Exploratory Data Analysis: Correlation

More than 2 numerical data points - look at scatter plots to look for relationships

How can we ~~quantify the~~ quantify the linear relationship between the variables?

Correlation Matrix - give strong or weak relationship between the numerical vars

- 1 - two numerical features perfectly correlated
- -1 - like y is equal to -x
- 0 - no linear relationship

Heatmap - use color to get direction (negative, positive) as well as strength

Pearson Correlation

$$P_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_i)^2}\sqrt{\sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

Variance $\sigma_x^2 = \frac{1}{n}\sum_{i=n}^{N}(x_i - \mu_x)^2$

Covariance between (x,y) $\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_m)(y_i - \mu_y)$

Correlation between (x,y) $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$