STAT 645: Assignment 3
Due Monday, September 14, 11:55pm Central

1. For the pollution data in "pollute_data.csv" [**Note:** There is at least one missing value in these data. Just remove any records with at least one missing value prior to doing the following analyses.]:

   (a) Based on the model

   $$\text{Mortality}_i = \beta_0 + \beta_1 \times \text{HCPot}_i + \epsilon_i, \tag{1}$$

   does there appear to be a significant linear relationship between HCPot and Mortality?

   (b) Make a scatterplot of HCPot versus Mortality. Do you notice anything unusual that might have impacted your model in 1(a) [**Hint:** You should ;-)]. Dig into the data and provide an explanation for any unusual features you notice.

   (c) Compare the California records to all others, in terms of each of the following:

      i. Percent of white-collar workers.
      ii. Median income.
      iii. Population per household.
      iv. Percent non-white residents.
      v. Mean July temperature.
      vi. Annual rainfall.

   (d) Write down the model for Mortality as a function of log(HCPot), as well as all variables from 1(c). **Note:** By "write down the model," I mean for you to write down an equation analogous to equation (1) above.

   (e) Interpret all coefficients in the model from 1(d).

   (f) Using the likelihood ratio test, test the null hypothesis that all coefficients other than that for log(HCPot) equal 0, in the model from 1(d). Test at $\alpha = 0.05$.

   (g) Using the likelihood ratio test, test the null hypothesis that the coefficients for percent white collar and percent non-white sum to zero. Test at $\alpha = 0.05$.

   (h) Report a 95% confidence interval for the sum of the coefficients for percent white collar and percent non-white. Interpret the result, and also test the hypothesis of (g) using this CI.

   (i) For the model that includes log(HCPot), as well as all variables from 1(c), identify any potential leverage or inluential points from this data.