Unless otherwise specified, assume all (p-value) thresholds to be 0.05, and all tests to be two-sided if that is an option. All calculations may be done with R or by hand unless otherwise specified; when a problem calls for something "by hand," feel free to use R as a calculator for basic operations like sums and means, and please show and explain your work as much as possible, using latex for displaying math whenever that would make it clearer.

Some hints for creating the simulated data:

- To created 100 observations where $y = 1 + 10x - \epsilon$ and x and ϵ have mean 0 and sd 1, you can do $x <- rnorm(100)$ and $y <- -1 + 10 * x + rnorm(100)$
- To create y as a function of 10 x variables, all normal with mean 0 and sd 1, where the coefficient on the first 5 variables is 1 and the second is 10 try **xmat <- matrix(rnorm(100*10),100,10)** and then use matrix multiplication to get y : **y <- xmat %*% c(rep(1, 5), rep(10, 5)) + rnorm(100)**.
- To create a factor that is a when$x < 0$ and b when $x >= 0$, a useful function is **cut** , eg. **myfactorvar <- cut(x, breaks = x(-Inf, 0, Inf), labels = c("a", "b"))** .You can also use a loop just like in the lesson.
- Finally, please use **set.seed(1)** somewhere at the start of your Rmarkdown so that your simulated data is consistent.

1. You roll five six-sided dice. Write a script in R to calculate the probability of getting between 15 and 20 (inclusive) as the total amount of your roll (ie, the sum when you add up what is showing on all five dice). Exact solutions are preferable but approximate solutions are ok as long as they are precise (10pts)
2. Create a simulated dataset of 100 observations, where x is a random normal variable with mean 0 and standard deviation 1, and $y = 0.1 + 2 * x + \epsilon$, where epsilon is also a random normal error with mean 0 and sd 1. (10pts)
    a. Perform a t test for whether the mean of Y equals the mean of X using R.
    b. Now perform this test by hand using just the first 5 observations. Please write out all your steps carefully.
    c. Assuming the mean and sd of the sample that you calculated from the first five observations would not change, what is the minimum total number of additional observations you would need to be able to conclude that the true mean μ of the population is different from 0 at the $p = 0.01$ confidence level?
3. Generate a new 100-observation dataset as before, except now $y = 0.1 + 0.2 * x + \epsilon$ (10pts)

a. Regress y on x using R, and report the results. Discuss the coefficient on x and its standard error, and present its 95% CI.

b. Use R to calculate the p-value on the coefficient on x from the t statistic for that coefficient as shown in the regression in 3a, and confirm that your p-value matches what is shown in 3a. What does this p-value represent (be very precise in your language here)?

c. Use R to calculate the p-value associated with the F statistic reported in your regression output. What does this test and its p-value indicate?

d. Using just the first five observations from your simulated dataset, calculate by hand the coefficient on x, its standard error, and the adjusted R2. Be sure to show your work, but you may use R for the simple math.

4. Now generate $y = 0.1 + 0.2 * x - 0.5 * x^2 + \epsilon$ with 100 observations(10pts)

   a.Regress y on x and x^2 and report the results. If x or x^2 are not statistically significant, suggest why.

   b.Based on the known coefficients that we used to create y, what is the exact effect on y of increasing x by 1 unit from 1 to 2?

   c.Based on the coefficients estimated from 4(a), what is the effect on y of changing x from -0.5 to -0.7?

5. now generate x_2 as a random normal variable with a mean of -1 and an sd of 1. create a new dataset where $y = 0.1 + 0.2 * x - 0.5 * x * x_2 + \epsilon$ and answer the following items. (20 pts)

   a.Based on the known coefficients, what is the exact effect of increasing x2 from 0 to 1 with x held at its mean?

   b.Regress y on x, x_2, and their interaction. Based on the regression-estimated coefficients, what is the effect on y of shifting x from -0.5 to -0.7 with x_2 held at 1?

   c.Regress y on x alone. Using the R2 from this regression and the R2 from 5(b), perform by hand an F test of the complete model (5b) against the reduced, bivariate model. What does this test tell you?

6. Generate a dataset with 300 observations and three variables: f, x1, and x2. f should be a factor with three levels, where level 1 corresponds to observations 1-100, level 2 to 101-200, and level 3 to 201-300. (Eg, f can be "a" for the first 100 observations, "b" for the second 100, and "c" for the third 100.) Create x1 such that the first 100 observations have a mean of 1 and sd of 2; the second 100 have a mean of0 and sd of 1; andt he third 100 have a mean of 1 and sd of 0.5. Create x2 such that the first 100 observations have a mean of 1 and sd of 2; the second 100 have a mean of 1 and sd of 1; and the third 100 have a mean of 0 and sd of 0.5. (Hint: It is probably easiest to create three 100-observation datasets first, and then stack them with rbind(). And make sure to convert f to a factor before proceeding.) (20pts)

   a.Using the k-means algorithm, perform a cluster analysis of these data using a k of 3 (use only x1 and x2 in your calculations; use f only to verify your results). Comparing your clusters with f, how many datapoints are correctly classified into the correct cluster? How similar are the centroids from your analysis to the true centers?

   b.Perform a factor analysis of this data using your preferred function. Using a scree plot and/or cumulative variance plot, how many factors do you think you should include? Speculate about how these results relate to those you got with the cluster analysis.

For the next questions use the Modified Massachussets Crimes dataset of 2019 available on the final canvas page ("mass_crimes_final.csv"). Modifying the dataframe object in R to perform your analysis correctly might be a part of the evaluation

7. Raise some hypothesis about the dataset, motivate it, filter the rows and columns (if needed), so that it can be tested using multiple regression. State all steps clearly and document your conclusions. (5pts).

8. Perform a WRONG regression using the dataset. A wrong regression is one that uses either inappropriate variables or other substantial errors, but that still results in a table and coefficients. Explain the results obtained and why they're not a proper application of the methods we learnt this semester? (5pts)

9. Perform either PCA or Clustering on the dataset. Present your results and conclusions into one or more paragraphs. (10pts)

10. Repeat the procedure chosen on question 8 but now transform the data to a new dataframe so that population is taken into account (*normally crime data is presented in X offenses by 100.000 habitants*). Show your results and compare them to the ones of question 9 with another paragraph (10pts Bonus)