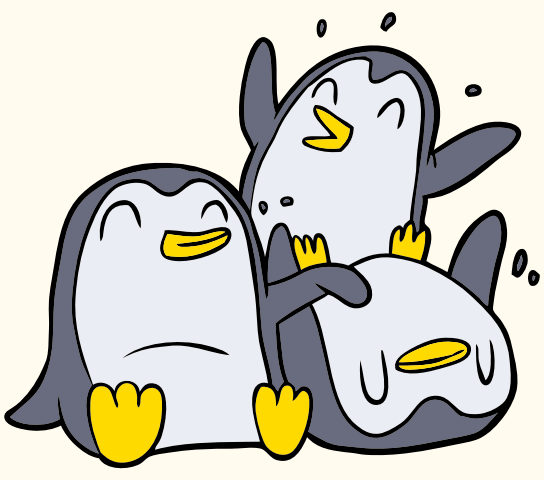
CHALLENGE

Sex of Penguins

01 PICK THE VARIABLES

In the problems we solve, we should be able to create a story behind every driver. If you dig long enough, you can find anything, but not necessarily what you need



02 HISTOGRAMS

Getting to know the variables is an important step. The histograms give you visibility on data distributions and hints towards the presence of outliers

03 BINARY VARIABLES

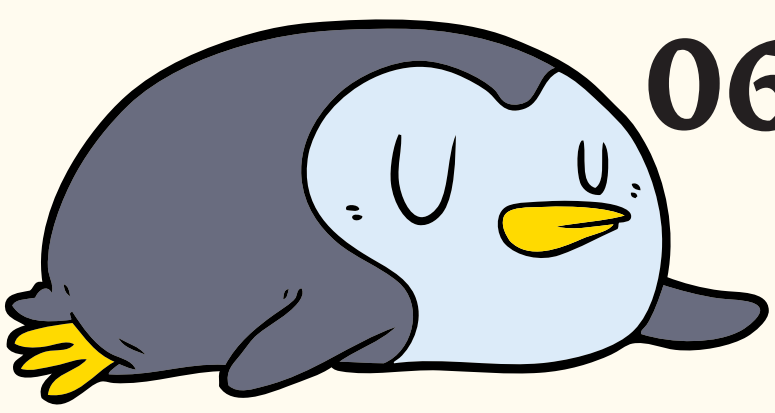
The Logistic Regression needs all inputs to be numeric. This means that we transform character/string variables into binary (0 or 1). However, please be wary of Dummy Variables

04 TRAINING AND TEST SET

Create a training and test set with an 80/20 split. This allows an unbiased assessment of the model

05 LOGISTIC REGRESSION

Now, you should be ready to apply the Logistic Regression to the Data. Don't forget to add a constant to X. You can also look at the summary, and interpret the output



06 ACCURACY ASSESSMENT

Look at the key metrics to understand whether your model is good at predicting the sex of penguins. That will give you more trust in the insights

Diogo Resende