Assessment 2:

1. Download the Datasaurus Dozen dataset (from here: https://www.openintro.org/data/index.php?data=datasaurus). You can use Inzight/on your own/use https://istats.shinyapps.io/Association_Quantitative/ / https://istats.shinyapps.io/LinearRegression/ .
    a. Calculate beta parameters in linear regressions and the correlations for all datasets. Hint: Do you need to calculate the beta parameters and correlations for each one separately?
    b. Make plots for all of them (don't use already existing plots) along with trend lines. Remember the "Subset by" functionality – exists only on Inzight to help make the scatterplot matrix.
    c. Given the correlations are so low, does that mean there is no relationship between variables for this dataset? What is an alternative technique that would help us to model relationships between variables here?

2. What is the maximum SSE (same as the sum of squares) you can achieve for the Quartet I in Seeing Theory in the section on Ordinary Least Squares while still having a non-zero slope (B1)? Include an image of the result (screenshot it) and the fitted values in the table. What Is the theoretical maximum? What is the minimum SSE you can achieve? Why is that minimum possible?

3. Get bootstrapped confidence intervals for the average (is the mean or median a better average to use?) homicide rate (only that one variable) across the states of the USA using the dataset you worked with from Assessment 1. Make a figure and report the 90% confidence intervals. Compare bootstrapped confidence intervals to the confidence interval when assuming the central limit theorem holds. You can use Inzight or https://istats.shinyapps.io/Boot1samp/ .

4. Demonstrate the central limit theorem for the mean of a dataset where the data itself is clearly not drawn from a bell curve (remember: non-natural datasets tend to not be Gaussian). Demonstrate first using an appropriate visualization that the data itself is visually not well fit by a bell curve/Gaussian distribution (you don't need to actually fit a distribution).

5. Build a machine learning system to distinguish between three classes (can be sound/image/pose) using Google's teacheable machine. How good is it? Run a validatation dataset test here: Present new sounds/images 20 times and generate a confusion matrix. Assess how many times your model gets it right, how often does it get a false positive? Can you give a reason why it might get it wrong when/if it does?

6. Suppose we know that a court of justice is good enough that 90 % of the guilty suspects are properly judged while, of course, 10% of the guilty suspects are improperly found innocent. On the other hand, innocent suspects are misjudged 1

% of the time. If the suspect was selected from a group of suspects of which only 5 % have ever committed a crime, and the court indicates that he is guilty, what is the probability that he is innocent? Is the court any good at giving good judgements?

7. Identify a recent case in the news where the Campbell-Goodhart law was demonstrated – this could be recent developments in machine learning/AI or in the policy decisions made by the government. Explain why the case you identified fits the Campbell-Goodhart law.

8. For the following variables from the NHANES dataset, make a plot and explain in 1 line what probability distribution may be appropriate to use for it:
    a. Age
    b. Marital Status
    c. Height of people older than 19
    d. Number of babies

9. Develop a distribution on seeing theory such that the samples for random variables from it yield something that fits:
    a. Normal distribution.
    b. Poisson distribution.