Katholieke
Universiteit
Leuven

Statistische Modellen & Data-analyse

Practicum 2

Tomas Fiers – r0380267

Juni 2017

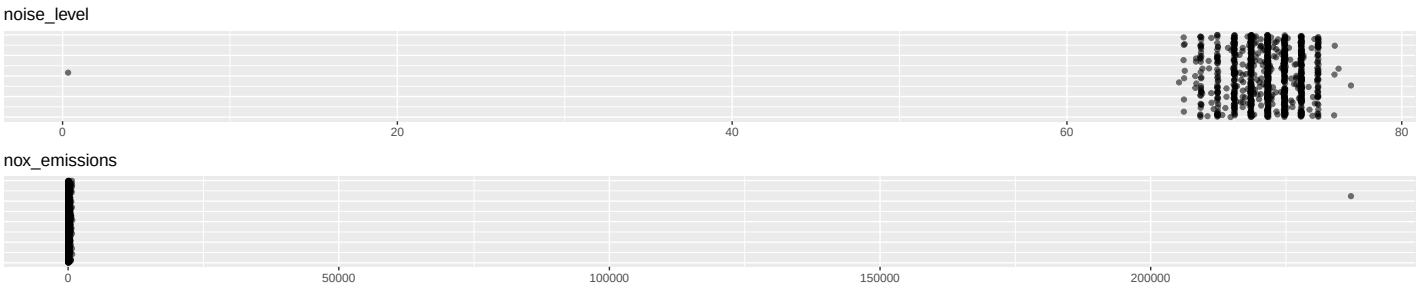

Figure 1: Univariate point plots of two variables with extreme outliers. Vertical jitter has been added to better show the distributions, by mitigating overplotting.

1 Linear regression

The complete dataset of 1500 cars was split into a 1000-car training set and a 500-car test set. We continue this analysis with the training set.

Explorative analysis & outlier removal

All variables were examined for univariate outliers. As shown in fig. 1, both `noise_level` and `nox_emissions` contain an extreme outlier. These are the *"Volkswagen Jetta (from NOV 06 Wk 45 >) – 1.4 TSI (170 PS) Sport"* with a `noise_level` of 0.3, and the *"Vauxhall Signum MY2008 – 3.0CDTi V6 24v with 16/17/18" wheel"* with a `nox_emissions` value of 237000, respectively. They were removed from the training set.

(no additional real outliers found – reference appendix, boxplot+points)

. (introduce figure)

Note the very strong correlations between the `urban_metric`, `extra_urban_metric`, `combined_metric`, and `co2` variables; and to a lesser extend the `engine_capacity` variable. This is in accordance with their meaning: the three "metrics" are all a measure of fuel consumption. And cars with larger engine volumes consume more liters of fuel, and every liter of fuel contains a fixed amount of CO_2.

appendix: boxplots of other variables (all variables, after outlier removal)

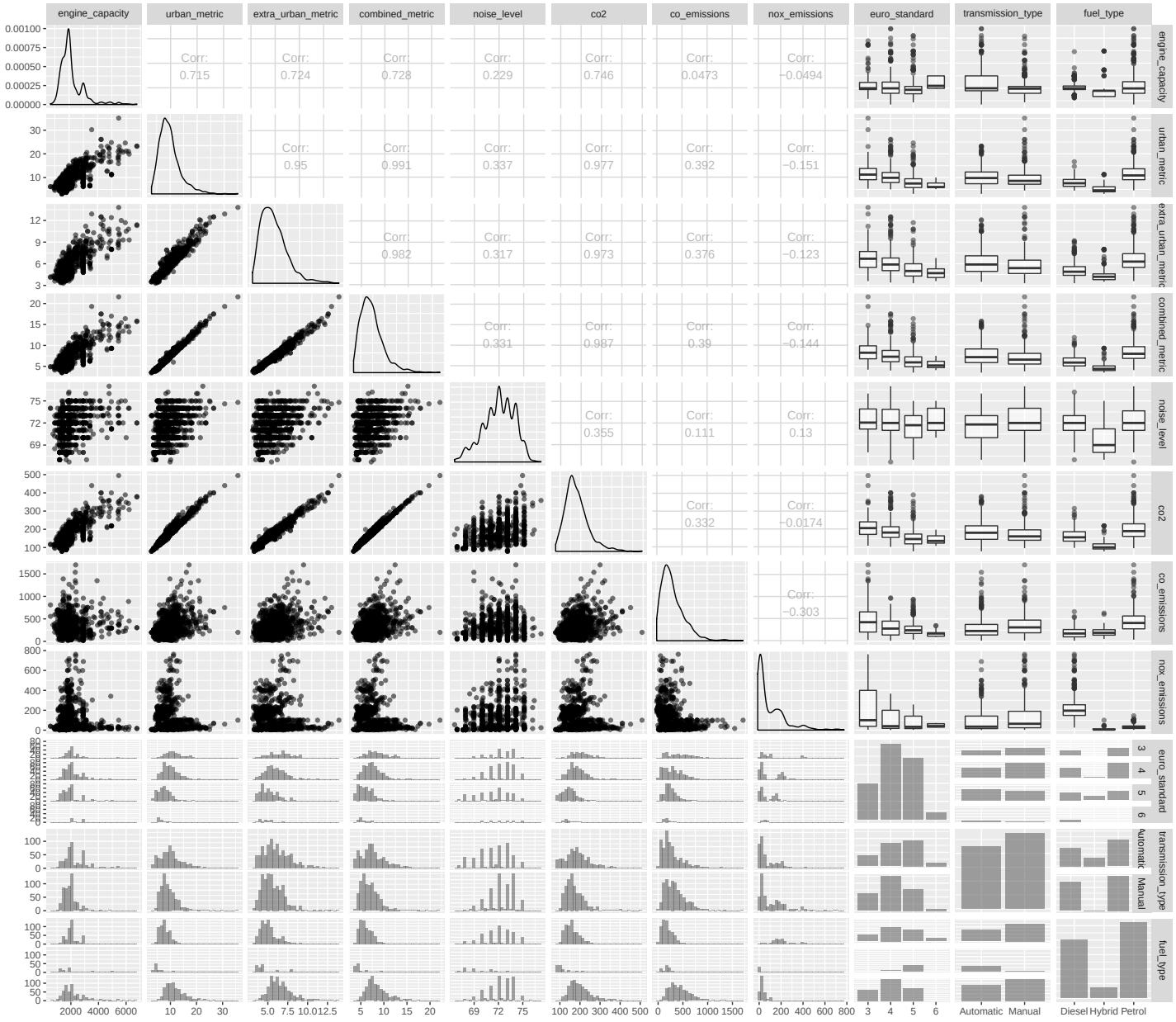

Figure 2: Uni- and bivariate distributions

split code in: - load & split data, make cols (init) - jitterbox outliers - outlier detection, removing - jitterbox all - pairs

Variable transformation

Variable selection

Linear model

Gauss-Markov conditions

2 Classification

Variable selection

Logistic model

Interpretation

Apparent error rate

Error rate on test set

Linear discriminant analysis

Quadratic discriminant analysis

Comparison

A Code

B Extra figures

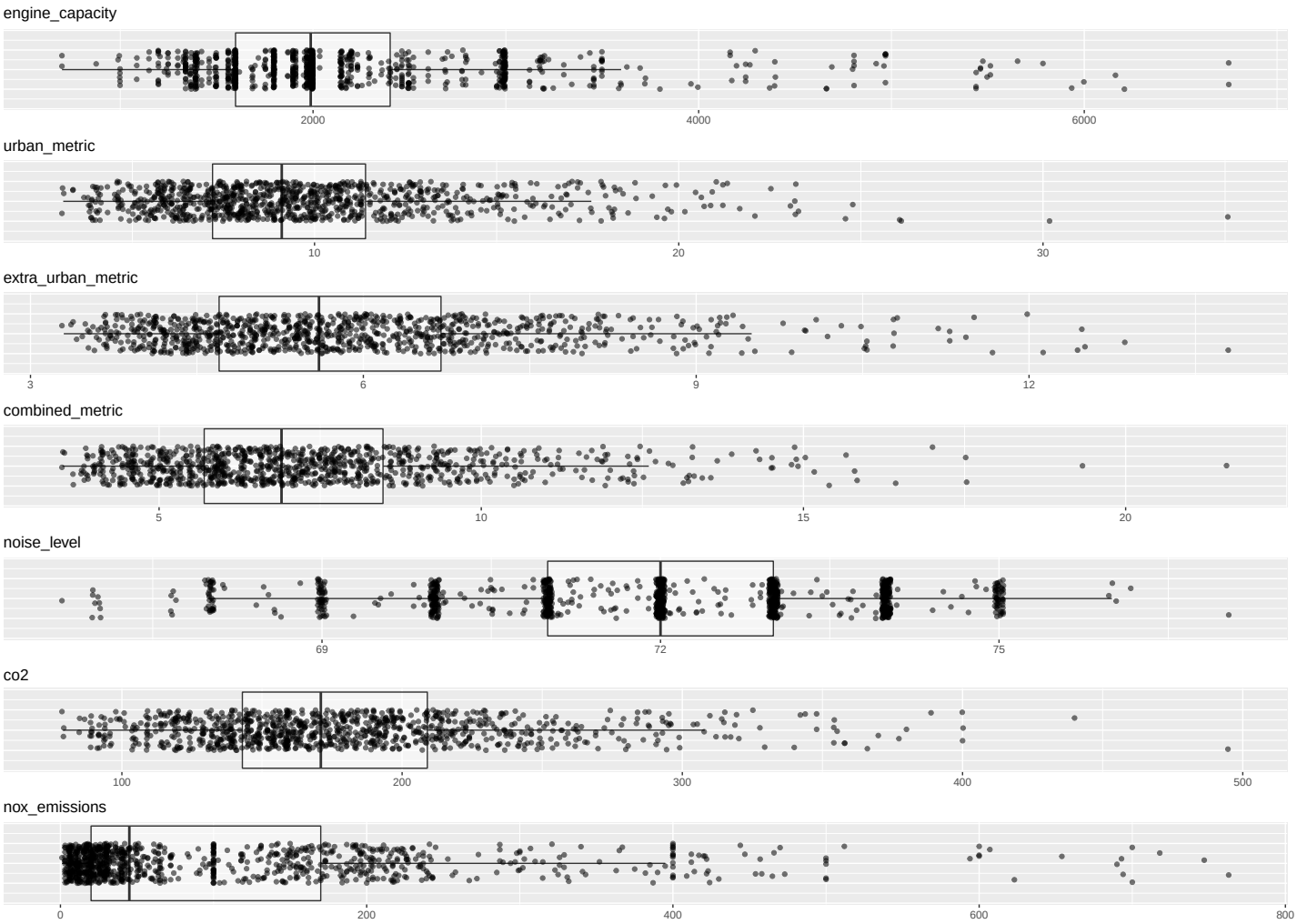

Figure 3: Univariate distribution of each continuous variable (after the two extreme outliers have been removed, as described in Explorative analysis & outlier removal). These are boxplots overlaid with point plots, where vertical jitter has been added to mitigate overplotting.