The dataset ST314ExamData_DA8.csv represents the midterm and final exam grades for students in the ST314 online and campus courses, for two previous terms. Use this data to complete a multiple linear regression analysis in R and answer the following questions.

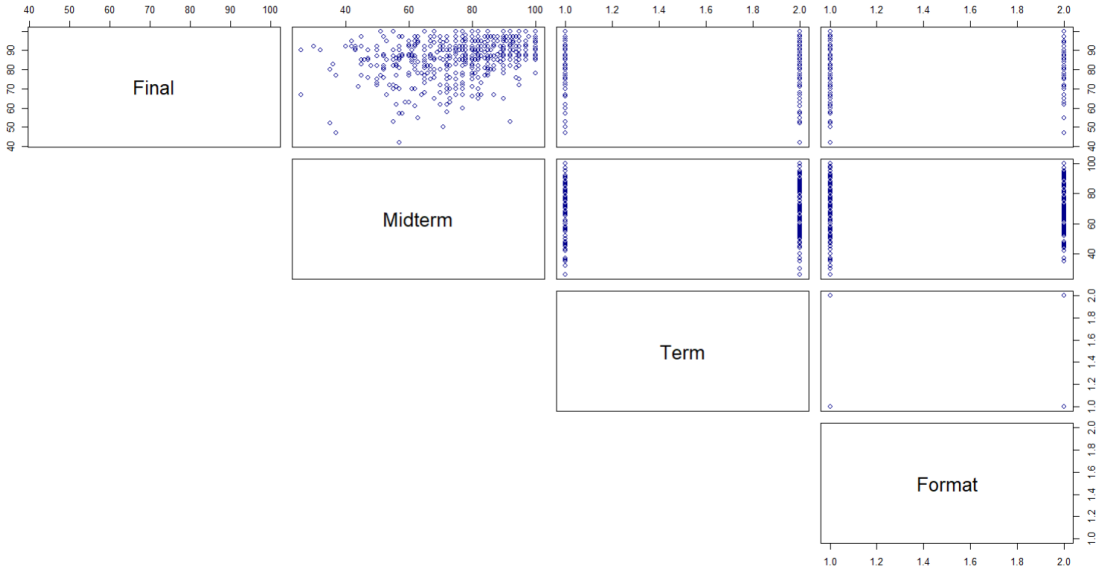
Note: the data set used in this assignment is different from last week's. Please make sure you are loading the csv file listed on the Data Analysis 8 Canvas page.

Final = Final Exam Score out of 100
Midterm = Midterm Exam Score out 100
Term = Has two levels Fall 2021 and Spring 2022
Format = Has two levels Campus and Online

Part 1. (6 points) Multivariate Visualization:

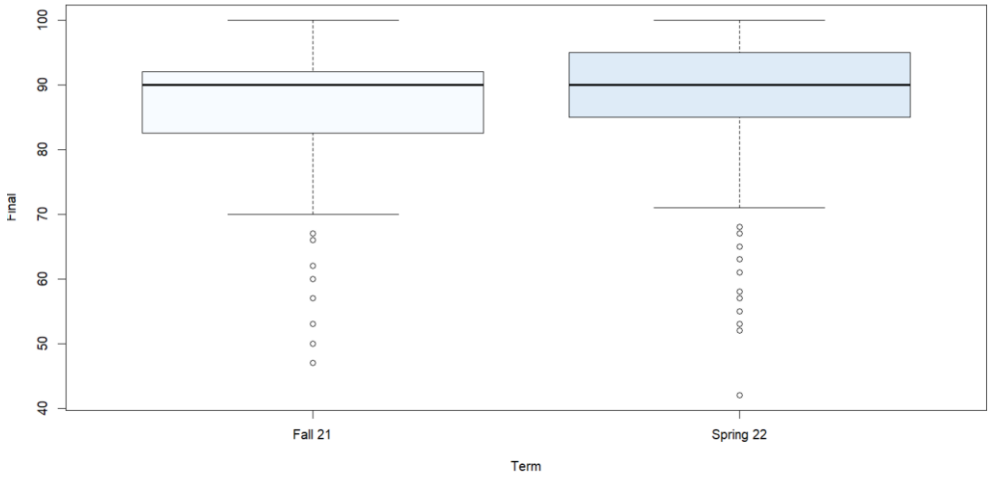
It is reasonable to consider that more than just midterm score may influence final exam score. Investigate the individual relationships between final exam score and the above explanatory variables. Use the R script Multivariate_Exam_Analysis.R to help you get started with the code.

a. Construct a scatterplot matrix including final and each of the explanatory variables.
   i. (1 point) Paste the plot.



   ii. (1 point) Do any of the variables have a visual relationship with Final?
   Midterm vs. Final: The plot shows a fairly strong positive correlation. As Midterm scores increase, Final scores seems to increase as well.

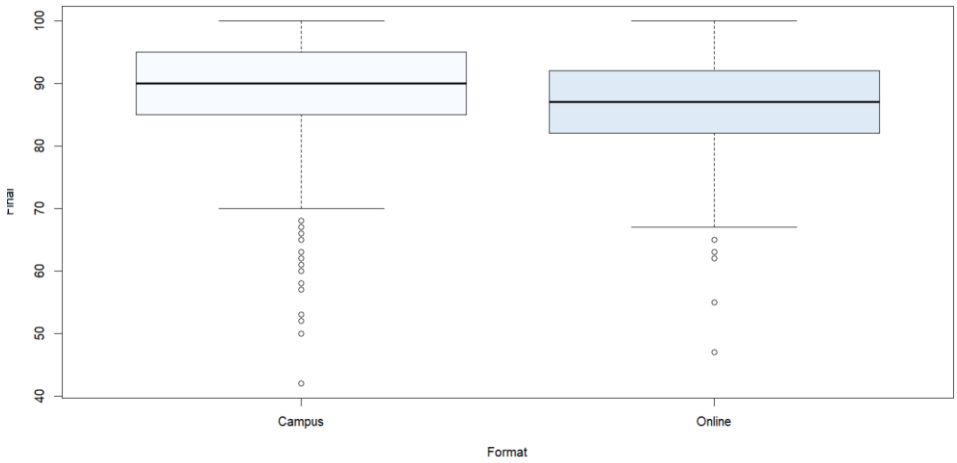
b. The scatterplot matrix is not all that helpful for the categorical variables Term and Format.
   i. Create a side by side boxplot that looks at the relationship between Term and Final.
      - (1 point) Paste your plot.



- (1 point) Describe the relationship. Visually does Term seem to have a relationship with Final?
  The overall distribution between both terms is fairly similar. However, Fall 21 shows more spread compared to Spring 22. Visually, term does not seem to have a relationship with the final score.

ii. Create a side by side boxplot that looks at the relationship between Format and Final.

- (1 point) Past your plot.



- (1 point) Describe the relationship. Visually does Format seem to have a relationship with Final?
  The distribution between different formats is pretty similar, with in Campus having a higher spread but also a higher median. Visually, format does not have a relationship with the final score.

Part 2. (7 points) Fit a Model

Fit a model that includes Term, Format and Midterm as explanatory variables for the response variable Final.

a. (1 point) Provide the R output of the model.

```
Call:
lm(formula = Final ~ Midterm + Term + Format)

Residuals:
    Min      1Q  Median      3Q     Max
-41.884  -4.032   1.806   5.700  17.442

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   70.0649     1.8253  38.386   <2e-16 ***
Midterm        0.2211     0.0219  10.094   <2e-16 ***
TermSpring 22  1.2171     0.6583   1.849   0.0649 .
FormatOnline  -1.6413     0.7507  -2.187   0.0291 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 717 degrees of freedom
Multiple R-squared:  0.1394,    Adjusted R-squared:  0.1358
F-statistic: 38.72 on 3 and 717 DF,  p-value: < 2.2e-16
```

b. (2 points) State the least squares regression equation of your model.
Final = 70.0649 + 0.2211 * Midterm + 1.217 * TermSpring22 − 1.6413 * FormatOnline

c. The model **without** the variables Term and Format has an adjusted R^2 value of 0.1274.

   i. (1 point) Does including the variables Term and Format improve the fit of the model?
   Since the adjusted R^2 value increases from 0.1274 to 0.1358 when Term and Format are included, this indicates that adding these variables improves the fit of the model.

   ii. (1 point) Interpret the adjusted R^2 value for the model that includes all three explanatory variables.
   Around 13% of the variability in the final exam scores can be explained by the model, which includes Midterm scores, TermSpring22, and FormatOnline. This provides a more accurate measure of the model fit.

Part 3. (10 Points) Model interpretation.

Note: Model Interpretation can get tricky when there is more than two levels in a factor. For example, Term has three levels instead of two. The R output will designate this as VariableLevel, like "TermSpring 22".

In the model, the coefficient for **TermSpring 22** is **1.2171** this means that ***while the other variables in the model are held constant, a student taking the exam in the spring 2022 will score 1.2171 points more on average than a fall 2021 student.*** We know to TermSpring 22 is compared to fall, because Fall 2021 is the variable not included in the output. Meaning, fall is represented when spring is at 0.

a. (2 point) Interpret each of the individual t tests by stating which variables are significant at 0.05, when the other variables are in the model.
Midterm: The p-value is <2e-16, which is much less than 0.05, so Midterm is significant.

TermSpring22: The p-value is 0.0649, which is greater than 0.05, so TermSpring22 is not significant.
FormatOnline: The p-value is 0.0291, which is less than 0.05, so FormatOnline is significant.

b. (2 points) Interpret in context $\beta_{FormatOnline}$ the coefficient for Format.
The coefficient for FormatOnline is -1.6413. This means that students taking the exam in an online format score on average 1.6413 points lower than students taking the exam in a non-online format. This is statistically significant.

c. (2 points) Interpret in context $\beta_{TermSpring22}$ the coefficient for TermSpring22.
The coefficient for TermSpring22 is 1.2171. This means that a student taking the exam in Spring 2022 scores on average 1.2171 points higher than a student taking the exam in Fall 2021. This is not statistically significant.

d. (2 point) Interpret in context $\beta_{midterm}$ the coefficient for the Midterm variable.
The coefficient for Midterm is 0.2211. This means that for every one-point increase in the midterm exam score, the final exam score is expected to increase by 0.2211 points on average. This is statistically significant.

e. (2 points) Calculate the 95% confidence interval for $\beta_{Midterm}$. Show work. Interpret the interval.

$$\text{CI} = 0.2211 \pm 1.96 \cdot 0.0219$$

Where 0.2211 is the estimated coefficient, 1.96 is the critical value for t-distribution for 95% confidence, and 0.0219 is the standard error of the coefficient.

$$0.2211 \pm 0.042924$$
$$(0.178176, 0.264024)$$

Part 4. (2 points) Prediction.

a. (2 points) Use the least squares regression equation to predict final exam score for a fall, online student with a midterm score of 85.
Using our equation before, we can plug in 85 for the midterm, 0 for TermSpring22, and 1 for FormatOnline.

$$\text{Final} = 70.0649 + 0.2211 \cdot 85 + 1.217 \cdot 0 - 1.6413 \cdot 1$$

Final = 87.2171

Gradescope Page Matching (2 points)
When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page "Submitting Assignments in Gradescope". Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item "Gradescope Page Matching".