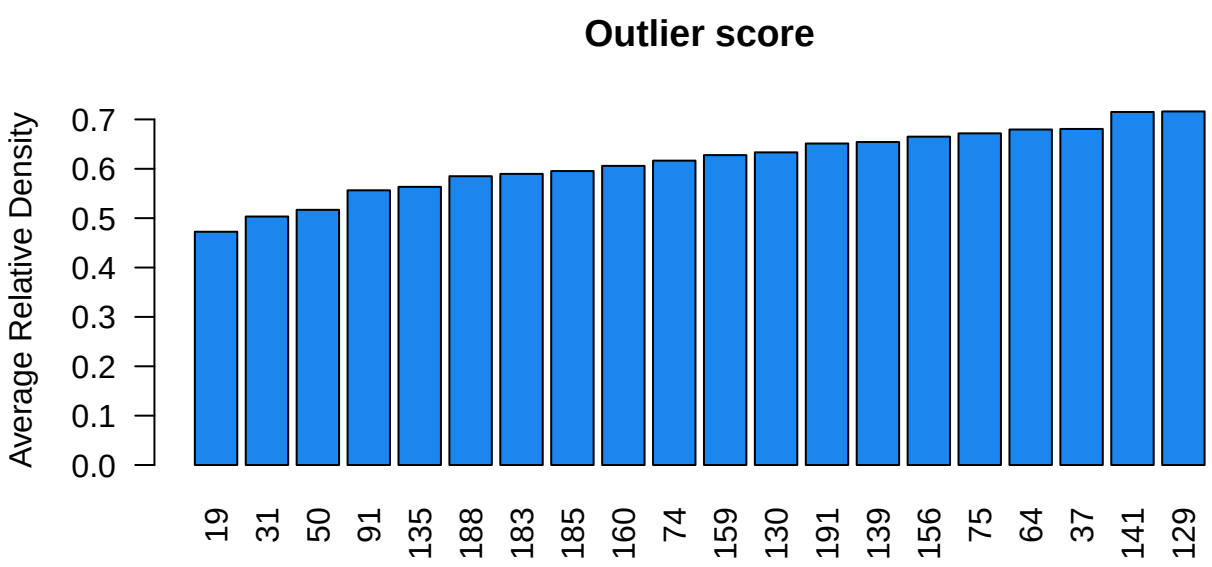

Figure 7: Outlier detection using KNN Average Relative Density.

Based on estimated densities, some possible outliers are: 19, 31, 50. A table of the top eight outliers is seen in Table 3.

Most of the predicted outlier cars seem to be 'natural outliers' in the meaning, that they differ significantly from the other cars, but this is not due to data error or something wrong, but simply because they are meant to be different. This includes observation 50, 130 and 73-75. However, it seems that there may be a typo in the bore ratio of observation 135. This is a Saab that is almost identical to the other Saabs, but it has a bore ratio exactly 1.00 below the bore of the other Saabs. This can be seen in Table 4.

Table 4: Comparison of all the Saabs in the data set. It seems that there is a typo in the bore of observation 135.

	symboling	make	body-style	bore	horsepower	highway-mpg	price
133	3	saab	hatchback	3.54	110	28	11850

Table 3: The eight data points with the lowest density as predicted using the KNN Average Relative Density Estimator. The presented points correspond to the eight left most bars in Figure 7

	symboling	make	body-style	bore	horsepower	highway-mpg	price
19	2	chevrolet	hatchback	2.91	48	53	5151
31	2	honda	hatchback	2.91	58	54	6479
50	0	jaguar	sedan	3.54	262	17	36000
91	1	nissan	sedan	2.99	55	50	7099
135	3	saab	hatchback	2.54	110	28	15040
188	2	volkswagen	sedan	3.01	68	42	9495
183	2	volkswagen	sedan	3.01	52	46	7775
185	2	volkswagen	sedan	3.01	52	46	7995

	symboling	make	body-style	bore	horsepower	highway-mpg	price
134	2	saab	sedan	3.54	110	28	12170
135	3	saab	hatchback	2.54	110	28	15040
136	2	saab	sedan	3.54	110	28	15510
137	3	saab	hatchback	3.54	160	26	18150
138	2	saab	sedan	3.54	160	26	18620

3 Association mining

This part of the report focuses on association mining. Many different endresults are possible depending on what is chosen in the process. Here, categorical variables are binarized using One-out-of-K-encoding, and continuous variables are binarized by splitting on the median, where observations equal to the median belong to the 'High'-group.

This leaves a data set consisting entirely of zeros and ones. Now, the number of high support item sets should be indirectly chosen through a minimum support value, and the number of association sets should be indirectly chosen through a minimum confidence value.

The minimum support is set to 80 %, which leads to 9 frequent item sets. These can be seen in Table 5. Furthermore, the minimum confidence is set to 80 %, which leads to 25 association rules with high confidence. These can be seen in Table 6.

Table 5: The item sets discovered using the Apriori algorithm with a support of at least 80 %.

Set #	Attributes	Support
1	aspiration.std	82 %
2	aspiration.std + engine-location.front	80 %
3	fuel-type.gas	90 %
4	fuel-type.gas + num-of-cylinders.high	88 %
5	fuel-type.gas + num-of-cylinders.high + engine-location.front	86 %
6	fuel-type.gas + engine-location.front	89 %
7	num-of-cylinders.high	98 %
8	num-of-cylinders.high + engine-location.front	96 %
9	engine-location.front	99 %

Table 6: The association sets discovered using the Apriori algorithm with a confidence of at least 80 %.

Rule #	Attributes	Support	Confidence
1	aspiration.std <- ∅	82 %	82 %
2	fuel-type.gas <- ∅	90 %	90 %
3	num-of-cylinders.high <- ∅	98 %	98 %
4	engine-location.front <- ∅	99 %	99 %
5	fuel-type.gas <- aspiration.std	96 %	79 %
6	aspiration.std <- fuel-type.gas	87 %	79 %
7	num-of-cylinders.high <- aspiration.std	97 %	80 %
8	aspiration.std <- num-of-cylinders.high	82 %	80 %
9	engine-location.front <- aspiration.std	98 %	80 %
10	aspiration.std <- engine-location.front	82 %	80 %
11	num-of-cylinders.high <- fuel-type.gas	97 %	88 %

Rule #	Attributes	Support	Confidence
12	fuel-type.gas <- num-of-cylinders.high	90 %	88 %
13	engine-location.front <- fuel-type.gas	98 %	89 %
14	fuel-type.gas <- engine-location.front	90 %	89 %
15	engine-location.front <- num-of-cylinders.high	99 %	96 %
16	num-of-cylinders.high <- engine-location.front	98 %	96 %
17	aspiration.std <- fuel-type.gas num-of-cylinders.high	87 %	76 %
18	fuel-type.gas <- aspiration.std engine-location.front	96 %	77 %
19	aspiration.std <- fuel-type.gas engine-location.front	87 %	77 %
20	num-of-cylinders.high <- aspiration.std engine-location.front	97 %	78 %
21	aspiration.std <- num-of-cylinders.high engine-location.front	81 %	78 %
22	engine-location.front <- fuel-type.gas num-of-cylinders.high	98 %	86 %
23	num-of-cylinders.high <- fuel-type.gas engine-location.front	97 %	86 %
24	fuel-type.gas <- num-of-cylinders.high engine-location.front	90 %	86 %
25	aspiration.std <- fuel-type.gas num-of-cylinders.high engine-location.front	86 %	75 %

The first 4 association rules are identical to the identified item sets with a only one item. The association rules, where one item implies another item, are doubled in the sense that if A implies B is a rule, then B implies A is another rule.

Note that some of the discovered association rules (5, 6, 17, 18, 19, 20, 21, 25) have a confidence below the minimum confidence.

Generally, the discovered association rules represent the most common cars, meaning having a front-located engine, four cylinders or above, no turbo and running on gas, not diesel.