The Basics of R-Squared 📏

- R-squared, aka the coefficient of determination, gauges how much of the response variable's variation is explained by the linear model.
- Formula: R-squared = Explained variation / Total variation

Mathematical Formula

Sum Squared Regression Error

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Sum Squared Total Error

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The Percentage Game 📊

- R-squared is always between 0% and 100%.
- 0% means your model explains none of the variability.
- 100% means your model explains all the variability.
- The higher, the better – indicating a snug fit of your model to the data.

The Catch with R-Squared 🤔

- Now, here comes the twist. Is it good to include as many independent variables as possible? Not quite.
- Adding variables, even if not meaningful, boosts R-squared. But do we want that? No!
- Here lies the conundrum - the basic problem with R-squared.

Enter Adjusted R-Square 🔄

- The savior! Adjusted R-Square penalizes the model for unnecessary variables.
- It's like a refined version of R-squared, ensuring that only meaningful variables contribute to the model's goodness of fit.

Adjusted R-Square Formula

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1-R^2)\cdot(n-1)}{n-k-1}\right)$$

Here:

- R2 (R-squared)is the coefficient of determination from the original model.
- n is the number of observations in the sample.
- k is the number of predictors in the model.

The Dilemma Solved! 🎉

- With Adjusted R-Square, we strike a balance. It considers the model's complexity and doesn't fall for the allure of adding variables just for the sake of it.
- Now, we have a reliable measure that values meaningful impact over unnecessary additions.