

## Formula sheet

### Summary statistics

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}, \quad r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Types of variables: categorical, quantitative, logical, date.

Descriptive words for univariate distributions:

- unimodal, bimodal, multimodal
- symmetric, right-skewed, left-skewed, uniform
- outliers

Descriptive words for bivariate distributions:

- shape: linear, non-linear, no relationship
- strength: weak, moderate, strong
- form: positive, negative

### Tidy data

Verbs: gather, spread, nest/unnest, separate/unite

### Wrangling data

Verbs: filter, arrange, select, mutate, summarise, group/ungroup

### Grammar of graphics

There are seven components of the grammar that define a data plot: DATA, AESTHETICS/MAPPINGS, GEOM, STAT, POSITION, COORDINATE, FACET.

Colour palettes: sequential, diverging, qualitative

### Models

Simple linear:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- $\varepsilon \sim N(\mu, \sigma)$
- Fitted values:  $\hat{Y} = b_0 + b_1 X$
- Residual:  $e = Y - \hat{Y}$
- Estimates:  $b_1 = r \frac{s_y}{s_x}$ ,  $b_0 = \bar{Y} - b_1 \bar{X}$
- $R^2 = 1 - \frac{\sum e^2}{\sum Y^2}$
- $MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(n-2)}$
- $RMSE = \sqrt{MSE}$
- $MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{(n-2)}$