MidTerm by **Kwasi Mensah**: I produced the material by myself, with help from the book and the professor.

1. (1 point) What are the lower bound and upper bounds of the (frequentist) 95% confidence interval of the mean difference?
    a. **The lower bound is -12.35187 and the upper bound is 14.30250**

2. (1 point) What is the point estimate of the mean difference?
    a. **The point estimate is derived by getting the difference between sample estimates: mean of x: 32.05410 and mean of y: 31.07879, so the point estimate of the mean difference may be about .97531.**

3. (1 point) Report the outcome of the null hypothesis significance test on the difference of means. Make sure to state the null hypothesis.
    a. **The t value is .14925, the degree of freedom is 31.048, and the p-value is .8823. When assuming there is an alpha threshold of .05, we would decide to fail to reject the null hypothesis, although our result show that the mean TDS in the treatment group will be lower than the mean TDS in the control group, which is the alternative hypothesis. This is due to the p-value being .8823, which is not less than or equal to .05. The 95% confidence interval is -12.35187 to 14.30250.**

4. (1 point) Report the lower and upper bounds of the 95% Highest Density Interval for the difference of means.
    a. **The boundary values calculated of the 95% Highest Density Interval or HDI ranges are from .0193 to 16.2. There is a 95% probability that the population mean difference between the two groups falls within the given range and has a mean value of 8.2.**

5. (1 point) Report the percentages of values in the posterior distribution of mean differences that are above zero and below zero.
    a. **97.2 % > 0 > 2.8%**

6. (5 points) Write a 1-2 paragraph technical report. The technical report should contain the detailed information that it would be important for other statisticians to know about the data, about the analytical results, about any anomalies you observed, and about how any such anomalies may have affected the reported results. You can cut and paste any of the graphics included above, as long as you provide a 2-3 sentence explanation of what the graphic means.

The testDF data frame has a total of 32 observations and 2 variables. Regarding the Control group the mean is 32.05, median =31.90, the minimum and maximum values are 30.27 & 34.53, and the 1st and 3rd quartiles are 31.25 & 32.81. The mean of the Treatment group is 31.079, the median value is 27.259, the 25th percentile of the Treatment group is 8.501 and the 75th percentile is 35.533. The minimum value of the Treatment group is 1.178 and its maximum value is 201.908, which represent an outlier in the data.

The null hypothesis test has t-value of .14925 and a p-value of .8823, which is the probability of the t-value will occur. Also due to the p-value not being less than or equal to .05 we will fail to reject the null hypothesis. The 95 % confidence interval are -12.35187 & 14.30250, the point estimate of the mean difference may be about .97531. The boundary values calculated of the 95% Highest Density Interval or HDI ranges are from .0193 to 16.2. There is a 95% probability that the population mean difference value is most likely 8.2, and the interval range does not cross over 0 so the analysis is statically significant.

7. 7. (5 points) Write a 1-2 paragraph report of the results of your analysis for presentation to the company's biologists and investors. This report should be in plain language, interpretable by non-statisticians. Make sure to integrate the Bayesian evidence, the frequentist confidence interval, and the results of the null hypothesis significance test. The biologists and investors need to decide what the startup should do next: The essential question they want to answer is whether or not the biofilm shows promise as an alternative to traditional filtering techniques. Use the results of these statistical analysis to provide them with guidance.

The summary function summarizes the values in a data frame, this shows us measures of dispersion, measures that describe the spread of data & measures of central tendencies, measures that show the approximate center of a distribution. The str function shows the structure of the data frame being examined. It shows us there are two groups being examined with 32 values each. The graph of the box plot shows that that the median TDS in the Control group is higher when compared to the Treatment group. The entire range of the Control group looks to overlap the Treatment group between the ranges of its median and the 3rd quartile. In addition, the minimum and maximum values in the Treatment group are significantly lower (for the minimum) and higher (for the maximum) values when compared to the Control group. Lastly it is shown that there is an outlier in the graph which is distinguished by the small transparent circle. The Graphs show that the Control group is more frequently higher in TDS levels compared to the treatment group. Treatment group data is more widely dispersed at the upper and lower levels compared to the Control group. Both the Control group and the Treatment1 histograms are more skewed to the left or positively skewed.

The 95 % confidence interval are -12.35187 & 14.30250, which tells us that is we do 100 replications we are 95 % sure that the population mean may have a chance of falling between the specified range and 5% sure that wouldn't fall within the range. The point estimate is derived by getting the difference between sample estimates: mean of x which is 32.05410 and mean of y which is 31.07879, so the point estimate of the mean difference may be about .97531. The confidence interval passes 0 which means that the results are not statically significant, and we will fail to reject the null hypothesis. Regarding the confidence interval its results show the long run possibilities, not about the accuracy of this confidence interval. Next, we ran an analysis using Bayes Theorem, which is the probability of an event, based on prior knowledge of conditions that might be related to the event. When we perform the Bayesian analysis shows that there is a 95% probability that the population mean difference between the two groups falls within the given "Highest Density Interval" or HDI ranges of .0193 to 16.2. The histogram shows that most likely the value of the mean is 8.2, although the results show that the analysis is statically significant, it is best to still run multiple iterations of the experiment but with larger sample sizes. The histogram also shows us that 97.2% of the mean differences in the distribution were positive (meaning that the biofilm treatment shows promise) and 2.8% were negative. This will help us have a stronger statistically significant results and help to reduce the size of the confidence interval ranges. As a result, we could conclude that the biofilm shows promise as an alternative to traditional filtering techniques.