ST 314 - Data Analysis 9

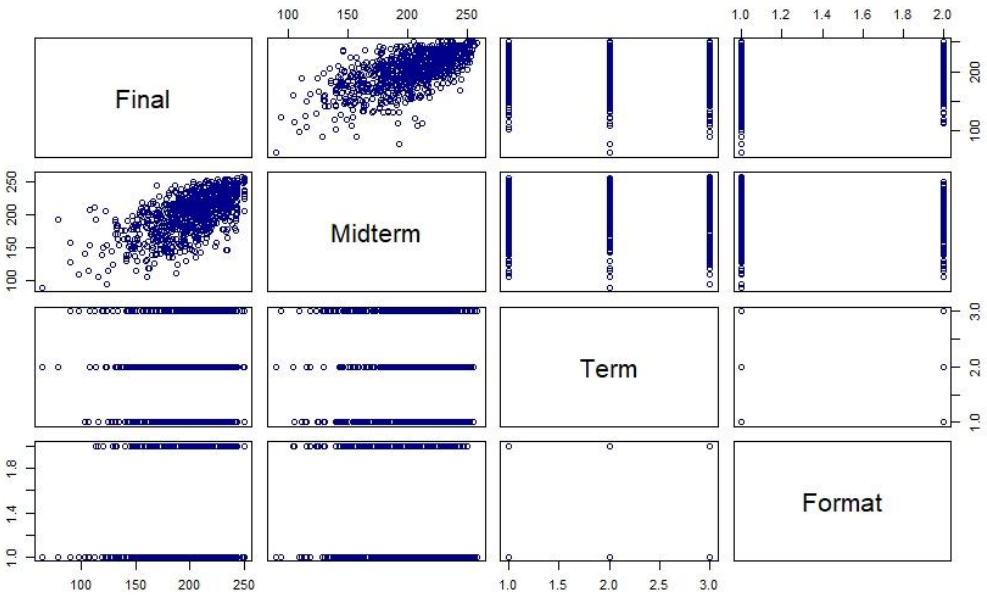
Ivan Timothy Halim

March 12, 2019

Part 1

a.

a.

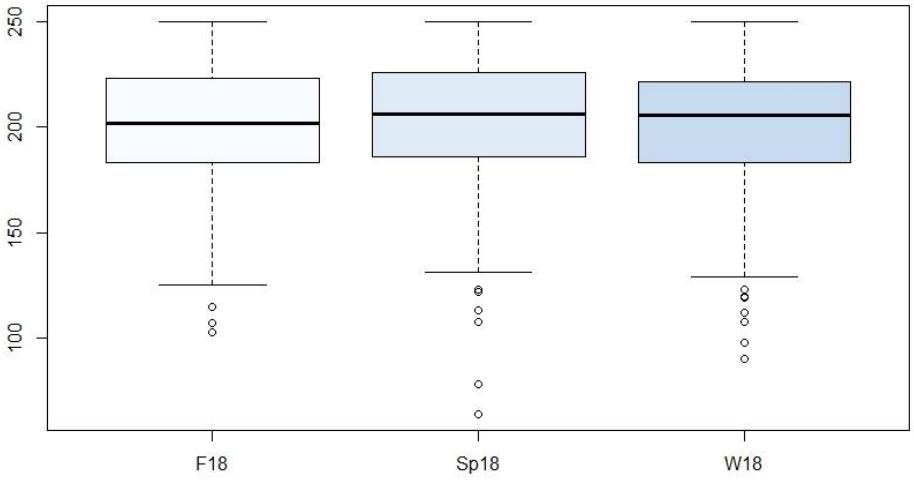


b. There seems to be a moderately strong, linear, positive relationship between Midterm scores and Final scores.

b.

a.

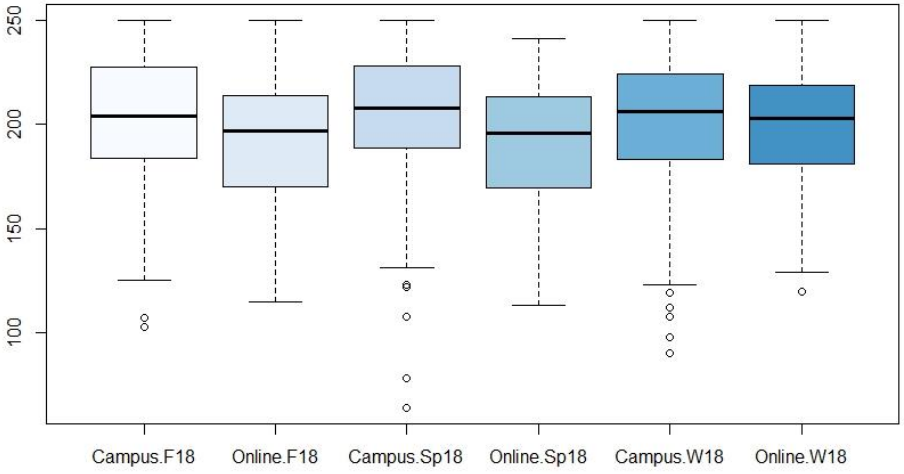
i.



ii. The boxplots have similar medians, interquartile range and upper-lower bound. The difference is in the outliers with Spring 2018 Term having the lowest outlier (about 50-60).

b.

i.



ii. The Campus final scores are slightly higher than the Online final scores, but the former has more outliers.

Part 2

a.

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

Residuals:
     Min       1Q   Median       3Q      Max
-114.779  -13.762    1.835   15.916   72.213

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  76.71561    5.44105  14.099  < 2e-16 ***
Midterm       0.63814    0.02649  24.086  < 2e-16 ***
FormatOnline -2.93799    2.02121  -1.454 0.146433
TermSp18     -7.09694    1.97595  -3.592 0.000348 ***
TermW18      -2.44671    2.17460  -1.125 0.260852
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.52 on 847 degrees of freedom
Multiple R-squared:  0.4155,    Adjusted R-squared:  0.4128
F-statistic: 150.6 on 4 and 847 DF,  p-value: < 2.2e-16
```

b. Model: $\hat{y} = 76.716 + 0.638x_1 - 2.938x_2 - 7.097x_3 - 2.447x_4$

where $\hat{y} =$ Final score

$x_1 =$ Midterm score

$x_2 =$ Format Online

$x_3 =$ Spring 2018 Term

$x_4 =$ Winter 2018 Term

c.

a. The fit of the model improved very slightly (the adjusted R-squared increases by 0.0086). However, we should be careful of artificially inflating the R^2 by adding explanatory variables that are not helpful predictors.

b. While accounting for sample size and the number of explanatory variables in the model, the explanatory variables in the model explain 41.28% of the variation in the final scores. The model is a poor fit.

Part 3

a. When the other variables are in the model,
   - Midterm score is significant with a t test stat = 24.086 and p-value $< 2 \times 10^{-16}$.
   - Format Online is not significant with a t test stat = -1.454 and p-value 0.1464.
   - Spring 2018 Term is significant with a t test stat = -3.592 and p-value 0.0003.
   - Winter 2018 Term is not significant with a t test stat = -1.125 and p-value 0.2608.

b. While all other variables are held constant, the Online final score is going to be 2.938 lower than the Campus final score.

c. While all other variables are held constant, the Spring 2018 Term final score is going to be 7.097 lower than the Fall 2018 Term final score.

d. While all other variables are held constant, for a one unit increase in the midterm score, the final score will increase by 0.638.

e. $95\%\ CI = 0.638 \pm 1.96(0.0265)$

$= (0.5861, 0.6901)$

The LSRL estimates that, while all other variables are held constant, for every one-point increase in midterm score, the final score will increase by 0.638 with a 95% confidence interval between 0.5861 and 0.6901.

Part 4

a. The final score for a winter, online student with a midterm score of 190 is,

$\hat{y} = 76.716 + 0.638(190) - 2.938(1) - 7.097(0) - 2.447(1)$

$= 193$

Extra credit

My midterm score is 218, so my final score would be,

$\hat{y} = 76.716 + 0.638(218) - 2.938(0) - 7.097(0) - 2.447(1)$

$= 213$