STATS 32: Practice with Manipulating Datasets

Kenneth Tay

Note: There are often multiple ways to answer each question.

1. Install and load the `MASS` package. Load the `nlschools` dataset.

2. How can we find a description of the `nlschools` dataset? Use some of the functions we learned to get a feel for the data.

3. How many students are there in the dataset?

4. How many students were there in each class? Which class had the most number of students?

5. Get a random sample of 10 rows from the dataset. (Hint: Look at the documentation for the `sample_n` function in the `dplyr` package.)

6. Make a scatterplot of IQ vs. lang.

7. There was a lot of overplotting in the previous scatterplot. Make the plot more readable by adding jitter and changing the alpha value of the points. What is the relationship between IQ and lang?

8. What is the correlation between IQ and lang?

9. Using the previous chart, color the points according to their value in the COMB column. Does the value of COMB affect the relationship between IQ and lang? What geometry could you add to the chart to check this?

10. Class size could also affect IQ. Create a data frame which shows the mean IQ and language test scores of students for each class size.