

## 1 EDA

The rounding in noise\_level is weird. (fig1)

Fuel\_type can be easily classified with other variables. (fig2)

There are a lot of linear relations between variables. (fig3)

## 2 PCA

The covariance matrix (fig 4) is unbalanced so PCA should be done scaled data.

The distribution of the orthogonal/score distances of the validation set are very similarly distributed as the ones of the train set. (fig 7)

## 3 clustering

HDBSCAN and hierarchical ward clustering deliver similar results and both find Fuel\_type. From the heatmap (fig 11) you can see the distinction between the distribution from cars with different fuel types.

## 4 linear regression

Because the variables suffer from multicollinearity we chose to for lars with lasso penalty. The lasso penalty means that we penalize the regression for big coefficients you could also think of it as constraintment on the coefficients. The lars part indicates a smart constrained optimization implementation of it.

The model that we chose satisfy all assumptions (fig ??, 15).

Doing an ANOVA with multicollinearity is tricky. We didn't do one but what we would do is clustering the variables and looking to the variance of these clusters.

## 5 classification

## 6 figures

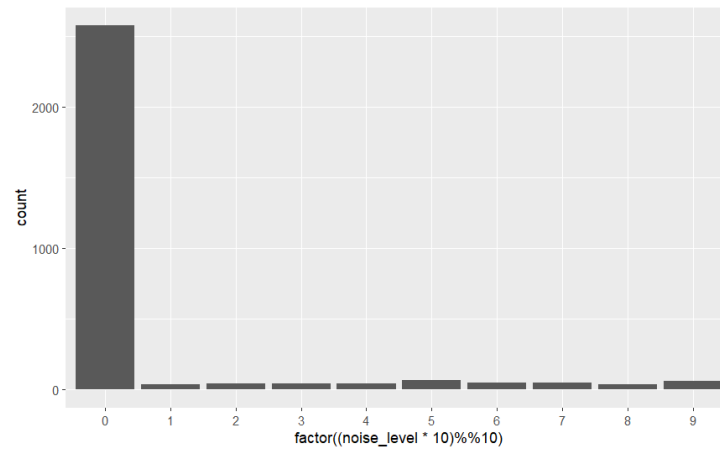


Figure 1: weird rounding noise

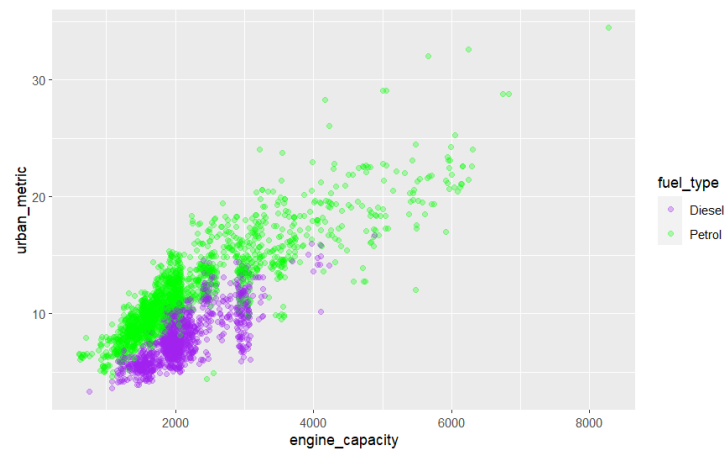


Figure 2: plt1

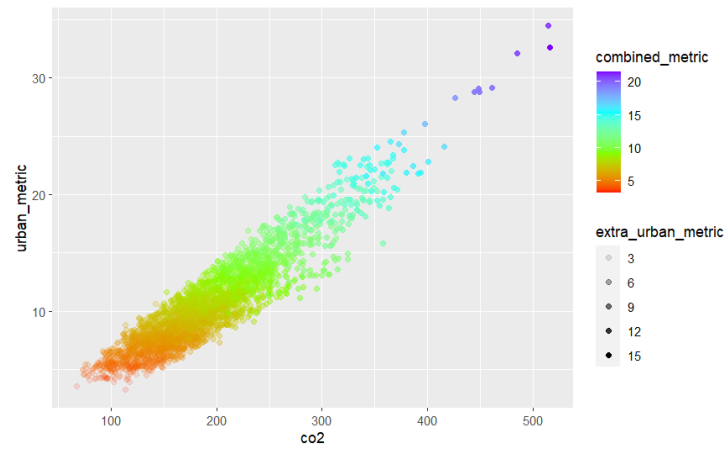


Figure 3: plt2

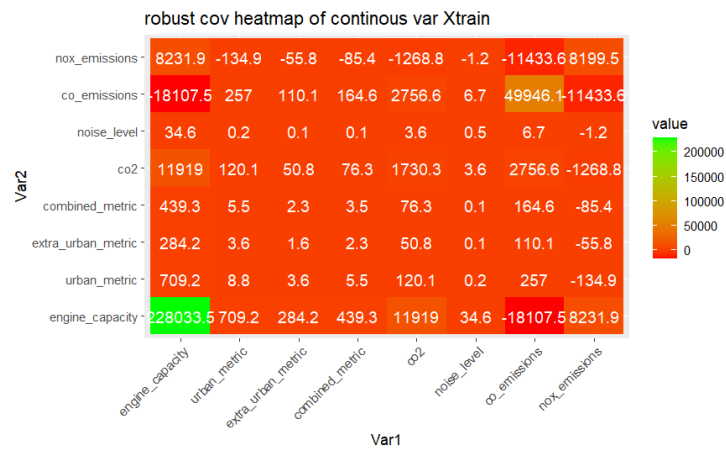


Figure 4: robust covariance

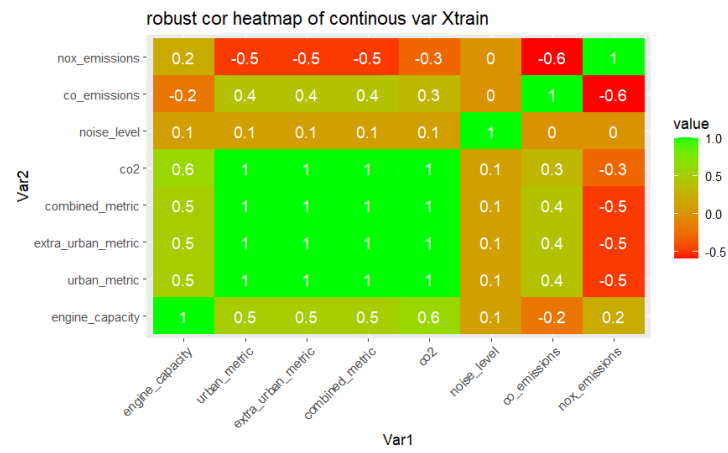


Figure 5: robust correlation

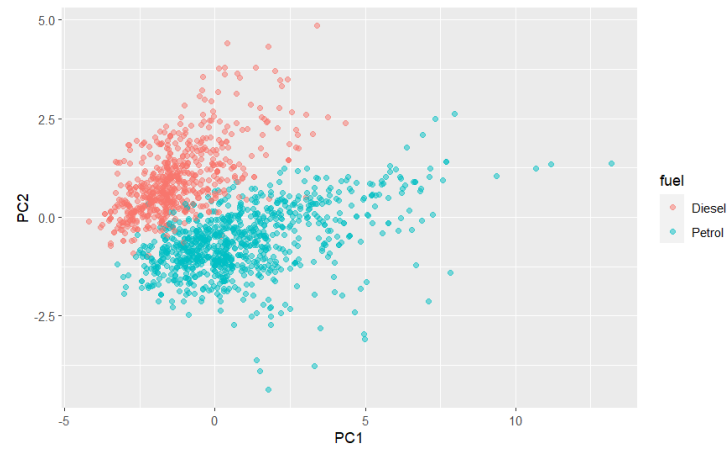


Figure 6: biplot

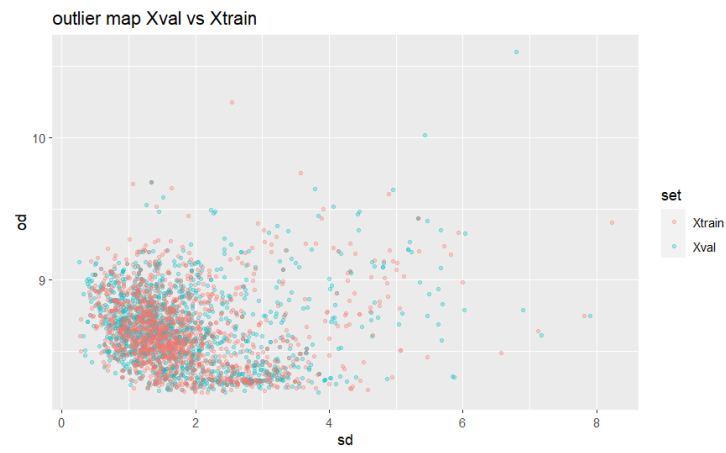


Figure 7: validation pca

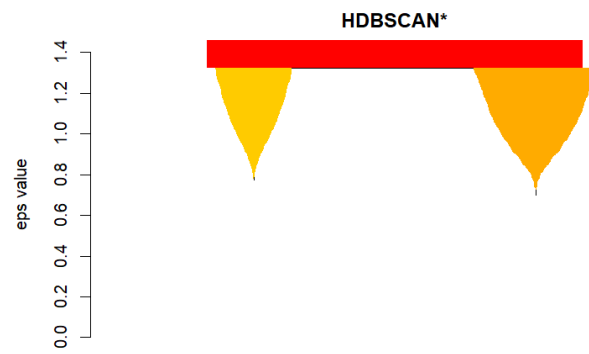


Figure 8: hdbbs1

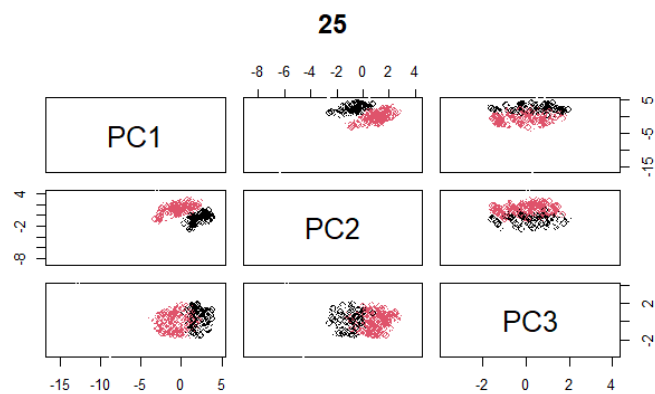


Figure 9: hdb2

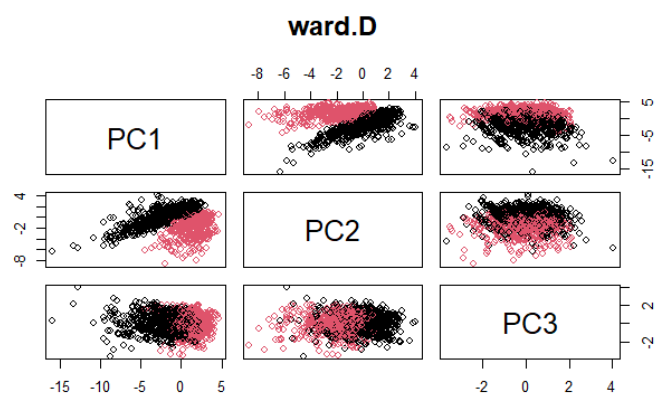


Figure 10: ward

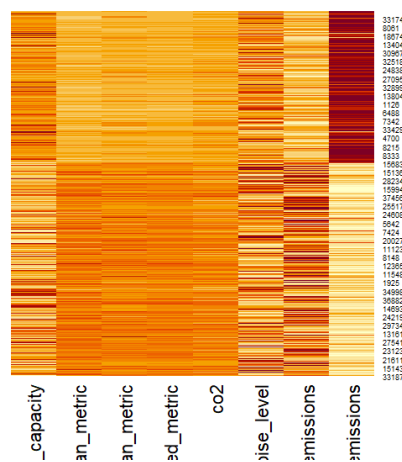


Figure 11: heatmap

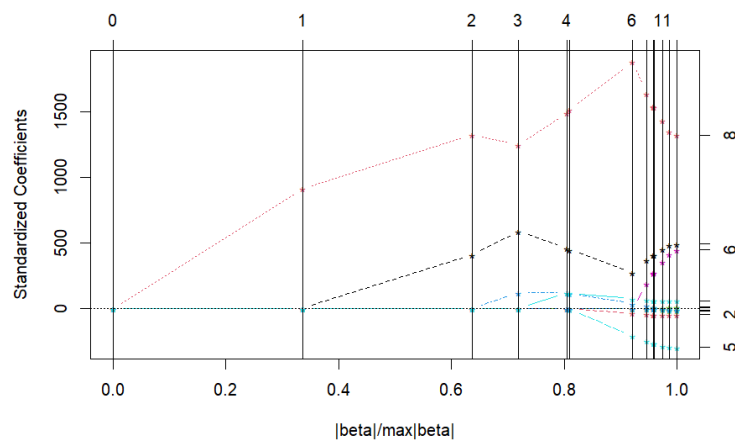


Figure 12: lars1

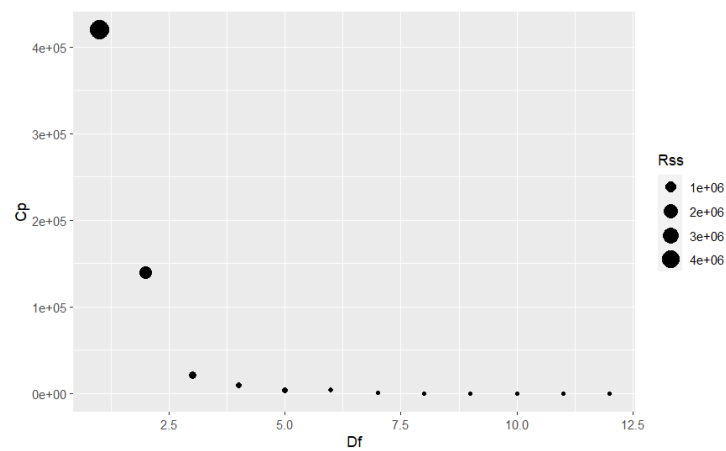


Figure 13: lars2

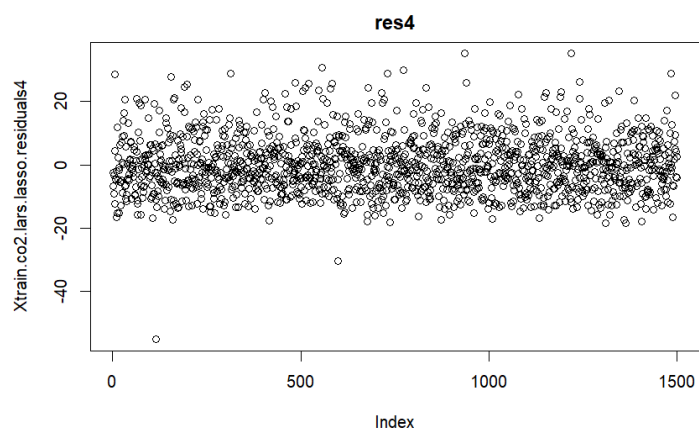


Figure 14: lars3



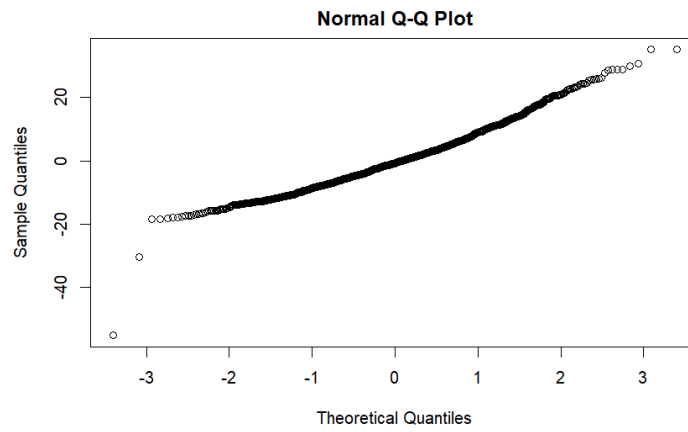


Figure 15: lars4

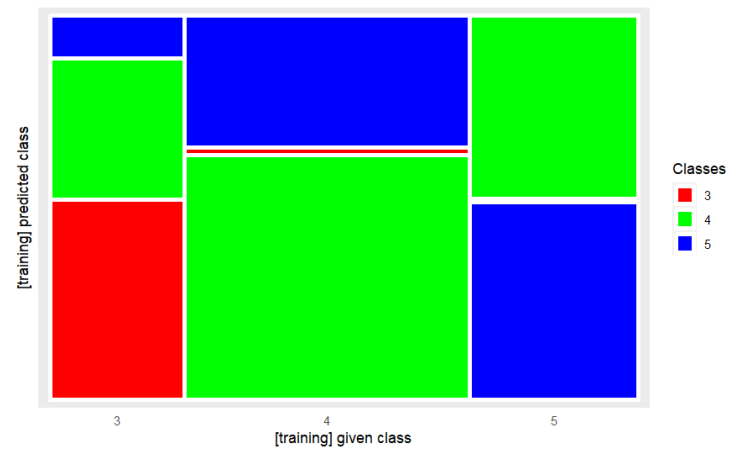


Figure 16: knn train



Figure 17: knn validation