What we already did, processing pipeline:

1. Nan check
2. Infinity check
3. Plot all hists -> noticed skewness (#, n\_tokens\_content)
4. Cubical root transform to skewed cols (beware of skew\_thresh and maybe not all colums you have to transform)
5. Finding and plotting columns with outliers (parameter – critical percentage of outliers)
6. Perform IQR filtering (remember – q\_from, q\_to, threshold percentage of outliers – all parameters)

Model preparation (13.05.22):

1. Building full model, plotted residual plots. Bad, R^2 ~ 0.11
2. Tried full model with log(Y) transform 🡪 residual plot became much better and R^2 became higher, up to 0.17

Model preparation (18.05.22)

1. Fixed train / validation / test set selection
2. Models:
   1. Full model (R^2 ~ 0.13)
   2. Poisson GLM (full model with log(Y) transform) 🡪 R^2 ~ 0.16
   3. Built best models after forward / backward selection (without log(Y) transform)
   4. Did the same as above (with log(Y) transform) -> R^2 ~ 0.16, 20 parameters
3. Compared models based on BIC, R^2 and test results