Hypothesis Tests of Relationship between Students' Study Patterns and their Academic Results

Code ▾

PUBLISHED
November 6, 2023

▶ Code

1 Executive summary

Exploiting collected data from a general survey across the DATA2X02 cohort of Semester 2, 2023, this report analyse and discuss certain behavioural patterns and factors that influence the students' study efficiency. Specifically, Weighted Average Mark (WAM) is chosen as the measure of learning performance, and its relationships with some potential related factors are then investigated using general data visualisation and hypothesis tests.

Figure 1.

2 General discussion about the data

2.1 Data source

The data was collected via an online survey, in particular a [Google Forms](#), being posted on DATA2X02's Ed platform in August 2023. Hence, the questionaire was accessible and responded by the students of DATA2X02 in Semester 2, 2023 academic year. Breaking down, this cohort has:

- 675 students in DATA2002;
- 84 students in DATA2902.

Figure 2.

As having similar participants on one or some factors might skew the results of a study (Emerson 2015) and the entire validity of the sampling technique is based on the use of an unbias sample (Drott 1969), it is critical to evaluate the credibility of the sample before making any conclusions or inferences from it. In order words, a random sample can reduce the influence of uncontrolled factors (Emerson 2015) and yield highly valid analysis results (Drott 1969).

Simple random sample, meanwhile, is obtained by choosing elementary units in search a way that each unit in the population has an equal chance of being selected (Mugo 2002). Based on common sense, it can be stated that this survey is collected from a random sample, as it was posted on the Ed channel which all current DATA2X02 students have access to hence each of them has equal opportunities to be included in the study. However, as this survey is not compulsory or is accounted to grades, there is no significant incentives for students to do it. This might result in the non-response bias - situations where due to non-response, certain types of survey respondents (students) are under-represented (Berg 2005). Besides, the fact that not all DATA202 students spend time checking the Ed platform/announcements thus not acknowledging about the existence of the survey might lead to convenience sampling or coverage errors - leaving certain groups of the DATA2X02 cohort.These potential, uncontrollable factors might end up make the sample not being random anymore.

In conclusion, while some useful insights can still be extracted from this data set, is it importance to take into consideration that **no absolute guarantee about random sampling** can be made, or responses themselves might **not** come from a **random sample** of DATA2X02 students.

Figure 3.

2.2 Data format and structure

The responses are then stored in a Comma Speparated Values (CSV) data file with totally:

▶ Code

```
 44  columns, and  312  rows.
```

Taking a closer look, as the first columns recorded the time stamp when the responses were submitted, the survey only consists of 43 questions varying from multiple choice, short answer, check boxes, and drop-down including:

▶ Code

```
1. How many units are you enrolled in this semester?
2. When it comes to assignments / due tasks do you:
3. How old are you?
4. Do you tend to lean towards saying "yes" or towards saying "no" to things
throughout life?
5. Have you taken one or more units of study from the Faculty of Arts and Social
Sciences?
6. Are you completing a major or minor in a subject area from the Faculty of Arts and
Social Sciences?
7. Have you read a novel this year?
8. Would you prefer to study at Fisher Library or SciTech Library?
9. Do you have private health insurance?
10. How many days in a week you normally consume sweets/chocolates/sugary drinks?
(Exclude Diet/Sugar Free Drinks & sweets)?
11. Do you pay rent?
12. What is your post code?
13. How many days do you go between haircuts (on average)?
14. What brand is your laptop?
15. You enter a public bathroom and find you're the only one there. There are three
urinals on the wall for you to choose from. Which do you choose?
16. You enter a public bathroom and there are three stalls to choose from. All three
are unoccupied. Which do you choose?
```

17. How many Weet-Bix would you typically eat in one sitting?
18. What is the average amount of money you spend each week on food/beverages?
19. Do you like pineapple on pizza?
20. What are your current living arrangements?
21. How tall are you?
22. How do you get to university?
23. How often would you say you feel anxious on a daily basis?
24. How many hours a week do you spend studying?
25. Do you work?
26. What is your favourite social media platform?
27. What is your gender?
28. How much sleep do you get (on avg, per day)?
29. What is your diet style?
30. Pick a number at random between 0 and 9
31. How do you like your steak cooked?
32. What is your dominant hand?
33. Which unit are you enrolled in?
34. On average, how many hours each week do you spend exercising?
35. How many hours a week (on average) do you work in paid employment?
36. Do you submit assignments on time?
37. Have you ever used R before starting DATA2x02?
38. What kind of role (active or passive) do you think you are when working as part of a team?
39.
How many hours do you spend on social media per day?
40. Which year of university are you currently in?
41. Which sports do you play most often?
42. What is your WAM?
43. What is your shoe size?

Hence, the collected data also varies in the types and format itself.

2.3 Potential biases

According to Lecture 02, bias is any factor that favours certain outcomes or responses, or influences an individual's responses.

As aforementioned, there might be some **non-response bias** as some students chose not to fill in the survey considering that this is time-consuming yet is not compulsory and there are no rewards or punishment of doing so. Besides, when it comes to the question **"What is your WAM?"**, it is possible that only students with "high" score or are confident about their academic performance chose to response. This might also leave certain groups out thus the sample might under-represent the population which also leads to a **non-response bias**.

There also might be **selection/sampling bias** given that the survey was posted only on Ed Discussion without any further notification which might end up leaving out students who rarely check Ed platform. Indeed, the survey only had 312 responses, approximately 41.11% of the whole co-hort (759 students). In other words, the respondents or the sample fails to represent the population - the whole DATA2X02 cohort of Semester 2, 2023.

Measurement or designed bias is noticeably potential due to several questions such as:

- **"How many hours a week do you spend studying?"**: Unlike working time or time spent on social media which have official tracking tools (screening time, work shifts, etc.), study time is much harder to be tracked. Having no subjective reference, the students might tend to either overestimate the time they study given this survey is released and managed by the lecturer.

- **"How often would you say you feel anxious on a daily basis?"**: The answer to this question might depend on students' mood at the moment they did the survey. For instance, if the student had a bad day or was waiting for an important event (interview, test result, etc.), they could be very nervous hence overestimate the answer. If he/she, otherwise, was completely happy on that day and had nothing to worried about, they could underestimate their answer. Thus, the responses might not yield an accurate average times of feeling anxious on a daily basis as expected.

2.4 Data shortcomings

Looking at the csv file, many questions need improved to generate more useful and easily analysed data. Specifically, response validation could have been used to restrict which type of data can the respondents can input. For example:

- **"What is your gender?"**: It is obvious that gender can be classified into certain groups such as Male, Female, Transgender, etc. Hence, using a multiple choice instead of a short answer question can automatically categorised the answers, which significantly support later analysis (grouping, aggregating calculation, etc.)

- Numerical-related questions such as **"How many hours a week do you spend studying?"**, **"How old are you?"**, etc.: Restricting the response type into numeric only and adding left-bounded condition (non-negativity) can prevent cases when respondents input vague or invalid values ($0 - 5$, -6, four, etc.)

In addition, some questions might need more detailed description or context-information in order to guide the question in the expected direction, for instance:

- **"What is your shoe size?"**: Given there are different shoe size systems such as US, EU, etc., this question can cause confusion to both respondents and analysts. Specifying the exact system being used would simply clarify this problem for both sides.

- **"How tall are you?"**: Similarly, as students from different countries might prefer using different units of height measurement, providing the exact expected units (ft, m, etc.) can unload a hge burden when it comes to data interpretation and analysis. Otherwise, analyst have to manual check each response and convert all of them into a common unit if wishing to use this variable.

3 Data wraggling

The data was imported, cleaned and analysed based on the guide of (Tarr 2023) using R version 4.3.1 (R Core Team 2023) within the Quarto environment for reproducibility. Some packages used within this report include:

- The **tidyverse** suite of packages, including readr, tidyr, dplyr and ggplot2 (Wickham 2023);

- The **gt** package was used to present the tables (Iannone et al. 2023);

- The **grid** package was used to work with graphical objects within R (R Core Team 2023) and the **gridExtra** packages was used for some functions to work with "grid" graphics (Auguie 2017).

The document was finally compiled using the Quarto publishing system (Allaire et al. 2022).

Firstly, the variables' names or columns's names were cleaned and shortened (Tarr 2023):

▶ Code

New	Old
timestamp	Timestamp
n_units	How many units are you enrolled in this semester?
task_approach	When it comes to assignments / due tasks do you:
age	How old are you?
life	Do you tend to lean towards saying "yes" or towards saying "no" to things throughout life?
fass_unit	Have you taken one or more units of study from the Faculty of Arts and Social Sciences?
fass_major	Are you completing a major or minor in a subject area from the Faculty of Arts and Social Sciences?
novel	Have you read a novel this year?
library	Would you prefer to study at Fisher Library or SciTech Library?
private_health	Do you have private health insurance?
sugar_days	How many days in a week you normally consume sweets/chocolates/sugary drinks? (Exclude Diet/Sugar Free Drinks & sweets)?
rent	Do you pay rent?
post_code	What is your post code?
haircut_days	How many days do you go between haircuts (on average)?
laptop_brand	What brand is your laptop?
urinal_position	You enter a public bathroom and find you're the only one there. There are three urinals on the wall for you to choose from. Which do you choose?
stall_position	You enter a public bathroom and there are three stalls to choose from. All three are unoccupied. Which do you choose?
n_weetbix	How many Weet-Bix would you typically eat in one sitting?

New	Old
food_budget	What is the average amount of money you spend each week on food/beverages?
pineapple	Do you like pineapple on pizza?
living_arrangements	What are your current living arrangements?
height	How tall are you?
uni_travel_method	How do you get to university?
feel_anxious	How often would you say you feel anxious on a daily basis?
study_hrs	How many hours a week do you spend studying?
work	Do you work?
social_media	What is your favourite social media platform?
gender	What is your gender?
sleep_time	How much sleep do you get (on avg, per day)?
diet	What is your diet style?
random_number	Pick a number at random between 0 and 9
steak_preference	How do you like your steak cooked?
dominant_hand	What is your dominant hand?
normal_advanced	Which unit are you enrolled in?
exercise_hrs	On average, how many hours each week do you spend exercising?
employment_hrs	How many hours a week (on average) do you work in paid employment?
on_time	Do you submit assignments on time?
used_r_before	Have you ever used R before starting DATA2x02?
team_role	What kind of role (active or passive) do you think you are when working as part of a team?
social_media_hrs	How many hours do you spend on social media per day?
uni_year	Which year of university are you currently in?
sport	Which sports do you play most often?
wam	What is your WAM?
shoe_size	What is your shoe size?

The visdat package ([Tierney 2017](#)) then was used to give a quick look at the missing data.

▶ Code
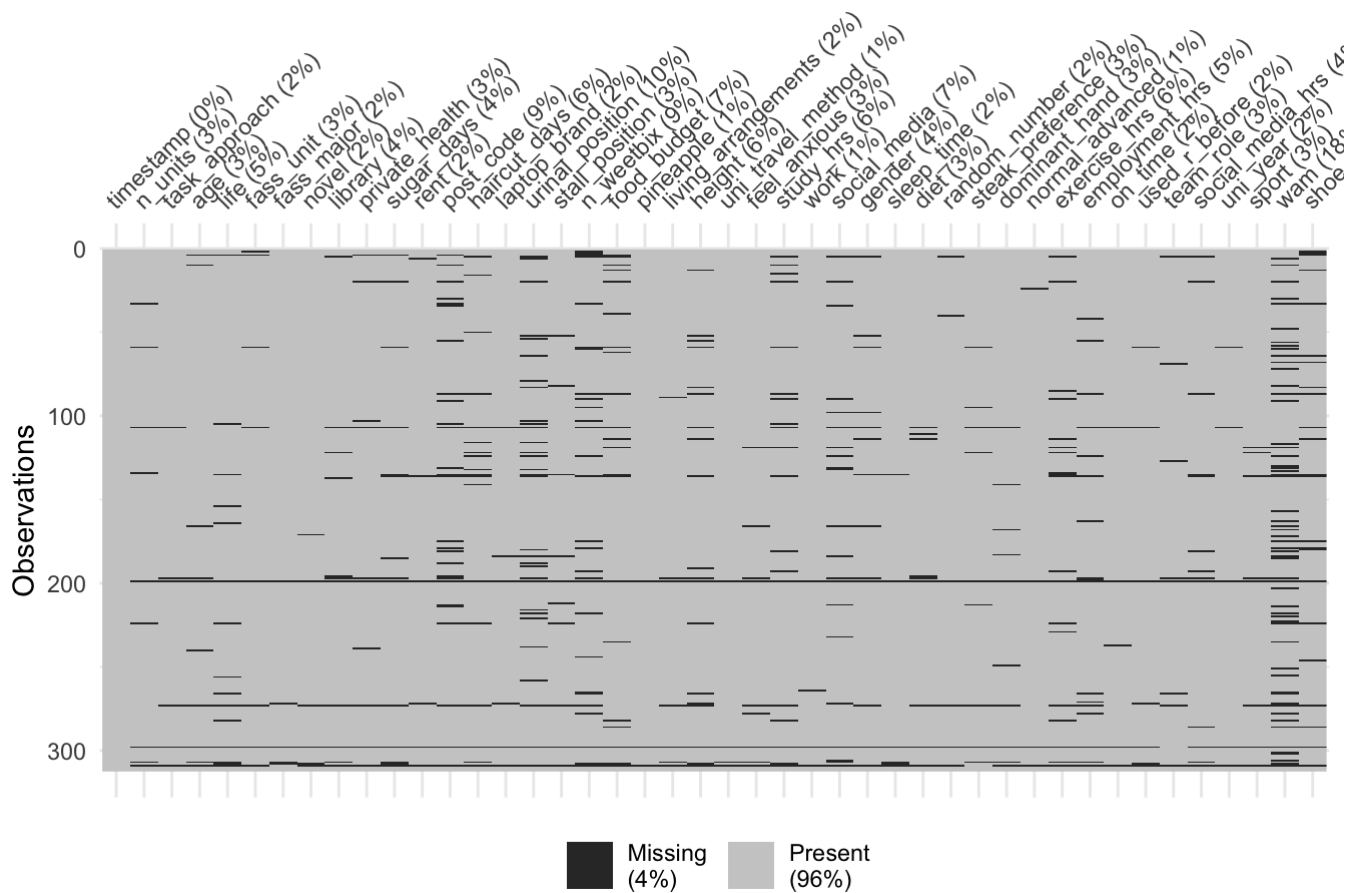
With the goal to understand the relationship of certain hehavioural patterns and external factors towards academic performance (WAM), extra cleaning steps needs to be taken at those variables for later analysis. Within the scope of this report, only four variables are investigated including:
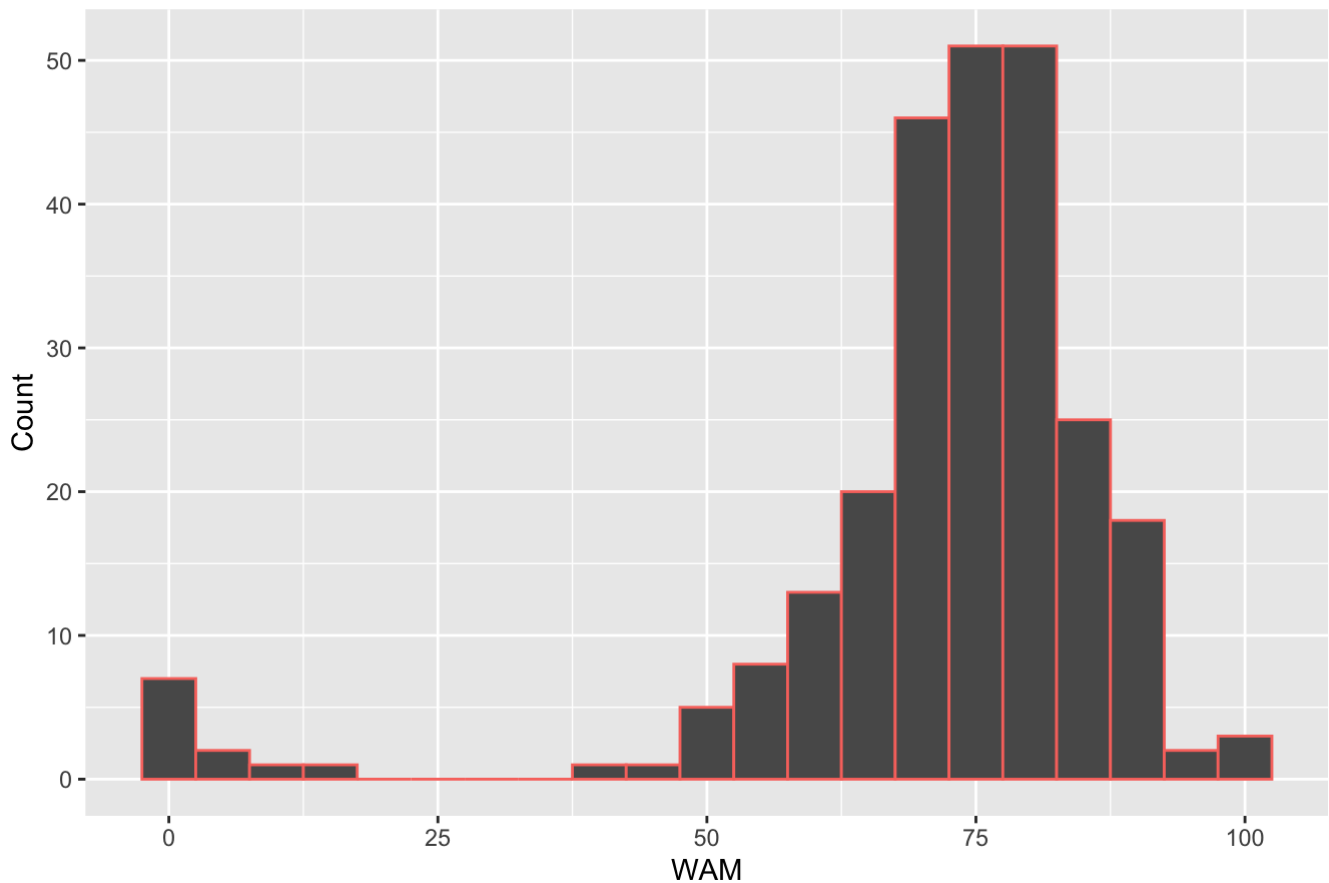
▶ Code

3.1 Weighted Average Mark (WAM)

First, a histogram can give a general distribution of the WAM.

▶ Code

WAM distribution

As can be seen, there are some noticeable outliers such as absolute (100), zero(0) and abnormaly low scores. Although it is still possible to achieve such high or low WAMs, this report eliminates them for the sake of preventing any uncontrollable error, skewness in the results. For later analysis, a new column is created categorising the WAM into result based on the University of Sydney's policy. Abbreviations are used instead of full text for quicker processing and interpretation, specifically: High Distinction (HD), Distinction (D), Credit (CR), Pass (P), Fail (F).

▶ Code

3.2 Times of feeling anxious per day

As this is a multiple choice question, response is restricted to certain possible values (integers from 1 to 10) thus there is no outliers or invalid values that need to be addressed. However, to better support later analysis, a new column is also created grouping the number of times feeling anxious on a daily basis into 3 categories. Specifically, Low $(0-3)$, Medium $(4-7)$, and High $(8-10)$.

▶ Code

3.3 Task approach

Similarly, this question is also multiple choice which leaves no problem with neither outliers nor invalid values. However, to better support later analysis, the three categories are grouped into two main groups only: students who start the assignment early, and who start late.

▶ Code

3.4 Role in a team

This is also a multiple choice question. Considering the wide range of values, a new column is also created grouping the level of being active/passive within a group into 3 categories: Passive $(0-3)$, Neutral $(4-6)$, and Active $(7-10)$.
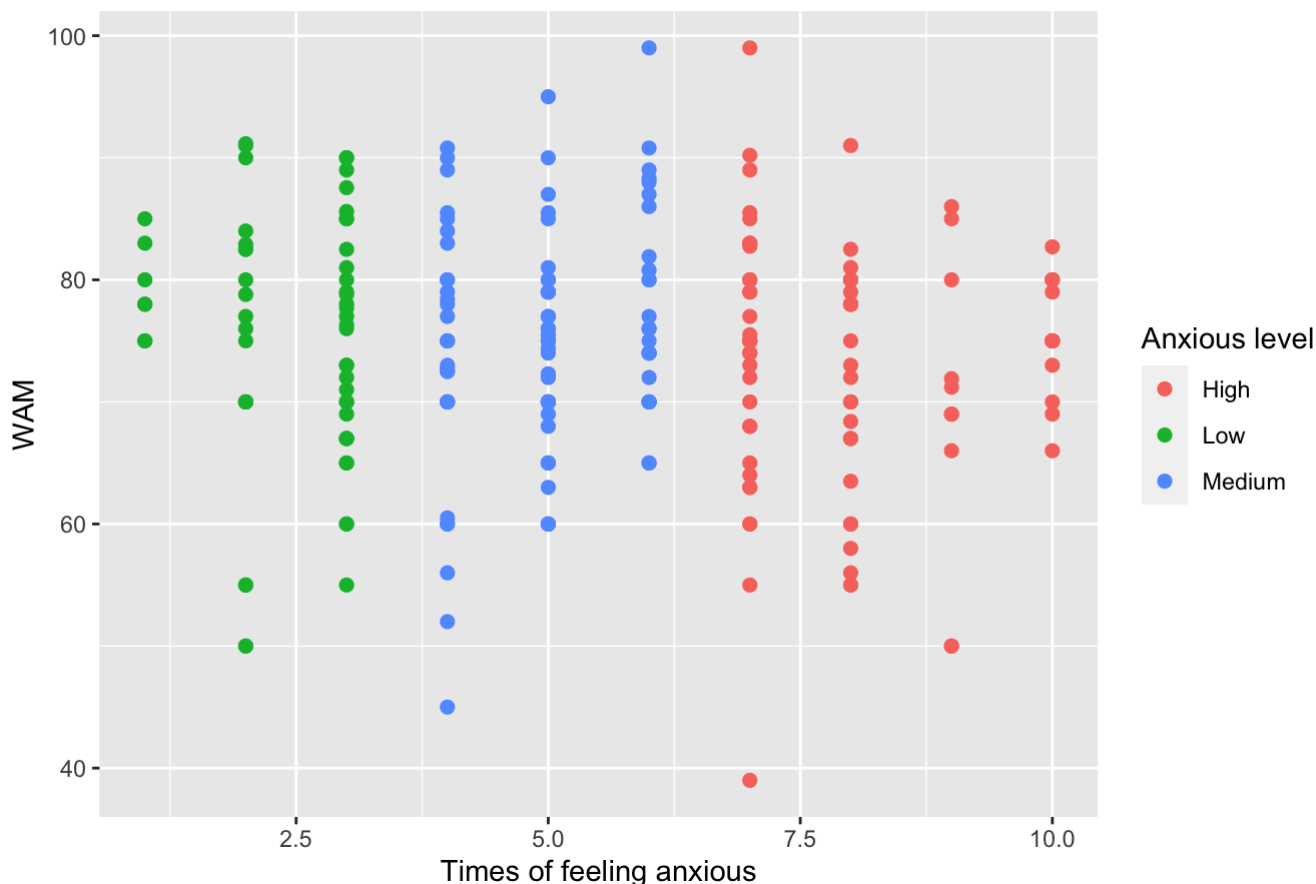
▶ Code

4 Tests and Results

Having all necessary data cleaned and ready, some hypothesis tests can now be conducted.

4.1 Is the WAM independent of Anxious Level?

A scatter plot is created to illustrate the distribution of WAM by the anxious level. It seems that extreme values of WAM usually happen at group with high anxious level. In other words, the higher anxious level is, the more widely spread WAM values are.

▶ Code

Distribution of WAM aross different anxious levels

To test whether WAM is independent of the number of times a student feeling anxious on a daily basis, a chi-squared test is conducted.

▶ Code

CR	D	F	HD	P
24	36	1	8	15
16	25	0	11	7
34	34	1	17	9

▶ Code

```
Warning in chisq.test(dat1): Chi-squared approximation may be incorrect
```

▶ Code

CR	D	F	HD	P
26.11765	33.52941	0.7058824	12.70588	10.941176
18.34454	23.55042	0.4957983	8.92437	7.684874
29.53782	37.92017	0.7983193	14.36975	12.373950

As can be seen, the three expected counts of the F (Fail) result are less than 5 which causes the warning. Thus, it can be collapsed into the P (Pass group) before conducting the test again:

▶ Code

```
	Pearson's Chi-squared test

data:  dat1
X-squared = 7.0903, df = 6, p-value = 0.3126
```

Hypothesis:

- H_0: WAM is independent of the anxious level
- H_1: WAM is not independent of the anxious level

Assumptions:

- Independent observations;
- $e_i = np_i \geq 5$, which is true after collapsing the columns above.

Test statistics:

$$T = \sum_{i=1}^{r} \sum_{i=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$$

Observed test statistic: $t_0 = 7.09$

p-value: $P(T \geq t_0) = P(\chi_6^2 \geq t_0) = 0.313$

Decision: As p-value is greater than 0.05, we do not reject the null hypothesis, or there is not enough evidence to conclude that WAM and the anxious level are dependent.

4.2 Is the WAM result of students who start assignment early higher than the ones who start late?

Firstly, a box plot is created to illustrate the distribution of WAM of two group of students who start assignment early and late. General observation indicates that the sooner a student start their assignment, the higher average WAM they have.
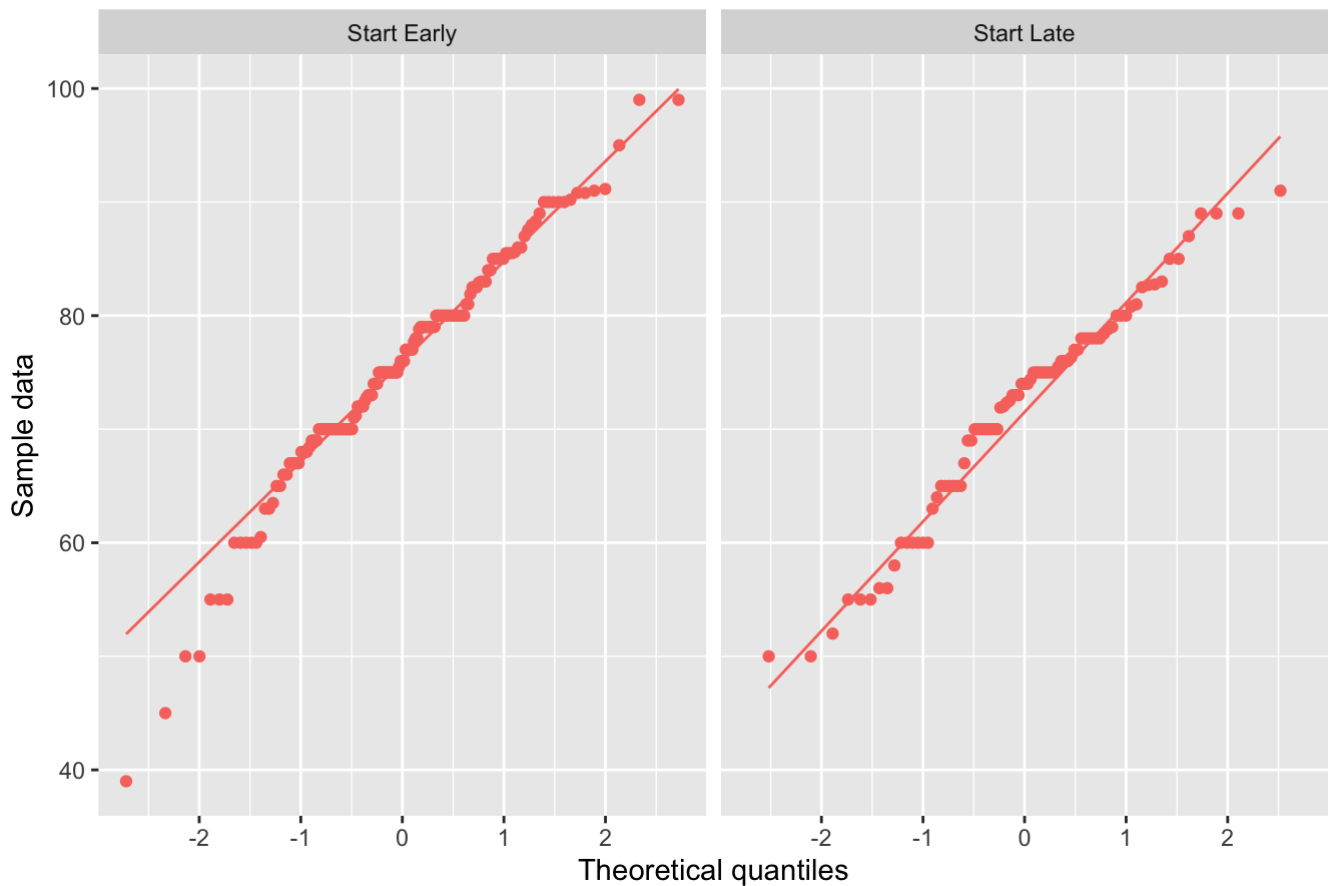
▶ Code
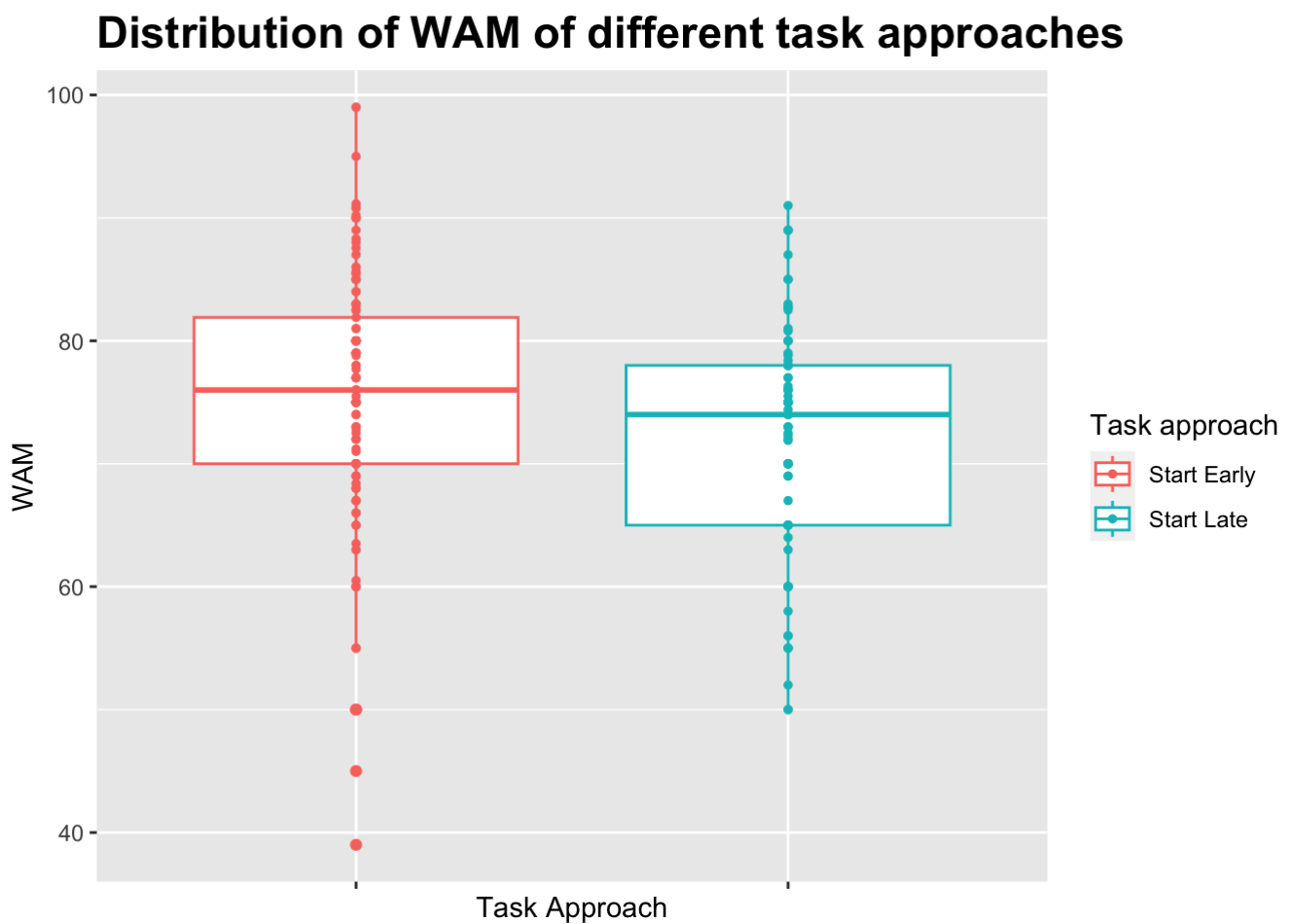
Two QQ plots are then plotted to check for the normality:

▶ Code

QQ plots

As can be seen, the QQ plot of the "Start Early" group shows a relatively curvable pattern, especially for values in the smaller half. In other words, the normality assumptions are in doubt thus an alternative is needed there. In this case, Wilcoxon rank-sum test is used instead as it relaxes normality assumptions of the t-test yet have more efficient use of data and hogher power than the sign test.

▶ Code

CR	D	F	HD	P
46	62	2	29	14
28	33	0	7	17

▶ Code

```
        Wilcoxon rank sum test with continuity correction

data:  start_early and start_late
W = 7918.5, p-value = 0.002676
alternative hypothesis: true location shift is greater than 0
```

Let μ_1 and μ_2 be the population means of WAM for students who start tasks early and who start late respectively.

Hypothesis:

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 > \mu_2$

Assumptions: Two groups of observation are independent and follow the same distribution but differ by a shift.

Test statistics: $W^+ = \sum_{i:D_i>0} R_i$

Observed test statistic: $w^+ = 7918.5$

p-value: $P(W^+ \geq w^+) = 0.003$

Decision: As p-value is less than 0.05, we reject the null hypothesis and conclude that the WAM of students who start their assignment early is significantly higher than their peers who start late.

4.3 Is the WAM result independent of their role in a team? (using Monte Carlo simulation)

To generally illustrate the distribution of WAM across different role in a team, a boxplot and a scatter plot are created as below. It can be observed that students who is active in a team have averagely higher WAM than those who present neutral and passive. Also, extreme values of WAM often happen within the group with middle level of being active (neutral).

▶ Code

Distribution of WAM based on student's roles in team
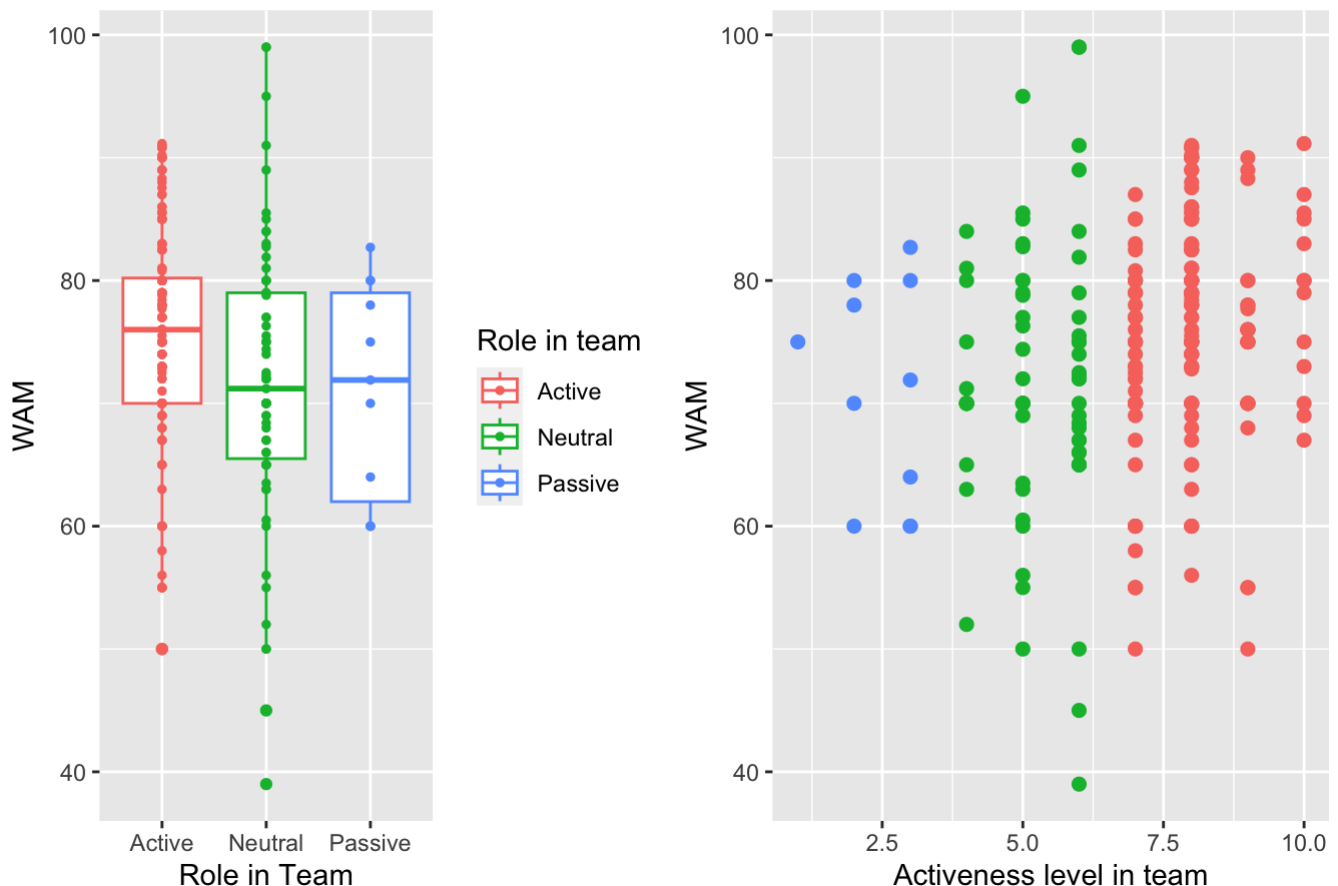
To test whether WAM is independent of the student's active level within a group, a chi-squared test is conducted.

▶ Code

CR	D	F	HD	P
43	71	0	29	17
29	19	2	7	10
2	5	0	0	4

▶ Code

```
Warning in chisq.test(dat3): Chi-squared approximation may be incorrect
```

▶ Code

CR	D	F	HD	P
49.747899	63.865546	1.34453782	24.201681	20.840336
20.831933	26.743697	0.56302521	10.134454	8.726891
3.420168	4.390756	0.09243697	1.663866	1.432773

As can be seen, as the expected counts of values in "F" and "Passive" group is less than 5, the assumption of Chi-squared Test is not met which might yield incorrect test result. To deal with this problem, a Monte Carlo simulation is conducted instead of normal method:

▶ Code

```
    Pearson's Chi-squared test with simulated p-value (based on 10000
    replicates)

data:  dat3
X-squared = 22.014, df = NA, p-value = 0.008999
```

Hypothesis:

- H_0: WAM is independent of the activeness level of student in a team;
- H_1: WAM is not independent of the activeness level

Assumptions:

- Independent observations;
- $e_i = np_i \geq 5$, which is true.

Test statistics:

$$T = \sum_{i=1}^{r} \sum_{i=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$$

Observed test statistic: $t_0 = 22.01$

p-value: $P(T \geq t_0) = P(\chi_8^2 \geq t_0) = 0.009$

Decision: As p-value is less than 0.05, we reject the null hypothesis and conclude that WAM is significantly dependent on the the level of being active within a group.

5 Conclution

Based on above analysis and testing, some noticeable insights about the dependence and correlation between WAM and some behaviourial patterns can be concluded:

- There is not enough evidence to conclude about the dependence of WAM on the anxious level of the students.

- Students who start doing assignment early have averagely higher WAM compared to their counterparts who start late.

- Students who are more active when working in a team have averagely higher WAM than those who play neutral and passive role.

However, it is critical to also take into consideration some limitations of the sampling process, which is discussed in Part 2, when intepreting and using those results.

---

6 Reference list

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. "Quarto." https://doi.org/10.5281/zenodo.5960048.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Berg, Nathan. 2005. "Non-Response Bias."

Drott, M Carl. 1969. "Random Sampling: A Tool for Library Research." *College & Research Libraries* 30 (2): 119–25.

Emerson, Robert Wall. 2015. "Convenience Sampling, Random Sampling, and Snowball Sampling: How Does Sampling Affect the Validity of Research?" *Journal of Visual Impairment & Blindness* 109 (2): 164–68.

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. 2023. *Gt: Easily Create Presentation-Ready Display Tables*. https://CRAN.R-project.org/package=gt.

Mugo, Fridah W. 2002. "Sampling in Research."

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tarr, Garth. 2023. "DATA2002 Assignment: Data Importing and Cleaning Guide." 2023. https://pages.github.sydney.edu.au/DATA2002/2023/assignment/assignment_data.html.

Tierney, Nicholas. 2017. "Visdat: Visualising Whole Data Frames." *JOSS* 2 (16): 355. https://doi.org/10.21105/joss.00355.

Wickham, Hadley. 2023. *Tidyverse: Easily Install and Load the Tidyverse.* https://CRAN.R-project.org/package=tidyverse.