# ETC1010-5510 Data Analysis Final Assessment

Siming Wu

In addition to the marks displayed in each question, an additional **15 marks** have been allocated for assessment of general coding style. **Please ensure that the report knits properly in HTML and all the R codes and outputs are visible in the final knitted report.**

## Undergraduate student: ETC1010 (85 marks)

### Answers Section A, B, C and D.

## Postgraduate student: ETC5510 (100 marks)

### Answers all questions. Section A, B, C, D and E.

```
# Libraries required for the analysis (you can add more if you want to)


library(dplyr)
library(ggplot2)
library(tidyverse)
library(knitr)

library(naniar)

library(rpart)
library(rpart.plot)
library(ggResidpanel)
```

### Section A: Data Wrangling (31 marks)

We are going to analyse student performance in Australia in the subjects of Mathematics, Science and Reading. You can find the data sets for this assessment in the folder called *data*.

The description of the variables are:

- **state**: state of Australia
- **schtype**: type of school: goverment (Gov), independent (Ind), or catholic (Catholic)
- **yr**: student's year level: year 10 (Y10), not year 10 (noY10)
- **outhours_study**: out-of-school study time per week (sum)
- **science_time**: science learning time per week (in minutes)
- **anxtest**: personal anxiety test (standardized score)

- **wealth**: family wealth (standardized score)

- **math**: mathematics score (0- 900)

- **science**: science score (0- 900)

- **read**: reading score (0- 900)

```
pisa<-read_csv("data/pisa2.csv")
```

**1. Read in the PISA2 data set ("pisa2.csv") and store it in a data object called `pisa`. Show the first 5 rows of the data frame. Take a closer look at the variable names and data type to better understand the data by using a function of your choice. [3m]**

```
## Rows: 14530 Columns: 10

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): state, schtype, yr
## dbl (7): outhours_study, science_time, anxtest, wealth, math, science, read

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
pisa[1:5,]
```

```
## # A tibble: 5 x 10
##   state schtype yr    outhours_study science_time anxtest  wealth  math science
##   <chr> <chr>   <chr>          <dbl>        <dbl>   <dbl>   <dbl> <dbl>   <dbl>
## 1 QLD   Gov     Y10               19          210  -0.152  0.0592  546.    590.
## 2 QLD   Gov     Y10                7          165   0.259  0.760   512.    557.
## 3 QLD   Gov     Y10               NA          210   2.55  -0.122   479.    569.
## 4 QLD   Gov     Y10               23          210   0.256  0.931   506.    529.
## 5 QLD   Gov     Y10                4          210   0.452  0.790   482.    504.
## # ... with 1 more variable: read <dbl>
```

```
pisa%>%select(c(state, schtype,yr,read))%>%group_by(state)%>%summarise(mean=mean(read))
```

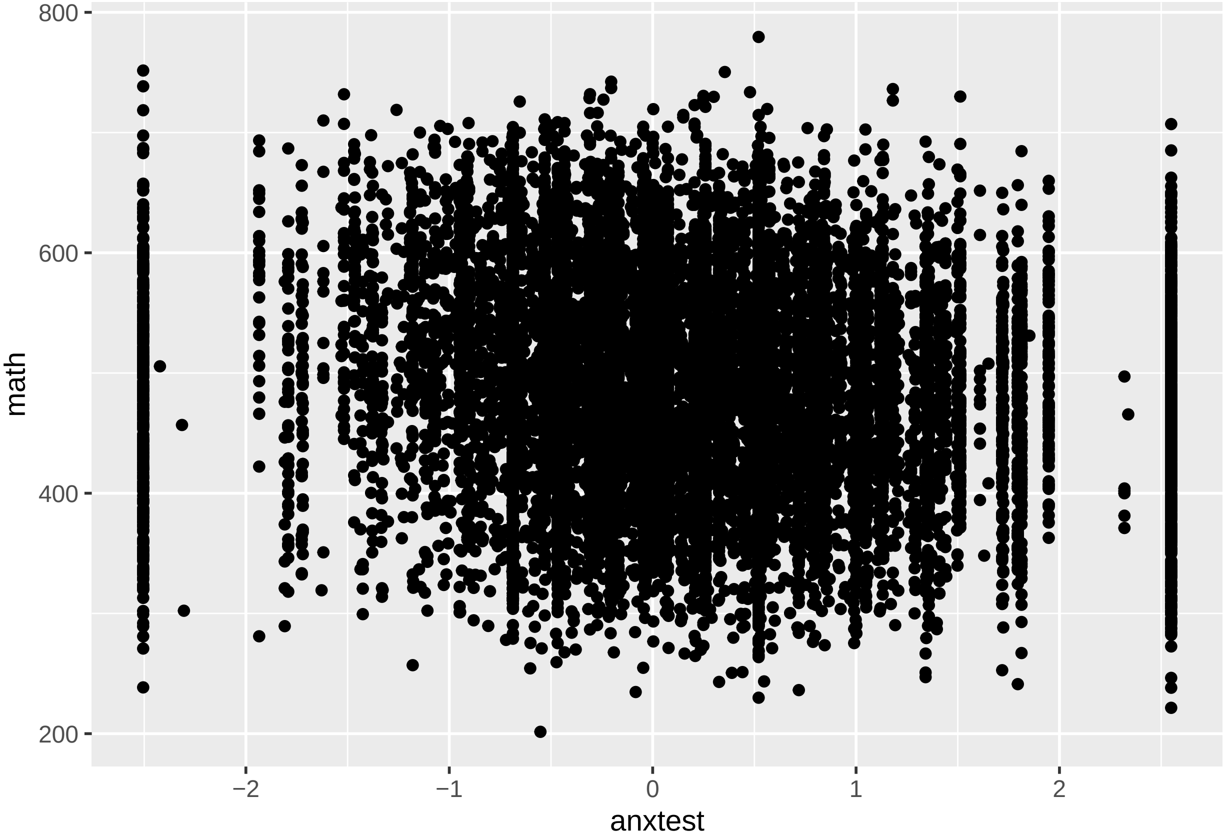
**2. Select only the variables `state`, `schtype`, `yr`, and `read`, then answer the following: Which state has the highest average score in Reading for year 10 students who are in government school, and what is the corresponding average score in Reading? [8m]**

```
## # A tibble: 8 x 2
##   state  mean
##   <chr> <dbl>
## 1 ACT    511.
## 2 NSW    493.
## 3 NT     464.
## 4 QLD    482.
## 5 SA     494.
## 6 TAS    469.
## 7 VIC    504.
## 8 WA     499.
```

```r
library(ggplot2)
ggplot(pisa,aes(x=anxtest,y=math))+geom_point()
```

**3. There are numerous studies suggesting that there is a relationship between students' anxiety level (`anxtest`) and their math performance (`math`). Based on the data in this study, is there any evidence suggesting of this relationship? Support your answer with one evidence (a number or a graph) with no more than 30 words. [5m]**

```
## Warning: Removed 632 rows containing missing values (geom_point).
```

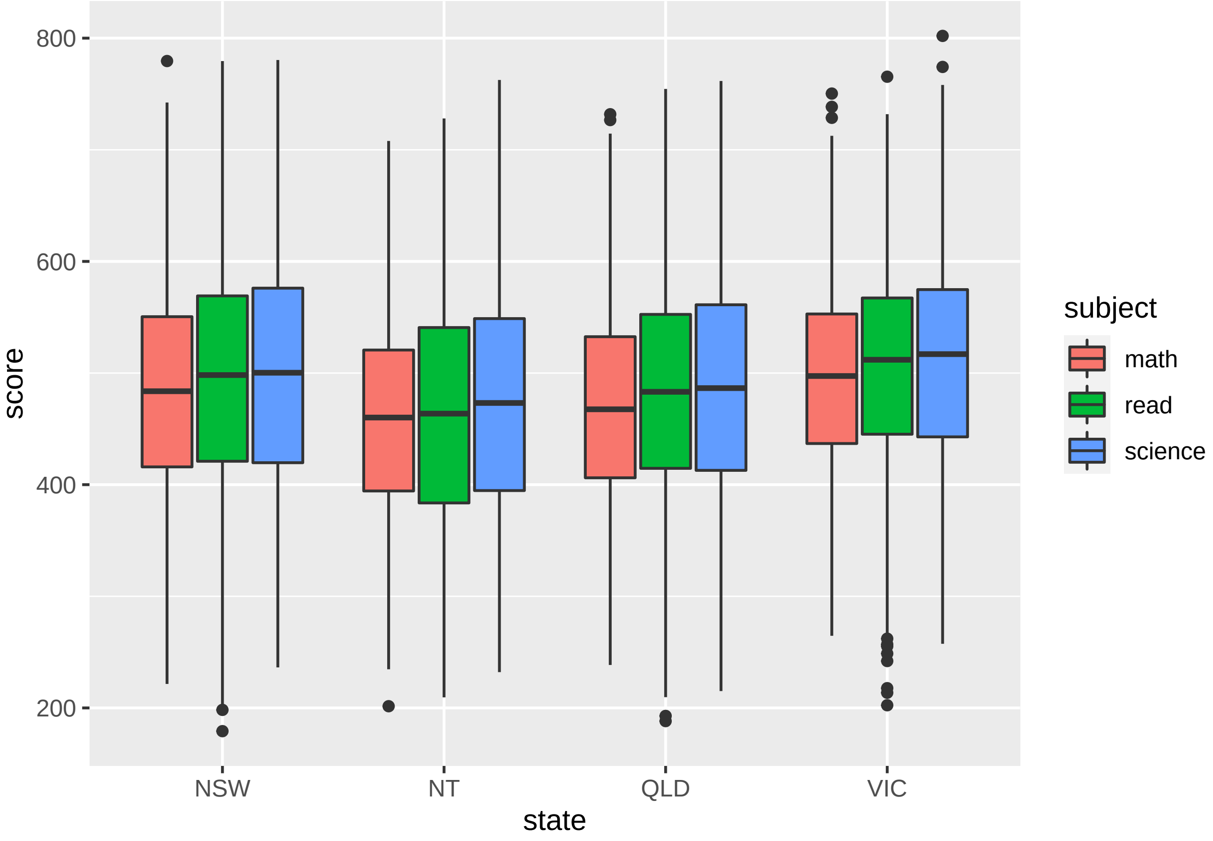


The plot from above doesn't show significant relationship between anxtest and math score. Most of the students performce anxtest from -2 to 2. The relation is not correlated.

```r
pisa_long <- pisa %>%
  select(state, read,math,science) %>%
  filter(state %in% c("VIC","QLD","NSW","NT")) %>%
  na.omit() %>%
  gather(key=subject, value=score, -state)
ggplot(pisa_long, aes(x=state, y=score, fill=subject)) + geom_boxplot()
```

**4. Let us determine whether the subjects' score distributions are different across states in Australia. First, select only these variables: `state`, `read`, `math`, and `science`. We will also focus only on 4 states in Australia, namely, `VIC`, `NSW`, `QLD`, `NT`. Next, remove any rows with missing values. Subsequently, transform the data into a long format so that a `subject` column contains 3 categories which are `read`, `math` and `science`. Store the new data frame in a data object named `pisa_long`. Finally, construct a relevant graph showing the distributions of subject scores for the four states in the three**

**subjects. What can you learn from this graph with no more than 50 words? [15m]**



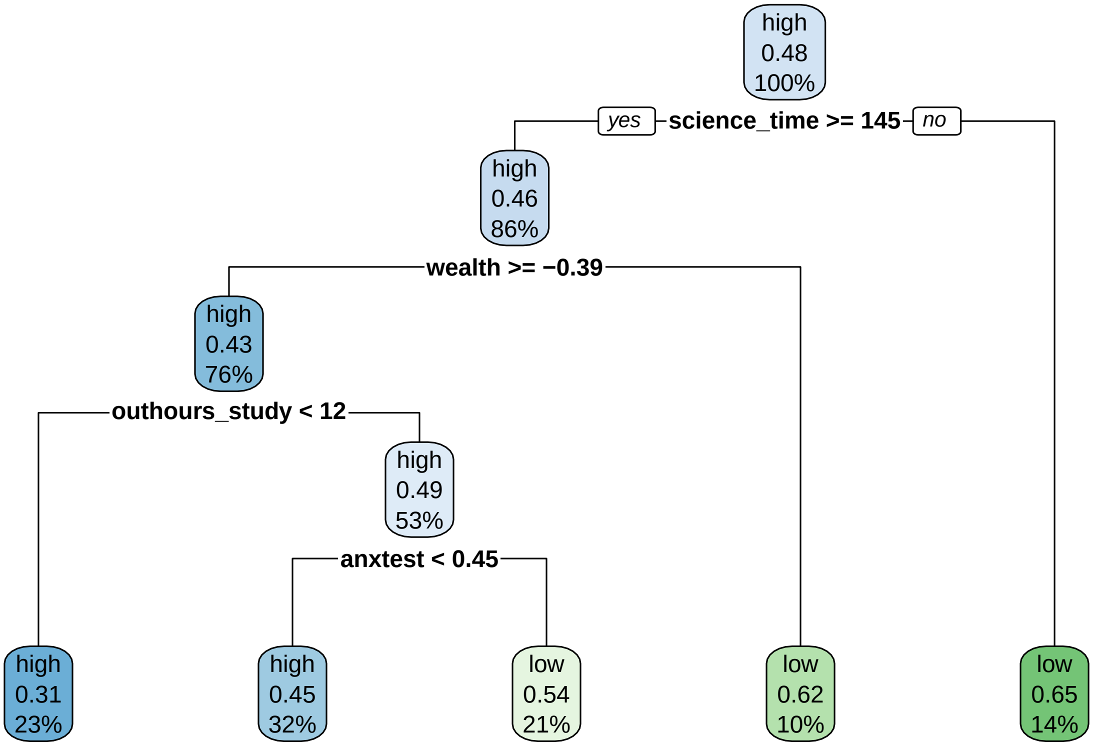
**Section B: Decision Tree (25 marks)**

```
pisa <- pisa %>%
  mutate(science_cat=ifelse(science<500, "low", "high"))
library(rpart)
library(rpart.plot)
rt <- rpart(science_cat~outhours_study+wealth+science_time+anxtest, pisa)
rt
```

**5. Group the `science` score into low and high such that those below 500 are `low` and greater or equal to 500 are `high` and name the variable as `science_cat`. Choose an appropriate decision tree model and use the new science variable (`science_cat`) as the dependent variable to answer the following question: What are the factors that influence the `science` score of `low` and `high` of the student by taking the following factors into consideration: `outhours_study`, `wealth`, `science_time`, and `anxtest`. Explain the characteristics of the student with high science score with no more than 50 words. [15m]**

```
## n=14126 (404 observations deleted due to missingness)
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 14126 6816 high (0.5174855 0.4825145)
##    2) science_time>=145 12134 5526 high (0.5445855 0.4554145)
##      4) wealth>=-0.39255 10729 4655 high (0.5661292 0.4338708)
##        8) outhours_study< 11.5 3246 1010 high (0.6888478 0.3111522) *
##        9) outhours_study>=11.5 7483 3645 high (0.5128959 0.4871041)
##         18) anxtest< 0.45005 4574 2069 high (0.5476607 0.4523393) *
##         19) anxtest>=0.45005 2909 1333 low (0.4582331 0.5417669) *
##      5) wealth< -0.39255 1405  534 low (0.3800712 0.6199288) *
##    3) science_time< 145 1992  702 low (0.3524096 0.6475904) *
```

```
rpart.plot(rt)
```



The most import predictor is the science time, when the science time is above 145, the predicted high category is 0.65, when the wealth is above -0.39, the probability that the student belongs to the high category is 0.43.

**6. Using the model constructed in 5., predict the `science_cat` of a student if a student spend the `science_time` of 160, has `anxtest` of 0.5, `wealth` of -0.2, `outhours_study` of 10. Also state it corresponding probability. Do not use any function or R command to answer this question. [5m]** The predicted science_cat is high, the probability is 0.69.

**7. Explain the 3 listed values/words in the node. What does the `low` and `high` represent? What does the second number represent? What is the meaning of the percentage value? [5m]** low and high represent the predict class, the second number represent the predicted probability of high, the percentage means the percentage of observations in the node.

### Section C: Text Mining (20 marks)

The data set "feedback.csv" contains the students' evaluation about this course ETC1010/ETC5510 in Semester 2 2021. Analyse the text data accordingly.

```
feedback<- read_csv("data/feedback.csv")
```

**8. Data Pre-processing. Read in the data. Create a table named `feedback` with 2 columns: `user` (simply assign a number to each comment to represent a single student) and `text` (the**

**feedback). *Hint*: For the user column, you can create a number from 1 up to the total number of comments. [5m]**

```
## Rows: 34 Columns: 1
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): comment
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
feedback
```

```
## # A tibble: 34 x 1
##    comment
##    <chr>
##  1 I feel confused and lost with little guidance.
##  2 Very informative and enjoyable
##  3 intrigued
##  4 Great for introduction for data analysis, but lots of works.
##  5 satisfied, helpful
##  6 It's a great course (especially how the assignments are run), Thankyou!!
##  7 R has so many commands! It looks easy but a bit more confusing than Python i~
##  8 Very interesting and practical!
##  9 it is a bit difficult but also helpful
## 10 The course is satisfying, though it can be tedious at times, due to the proc~
## # ... with 24 more rows
```

```
feedback <- data.frame(user=1:nrow(feedback), text=feedback$comment)
```

```

library(tidytext)
feedback_word <- feedback %>%
  unnest_tokens(word, text)
head(feedback_word, 5)
```

**9. Split the text column into tokens by word. Name the new data object `feedback_word` and display the first 5 rows. [3m]**

```
##   user     word
## 1    1        i
## 2    1     feel
## 3    1 confused
## 4    1      and
## 5    1     lost
```

```

final_feedback <- feedback_word %>%
  anti_join(stop_words)
```

**10. Remove all the stop words (use the "smart" lexicon) from `feedback_word`. Name this new data set as `final_feedback`. Display the first 5 rows of `final_feedback`. [4m]**

```
## Joining, by = "word"
```

```
head(final_feedback, 5)
```

```
##   user        word
## 1    1        feel
## 2    1    confused
## 3    1        lost
## 4    1    guidance
## 5    2 informative
```

```

install.packages("textdata")
```

**11. How the students feel about the course of ETC1010 in general? [use an appropriate sentiment lexicon that can express feelings such as joy, surprise, positive and etc.] Build a table to show the sentiment categories together with the number of words found. What are the few *top* sentiments that have the highest count and explain the findings with no more than 50 words? [8m]**

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(textdata)
```

```
#final_feedback %>% left_join(get_sentiments("nrc")) %>%
 # group_by(sentiment) %>% summarise(count=n()) %>% arrange(desc(count)) %>%
 # na.omit() %>% head()
```

### Section D: Social Network Analysis (9 marks)

Karate Club Network data set is used to study the community structure of the social network between the karate club members. The data set was popularised by Zachary in his paper "An Information Flow Model for Conflict and Fission in Small Groups" published in 1977. There are 34 members in the karate club. This study is to find out whether the members interacted with each other outside the club. The `unit` variable represents their working background. Answer the following questions:

**12. What are nodes and edges in this network analysis? Explain with no more than 30 words. [4m]** The nodes are the 34 members, the edges are the interactions with each other outside the clubs.

**13. Based on the R commands below, what can you learn from the 2 social network graphs? Explain with no more than 50 words [Hint: Irrelevant answers will have mark deducted]. [5m]**

### Section E: (Postgraduate Students ONLY) Linear Regression (15 marks)

Continue from Section A above, answer the following questions.

**14. Estimate a linear regression model using the variables `state`, `schtype`, `outhours_study`, `science_time`, `anxtest`, `wealth` to predict the `science` score (0-900). Present the results in a tidy format. Briefly explain the variable(s) that has/have significant impact on the science score with less than 100 words. [7m]**

**15. Examine the model adequacy and explain the model fit. Do the model assumptions hold? Do you observe any outlier? Provide evidence and explanation with less than 30 words. [5m]**

**16. Explain the difference between regression tree and linear regression model. [3m]**