# Models

Caret::Train

Rstanarm::stan\_lm

# Validation/Testing Methods

Rstanarm::kfold

* a matrix containing point estimates and standard errors of the expected log pointwise predictive density ("elpd\_kfold"), the effective number of parameters ("p\_kfold", always NA) and the K-fold information criterion "kfoldic" (which is -2 \* elpd\_kfold, i.e., converted to the deviance scale).

Rstanarm::waic()

Anova()

Pipe model into summary, can extract values like r.squared from them.

# EDA Method

* Histogram of distribution of each variable
* Scatterplot of each variable against the response (with geom\_smooth trendline)
* Correlation Heatmap
* For binary variables, histogram of both with fill used to denote which is event and which is non-event (OR even better, violin plots and frequency plots)
* Summary
* Glimpse
* Could use pivot\_longer where you examine the boxplot of each of the inputs, grouped by/colored by/filled by the outcomes.

# Notes about data

* All variables in train\_x data are only weakly correlated with each other at best, but a number of the variables in train\_v are ***very*** *correlated with each other.*

Remember to join train\_x with train\_outputs.

A picture containing text, outdoor, crowd, crowded

Description automatically generated

1. **Check out ION example model from week 4. The inputs are highly correlated there. Prof shows how to extract features via principle component analysis to create uncorrelated features to use in the models.**