Exercise 1:

(a) Which of the following statement(s) apply to feature effect methods?

(i) The value of the PDP at a point x_j, corresponds to the point-wise average of the values of the ICE curves at this point.

(ii) The PDP of a feature provides information about possible interaction effects of the feature.

(iii) ICE curves of a feature for multiple data points provide information about possible interaction effects of the feature with others.

(iv) If we center the ICE/PDPs for categorical features, the expected changes always refer to a selected reference category.

(v) ALE plots are based on conditional distributions, PDPs on marginal distributions.

(vi) If features are uncorrelated, ALE plots are equal to PDPs.

(vii) ALE plots are faster to compute than PDPs.

(b) You fitted a model that should predict the value of a property depending on the number of rooms and square meters. You want to compute feature effects using the following methods: PDP, M-plots and ALE plots. Which of the following strategies reflect which method?
The feature effect for a 30 m^2 corresponds to...

a) ... what the model predicts on average for flats that also have around 30 m^2, e.g., 28 m^2 to 32 m^2.

b) ... how the model predictions changes on average when flats with 28 m^2 to 32 m^2 have 32 m^2 vs. 28 m^2.

c) ... what the model predicts on average if all properties in the dataset have 30 m^2.

Exercise 2:

In exercise 2 on sheet 1 you received a dataset with 11 observations and two features:

	1	2	3	4	5	6	7	8	9	10	11	$\sum_{i=1}^{n}$
y	-7.90	-6.08	-3.74	-1.18	-1.23	-0.55	0.05	0.88	4.74	2.93	2.55	-9.53
x1	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	0
x2	0.95	0.65	0.40	0.07	0.06	0.02	0.02	0.14	0.34	0.60	0.98	4.23

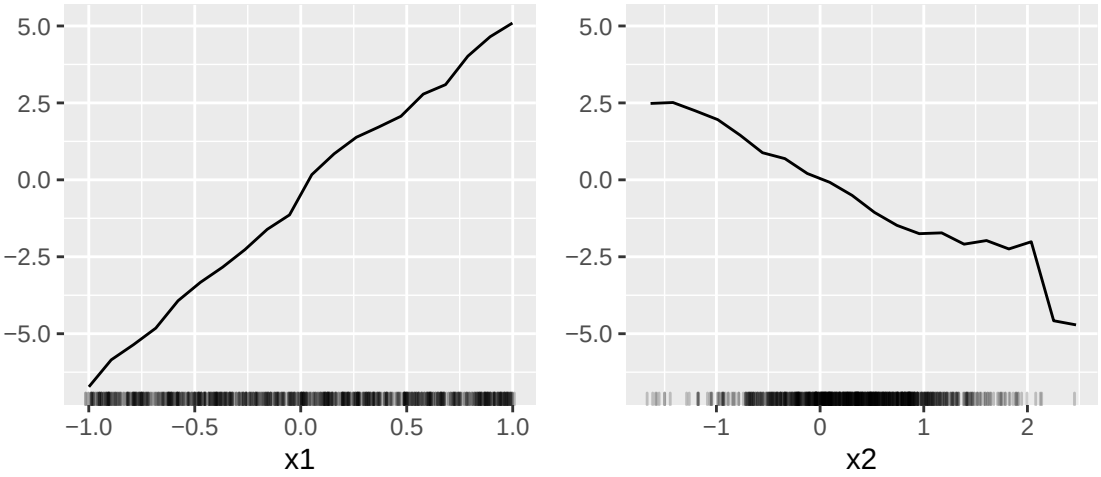
The last column corresponds to the sum of values of each row.

Instead of 11 data points you now received a dataset with 1000 data points from the same data generating process as above. You fitted a linear model to the data $\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}_1 + \hat{\beta}_2\mathbf{x}_2 + \hat{\beta}_3\mathbf{x}_1\mathbf{x}_2$ The following plots show the PDP (first row) and ALE (second row) for x_1 and x_2.

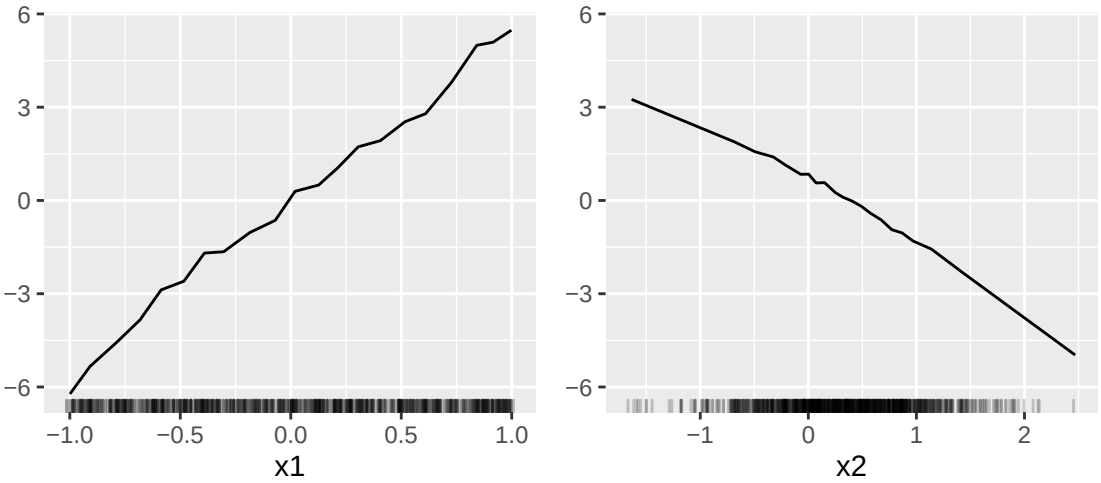
(a) Interprete the plots with respect to the feature effect of x_1 and x_2.

(b) Would you rather trust the PDP or ALE plot? Give reasons for your decision.

PDP


5.0
2.5
0.0
−2.5
−5.0
−1.0
−0.5
0.0
0.5
1.0
x1
5.0
2.5
0.0
−2.5
−5.0
−1
0
1
2
x2

ALE


6
3
0
−3
−6
−1.0
−0.5
0.0
0.5
1.0
x1
6
3
0
−3
−6
−1
0
1
2
x2