Exercises

1. Create a scatterplot of faceoff\_win\_pct against win\_loss\_pctg with a least squares regression line. What is the trend of this relationship?
2. Fit a linear model with faceoff\_win\_pct predicting win\_loss\_pctg. Is faceoff\_win\_pct an effective predictor? Write down the equation of the model. What is the R-squared value?
3. Create another scatterplot of faceoff\_win\_pct against win\_loss\_pctg but with separate least squares regression lines for the Mens and Womens Divisions. How do the two lines compare to each other?
4. Write the appropriate code to make Division an indicator variable with 1 representing “Mens” and 0 representing “Womens”.
5. Fit a model that would allow us to have different intercepts for the Mens and Womens divisions.
6. Approximately what is the value of the coefficient associated with the indicator variable? What else do you notice about this model? Why does this make sense?
7. Now, fit a model that would allow us to have different intercepts **and** slopes for the Mens and Womens divisions.
8. Use the model to estimate the equation for each of the two divisions. For each, interpret the estimated slopes and intercepts.
9. Is the model using two completely separate lines (i.e., different intercepts and slopes) better than the single model ignoring Division entirely?
10. At approximately what point do the lines cross in the plot of the model? Why do you think that is the case?
11. Filter the data to make separate sets for the Mens and Womens divisions. Now, calculate the correlation between shots\_per\_game and faceoff\_win\_pct for the men and women respectively. Which is higher? What conclusions can you draw?
12. Using the filtered data sets again, calculate the mean shot\_pctg (total goals divided by total shots) of Men and Women respectively. Which is higher? How can this add to our concusions?