**Part 1 – simulating correlations and performing QPRs**

1. Write a function “correlation\_sim”. The function gets as input the number of pairs (N). The function then:
   1. Draws one sample of N subjects, from a normal distribution with a mean of 100 and sd of 15 and saves it in a variable named x1.
   2. Draws a second sample of N subjects, from a normal distribution with a mean of 100 and sd of 50 and saves it in a variable named x2.
   3. Creates a variable named y, where y=x1+x2
   4. Returns x1 and y in a dataframe.
2. Calculate and examine the significance of the correlation between the two simulated variables (x1 and y) you have drawn with correlation\_sim. Use N=70 and a seed of 1 (set.seed(1) before you run – the same data should be received every time) to get identical results each time. Report Pearson’s r, degrees of freedom and p-value.
3. Divide the sample into four groups by the three quartiles of x1 (25, 50, 75%. You can use the “quantile” function). Samples below percentile 25% (including) will be rated as “a”, between 25%-50% will be rated as “b”, and so forth. Save the group in a new column in your data frame.
4. Conduct a One-way ANOVA on the groups created in question 3, with y as the DV. Report the results.
5. Plot a scatter plot (points) of x1 against y, and another scatter plot of the quartile groups against y.
6. Based on your results, how does splitting the data into groups affects the power of your study, compared to using the continuous measure?
7. Discuss - will this result generalize, in the sense we can always count on your answer to the previous question? Do you think it reflects a general trend in splitting a continuous variable into groups? Suggest a way to check this.
8. Change the function so it only returns x1, x2, without y. What would be the expected value of the correlation? Briefly explain why.
9. Sample the data with x1 and x2 (what you had at the end of question 8), with the same seed from before. Perform the following operations (those are QPRs! Do not do this on your real-life data) on the original data, and save the p-value from a 1-tailed correlation test each time (you can use the argument “alternative=’greater’ ” in the cor.test function):
   1. No changes are made
   2. Remove outlier pairs which are more than 1.5 SDs from the mean of x1
   3. Remove outlier pairs which are more than 1.5 SDs from the mean of x2
   4. Remove outlier pairs which are more than 2.5 SDs from the mean of x1
   5. Remove outlier pairs which are more than 2.5 SDs from the mean of x2
   6. Remove the pairs highest and lowest value from x1 (hint: subset for anything except the maximum & minimum value)
10. Which was the lowest p-value? What is the problem with choosing it as the p-value reflecting the result?
11. Repeat question 9 above for 2,000 iterations (without a constant seed), saving the lowest p-value each time. What are the chances that relying on the above QPRs will result in a type I error in your sample? Explain why briefly.

**Part 2**

1. Load the data “Expt1\_OverallRecognition.csv”. The data is already cleaned and ready for testing. The data is taken from the study:

Denis, D., Sanders, K. E., Kensinger, E. A., & Payne, J. D. (2022). Sleep preferentially consolidates negative aspects of human memory: Well-powered evidence from two large online experiments. *Proceedings of the National Academy of Sciences*, *119*(44), e2202657119.  
  
The study presented 2 groups of subjects with stimuli and tested their memory later. One group slept during this period (overnight), while the other stayed awake (during the day).

There were 3 factors serving as individual variables (type of stimuli: object/background, valence: negative/neutral, and group: sleep/wake). We will not analyze the background images (omitted from dataset) and address only the neutral/negative and sleep/wake factors.

Each subject has 2 rows – one for each stimulus type – and was assigned to one group (sleep/wake).

Relevant columns are:

- *delay\_cond* – whether the subject slept between the encoding and retrieval sessions

- *valence* – “neg” for negative, “neu” for neutral stimuli

- *id* - subject identity

- *value* – calculated memory score (the DV)

The hypothesis was that sleep would enhance memory for negative stimuli more than it would for neutral stimuli. This means that the difference between the sleep and wake group was hypothesized to be positive (higher memory for sleep group) and bigger for negative stimuli, compared to neutral stimuli.

1. Run the relevant ANOVA to test the hypothesis and report the results, including the relevant effect size for the effect.
2. What was the noise factor (the denominator in the F test, MS representing the error) used for each of the three calculated effects and why was it chosen?
3. Conduct the relevant planned contrasts to test for the specific hypothesized differences.
4. Assume there was no a-priori hypothesis, and the study was exploratory. Run (4) but with Tukey correction.
5. Discuss the differences between the results of the post hoc and planned analysis. What were the differences between the two kinds of tests? (Up to 50 words)
6. Plot an informative graph that shows the result.
7. Explain the results in simple terms.