PPAS Challenge Questions

For this challenge, we’ll be using the “USArrests” and “state.x77” datasets in base R’s “datasets” package. These datasets include information about demographics and crime in the United States in the 1970s on a state-by-state basis.

There are hints at the end of this document as to which functions will help with each question, and never be ashamed to use Google and/or StackExchange to help.

* 1. Take a look at R’s documentation of these datasets to familiarize yourself with them. Look at data summaries and histograms to get a sense for the distribution of values.
     1. Are both datasets of the class “data.frame”? You’ll want to make sure they both are.
  2. Join the information from the two data frames together into a single data frame, matching by state.
  3. You’ll want to make sure that the names of your columns make sense, and that no two columns have the same name.
  4. Create a correlation matrix of all the numeric columns. Later in modeling, it will be important to know which variables are correlated to each other.

Crime statistics from the “USArrests” data frame come from 1973, while the murder rates from the “state.x77” data frame are from 1976. All other variables from the “state.x77” data frame are from before 1976. I smell a predictive modeling project!

* 1. Create a pivot table that splits observations into five groups of ten, ordered by 1973 murder rates, and then calculates average 1976 murder rates within each group.

If you’re feeling especially bold, pick three new variables from the dataset, split each one into two groups by ordered value, and then calculate average 1976 murders rates in each of the eight group combinations.

* 1. Create a linear regression model to predict murder rates in 1976 information from previous years. Feel free to use any predictor variables that make sense, but be sure to include murder and assault rates in 1973 in order to answer later parts of this question.
     1. Notice that two of the column names from the original state.x77 dataset have spaces. This creates problems in fitting a linear model if you want to use those variables. Change those variable names so that they don’t have spaces.
     2. Check the normality of the response variable using a quantile-quantile plot, and/or find a statistical hypothesis test for normality.
     3. Fit a model to predict murder rates by state in 1976, using at least 1973’s assault and murder rates, and then anything else you think might be predictive. Look at the model summary.
  2. Note the high correlation between assault rates in 1973 and murder rates in 1973 in your correlation matrix from earlier. One of those variables is likely to be statistically *in*significant in your linear model.
     1. Think about what is happening here, and what we can do to clarify effects in a linear model. Implement your idea as part of your best model.
     2. Arrive at a best model, and check the residual plots for any funny business.

# Hints

1. ?, summary, hist, as.data.frame
2. left\_join, mutate, rownames, cbind
3. names
4. cor, select
5. summarize, group\_by, ntile, cut
6. names, qqnorm, qqline, shapiro.test, lm, summary
7. lm, resid, plot, hist, summary