Exercise 1: Quiz

(a) Which of the following statement(s) is/are correct?

(i) Interpretation methods are mainly needed to better explain real world phenomena.

(ii) Model-agnostic methods need access to gradients to explain a model.

(iii) In IML we distinguish between global IML methods, which explain the behavior of the model over the entire feature space, and local IML methods, which only explain the prediction of individual observations.

(iv) We can also draw conclusions about feature importance from feature effect methods.

(v) Technically, correlation is a measure of *linear* statistical dependence.

(vi) Features that have an equal feature effect are correlated.

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

(c) 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...

(i) ... what the model predicts on average for flats that also have around 30 m^2, for example, 28 m^2 to 32 m^2.

(ii) ... 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.

(iii) ... what the model predicts on average when all properties in the dataset have 30 m^2.

Exercise 2: Application
Make yourself familiar with the iml package in R and the method FeatureEffects. Answer the following questions for your chosen dataset

(a) Create PDP and ALE plots for at least 6 features of your dataset you are most interested in how they affect your target variable. Can you see any differences between PDP and ALE plots? Can you explain those differences?

(b) Are your chosen features interacting with any other features? Does your chosen visualization technique show you with which other features they are interacting?

(c) Based on the PDPs: Which of the 6 features would you consider most and which one least important?