Business Analytics (CGMAO2003E) - Written sit-in (UC)

Copenhagen Business School

Department of Operations Management

Name of study: MSc EBA in General Management and Analytics

Course name: Business Analytics

Exam type: Ordinary Exam

Examination type: Written sit-in — 4 hours

Contents

Section 1 2
Section 2 3
Section 3 6
Section 4 7
Section 5 11
Section 6 12

Section 1

Question for Word file 1.1. Which variable(s) are wrongly formatted after importing the data into RStudio? Why?

```
> str(df)
spc_tbl_ [1,850 × 8] (S3: spec_tbl_df/tbl_df/tbl/data.fram
e)
 $ course_id       : chr [1:1850] "0001" "0002" "0003" "000
4" ...
 $ course_type     : chr [1:1850] "classroom" "classroom"
"online" "online" ...
 $ year            : num [1:1850] 2018 2020 2016 2013 2019
...
 $ enrollment_count: num [1:1850] 165 175 257 251 165 246 2
41 261 261 180 ...
 $ pre_score       : chr [1:1850] "28.14" "79.68" "57.24"
"97.67" ...
 $ post_score      : num [1:1850] 73 86 80 75 76 76 84 88 8
7 79 ...
 $ pre_requirement : chr [1:1850] "Beginner" "None" NA "Beg
inner" ...
 $ department      : chr [1:1850] "Science" "Science" "Math
ematics" "Technology" ...
 - attr(*, "spec")=
  .. cols(
  ..   course_id = col_character(),
  ..   course_type = col_character(),
  ..   year = col_double(),
  ..   enrollment_count = col_double(),
  ..   pre_score = col_character(),
  ..   post_score = col_double(),
  ..   pre_requirement = col_character(),
  ..   department = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

Pre_score is wrongly formatted, as it should be numerical, but it is set to character. It should be numerical because it is an actual numeric value tied to the variable, and not categorized into different categories for different scores.

Question for Word file 1.2. Copy the full console output of the data frame structure and paste it into your Word document.

```
> str(df)
tibble [1,850 × 8] (S3: tbl_df/tbl/data.frame)
 $ course_id       : chr [1:1850] "0001" "0002" "0003" "0004" ...
 $ course_type     : Factor w/ 2 levels "classroom","online": 1 1 2 2 1 2 2 2 2 1 ...
 $ year            : num [1:1850] 2018 2020 2016 2013 2019 ...
 $ enrollment_count: num [1:1850] 165 175 257 251 165 246 241 261 261 180 ...
 $ pre_score       : num [1:1850] 28.1 79.7 57.2 97.7 39.8 ...
 $ post_score      : num [1:1850] 73 86 80 75 76 76 84 88 87 79 ...
 $ pre_requirement : Factor w/ 3 levels "Beginner","Intermediate",..: 1 3 3 1 1 1 1 3 3 1 ...
 $ department      : Factor w/ 4 levels "Engineering",..: 3 3 2 4 3 2 3 4 4 1 ...
```

Section 2

Question for Word file 2.1. What is the percentage of online vs. classroom courses?

```
classroom    online
      475      1375
> 1375/(475+1375)
[1] 0.7432432
```

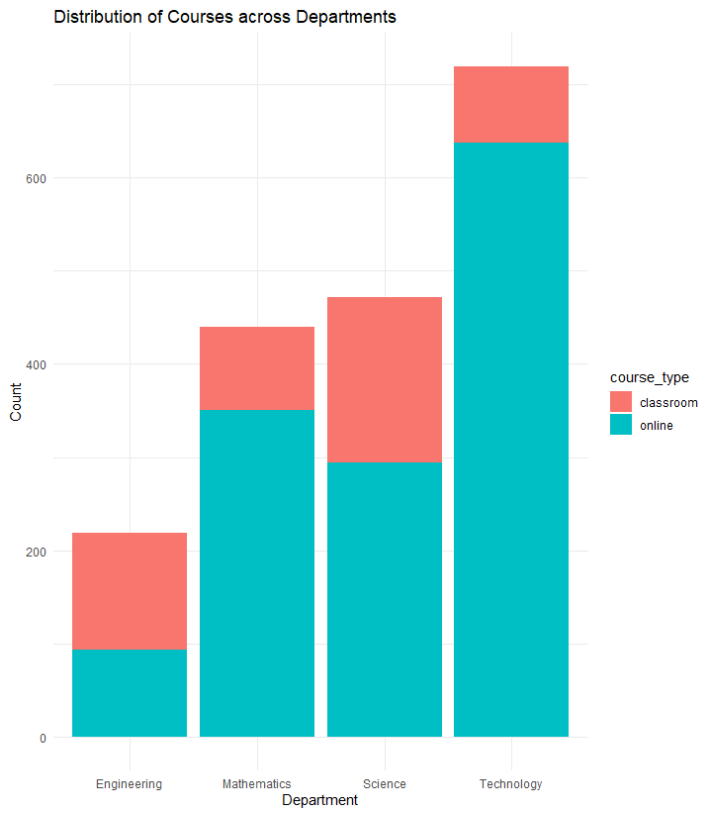
The percentage of online vs. classroom courses is 74.32%.

Question for Word file 2.2. Copy the chart into your report and answer the following:

- Inspect the graph. Which department offers the most courses?
- Inspect the graph. Which department has the lowest proportion of online courses?

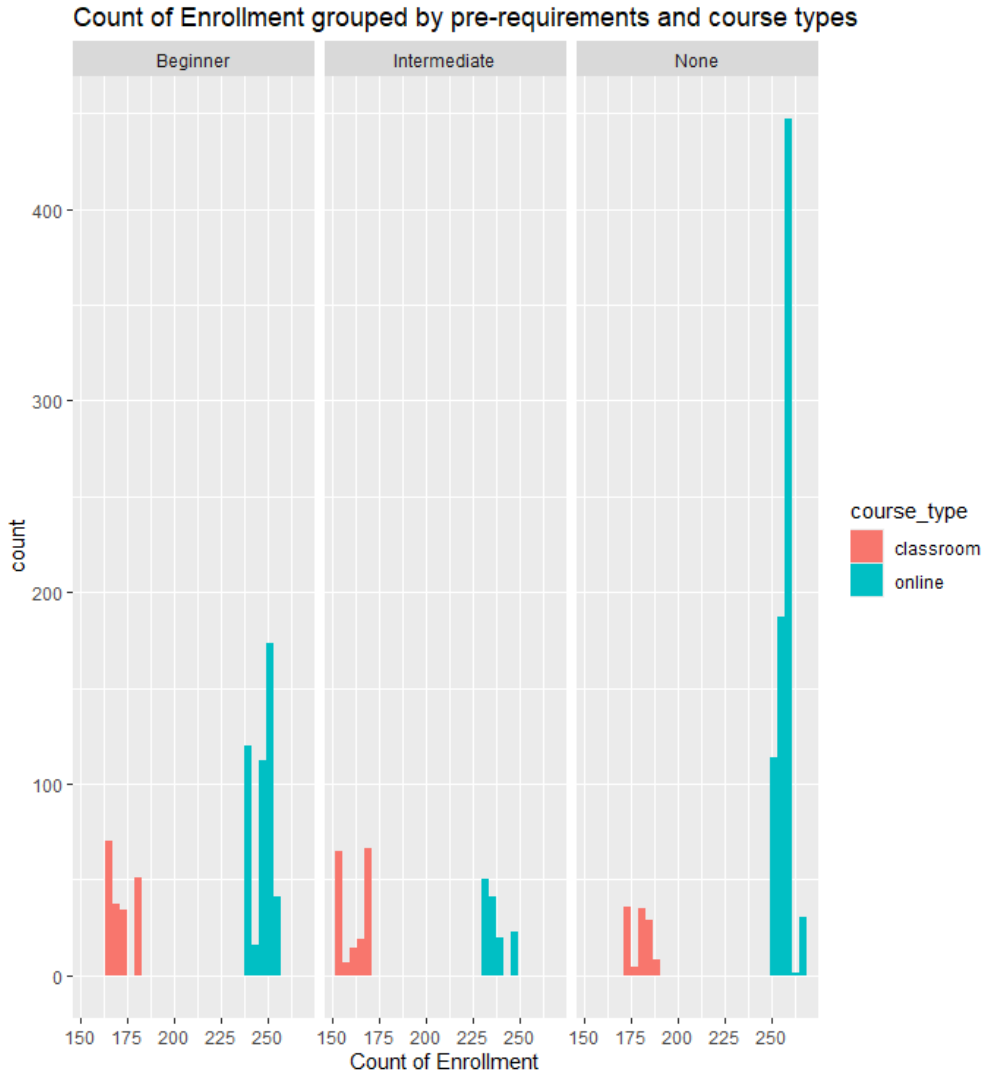
The department that offers the most courses is technology with a little over 700 courses, where most of them are online.

The department that offers the lowest proportion of online courses is Engineering, with the amount being ~90.



Question for Word file 2.3. Copy the chart into your report and answer the following:

- Within each pre-requirement level (None, Beginner, Intermediate), compare the `enrollment_count` across course type?



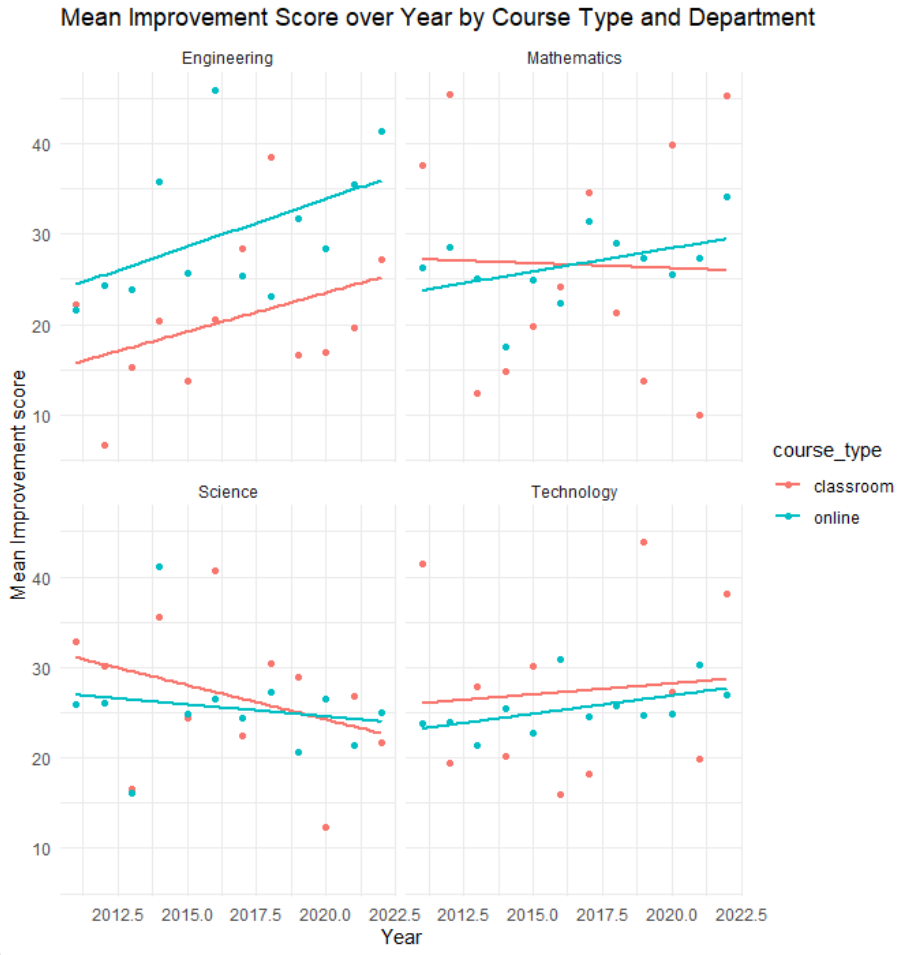
With courses that have no pre-requirements (None), we can see that it has the highest count of enrollment, with the enrollment count between ~250 – 265 happens of a frequency of ~450 times. The highest frequency in the "None" pre-requirement group, is online courses.

The Beginner pre-requirement also has some frequencies, with online courses with count of enrollment as ~235 – 255 with frequencies at ~175.

The Intermediate pre-requirement has the lowest amount of frequencies distributed across counts of enrollment, which could make sense since it has the most difficult pre-requirement. Here the frequency sits at ~60, with the counts of enrollment ranging from 150-~175 for classroom courses and ~225 - ~250 for online courses.

Question for Word file 2.4. Copy the chart into your report and answer the following:

- In which department is the difference in `mean_improvement` between course type (online vs. classroom) most pronounced?
- Which departments experience a negative trend in `mean_improvement` over time?



From the plot, with added regression line, it shows that Engineering has a positive trend regarding mean improvement scores over the years 2012.5 – 2022.5. This is a very positive thing, since it clearly shows that over time the students improve. It also shows that students enrolled in Engineering online courses have the highest mean improvement score.

The departments that suffer a negative trend are Science and Mathematics. For Science, there is a negative trend line for both classroom courses and online, with classroom courses showing the most negative trendline over the years. For Mathematics, the negative trendline only applies to the classroom course type, but it has to be said that it shows not a severe negative trend, but a negative trend nevertheless.

Section 3

Question for Word file 3.1. Copy your stated hypotheses and the console output of the paired t-test into your report. Then answer the following:

- Discuss the p-value.
- What does the `mean of the differences` value represent in the output?
- Why is a paired t-test appropriate on this specific occasion?

$$H_0: \mu_{post} = \mu_{pre}$$

$$H_1: \mu_{post} \neq \mu_{pre}$$

```
        Paired t-test

data:  df$post_score and df$pre_score
t = 45.187, df = 1849, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 24.58490 26.81587
sample estimates:
mean difference
       25.70038
```

From the paired t-test, the p-value < 0.05, t = 45.187 and df = 1849.

95% confidence interval: $[24.5849\ ;\ 26.81587]$

Since the p-value is below 0.05, we can reject the null hypothesis and conclude that there is a statistically significant difference between the group means. The p-value represents the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true. In this context, it tells us how likely it would be to see the observed difference in average scores between post scores and pre scores if there were no difference.

The t-value tells us that the observed difference in means is 45,187 standard errors away from zero. A larger absolute t-value means more evidence against the null.

The 95% confidence interval does not contain 0, which supports the conclusion of significant difference. It means we are 95% confident that the true difference in means lies between 24.5849 and 56.81587.

The difference in mean is 25.70038, indicating that the difference in means is not zero, which means there is statistically significant difference in the average scores for post_score and pre_score segments.

The paired t-test is appropriate on this specific occation because we want to compare before/after or matched values from same subjects.

Section 4

Question for Word file 4.1. Copy the residual plot and the summary output of both models into your report and answer the following:

- Which model better explains the variance in post_score?
- In your full model, identify the significant predictor(s). Interpret the sign and effect size.
- Briefly comment on the residual diagnostics. Do the residuals look roughly randomly distributed, or is there a visible pattern that would concern you?

```
Call:
lm(formula = post_score ~ course_type + year + enrollment_count +
    pre_score + pre_requirement + department, data = df_regression)

Residuals:
     Min       1Q   Median       3Q      Max
-20.8570  -4.4729  -0.2554   4.4684  26.6398

Coefficients:
                           Estimate Std. Error t value
(Intercept)                -9.630392 131.985574  -0.073
course_typeonline          15.050036  37.043834   0.406
year                        0.056411   0.046203   1.221
enrollment_count           -0.202565   0.487258  -0.416
pre_score                   0.154934   0.005716  27.107
pre_requirementIntermediate -2.831187   4.895299  -0.578
pre_requirementNone         2.502011   4.894500   0.511
departmentMathematics      -2.110426   4.860324  -0.434
departmentScience          -2.963531   7.329386  -0.404
departmentTechnology       -0.924732   2.533725  -0.365
                          Pr(>|t|)
(Intercept)                  0.942
course_typeonline            0.685
year                         0.222
enrollment_count             0.678
pre_score                   <2e-16 ***
pre_requirementIntermediate  0.563
pre_requirementNone          0.609
departmentMathematics        0.664
departmentScience            0.686
departmentTechnology         0.715
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.824 on 1840 degrees of freedom
Multiple R-squared:  0.2894,    Adjusted R-squared:  0.2859
F-statistic: 83.26 on 9 and 1840 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = post_score ~ pre_score * course_type * pre_requirement,
    data = df_regression)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1525  -4.5375  -0.2468   4.4425  26.5049

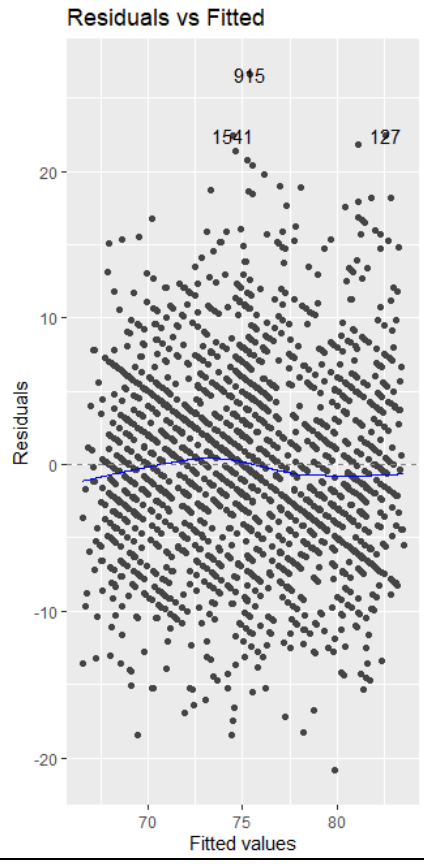
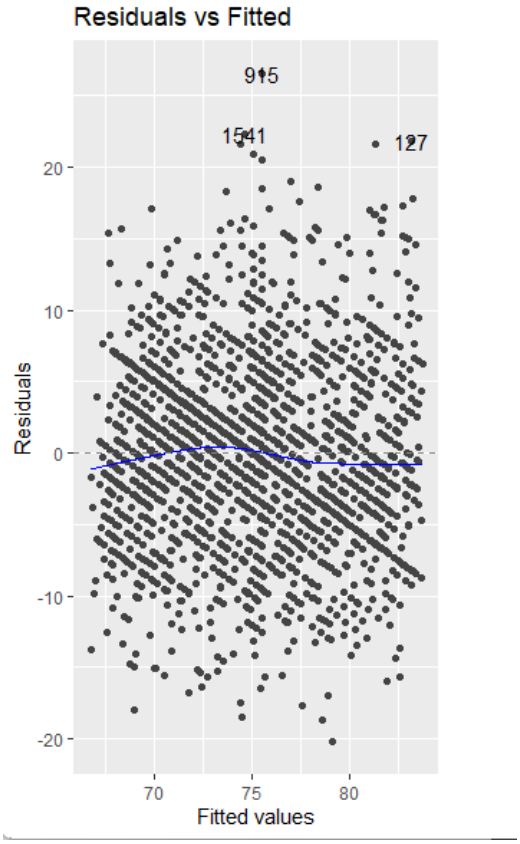
Coefficients:
                                                          Estimate
(Intercept)                                               67.93987
pre_score                                                  0.14053
course_typeonline                                         -0.88438
pre_requirementIntermediate                                0.01878
pre_requirementNone                                        1.17981
pre_score:course_typeonline                                0.02393
pre_score:pre_requirementIntermediate                      0.01223
pre_score:pre_requirementNone                             -0.01952
course_typeonline:pre_requirementIntermediate             -0.49218
course_typeonline:pre_requirementNone                     -0.65439
pre_score:course_typeonline:pre_requirementIntermediate -0.05086
pre_score:course_typeonline:pre_requirementNone            0.01687
                                                          Std. Error
(Intercept)                                                  1.02031
pre_score                                                    0.01800
course_typeonline                                            1.20253
pre_requirementIntermediate                                  1.48955
pre_requirementNone                                          1.67286
pre_score:course_typeonline                                  0.02126
pre_score:pre_requirementIntermediate                        0.02599
pre_score:pre_requirementNone                                0.02915
course_typeonline:pre_requirementIntermediate                1.97912
course_typeonline:pre_requirementNone                        1.85878
pre_score:course_typeonline:pre_requirementIntermediate      0.03532
pre_score:course_typeonline:pre_requirementNone              0.03248
                                                          t value
(Intercept)                                                66.587
pre_score                                                   7.805
course_typeonline                                          -0.735
pre_requirementIntermediate                                 0.013
pre_requirementNone                                         0.705
pre_score:course_typeonline                                 1.125
pre_score:pre_requirementIntermediate                       0.470
pre_score:pre_requirementNone                              -0.670
course_typeonline:pre_requirementIntermediate              -0.249
course_typeonline:pre_requirementNone                      -0.352
pre_score:course_typeonline:pre_requirementIntermediate    -1.440
pre_score:course_typeonline:pre_requirementNone             0.520
```

```
                                                          Pr(>|t|)
(Intercept)                                               < 2e-16
pre_score                                                 9.91e-15
course_typeonline                                            0.462
pre_requirementIntermediate                                  0.990
pre_requirementNone                                          0.481
pre_score:course_typeonline                                  0.261
pre_score:pre_requirementIntermediate                        0.638
pre_score:pre_requirementNone                                0.503
course_typeonline:pre_requirementIntermediate                0.804
course_typeonline:pre_requirementNone                        0.725
pre_score:course_typeonline:pre_requirementIntermediate      0.150
pre_score:course_typeonline:pre_requirementNone              0.603

(Intercept)                                               ***
pre_score                                                 ***
course_typeonline
pre_requirementIntermediate
pre_requirementNone
pre_score:course_typeonline
pre_score:pre_requirementIntermediate
pre_score:pre_requirementNone
course_typeonline:pre_requirementIntermediate
course_typeonline:pre_requirementNone
pre_score:course_typeonline:pre_requirementIntermediate
pre_score:course_typeonline:pre_requirementNone
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.799 on 1838 degrees of freedom
Multiple R-squared:  0.2952,    Adjusted R-squared:  0.291
F-statistic: 69.99 on 11 and 1838 DF,  p-value: < 2.2e-16
```

```
> full_model %>%
+   glance()
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic   p.value    df logLik    AIC    BIC deviance df.residual  nobs
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>  <dbl>  <dbl>  <dbl>    <dbl>       <int> <int>
1     0.289         0.286  6.82      83.3 8.52e-130     9 -6173. 12368. 12428.   85678.        1840  1850
> interaction_model %>%
+   glance()
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic   p.value    df logLik    AIC    BIC deviance df.residual  nobs
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>  <dbl>  <dbl>  <dbl>    <dbl>       <int> <int>
1     0.295         0.291  6.80      70.0 4.06e-131    11 -6165. 12356. 12428.   84977.        1838  1850
>
```

Full_model	Interaction_Model
	

The residuals vs Fitted plot checks linearity and homoscedasticity, or constant variance.

From the plot we can see that both fitted models look very similar, both the favors the interaction model slightly. The residuals look roughly randomly distributed, but have the most observations between y = -10 and 10.

The adj.r.squared for the full model equals 0.286 and the interaction model equals 0.291, indicating the interaction model serves better. But it is noted that the adj.r.squared is not that high, indicating that the models does not explain the post scores very well, but only 28.6% and 29.1%.

We can see that both models have F > 60, indicating that it is good models, and the full model having F = 83-26 > interaction model F = 69.99.

Full model

The only significant predictor is the pre_score variable, having an extremely low p-value. Beta is ~0.15, meaning higher pre scores increase post scores.

Section 5

Question for Word file 5.1. Copy the logistic regression confusion matrix into your report and answer the following:

- Briefly comment on the model's predictive performance.
- Discuss what the asymmetry between specificity and sensitivity means in the context of the university's problem. Which type of courses is the model good at identifying, and which type does it frequently miss?
- Briefly suggest one change to further improve the model.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  239  61
       Yes  29  40

               Accuracy : 0.7561
                 95% CI : (0.709, 0.7991)
    No Information Rate : 0.7263
    P-Value [Acc > NIR] : 0.109199

                  Kappa : 0.3194

 Mcnemar's Test P-Value : 0.001084

            Sensitivity : 0.8918
            Specificity : 0.3960
         Pos Pred Value : 0.7967
         Neg Pred Value : 0.5797
             Prevalence : 0.7263
         Detection Rate : 0.6477
   Detection Prevalence : 0.8130
      Balanced Accuracy : 0.6439

       'Positive' Class : No
```

- The model gives good accuracy of 75.61%, indicating that it classifies a good majority of observations correctly.
- Sensitivity = 0.8918: The model detects nearly all actual positives, minimizing false negatives.
- Specificity = 0.3960: The model correctly identifies ~40% of the actual negatives, but has a majority of false positives.
- Precision (pos pred value) = 0.7967: When the model predicts a high performer, it is correct ~80% of the time.
- Kappa = 0.3194: Indicates a good agreement beyond chance
- Balanced Accuracy = 0.6439: Useful when class distribution is imbalanced, this value is over 0.5 but could be higher. This is due to specificity.

The asymmetry in sensitivity and specificity means that the model does not always predict correctly the high performers, which could lead to misplacement in categories when the university looks at the data.

The model performs decent, especially in detecting high performers. To improve the model, it could maybe include gender-based insights to get a better overview of who the actual high-performers are, and could also include age, demographics etc. However, ethical considerations around different genders and demographics should be noted.

Section 6

Question for Word file 6.1. Copy the random forest confusion matrix and the variable importance plot into your report, then answer the following:

- Briefly comment on the model's predictive performance.
- Compare the random forest with the logistic regression. Briefly explain why a more flexible model (random forest) does or does not outperform a simpler one (logistic regression) on this dataset.
- Identify the two most important predictors of `high_performer` using the variance importance plot.

Confusion Matrix	Variable Importance plot

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  236  66
       Yes  32  35

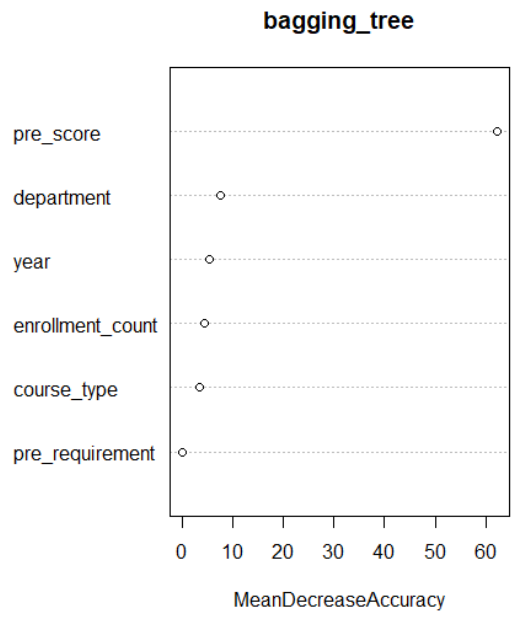
               Accuracy : 0.7344
                 95% CI : (0.6862, 0.7788)
    No Information Rate : 0.7263
    P-Value [Acc > NIR] : 0.3883031

                  Kappa : 0.2537

 Mcnemar's Test P-Value : 0.0008576

            Sensitivity : 0.8806
            Specificity : 0.3465
         Pos Pred Value : 0.7815
         Neg Pred Value : 0.5224
             Prevalence : 0.7263
         Detection Rate : 0.6396
   Detection Prevalence : 0.8184
      Balanced Accuracy : 0.6136

       'Positive' Class : No
```



Interpreting result:

Confusion Matrix:

- Accuracy: 0.7344 → Model correctly predicted ~73% of outcomes
- Kappa: 0.2537 → Model moderate agreement beyond chance
- Sensitivity: 0.8806 → Model correctly identified 88% of high performers
- Specificity: 0.3465 → ~35% of non-high performers were correctly identified
- Balanced Accuracy: 0.6136 → Moderate performance across both classes

The logistic regression has higher accuracy: 0.7561 > 0.7344, but not by a lot. Furthermore, the kappa is also higher for the logistic regression model, giving kappa of 0.3194 compared to random forest kappa of 0.2537. This is by quite some margin. The sensitivity and specificity for the logistic regression model are also higher, with values of 0.8918 > 0.8806 and 0.3960 > 0.3465, respectively. The balanced accuracy is therefore also higher for the logistic regression model.

Overall, the logistic regression model has better indicators, ultimately leading to being the overall better model in this case.

Variable Importance Plot

Higher value equals more impact on prediction accuracy. In the plot, we can see that pre_score and department, respectively, have the highest values (~60 and ~10) and therefore are the most important predictors of high performers.