Question 1

Suppose we have a feature with all the values between 0 and 1 except few outliers larger than 1. What can help us to decrease outliers' influence on non-tree models?

Correct answers:

- Apply rank transform to the features. Yes, because after applying rank distance between all adjacent objects in a sorted array is 1, outliers now will be very close to other samples.
- Apply np.log1p(x) transform to the data. This transformation is non-linear and will move outliers relatively closer to other samples.
- Apply np.sqrt(x) transform to the data. This transformation is non-linear and will move outliers relatively closer to other samples.
- Winsorization. The main purpose of winsorization is to remove outliers by clipping feature's values.

Incorrect answers:

- StandardScaler. No, despite feature will be scaled, relative distances between outliers and other values still will be huge.
- MinMaxScaler. No, despite feature will be scaled, relative distances between outliers and other values still will be huge.

Question 2

Suppose we fit a tree-based model. In which cases label encoding can be better to use than one-hot encoding?

Correct answers:

- When categorical feature is ordinal. Correct! Label encoding can lead to better quality if it preserves correct order of values. In this case a split made by a tree will divide the feature to values 'lower' and 'higher' that the value chosen for this split.
- When we can come up with label encoder, that assigns close labels to similar (in