Questions to answer:

1. Reasons for bimodal distribution
2. Are linear models appropriate for this regression?
3. What are the most important parameters that affect number of shares?

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:

1. Building full model (**try log transform of Y variable, shares are highly skewed!**)