STA302 Final Project

Janhavi Agarwal(1005786534), Jiyun (Lyla) Won(1004869476), Joshua Cuevas(1007334216),
Tze Sie Jercy Tiong(1005368772)

2022-08-22

Abstract

Introduction

Throughout the first 4 weeks of STA302, data has been collected from students about how much time they spent each week studying, how much time they spend each week thinking about COVID and how much time each week they spend on miscellaneous activities. At the end of the fourth week, the course's term test took place, and each students's score was added to this dataset.

The variables collected could certainly be factors in the performance of students on the term test. According to a study conducted at Troy University, some other factors can be time taken to complete a test, student seating location and perception of test difficulty. Other variables that were collected by this study include familiarity of students with programming language R and attendance of office hours throughout the semester.

Purpose of the project

The report aims to study the relationship between the student's performance on the STA302 term test and the other predictor variables, including Studying n, COVID n, Miscellaneous n, OH, and Familiarity. The report intends to distinguish the predictors with the highest correlation with the term test results through the regression model.

The model development will help identify the overall student progress and factors influencing performance. This will provide effective solutions for professors and students in determining possible improvements in future course development that can maximize students' performance. With a better understanding of the influential factors of the term test performance, students can strategically plan their future studies to minimize the negative influential factors and build an effective study plan.

The Data

The data was read into the report using the `tidyverse` package of R. This section describes the process of cleaning the data and the variables we are working with.

Description of variables

The dataset contains 15 variables, with the Term Test being the response variable. The remaining 14 variables are predictor variables. The table below summarizes the variables being processed in this analysis.

The quantitative data Studying n, COVID n, and Miscellaneous n were collected weekly, so they are categorized into four variables. Then there are two qualitative data Office Hour and Familiarity.

Variable	Description	Type
Term Test	Students performance on term test (score out of 55 + 1 bonus)	Ordinal
Studying	Hours spent on studying for STA302 during week 1	Continuous
Studying2	Hours spent on studying for STA302 during week 2	Continuous
Studying3	Hours spent on studying for STA302 during week 3	Continuous
Studying4	Hours spent on studying for STA302 during week 4	Continuous
COVID	Hours spent on thinking about COVID during week 1	Continuous
COVID2	Hours spent on thinking about COVID during week 2	Continuous
COVID3	Hours spent on thinking about COVID during week 3	Continuous
COVID4	Hours spent on thinking about COVID during week 3	Continuous
Miscellaneous1	Hours spent on miscellaneous activities during week 1	Continuous
Miscellaneous2	Hours spent on miscellaneous activities during week 2	Continuous
Miscellaneous3	Hours spent on miscellaneous activities during week 3	Continuous
Miscellaneous4	Hours spent on miscellaneous activities during week 4	Continuous
Office Hour (OH)	Frequency of office hour attendance	Categorical
Familiarity	Student's familiarity to the course, STA302	Categorical

Data Cleaning process

The data had to be cleaned to include a variable that calculates the total hours spent studying before the midterm, total hours spent on miscellaneous activities and the total hours spent thinking about COVID19. To do this, R's `rowSums` functionw was utilized.

Exploratory Data Analysis

Exploring each independent variable individually

Histograms have been used to show each continuous variable, ie, hours spent studying each week, hours spent thinking about COVID each week, hours spent on miscellaneous activities each week.

Histograms

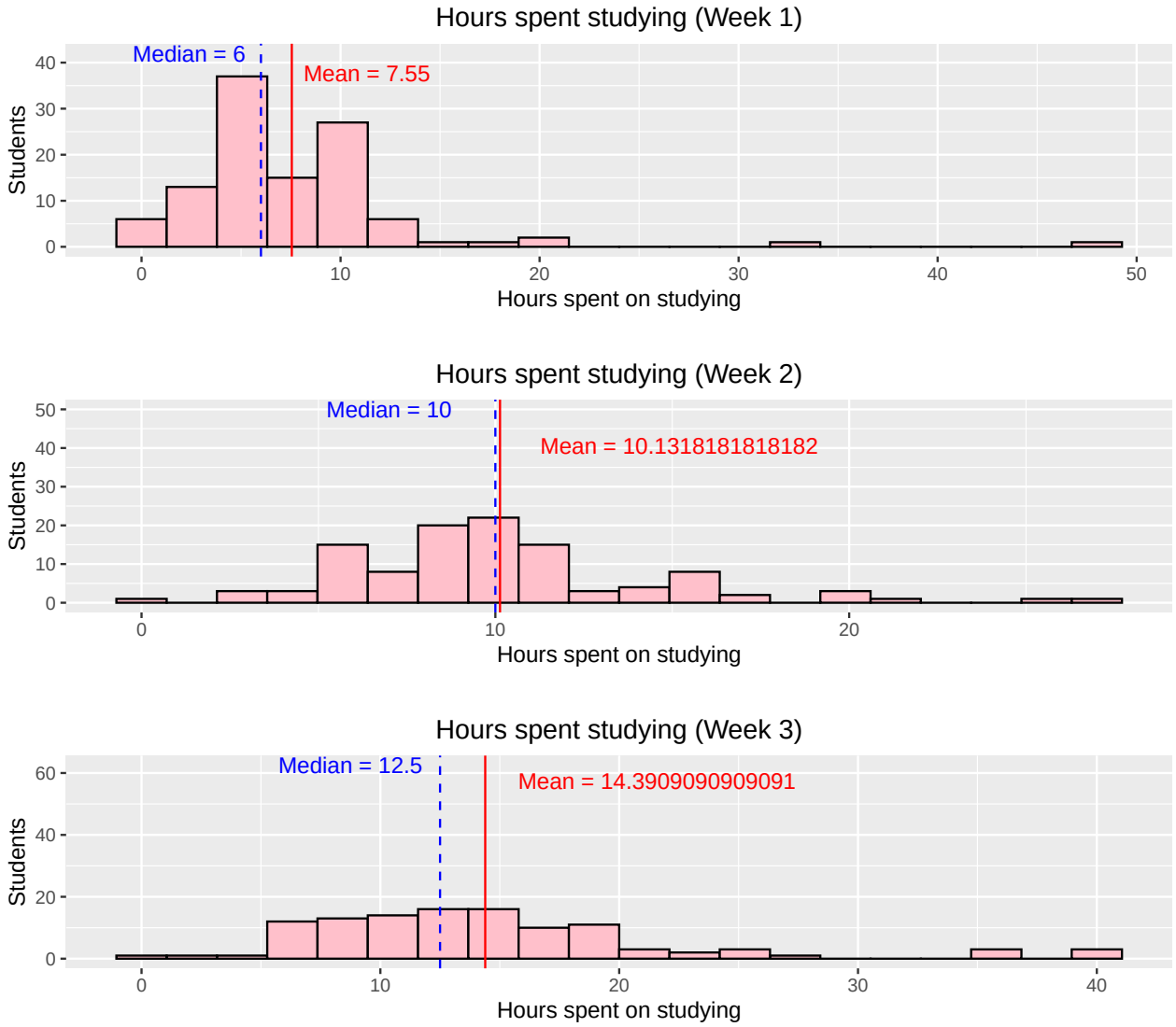
The distribution of continuous predictor variables, Study, COVID, and Miscellaneous, were presented using histograms. Figure 1-1 represents the hours spent studying each week, Figure 1-2 represents the hours spent thinking about COVID, and Figure 1-3 illustrates the hours spent on miscellaneous activities. The mean and median were indicated by the solid red line and the blue dashed line, respectively.

The histogram displaying the distribution of the students' hours spent studying during weeks 1-4 (figure 1-1) varies slightly. All the weeks tend to reveal a right-skewed model, with weeks 1 and 4 being the most extreme. According to the right-skewed data distribution, this indicates that there were more students that spent less than 20-30 hours studying for the STA302 course. The median values were calculated to be 6, 10, 12.5, and 16 hours across four weeks, which is significantly low compared to the range of hours (approx. 0-60 hours). Moreover, it is significant to note that the median and mean hours spent studying per week show an increase every week. Knowing that the term test was on week 4, this increasing pattern can contribute to the assumption that students tend to study more as the term test approaches.

The histogram for hours spent thinking about COVID during weeks 1-4 (figure 1-2) also reveals the right-skewed distribution with two or three anomalies where extremely more considerable hours were recorded. The medians calculated for each week were constant (1 hour) throughout, and the mean values were also shown to be within the range of 1-2.5 hours. Overall, the histogram seems to represent low hours spent on thinking about COVID, which is plausible since the course was flexible with students' choice in learning online and in person. The abnormally extreme variables can be explained by vaguely assuming some students tested positive for COVID and spent more time thinking about COVID.

The histogram of the hours spent on miscellaneous activities (figure 1-3) denoted right-skewed distribution with the same median value of 20 over weeks 2 to 3. Students had fewer hours spent on miscellaneous in the first week, which can be reasoned by generalizing that most students tend to be more motivated to study when the new semester begins. Also, many students have summer jobs, so as summer jobs began, the hours on miscellaneous activities might have been influenced (such as abnormally large hours) and increased the overall mean of each week.

Figure 1-1: Hours spent studying each week



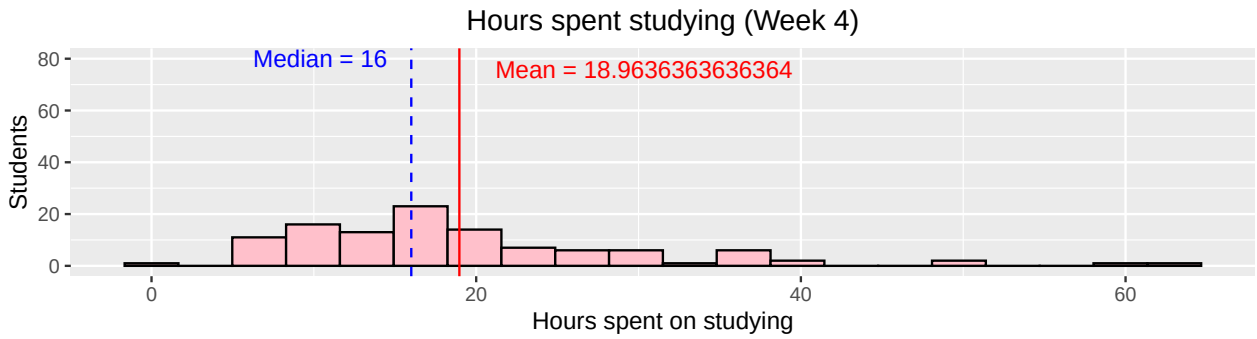
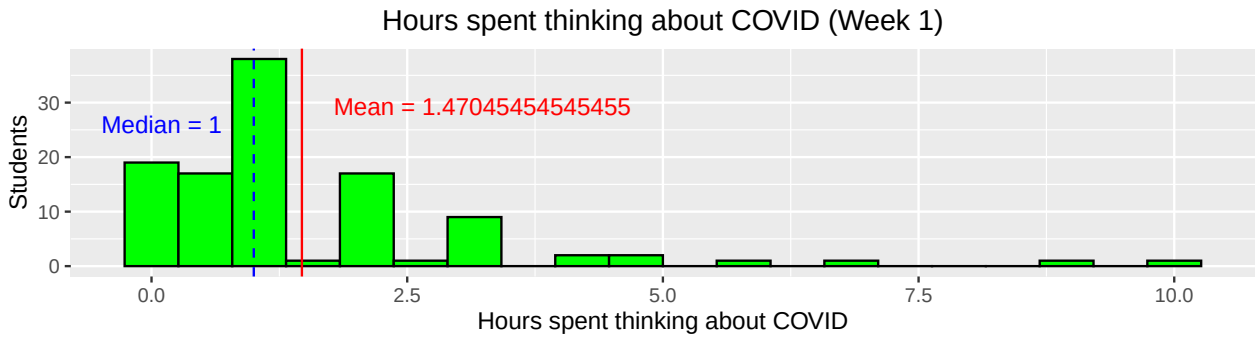
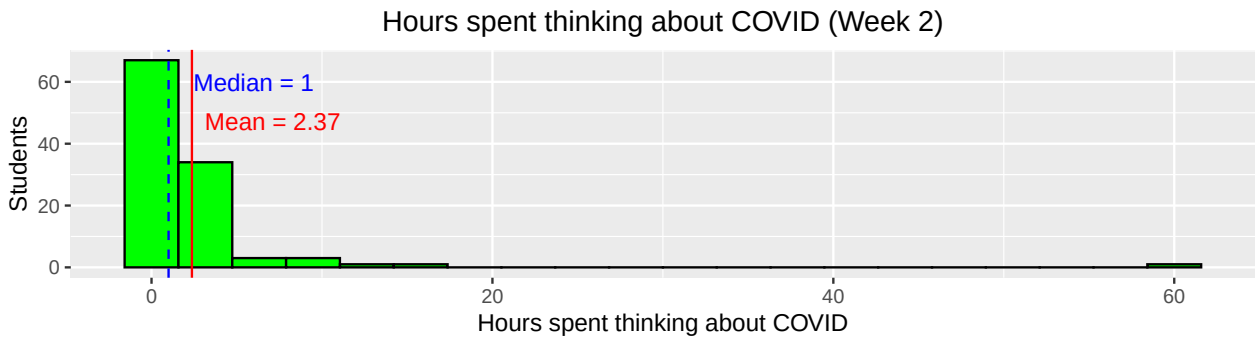
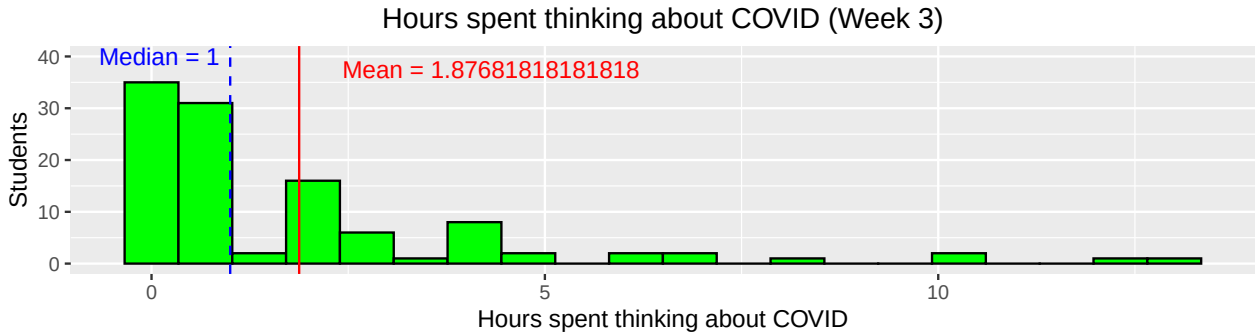


Figure 1-2: Hours spent thinking about COVID







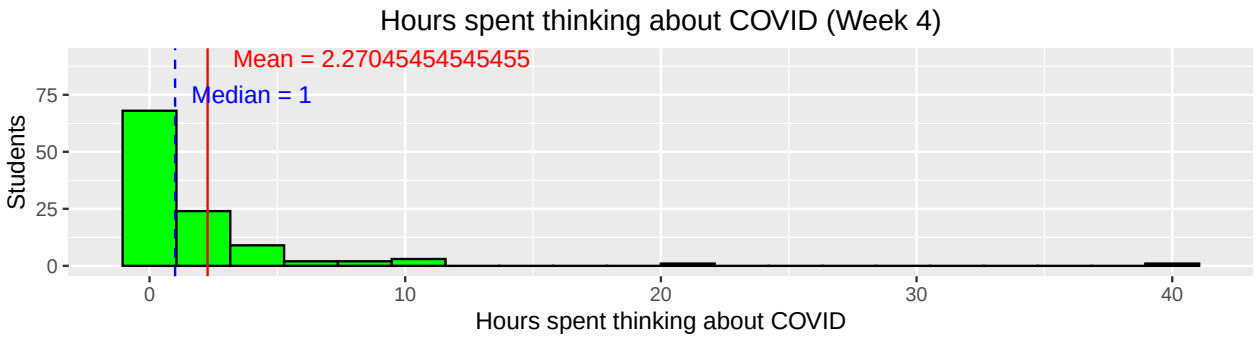
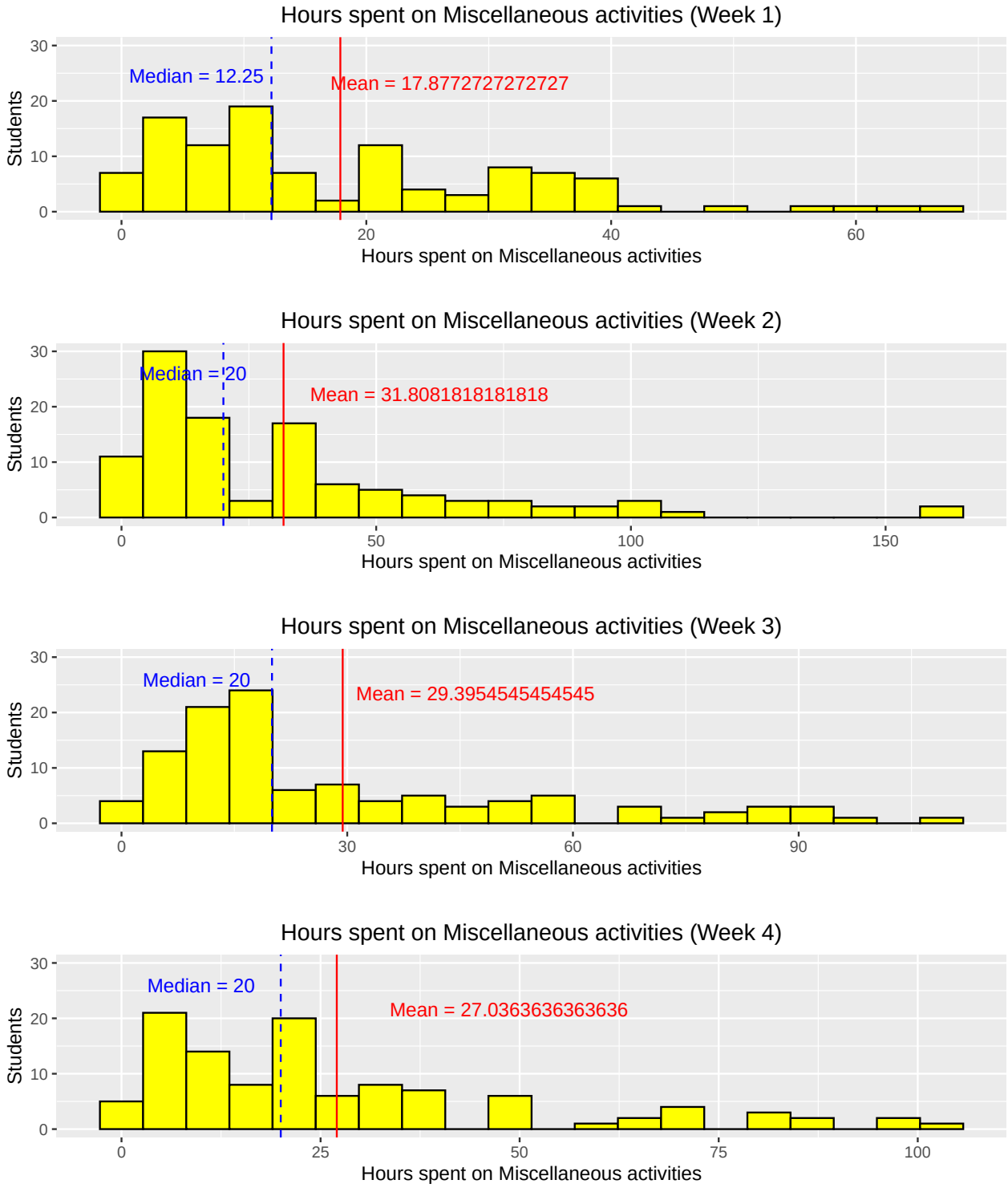

Hours spent thinking about COVID (Week 4)
Mean = 2.27045454545455
Median = 1
Students
75
50
25
0
0
10
20
30
40
Hours spent thinking about COVID

Figure 1-3: Hours spent on Miscellaneous activities



Boxplots

After examining the distribution of students' hours spent on three predictor variables, violin boxplots are implemented to more thoroughly visualize the side-by-side comparison of each week's collected data. The box plot was plotted for all continuous variables, including Study (figure 2-1 and 2-2), COVID (figure 2-3 and 2-4), and Miscellaneous (figure 2-5 and 2-6).

According to the histogram plotted above, the extreme outliers were easily noticeable because of their right-skewed distribution. Likewise, the boxplot data evidently displayed the outlying values and disrupted the effective visualization of the boxplot spread. The histogram data plot exploration revealed that these outliers were insignificant since these were abnormal data points from one or two students in exceptional cases like getting tested positive for COVID (extreme hours on covid) or getting a summer job (extreme hours on miscellaneous). Thus, outliers were excluded from the dataset and re-graphed.

Correspondingly, the violin boxplot also demonstrates the distribution concentration by the shape of the violin, where more concentration is represented by a broader outer form (used to visualize the median). A solid black line was added for each plot to present the trend of students' data changes over the weeks. Interestingly, all the boxplots suggest a somewhat positive increasing trendline over the four weeks. Although most plots have a minor upper-slant line, the study hours show an evident positive trend from week 1 to week 4. As mentioned in the histogram examination, this is assumed to have occurred due to students' motivation status over the summer. However, ironically, the trendline for figure 2-6, hours spent on miscellaneous, also displays a slightly increasing trend over the weeks. Therefore, there is a need to investigate further the relationship between the hours studied and hours spent on miscellaneous activities. This is done by developing a scatterplot in the next section of the exploratory data analysis.

Figure 2-1 : Violin boxplot on hours spent on study

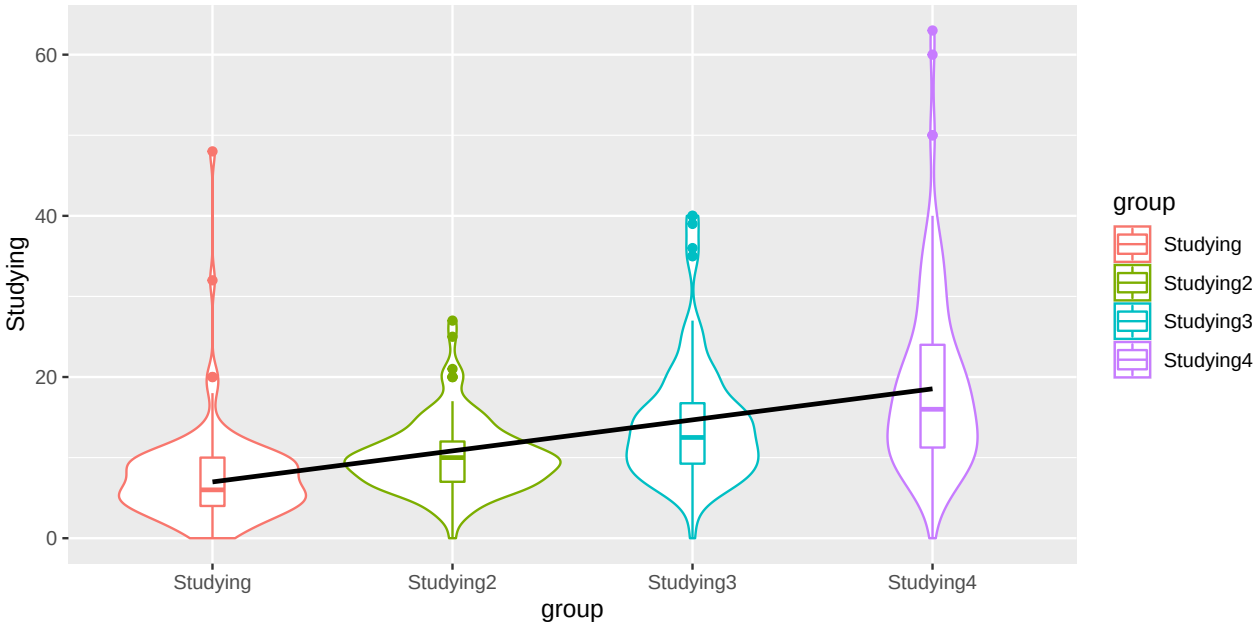


Figure 2-2 : (OUTLIERS REMOVED) Violin boxplot on hours spent on study

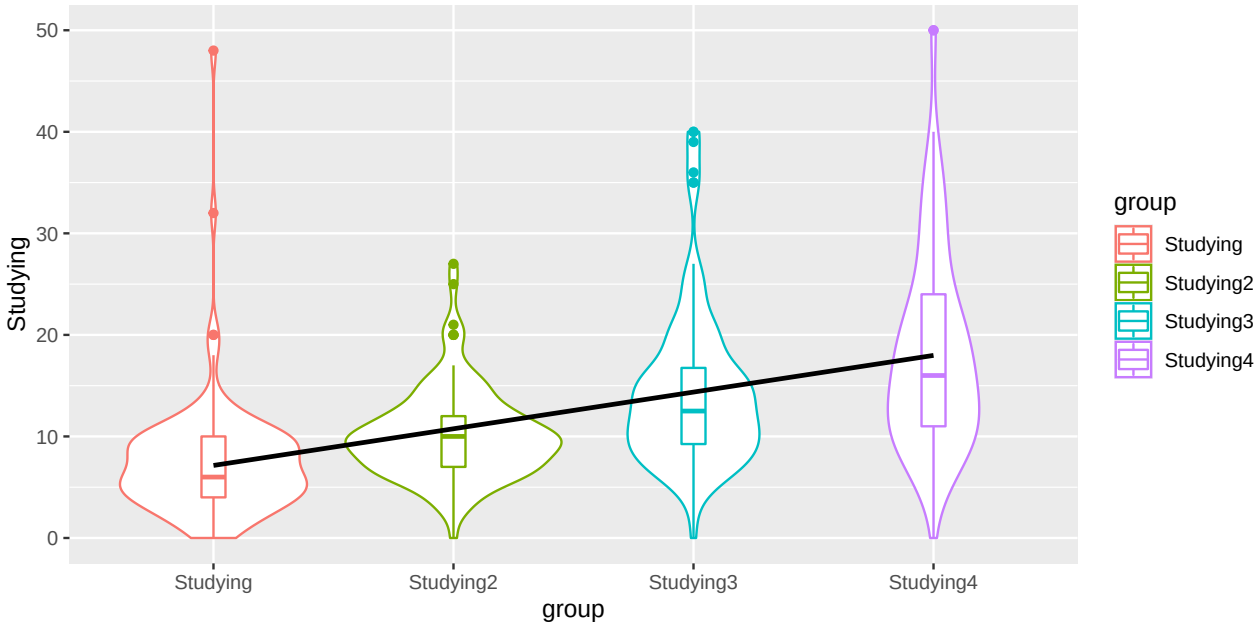


Figure 2-3 : Violin boxplot on hours spent thinking of COVID

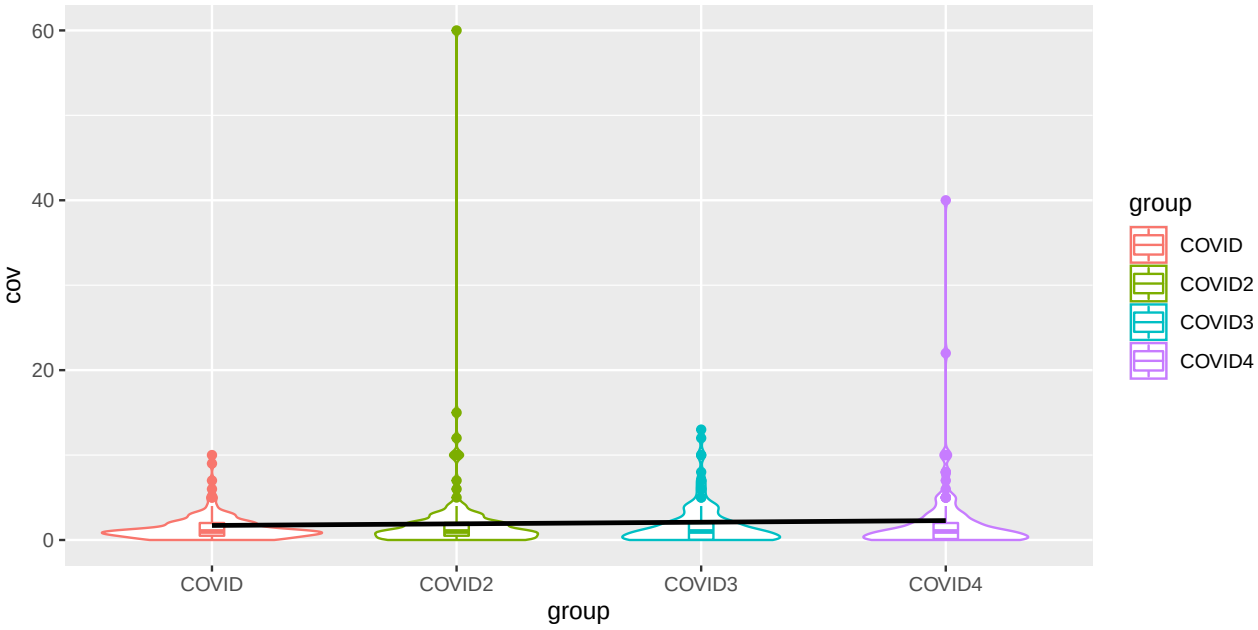


Figure 2-4 : (OUTLIERS REMOVED) Violin boxplot on hours spent thinking of COVID

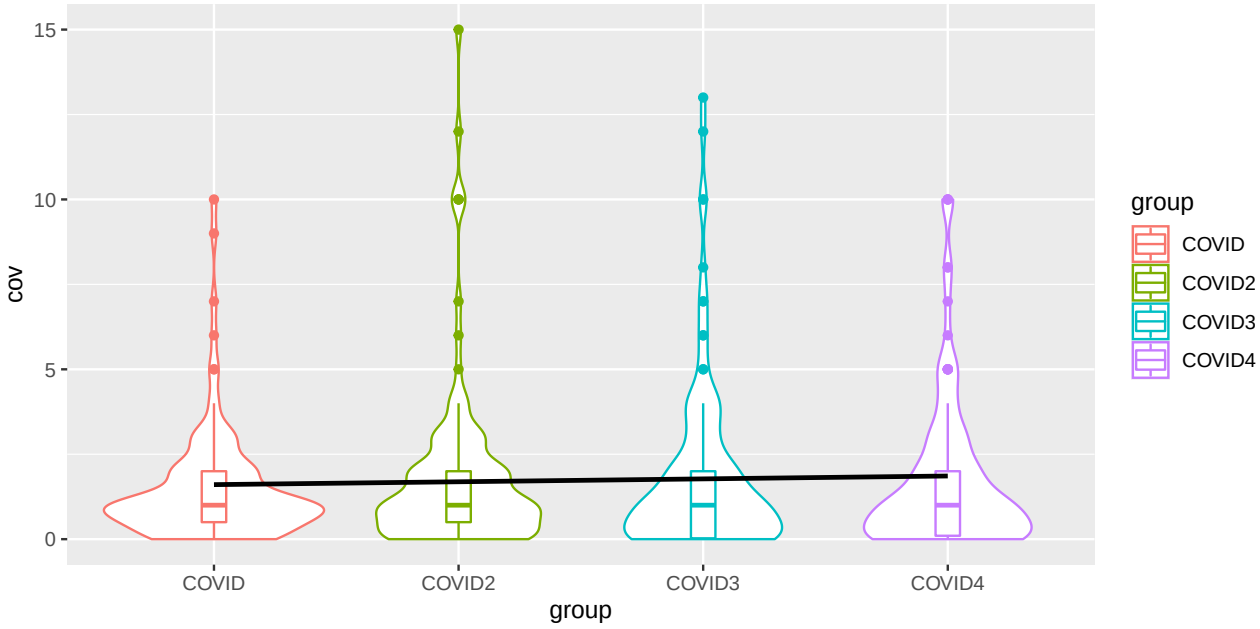


Figure 2-5 : Violin boxplot on hours spent on Miscellaneous activities

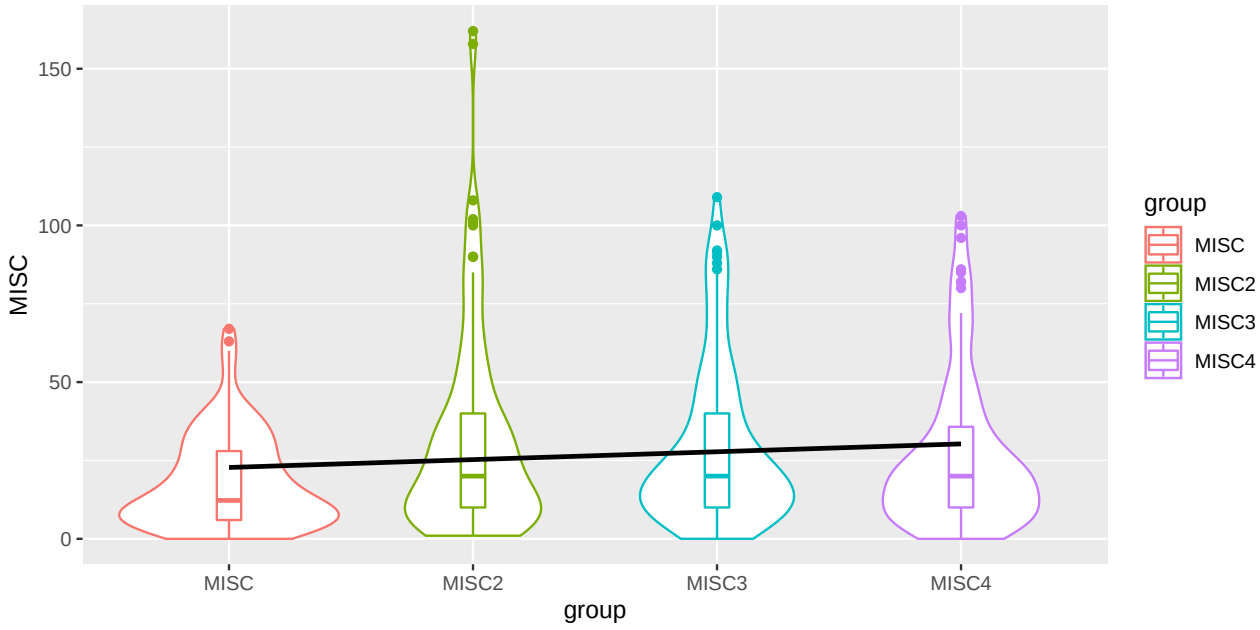
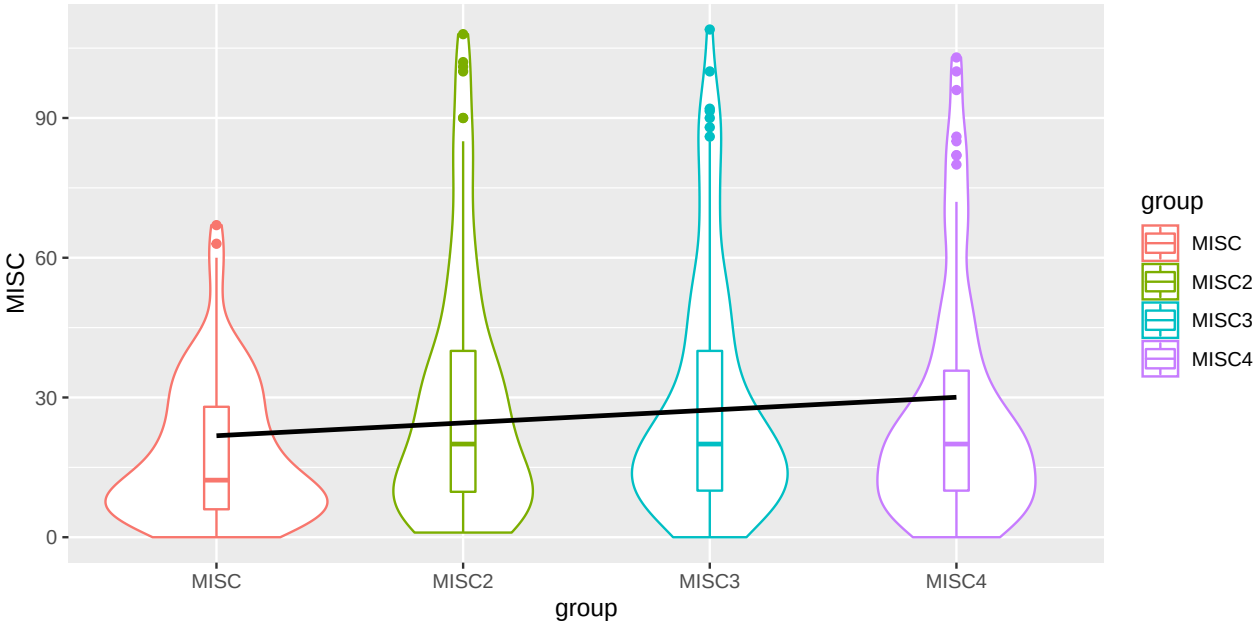


Figure 2-6 : (OUTLIERS REMOVED) Violin boxplot on hours spent on Miscellaneous activities



Scatterplots for aggregated variables

Now, the relationship between the variables is investigated using a scatterplot. The qualitative variables (OH and MICS) are also used to visualize the data points into categories further to analyze their potential correlation.

The figure 3 plots are graphed based on the categorical variable OH to distinguish how the frequency of office hour attendance has influenced the total hours spent on study, miscellaneous activity, and thinking about COVID. According to figures 3-1, the student's total hours spent studying did not seem to have been influenced by the total hours of covid related thoughts. Although most students spent less than 20 hours thinking about covid, there appears to be a slightly higher concentration of students who have never gone to the office hours thinking less of covid. The most extreme data points on covid_total that exceeded over 40 hours were both students who went to the office hours once a week. On the other hand, figure 3-3 demonstrates that students generally spent far more hours in miscellaneous activities than studying. These data points are not the best to produce valid reasoning, but it can be vaguely assumed that since the students who regularly went to in-person office hours thought significantly about COVID, this could be because they are more exposed to contact with a group of people. However, the data overall indicates most students do not spend much time concerning covid and that it did not influence their hours spent on studying and miscellaneous activities. Additionally, figure 3-2 plot shows an unexpected positive relationship between hours spent on study and hours spent on miscellaneous.

Figure 3-1: Relationship between STUDY and COVID with categorical variable OFFICE HOUR

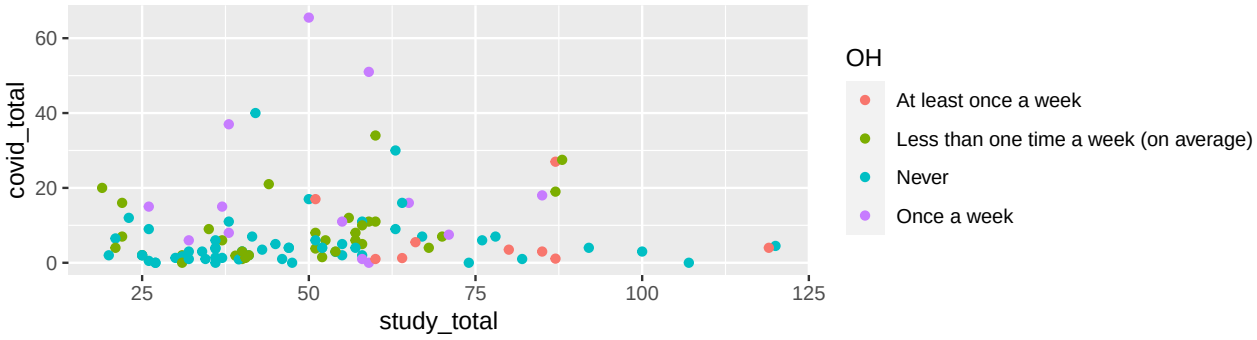


Figure 3-2: Relationship between STUDY and MISC with categorical variable OFFICE HOUR

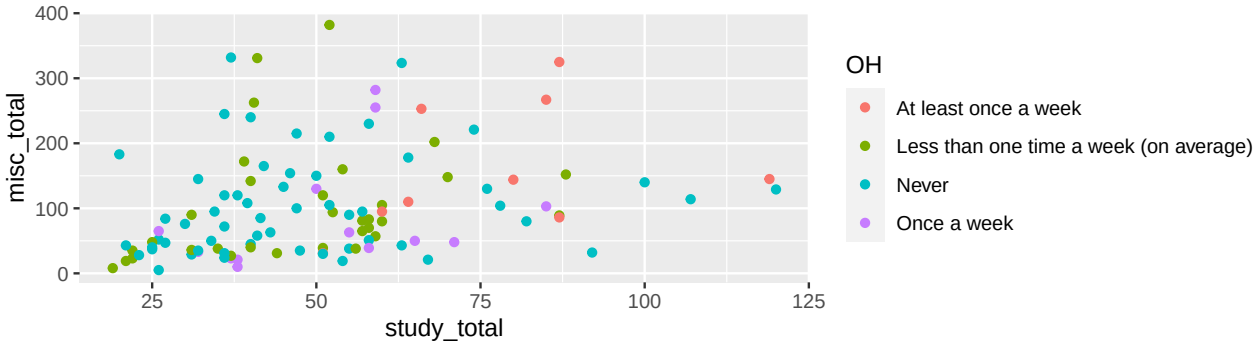


Figure 3-3: Relationship between MISC and COVID with categorical variable OFFICE HOUR

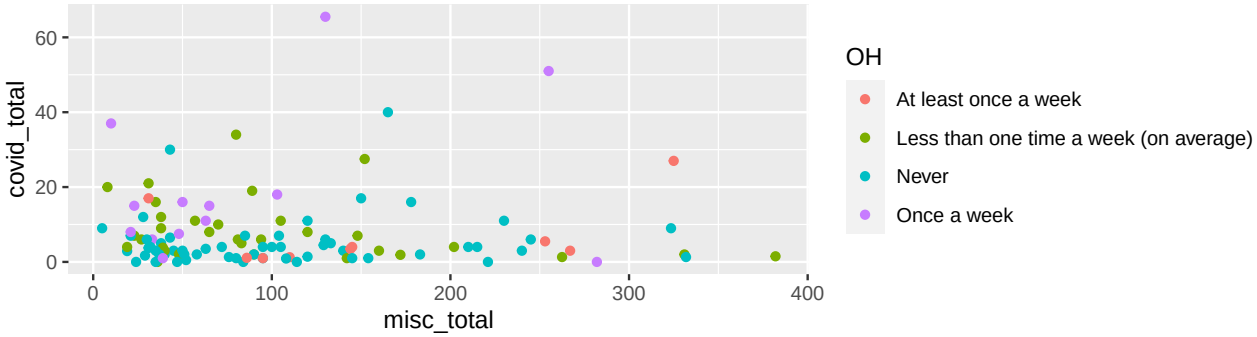


Figure 4 scatter plots explore the relationship between study, misc., COVID hours in consideration of the categorical variable, familiarity. The distribution of plots is identical to figure 3 plots since the x and y variables being examined are the same. Despite this, figure 4 uses familiarity as the qualitative variable to categorize the students. Observing the data closely, all the categories of familiarity seem to be randomly distributed, with no specific pattern being demonstrated in all three scatter plots. Since this categorical variable is highly subjective to one's standard, it is considered a dataset with low reliability.

Figure 4-1: Relationship between STUDY and COVID with categorical variable FAMILIARITY

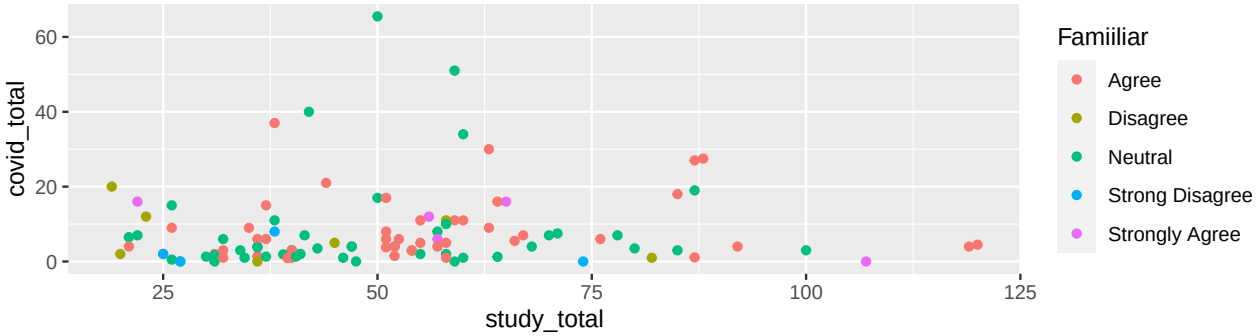


Figure 4-2: Relationship between STUDY and MISC with categorical variable FAMILIARITY

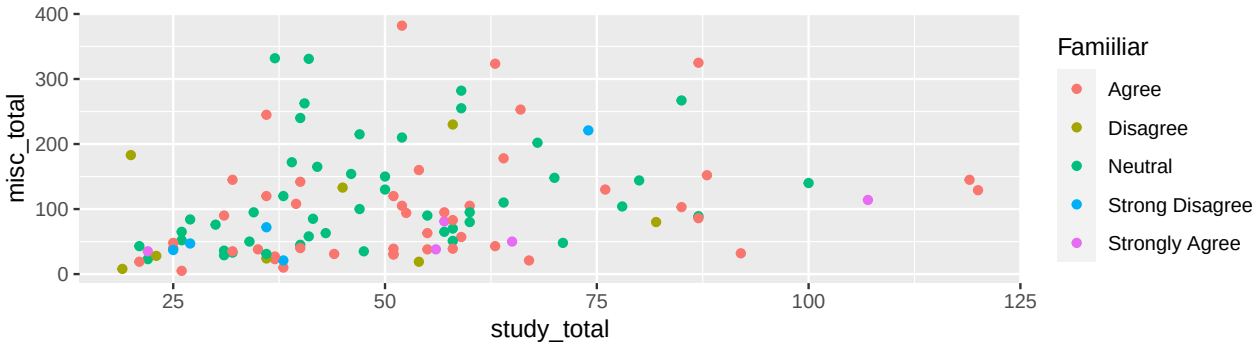


Figure 4-3: Relationship between MISC and COVID with categorical variable FAMILIARITY

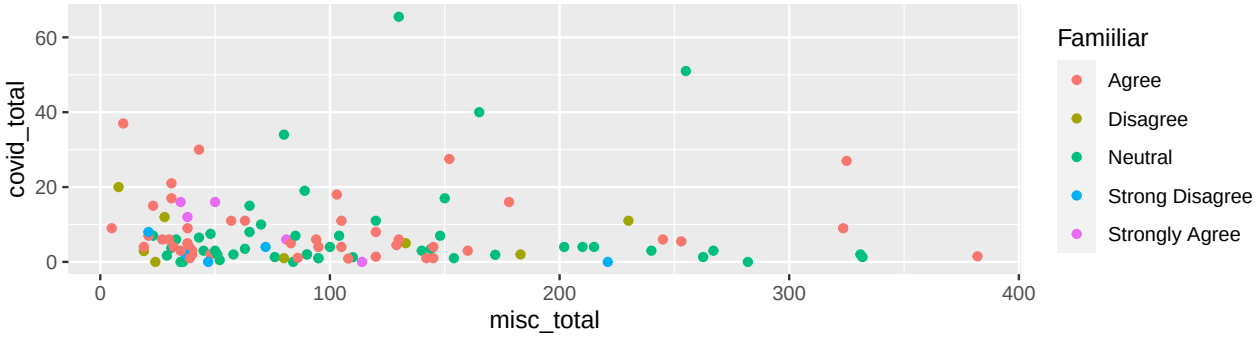


Figure 5 analyzes the correlation between students' study hours per week and the overall term test result with the categorical variable office hour. This observation is critical since study hours and attending office hours are the direct influential factors of test performance. For example, in figure 5-1, the data on week 1 is self-evident with low study hours since there isn't much course content to study in the first week. However, according to the data points from weeks 2 to 4, it can be seen that all of the students that attended office hours at least once a week studied at least 10+ hours. From this, it can be assumed that students who put in the effort to participate in office hours tend to be more passionate about this course and study more. Nonetheless, there seem to be no influential changes that office hour attendance and study have had on the term test performances of students.

Figure 5-1: Relationship between STUDY and TERM TEST with categorical variable OH

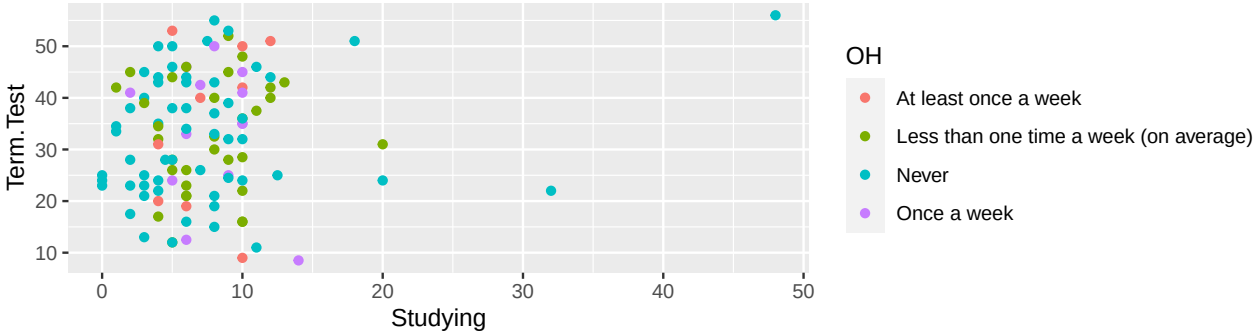


Figure 5-2: Relationship between STUDY2 and TERM TEST with categorical variable OH

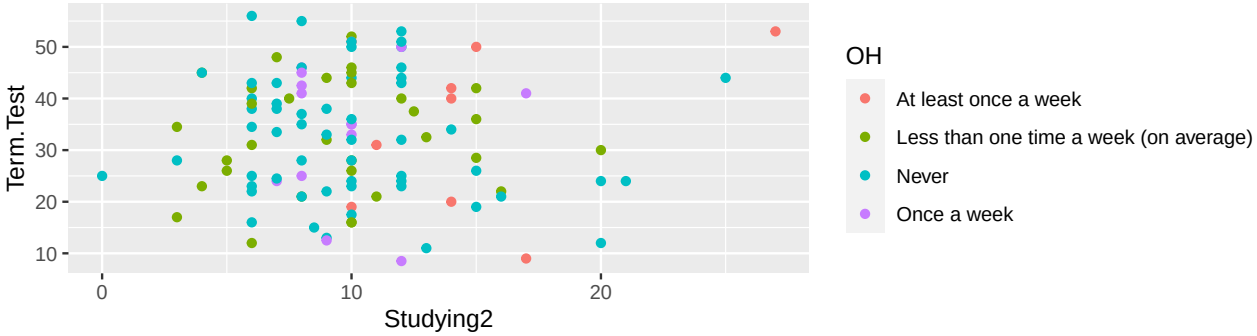


Figure 5-3: Relationship between STUDY3 and TERM TEST with categorical variable OH

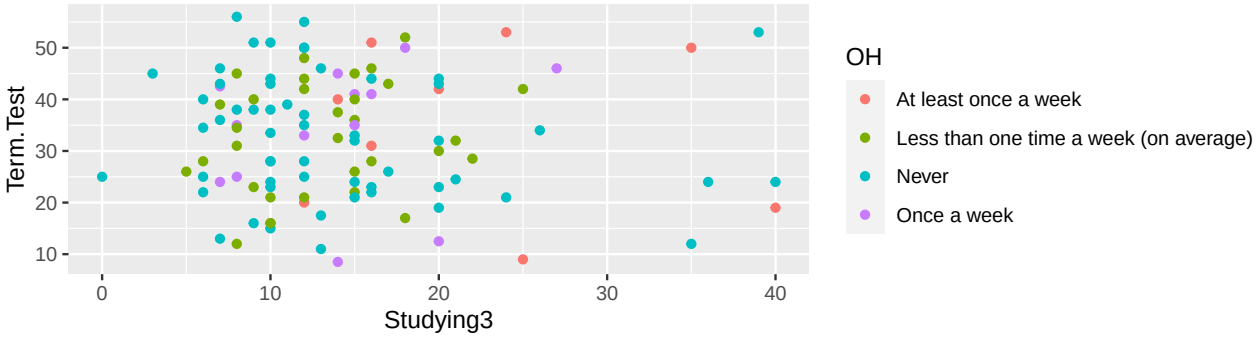


Figure 5-4: Relationship between STUDY4 and TERM TEST with categorical variable OH

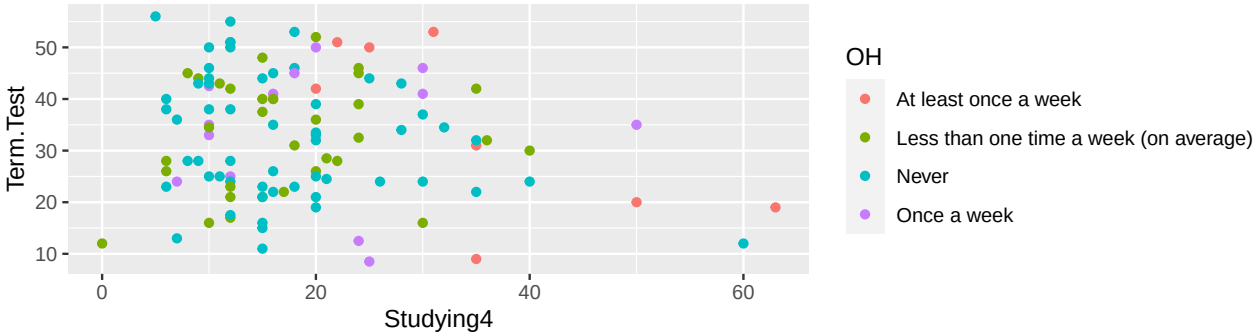


Figure 6 examines the correlation between the study and term test with familiarity data to categorize the students. The relationship between term tests and studying was discussed previously, and concluded that there seems to be no strong correlation between them. Likewise, the categorical familiarity values are an abstract data set that could vary depending on individuals, so it is difficult to detect any relationship between the two quantitative variables plotted.

Figure 6-1: Relationship between STUDY and TERM TEST with categorical variable FAMILIARITY

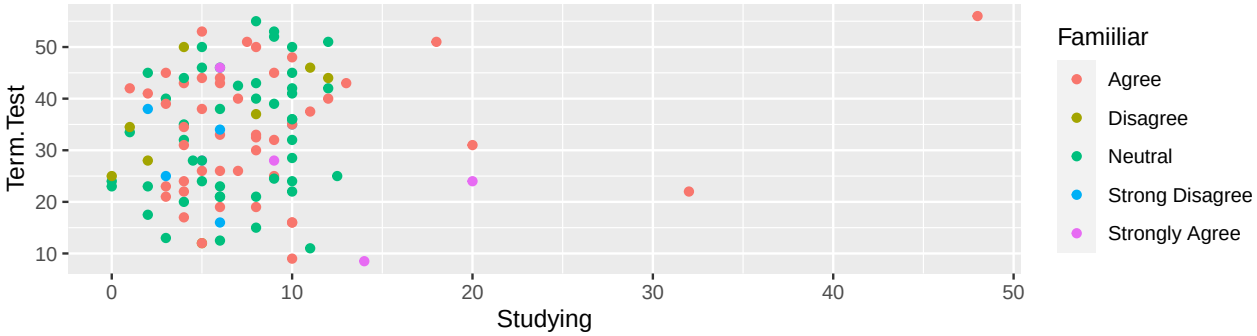


Figure 6-2: Relationship between STUDY and TERM TEST with categorical variable FAMILIARITY

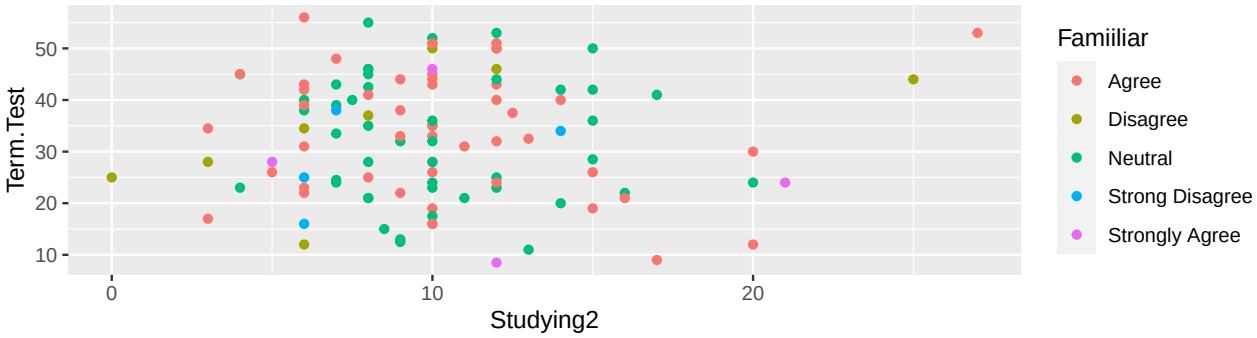


Figure 6-3: Relationship between STUDY and TERM TEST with categorical variable FAMILIARITY

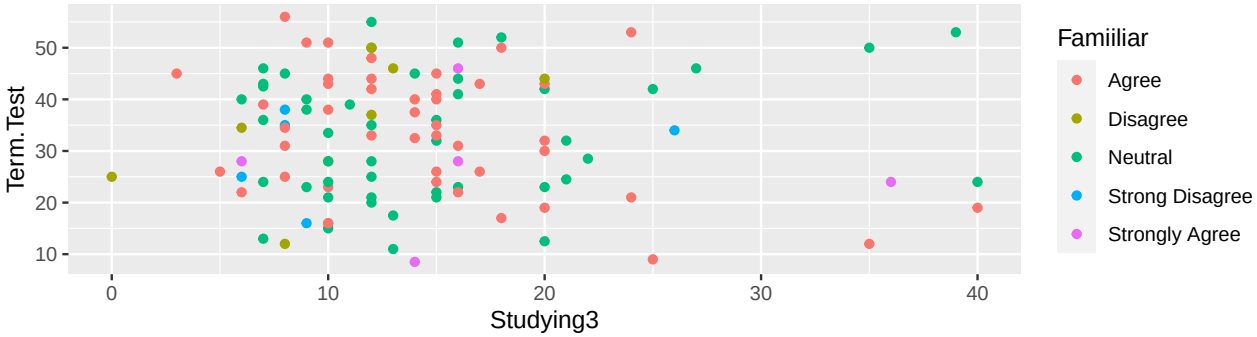
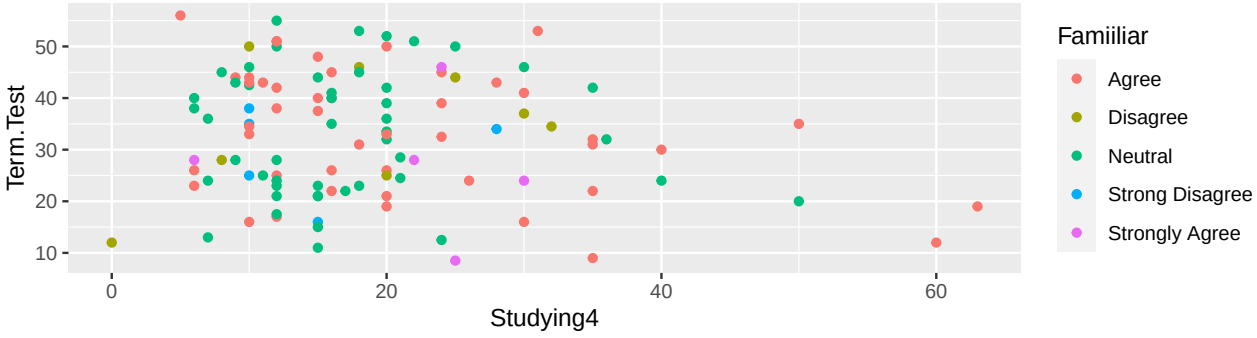


Figure 6-4: Relationship between STUDY and TERM TEST with categorical variable FAMILIARITY



Pairwise Scatterplots

Pairwise scatterplots were used to explore the interactive relationship between all variables based on each predictor and reference variable. Figure 7 represents the relationship of weekly continuous predictor variables to the term test result. Figure 7-1 compares the interaction between term test and studying, figure 7-2 compares term test and Miscellaneous, and figure 7-3 on the interaction between term test and covid. As demonstrated in the histogram and boxplot, the same pattern can also be detected from the pairwise scatterplot. The hours spent studying seem to not correlate with the term test result, but the pattern of increasing study hours over the weeks is still evident. The same applies to figures 7-2; the hours spent on miscellaneous activity did not necessarily influence the overall term test performance distribution.

Figure 8 visualizes the weekly pairwise scatterplot on four variables (term test, covid, studying, and miscellaneous) to observe their relationship. Generally, these variables lead to an assumption that as most hours are spent on miscellaneous and fewer hours on studying, the term test mark will decrease. In contrast to the hypothesis, the scatterplots calculated indicate no such correlation and, in fact, reveal the opposite pattern: scatterplot of The correlation between miscellaneous and term test represents slight positive relation, meaning students that spent more time on non-study related activities performed better on the midterm. Even the covid data plots are highly concentrated near less than 10 hours, so covid was not the students' important factor that could potentially influence their study performance unless they tested positive.

Figure 7-1: Pairwise scatterplots between term test results and studying hours each week

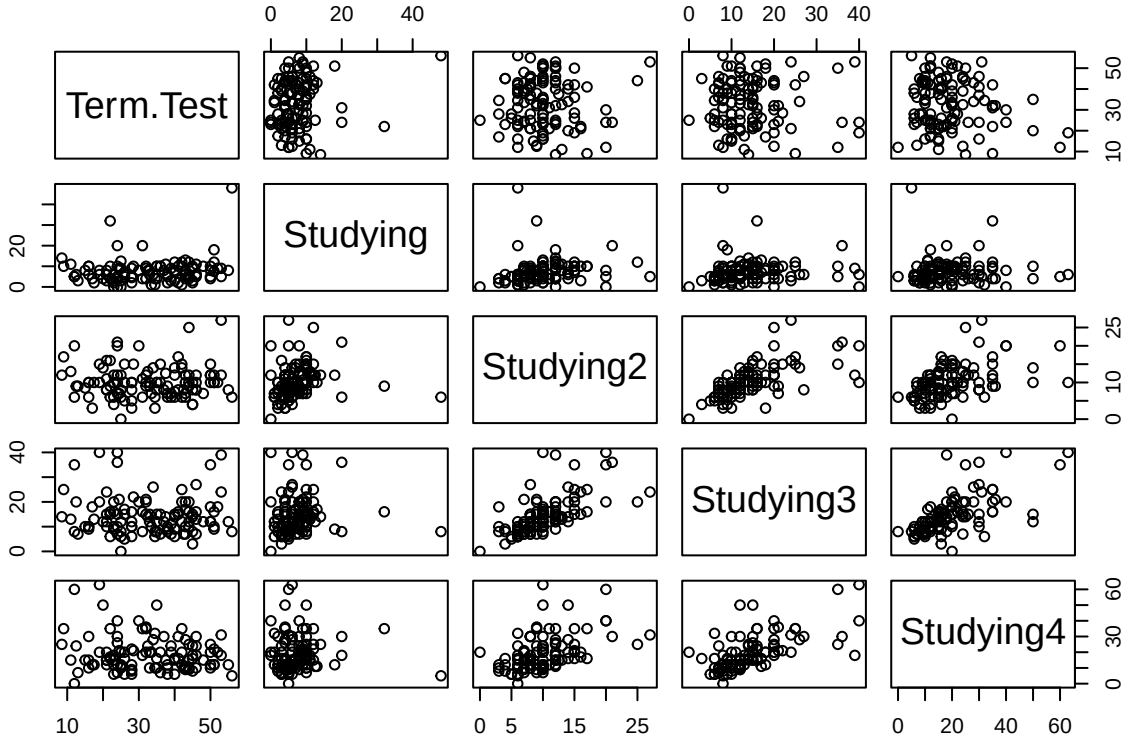


Figure 7-2: Pairwise scatterplots between term test results and miscelleneous hours each week

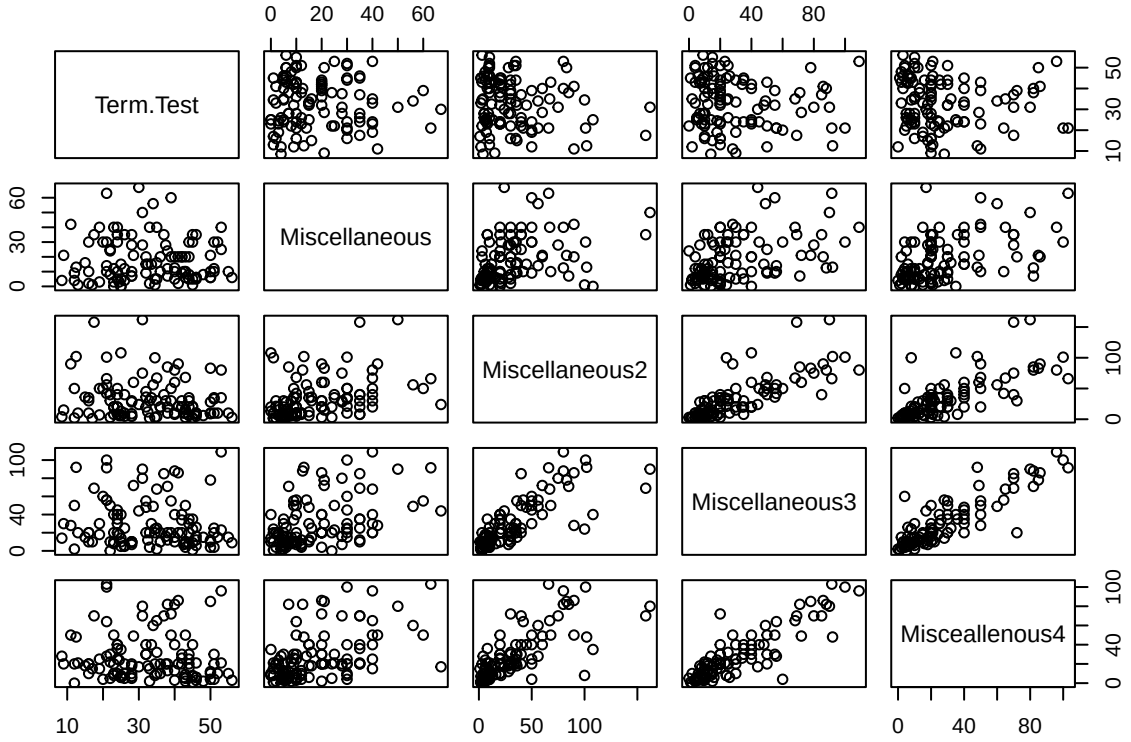
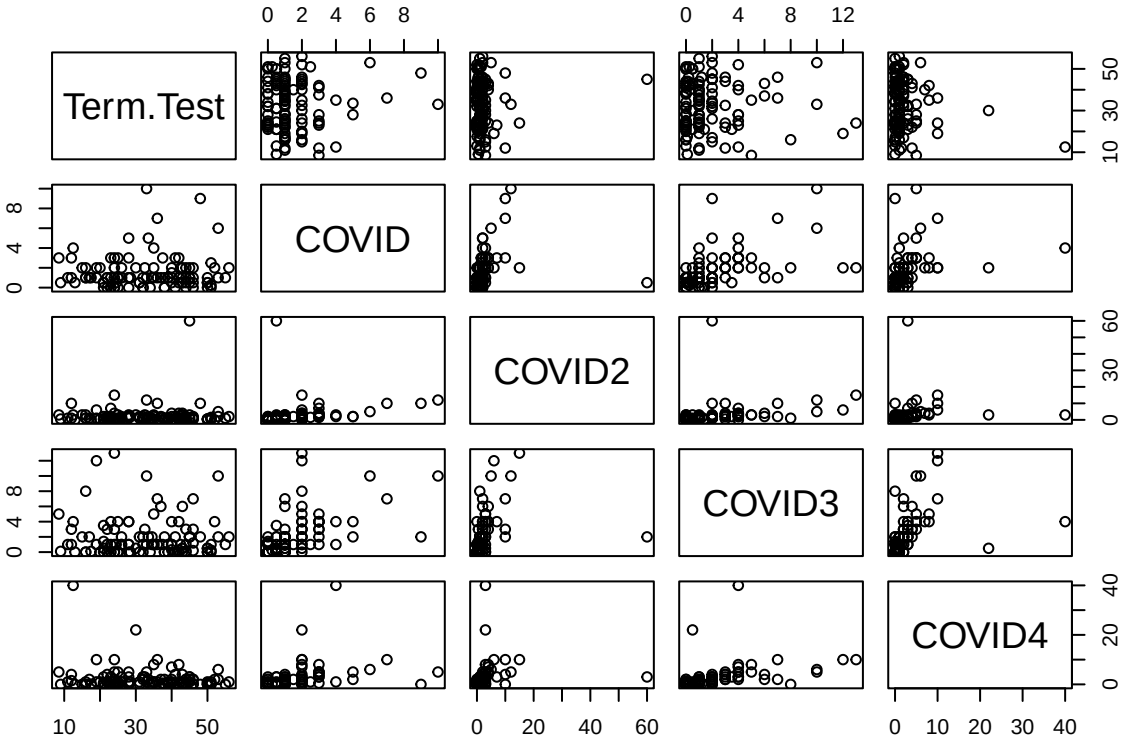


Figure 7-3: Pairwise scatterplots between term test results and COVID thinking hours each week



Pairwise scatterplots to see how studying, miscellaneous activities and thinking about COVID relate each week with term test scores

Figure 8-1: Week 1

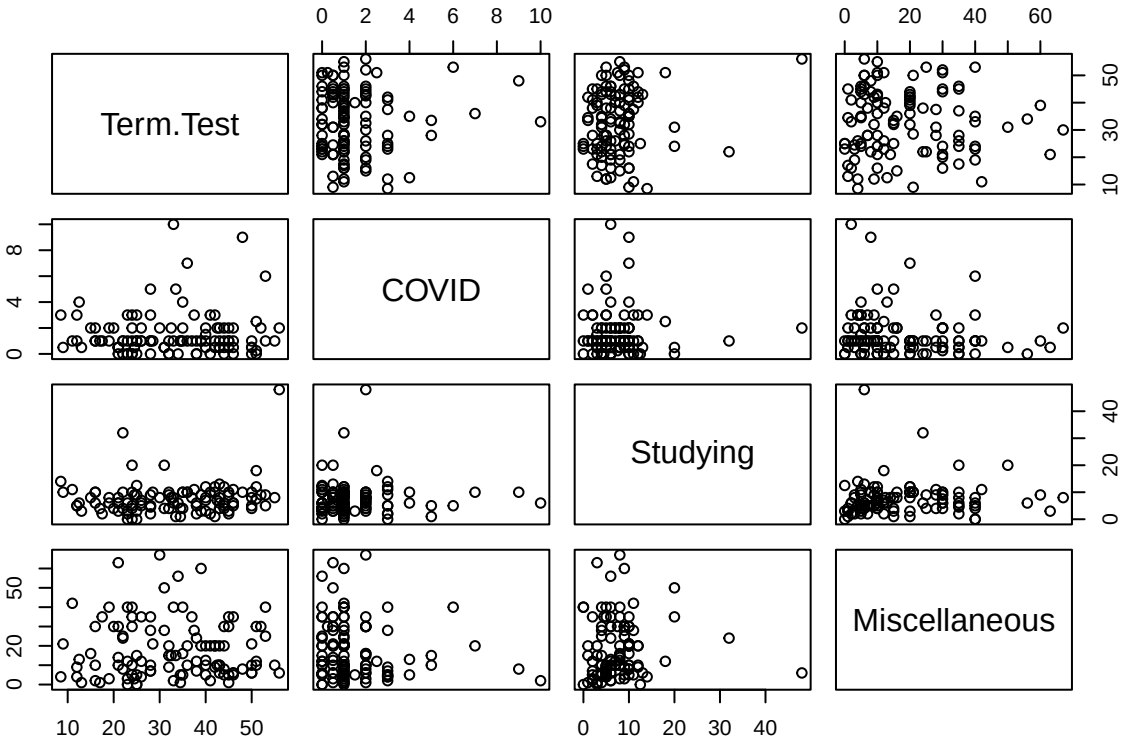


Figure 8-2: Week 2

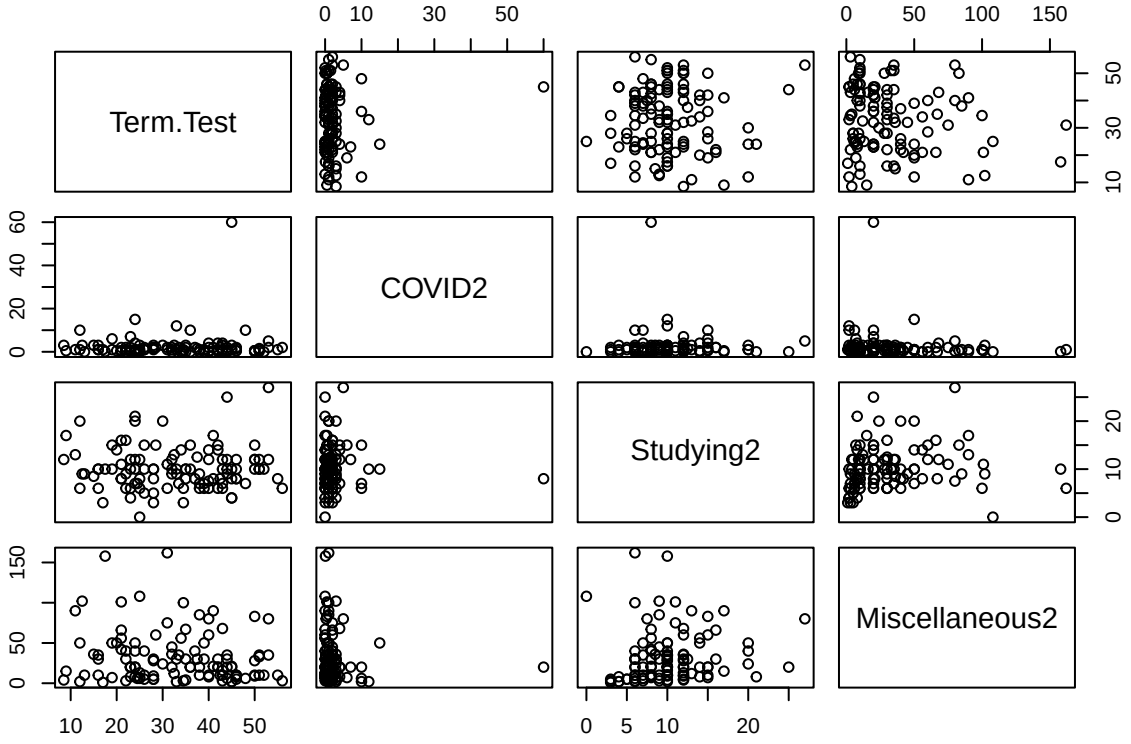


Figure 8-3: Week 3

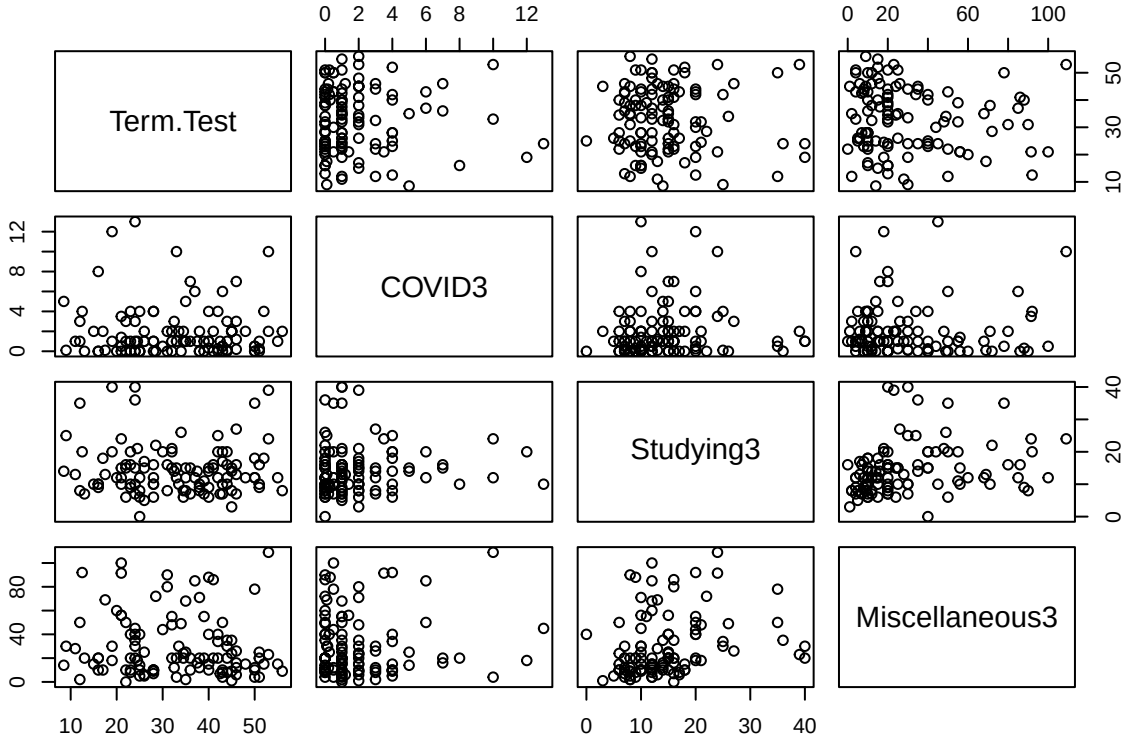
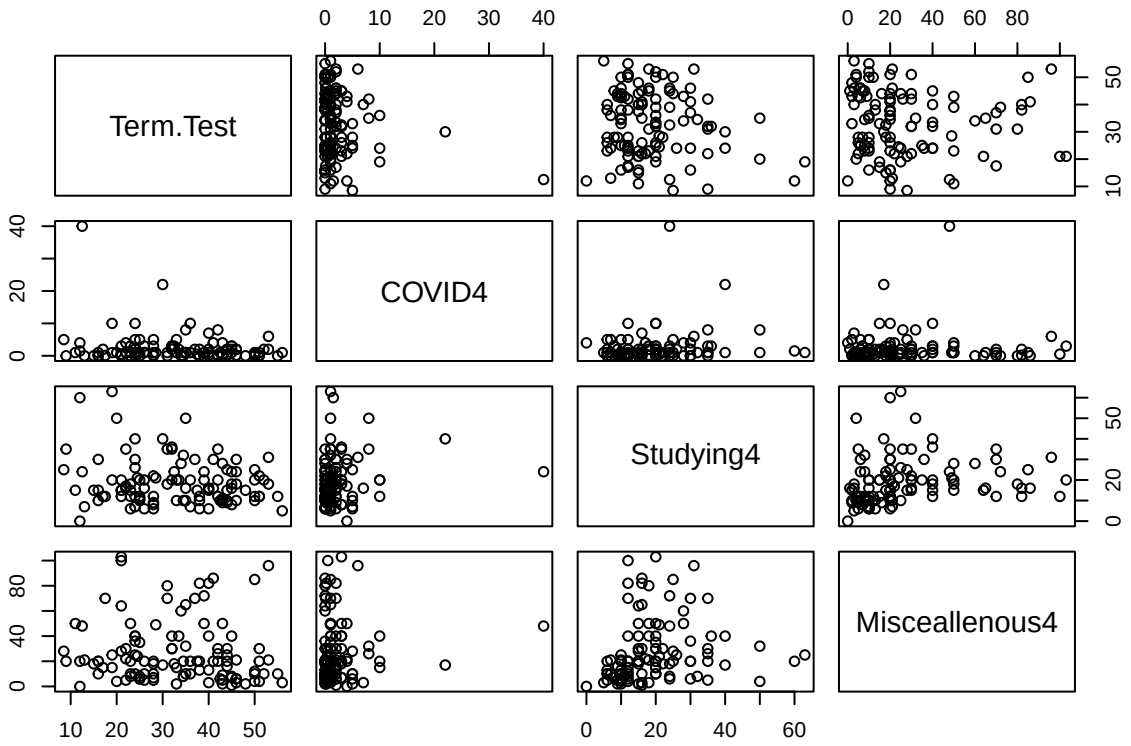


Figure 8-4: Week 4



Creating Correlation and Covariance Tables

Covariance and correlation tables for all the quantitative variables totaled for each week

Covariance Table:

Variable	Total Study Time	Total Miscellaneous Time	Total COVID time	Term Test Grade
Total Study Time	455.26013	494.86405	19.295863	-12.358590
Total Miscellaneous Time	494.86405	7134.39263	18.673654	-112.256168
Total COVID Time	19.29586	18.67365	112.021201	-5.569085
Term Test Grade	-12.35859	-112.25617	-5.569085	142.518953

Correlation Table:

Variable	Total Study Time	Total Miscellaneous Time	Total COVID time	Term Test Grade
Total Study Time	1.00000000	0.27458549	0.08544453	-0.04851799
Total Miscellaneous Time	0.27458549	1.00000000	0.02088818	-0.11132569
Total COVID Time	0.08544453	0.02088818	1.00000000	-0.04407551

Variable	Total Study Time	Total Miscellaneous Time	Total COVID time	Term Test Grade
Term Test Grade	-0.04851799	-0.11132569	-0.04407551	1.00000000

Model Development

Developing Initial Model When the graphs of studying time in different weeks against term test marks were first plotted, the patterns show uniform distributions. Additionally, the data shows that some students can study for 60 hours and only get a passing grade while other students can study for 20 hours and get a higher mark. Similarly, when the other predictor variables (i.e. miscellaneous hours, covid hours, office hour attendances, familiarity of the course) are plotted separately against the term test marks, insignificant results are shown for the data to conclude anything about linear regression relationship between the predictors and the response. Therefore, some adjustments are needed to create the initial model instead of using seemingly unrelated variables.

Categorizing students

Since the data showed a uniform distribution against studying time and term mark, it's clear some students have more effective studying than others. The initial thought to improve the dataset was to categorize the students, so at least there would be cleaner linear relationships shown between the categorized studying time and term test marks. One way to categorize students was to isolate the data for the students who said they were comfortable with the course information. The intention of this was to isolate the quality of the student in terms of STA302 knowledge. However, no relation was found at all. Furthermore, the relationship helps to conclude that optimism bias between students creates an unreliable data set for familiarity. Next, categorizing students by the frequency of office hour attendances was attempted and plotted with their studying time against the term test mark. Intuitively, students who went to office hours may have had an advantage in effective studying as they would have a better idea of the concepts to study for. Additionally, the smartest students will typically attend office hours to ensure they have a firm grasp on the material. Isolating for students who went to office hours allows us to isolate a group of students who may study equally effectively. When plotting the isolated data for the study hours of students against their term mark for students who went to office hours once a week, an almost positive polynomial relationship. Looking at the students who never went or went less than once a week, this polynomial relationship collapses. Since the data points for students who went once a week or more only made up less than a quarter of all of our available data, the method of categorizing students from the given data should be thought beyond only limiting at the interaction between students' studying time (in each of the 4 weeks) and another predictor variable. In this case, the idea for the model was to include all the predictor variables and interactions between them, and then compare against the term test mark.

Deletion of COVID n, Familiarity, and Office Hours Predictors

Although a perfect model would include collected data of all the predictor variables and the interactions between them, the reports of familiarity, frequency of going to office hours and data of the hours which people spent thinking about COVID were unconsidered in the proposed initial model. First of all, the relationships between variables of COVID thinking time, familiarity of the course, frequency of office hour attendances, and the term test mark are insignificant as shown in the exploratory data analysis section because they have no visible relationship against the response variable (term test marks). Secondly, the familiarity is removed because of the existence of subjectivity data bias in the data collection process. Students reported how familiar they were with STA302 knowledge before the term test, but there was a high possibility that students misjudged their understanding of the course. Basically, students could be overconfident, or they could underestimate themselves, making this variable subject to bias and therefore too inconsistent to be taken into consideration. Next, the amount of time that students took to think about COVID was taken away from the consideration as well because intuitively people thinking about COVID would not influence

much on the term test mark they get. There might be a relationship between them if students think about COVID for a sufficient amount of time, for example 20 hours, because this could represent that the student caught COVID. However, from the histogram and box-plot in exploratory data analysis, it is obvious that there is an insufficient amount of data (only three students think about COVID for longer than 20 hours in four weeks) to support the representation of students getting COVID. Last but not the least, the attendance of office hours is taken out of consideration because the reason to attend office hours is totally different for each person, for example one student may take office hours at the last minute because they are struggling but there are many students who go to office hours only to attain a firm grasp on the material to guarantee a high mark in the course. Additionally, people who study better individually could choose not to attend office hours but still get high marks. Overall, it is better to remove these three predictor variables as they are too unstable to be a prediction of the response variable.

Improving the Initial Model

After removing the unwanted predictor variables, a linear regression model with the response variable as test marks and the predictors as studying per week and extracurriculars per week is produced. Additionally, the interaction between each of the predictor variables are included as well in order to control for the collinearities between the variables, and overcome the possible collinearities later while carrying out the backward elimination method. In order to successfully categorize the predictors, the choice of multi-linear regression model is made for the initial model. Also, as the intersection (or interaction) variables are added to overcome the problems of multi-collinearities, the amount of beta estimates in this multi-linear regression model increases exponentially. The initial second order, eight predictor variable multi-linear regression model used is summarized below.

The Initial Model

Variable	Estimate	Std. Error	t value	Pr(>
(Intercept)	21.1970226	8.4221758	2.517	0.0143 *
Studying	-0.3231040	1.3973585	-0.231	0.8179
Studying2	0.6676936	1.5636652	0.427	0.6708
Studying3	-0.1905920	1.1101385	-0.172	0.8642
Studying4	1.9272373	0.7733660	2.492	0.0153 *
Miscellaneous	0.7624838	0.3999584	1.906	0.0610 .
Miscellaneous2	-0.6372560	0.4267838	-1.493	0.1402
Miscellaneous3	-0.1971204	0.5892248	-0.335	0.7390
Misceallenous4	-0.2504582	0.6168768	-0.406	0.6861
I(Studying * Studying3)	0.0450728	0.0960846	0.469	0.6406
I(Studying * Studying4)	0.0041504	0.0424564	0.098	0.9224
I(Studying * Studying2)	-0.0977821	0.1491825	-0.655	0.5145
I(Studying2 * Studying3)	-0.0197413	0.1481213	-0.133	0.8944
I(Studying3 * Studying4)	0.0282476	0.0290833	0.971	0.3350
I(Studying2 * Studying4)	-0.1571206	0.0981474	-1.601	0.1143
I(Miscellaneous * Miscellaneous2)	-0.0216925	0.0087857	-2.469	0.0162 *
I(Miscellaneous2 * Miscellaneous3)	-0.0025590	0.0104239	-0.245	0.8068

Variable	Estimate	Std. Error	t value	Pr(>
I(Miscellaneous3 * Misceallenous4)	-0.0034136	0.0128030	-0.267	0.7906
I(Miscellaneous2 * Misceallenous4)	0.0152893	0.0119049	1.284	0.2036
I(Miscellaneous * Misceallenous4)	-0.0048088	0.0155090	-0.310	0.7575
I(Miscellaneous * Miscellaneous3)	0.0218751	0.0149977	1.459	0.1495
I(Miscellaneous^2)	-0.0002105	0.0076165	-0.028	0.9780
I(Miscellaneous2^2)	-0.0010068	0.0023878	-0.422	0.6747
I(Miscellaneous3^2)	-0.0069918	0.0116794	-0.599	0.5515
I(Misceallenous4^2)	0.0032586	0.0061121	0.533	0.5958
I(Studying^2)	0.0099070	0.0208313	0.476	0.6360
I(Studying2^2)	0.1170590	0.0814124	1.438	0.1553
I(Studying3^2)	-0.0050660	0.0340553	-0.149	0.8822
I(Studying4^2)	-0.0157370	0.0145471	-1.082	0.2833
I(Studying * Miscellaneous)	0.0060204	0.0380086	0.158	0.8746
I(Studying * Miscellaneous2)	-0.0046975	0.0213582	-0.220	0.8266
I(Studying * Miscellaneous3)	0.0697122	0.0548176	1.272	0.2080
I(Studying * Misceallenous4)	-0.0316169	0.0589049	-0.537	0.5933
I(Studying2 * Miscellaneous)	0.0387050	0.0498809	0.776	0.4406
I(Studying2 * Miscellaneous2)	-0.0031137	0.0428503	-0.073	0.9423
I(Studying2 * Miscellaneous3)	0.0755887	0.0571122	1.324	0.1903
I(Studying2 * Misceallenous4)	-0.0662715	0.0682564	-0.971	0.3352
I(Studying3 * Miscellaneous)	0.0150497	0.0364637	0.413	0.6812
I(Studying3 * Miscellaneous2)	0.0498220	0.0312500	1.594	0.1157
I(Studying3 * Miscellaneous3)	-0.0777485	0.0387462	-2.007	0.0490 *
I(Studying3 * Misceallenous4)	0.0098106	0.0280987	0.349	0.7281
I(Studying4 * Miscellaneous)	-0.0566401	0.0241911	-2.341	0.0223 *
I(Studying4 * Miscellaneous2)	0.0073394	0.0161141	0.455	0.6503
I(Studying4 * Miscellaneous3)	0.0178784	0.0186933	0.956	0.3424
I(Studying4 * Misceallenous4)	0.0096597	0.0163914	0.589	0.5577

R-code returns the statistics of the chosen initial models, which includes the coefficients, standard errors, t-values and p-values of the predictors. For the first column on the left is the title name of the predictors, which just like the graphs shown in the exploratory data analysis the names are adjusted from the data collection process so that the model does not look messy. Studying and miscellaneous means time spent studying and time spent on miscellaneous activities respectively, the numbers at the end of the name represents the week of data collection (if there is no number behind the name then it means week 1). The coefficients represent the mean change against the term test marks with the other predictor variable being held constant, in other words it represents the average change of each possible category of student. Last but not the least, the I(x1,x2) indicates the interaction between the two predictor variables and how the interaction is related to the term test marks.

Backward Elimination Method

The backward elimination method is chosen to implement and improve the initial model in order to find the final model. We are able to use this method effectively because we minimized predictor variables when creating our initial model through empirical reasoning. Therefore, comparing to forward method or the method of comparing the adjusted r between each one of the predictor variables where excessive amount of single regression models have to be made for the analysis, backward elimination selects the data efficiently by deleting the insignificant predictors one by one and reach our final model directly and transparently.

The backward elimination method starts by looking at p-values of the initial model, as it is obvious that different predictor variables contain different p-values. Firstly the predictor that has the highest p-value was removed. This was repeated until no p-values exceeded 0.05 and had less than * or a 95% significance level. In the case of the initial model, the interaction term between the time spent on miscellaneous activities and itself in week one had the highest p-value (0.978) which also exceeds over 0.5. The first iteration towards a final model is seen below.

1st iteration of removing highest p value until all <0.05 (removed interaction term between misc1 and misc1), Second Model

Variable	Estimate	Std. Error	t value	Pr(>
(Intercept)	21.2175678	8.3255637	2.548	0.0132 *
Studying	-0.3277725	1.3765733	-0.238	0.8125
Studying2	0.6683734	1.5515912	0.431	0.6680
Studying3	-0.1936831	1.0960986	-0.177	0.8603
Studying4	1.9279914	0.7670116	2.514	0.0144 *
Miscellaneous	0.7622154	0.3968022	1.921	0.0591 .
Miscellaneous2	-0.6381537	0.4223126	-1.511	0.1355
Miscellaneous3	-0.1957699	0.5827338	-0.336	0.7380
Misceallenous4	-0.2518103	0.6102616	-0.413	0.6812
I(Studying * Studying3)	0.0453730	0.0947433	0.479	0.6336
I(Studying * Studying4)	0.0040306	0.0419138	0.096	0.9237
I(Studying * Studying2)	-0.0972645	0.1468782	-0.662	0.5101
I(Studying2 * Studying3)	-0.0198340	0.1469581	-0.135	0.8931
I(Studying3 * Studying4)	0.0282820	0.0288358	0.981	0.3303
I(Studying2 * Studying4)	-0.1572484	0.0972935	-1.616	0.1108
I(Miscellaneous * Miscellaneous2)	-0.0216698	0.0086808	-2.496	0.0151 *
I(Miscellaneous2 * Miscellaneous3)	-0.0025643	0.0103430	-0.248	0.8050
I(Miscellaneous3 * Misceallenous4)	-0.0034368	0.0126784	-0.271	0.7872
I(Miscellaneous2 * Misceallenous4)	0.0152575	0.0117590	1.298	0.1990
I(Miscellaneous * Misceallenous4)	-0.0049021	0.0150217	-0.326	0.7452
I(Miscellaneous * Miscellaneous3)	0.0218219	0.0147603	1.478	0.1441
I(Miscellaneous2^2)	-0.0009945	0.0023284	-0.427	0.6707

Variable	Estimate	Std. Error	t value	Pr(>
I(Miscellaneous3^2)	-0.0069431	0.0114576	-0.606	0.5466
I(Miseallenous4^2)	0.0032855	0.0059883	0.549	0.5851
I(Studying^2)	0.0099156	0.0206707	0.480	0.6330
I(Studying2^2)	0.1171740	0.0806883	1.452	0.1512
I(Studying3^2)	-0.0050315	0.0337740	-0.149	0.8820
I(Studying4^2)	-0.0156643	0.0141991	-1.103	0.2740
I(Studying * Miscellaneous)	0.0060101	0.0377179	0.159	0.8739
I(Studying * Miscellaneous2)	-0.0045536	0.0205570	-0.222	0.8254
I(Studying * Miscellaneous3)	0.0696826	0.0543907	1.281	0.2046
I(Studying * Miseallenous4)	-0.0318735	0.0577268	-0.552	0.5827
I(Studying2 * Miscellaneous)	0.0383017	0.0473362	0.809	0.4213
I(Studying2 * Miscellaneous2)	-0.0029605	0.0421675	-0.070	0.9442
I(Studying2 * Miscellaneous3)	0.0757139	0.0564997	1.340	0.1848
I(Studying2 * Miseallenous4)	-0.0663752	0.0676352	-0.981	0.3300
I(Studying3 * Miscellaneous)	0.0151807	0.0358798	0.423	0.6736
I(Studying3 * Miscellaneous2)	0.0497629	0.0309398	1.608	0.1125
I(Studying3 * Miscellaneous3)	-0.0779713	0.0376106	-2.073	0.0421 *
I(Studying3 * Miseallenous4)	0.0100388	0.0266551	0.377	0.7077
I(Studying4 * Miscellaneous)	-0.0568242	0.0230801	-2.462	0.0164 *
I(Studying4 * Miscellaneous2)	0.0072800	0.0158491	0.459	0.6475
I(Studying4 * Miscellaneous3)	0.0178756	0.0185510	0.964	0.3388
I(Studying4 * Miseallenous4)	0.0097630	0.0158384	0.616	0.5397

Continues the same process until all the predictors (and the interaction of predictors) contain p-values < 0.5. That is when the final model of this research is reached.

The Final Model

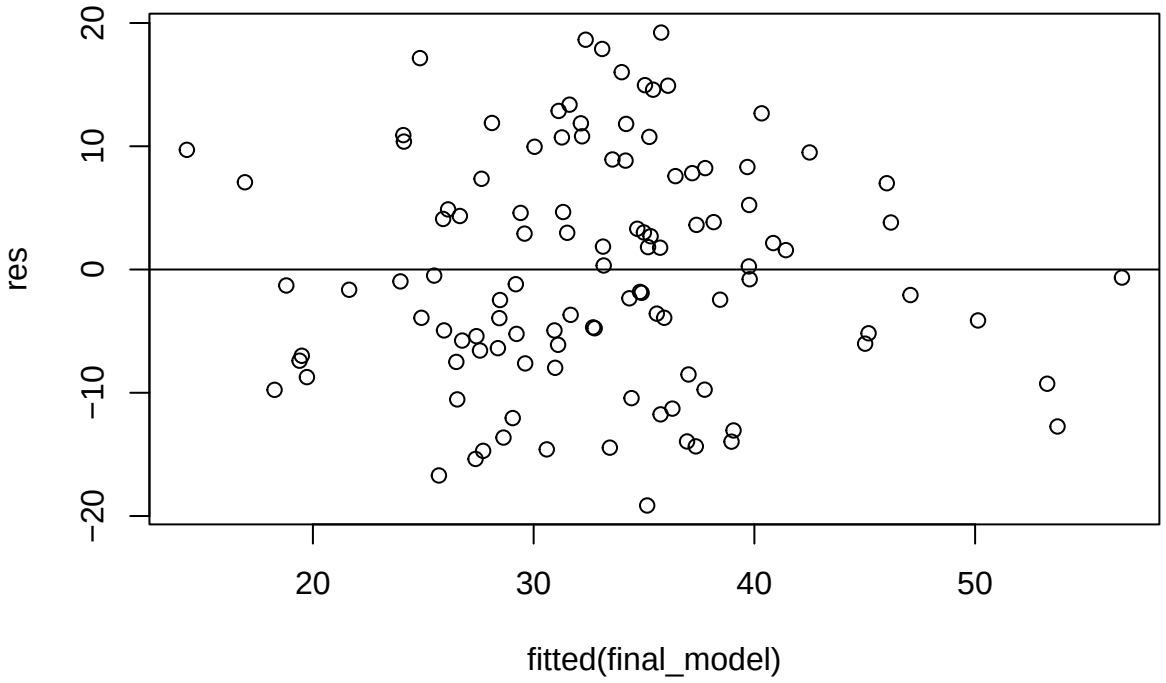
Variable	Estimate	Std. Error	t value	Pr(>
(Intercept)	19.917103	4.134334	4.817	5.78e-06 ***
Studying4	1.663136	0.379718	4.380	3.17e-05 ***
Miscellaneous	0.800071	0.217471	3.679	0.000396 ***
Miscellaneous2	−0.611032	0.123671	−4.941	3.52e-06 ***
I(Studying * Studying2)	−0.084651	0.034590	−2.447	0.016315 *
I(Studying2 * Studying4)	−0.173602	0.031839	−5.453	4.23e-07 ***
I(Miscellaneous * Miscellaneous2)	−0.021017	0.004710	−4.462	2.32e-05 ***

Variable	Estimate	Std. Error	t value	Pr(>
I(Miscellaneous2 * Misceallenous4)	0.011038	0.002480	4.452	2.41e-05 ***
I(Miscellaneous * Misceallenous4)	0.013797	0.005434	2.539	0.012816 *
I(Miscellaneous3^2)	−0.009222	0.002282	−4.042	0.000111 ***
I(Studying^2)	0.014218	0.006223	2.285	0.024658 *
I(Studying2^2)	0.183627	0.039721	4.623	1.24e-05 ***
I(Studying * Miscellaneous3)	0.036260	0.009542	3.800	0.000261 ***
I(Studying2 * Miscellaneous3)	0.077177	0.020034	3.852	0.000218 ***
I(Studying2 * Misceallenous4)	−0.072848	0.018909	−3.852	0.000217 ***
I(Studying3 * Miscellaneous2)	0.040709	0.009627	4.229	5.59e-05 ***
I(Studying3 * Miscellaneous3)	−0.043003	0.012288	−3.500	0.000723 ***
I(Studying4 * Miscellaneous)	−0.019604	0.006598	−2.971	0.003794 **
I(Studying4 * Miscellaneous3)	0.014171	0.006473	2.189	0.031154 *

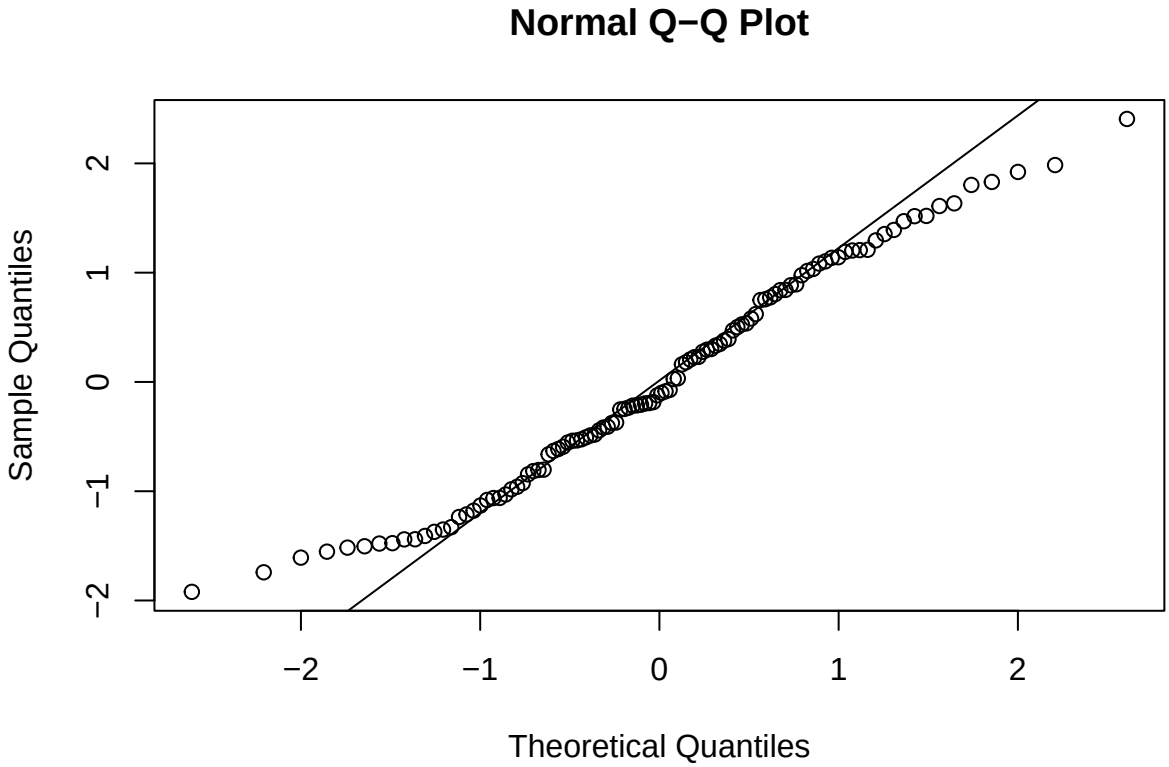
The reason that the p-values that are higher (and highest) than 0.05 are eliminated for each step of selection is because in this way the predictors that are insignificant could be removed from the final model; more specifically, p-values 0.05 indicates that the estimates of the relationship between each of the predictors and the term test marks is 95% significant as shown by the coefficients. For example, the final model is 95% confident that the average increase of studying 1 hour on week 4 would increase the term test mark by approximately 1.663 when the miscellaneous time spent on week 1, 2 and the interaction variables are constant. A Similar interpretation could be reached for the relationship between miscellaneous time spent on week 2 and term test mark. A negative relationship is found when an increase of 1 hour of miscellaneous time spent decreases the term test mark by -0.611. However, note that these three predictors are only significant enough to be used to predict the term test marks when the interactions between the predictors are taken into consideration. The visual posted above summarizes the interpretation and statistics of the chosen final model. The code of the initial model and final model are posted in the Appendices section, and the selection process of backward elimination is done manually. The individual steps of removing the variables starting from the initial model are listed in comments, read left to right then top to down.

Final Model Diagnostics

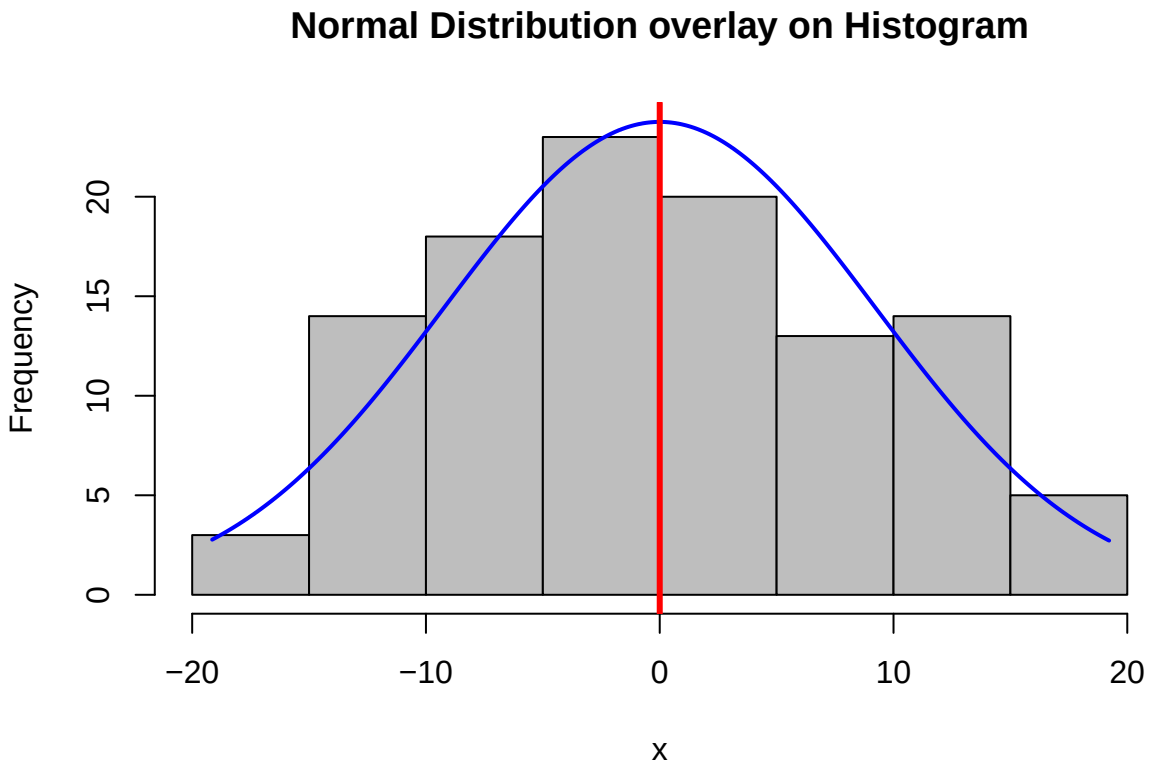
Residuals vs fitted values of final model



Standard residuals in Q-Q plot

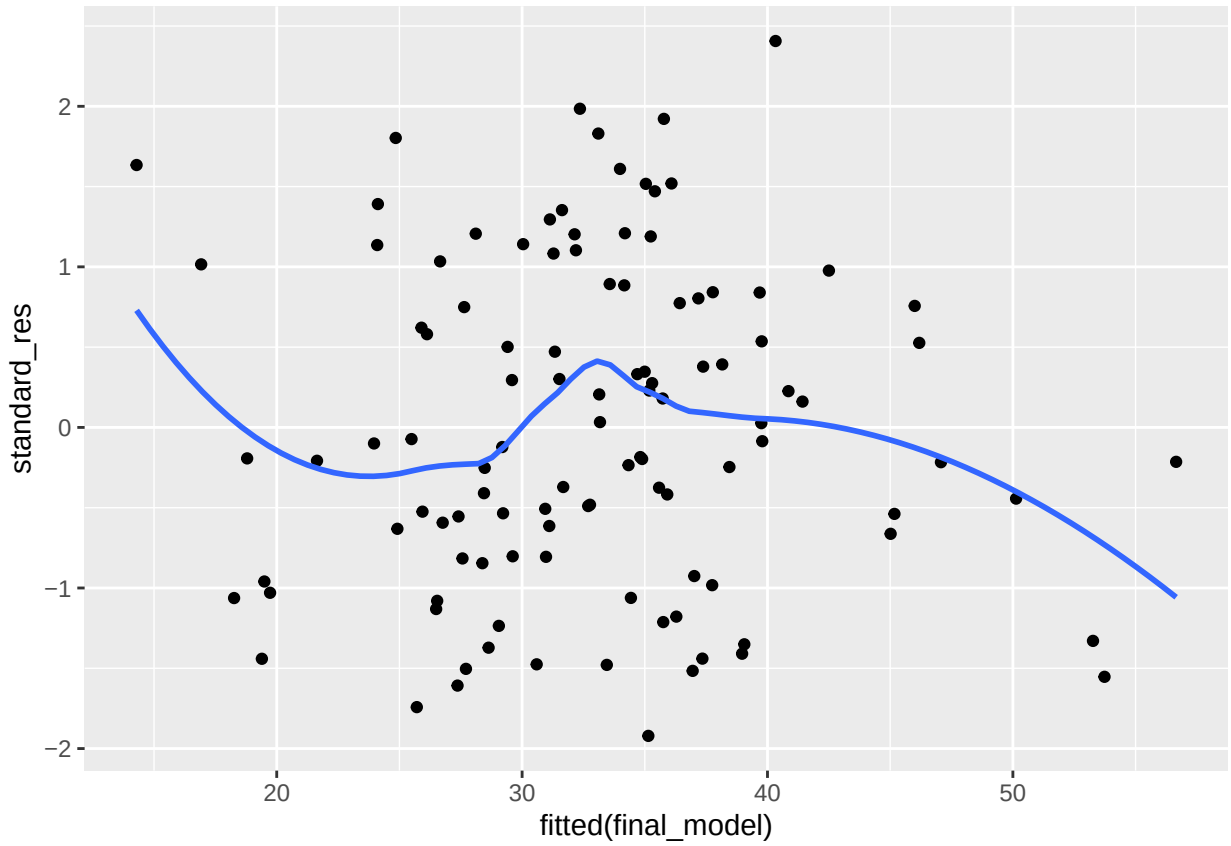


Histogram of residuals

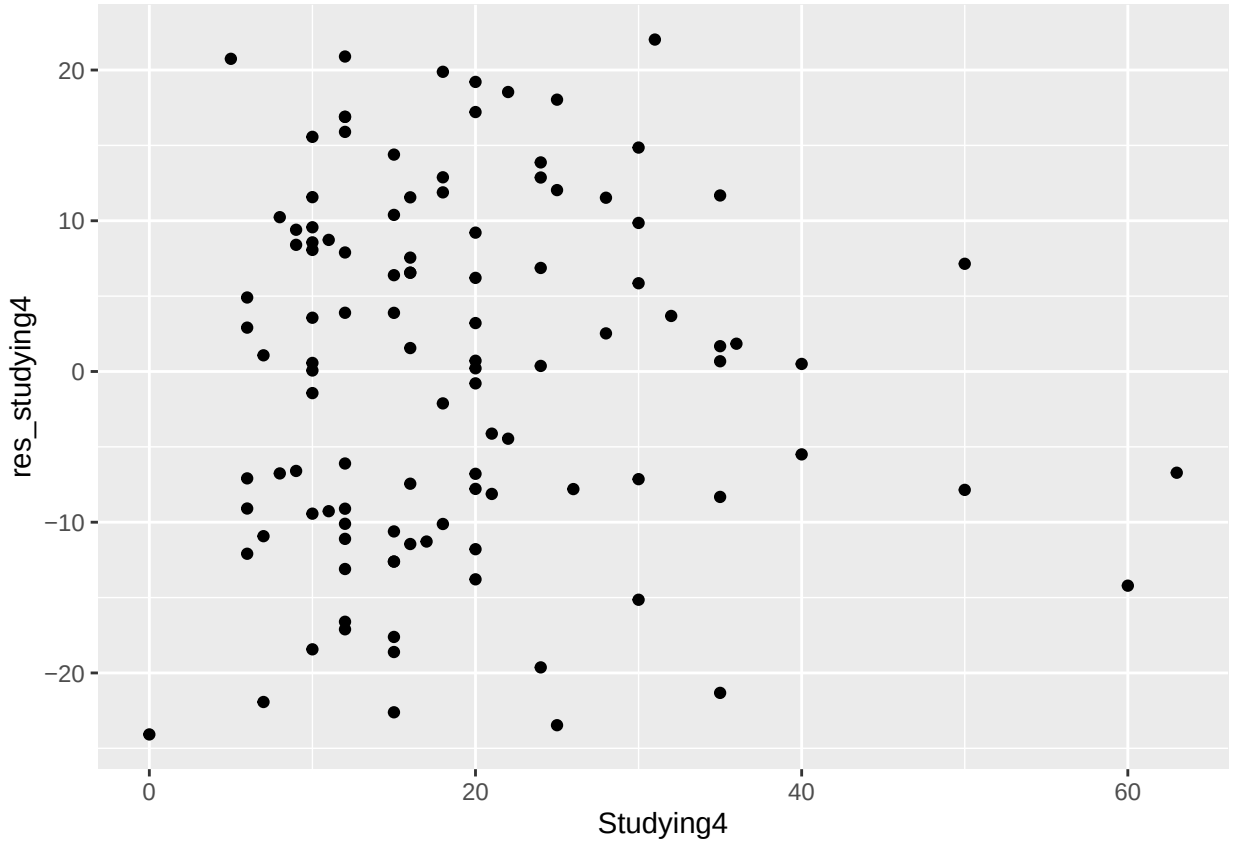


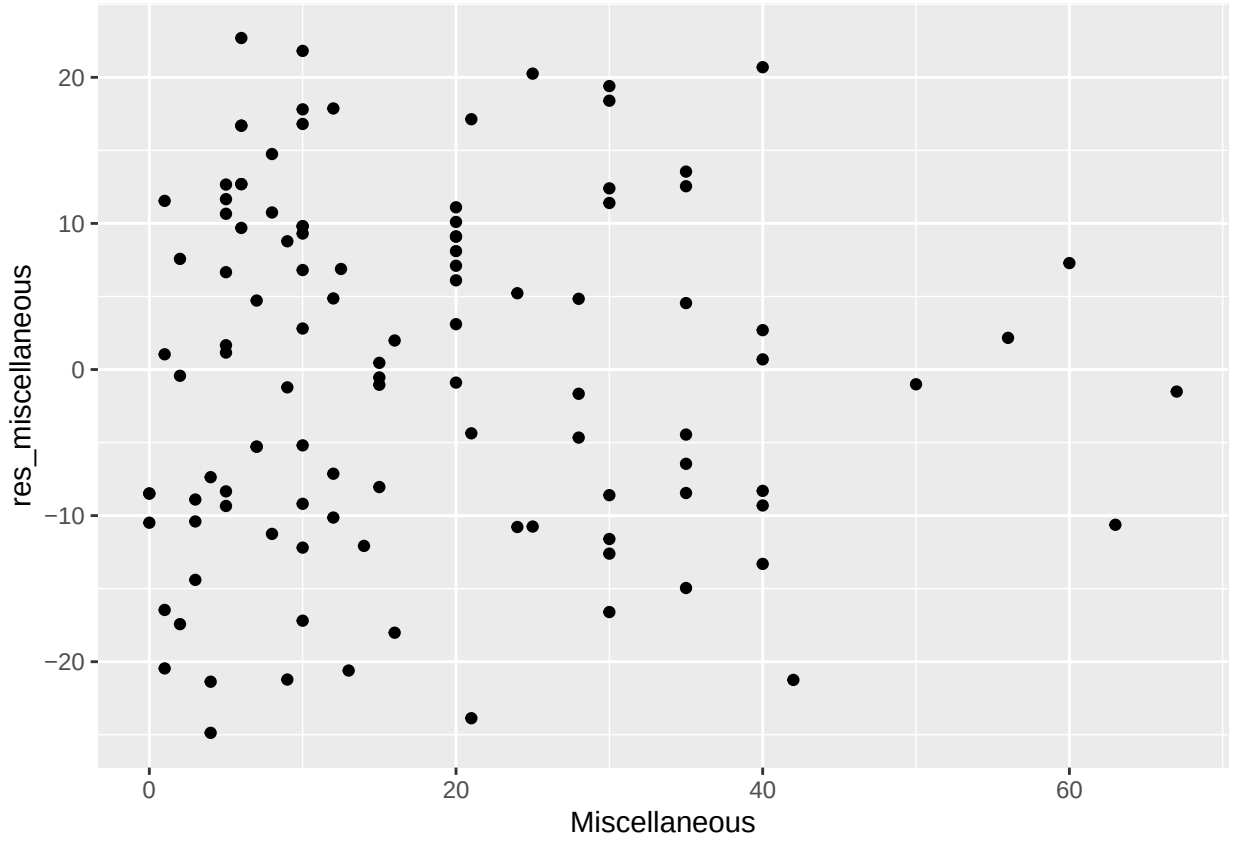
Standardized residuals vs fitted Term test mark

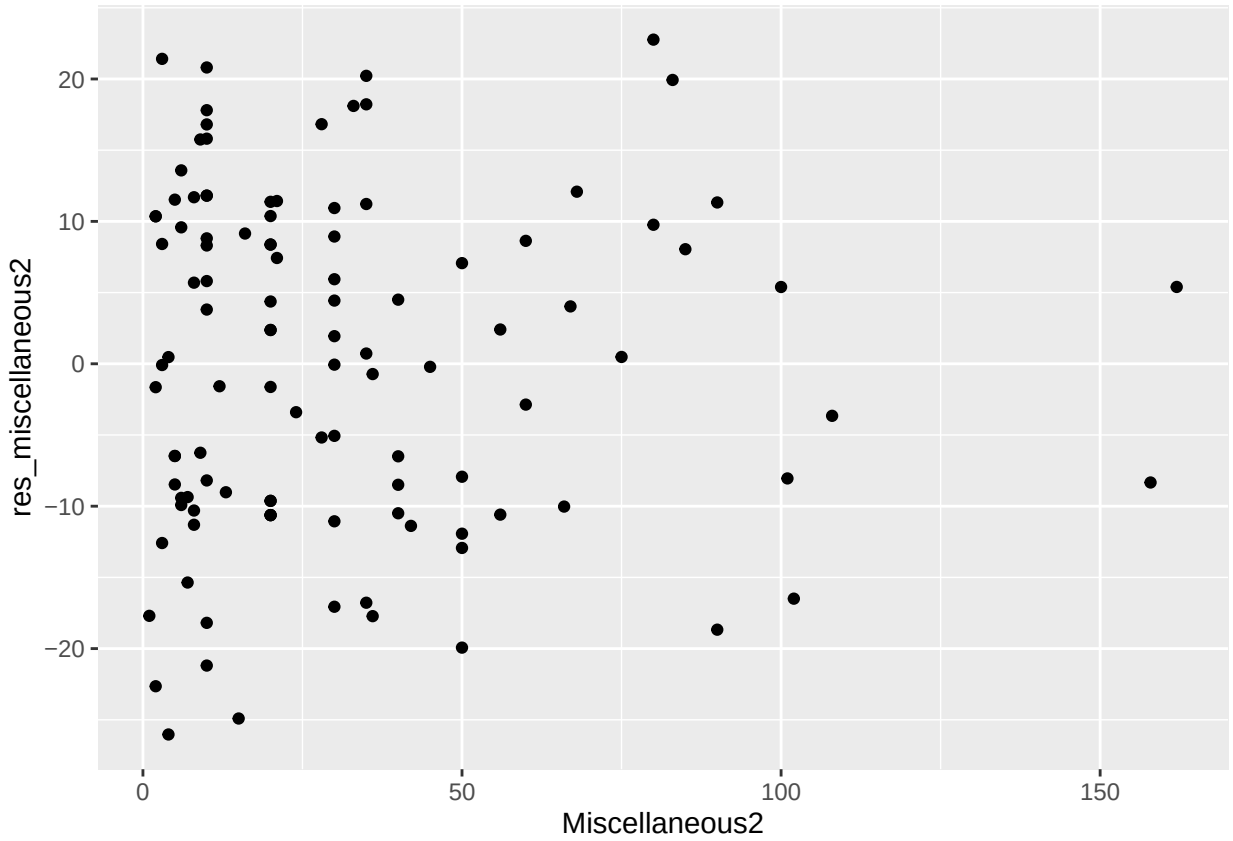
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

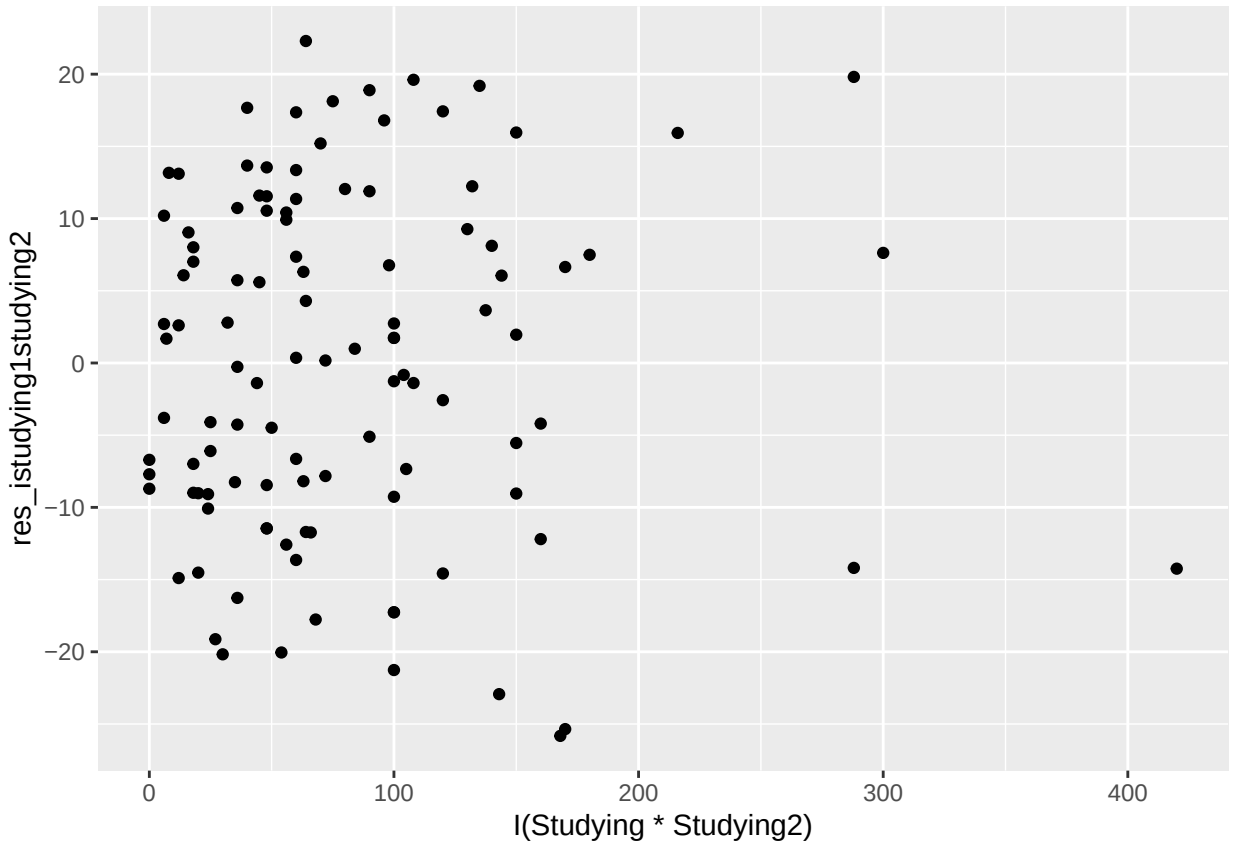

standard_res
2
1
0
−1
−2
20
30
40
50
fitted(final_model)

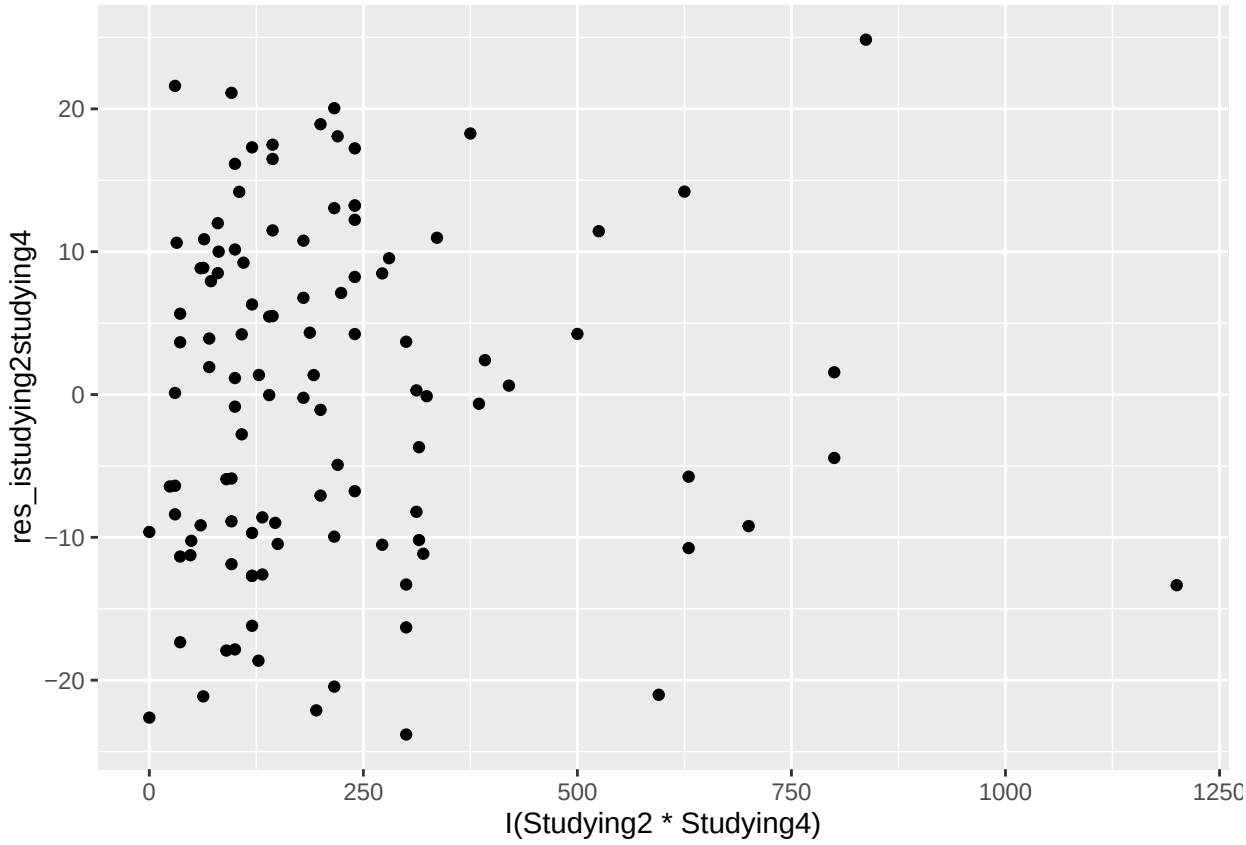
The next 18 figures represent residuals against the variables used in the final model

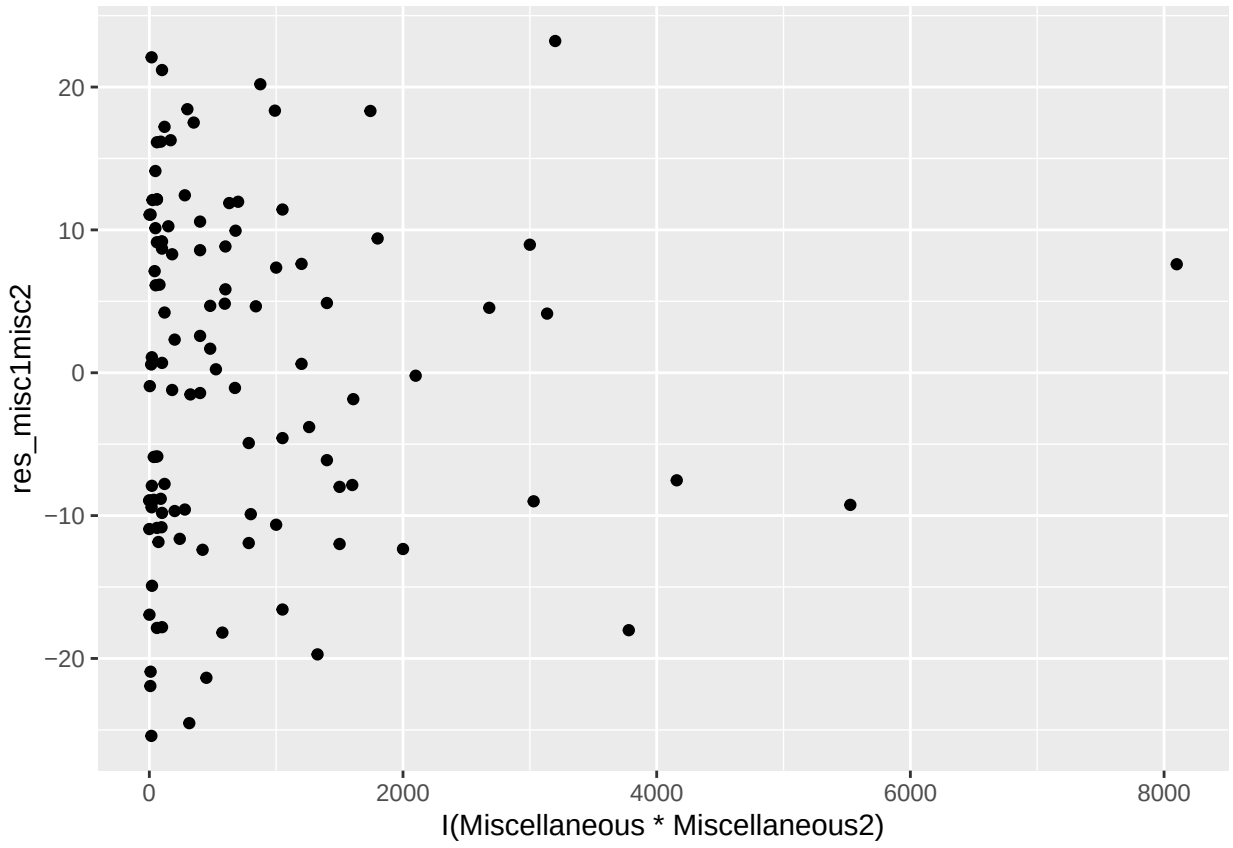


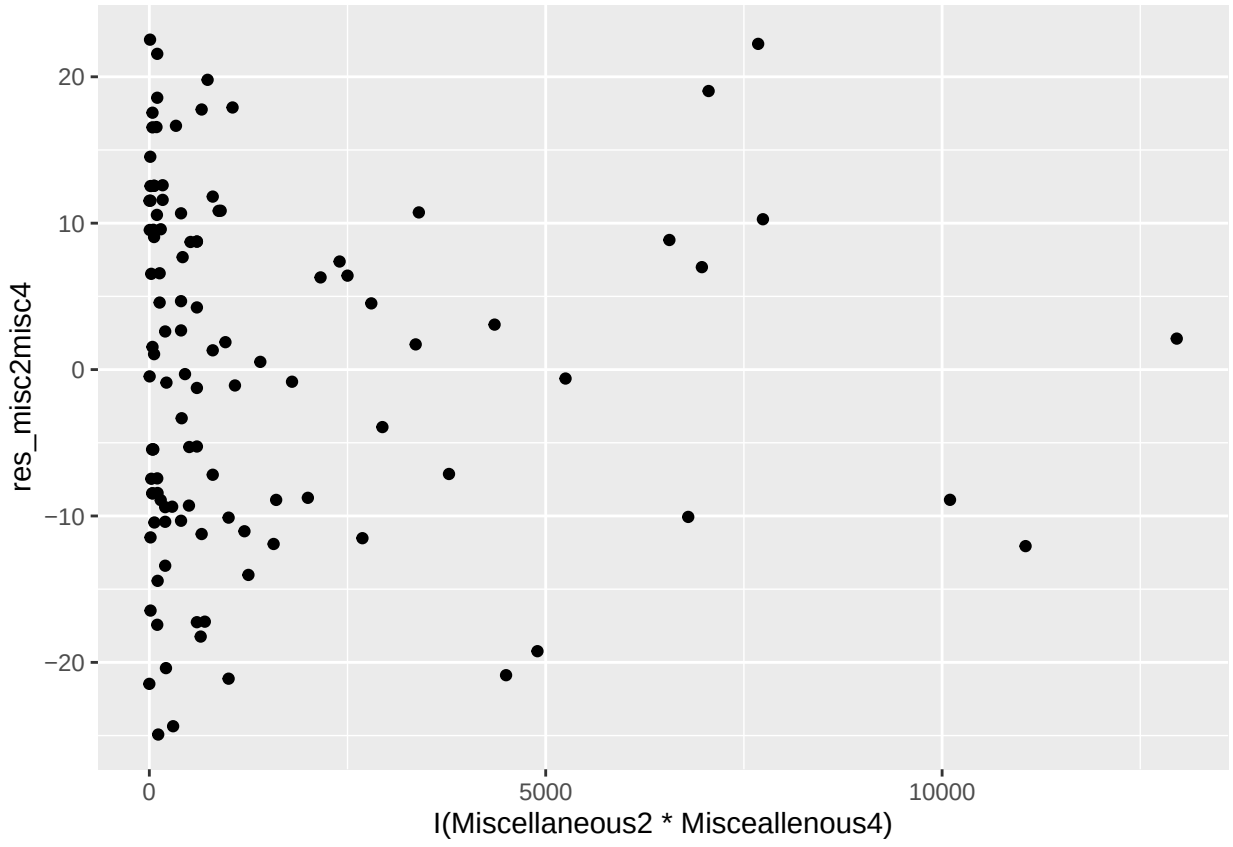

res_miscellaneous
20
10
0
−10
−20
0
20
40
60
Miscellaneous

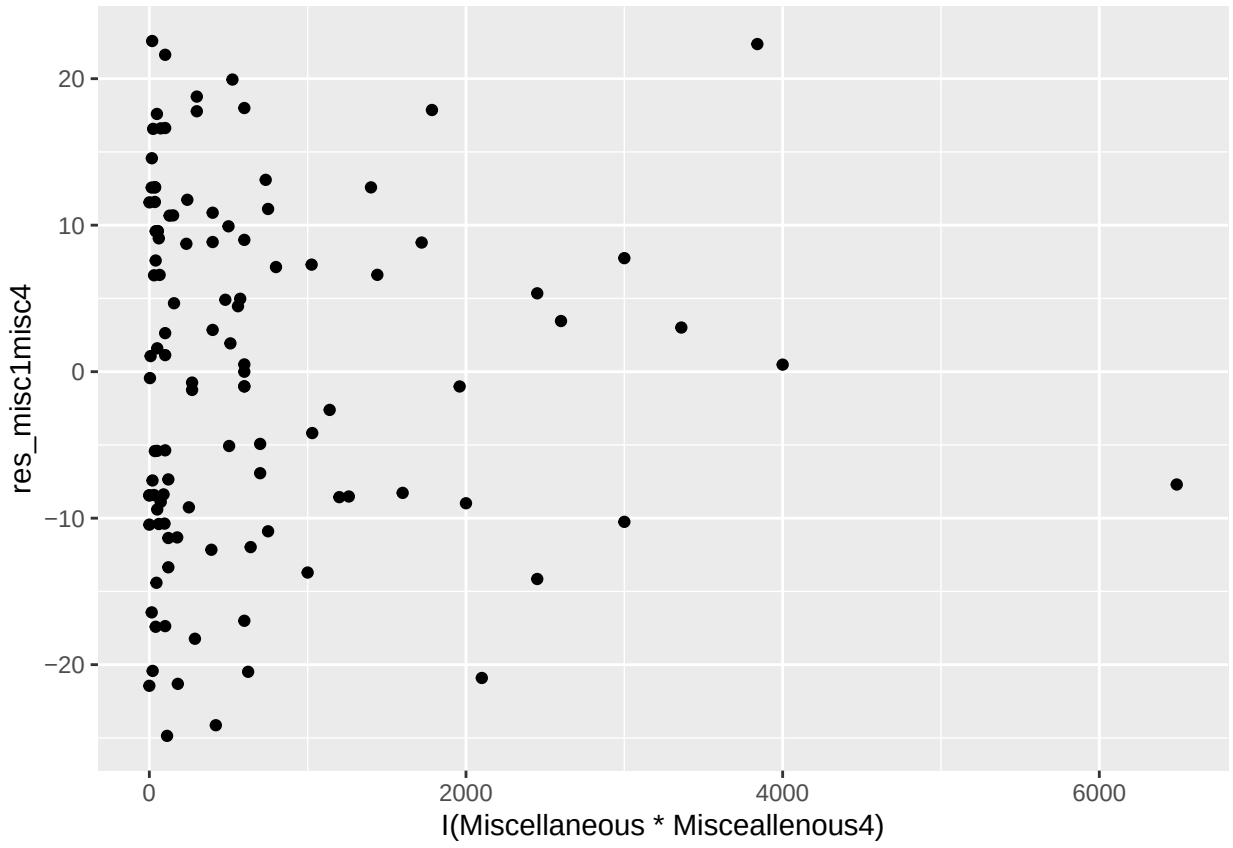

res_miscellaneous2
20
10
0
−10
−20
0
50
100
150
Miscellaneous2

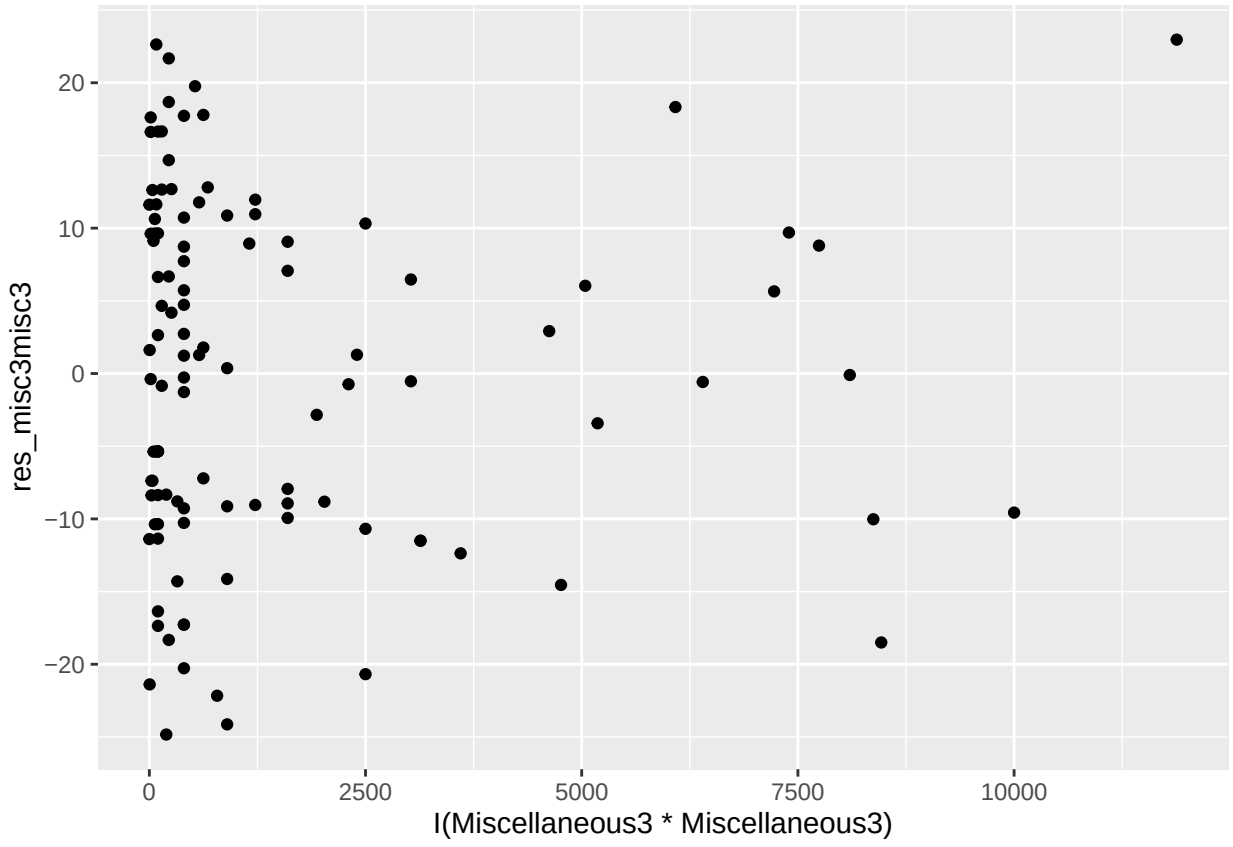

res_istudying1studying2
20
10
0
−10
−20
0
100
200
300
400
I(Studying * Studying2)

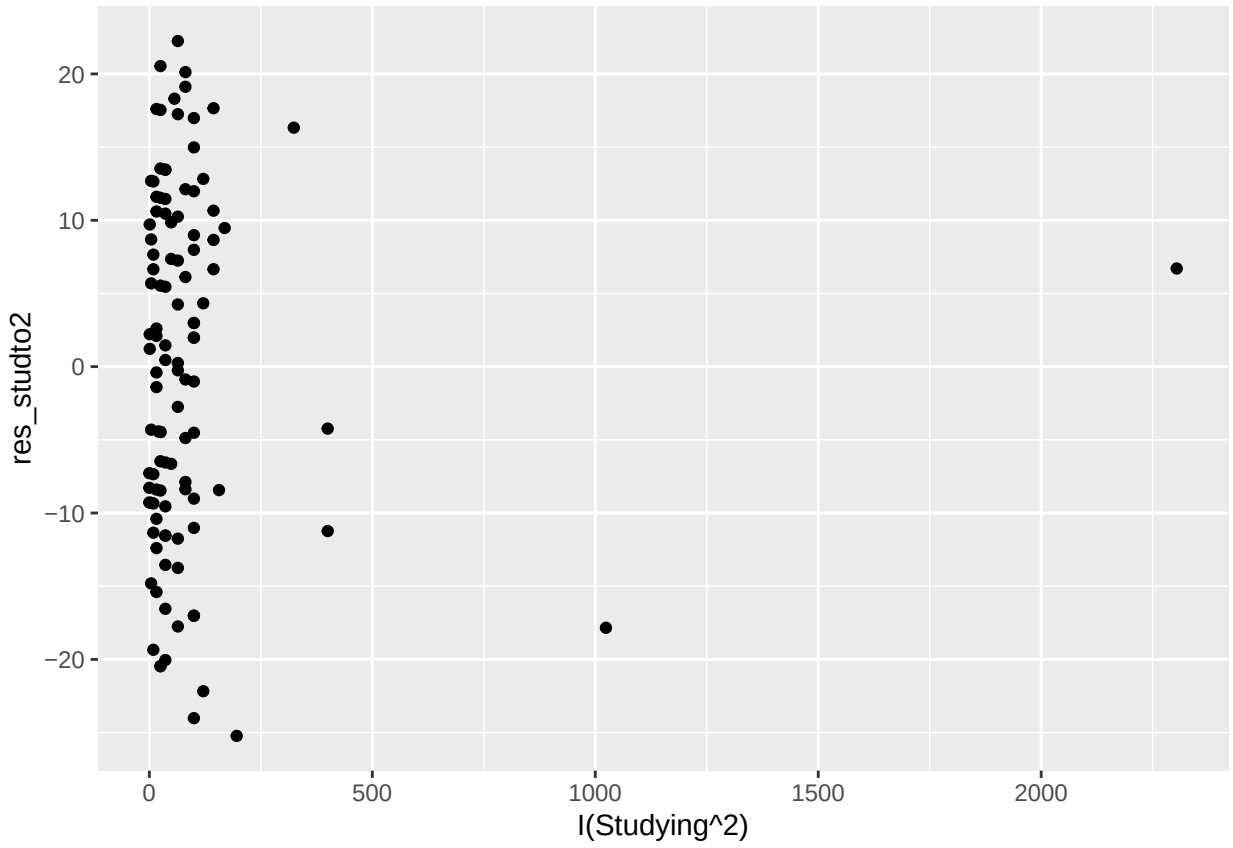

res_istudying2studying4
20
10
0
−10
−20
0
250
500
750
1000
1250
I(Studying2 * Studying4)

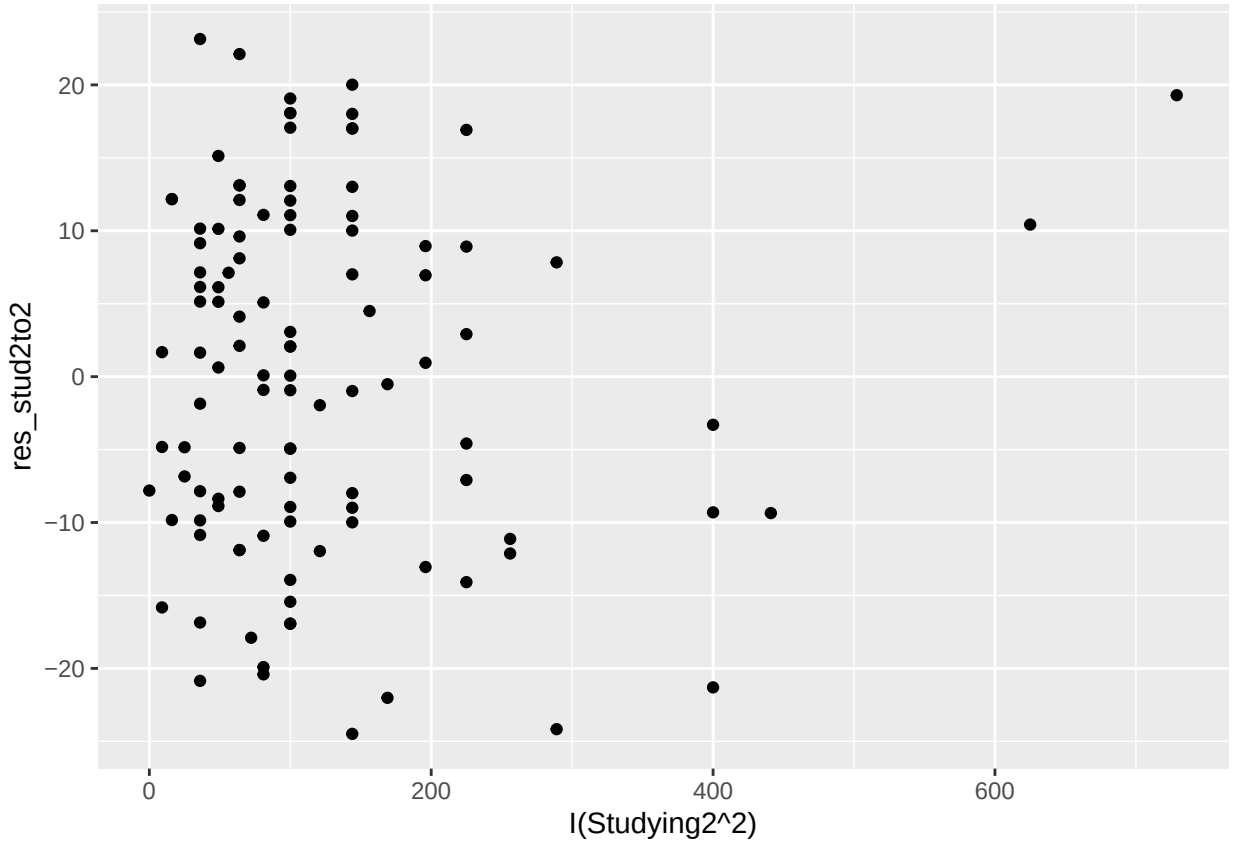

res_misc1misc2
20
10
0
−10
−20
0
2000
4000
6000
8000
I(Miscellaneous * Miscellaneous2)

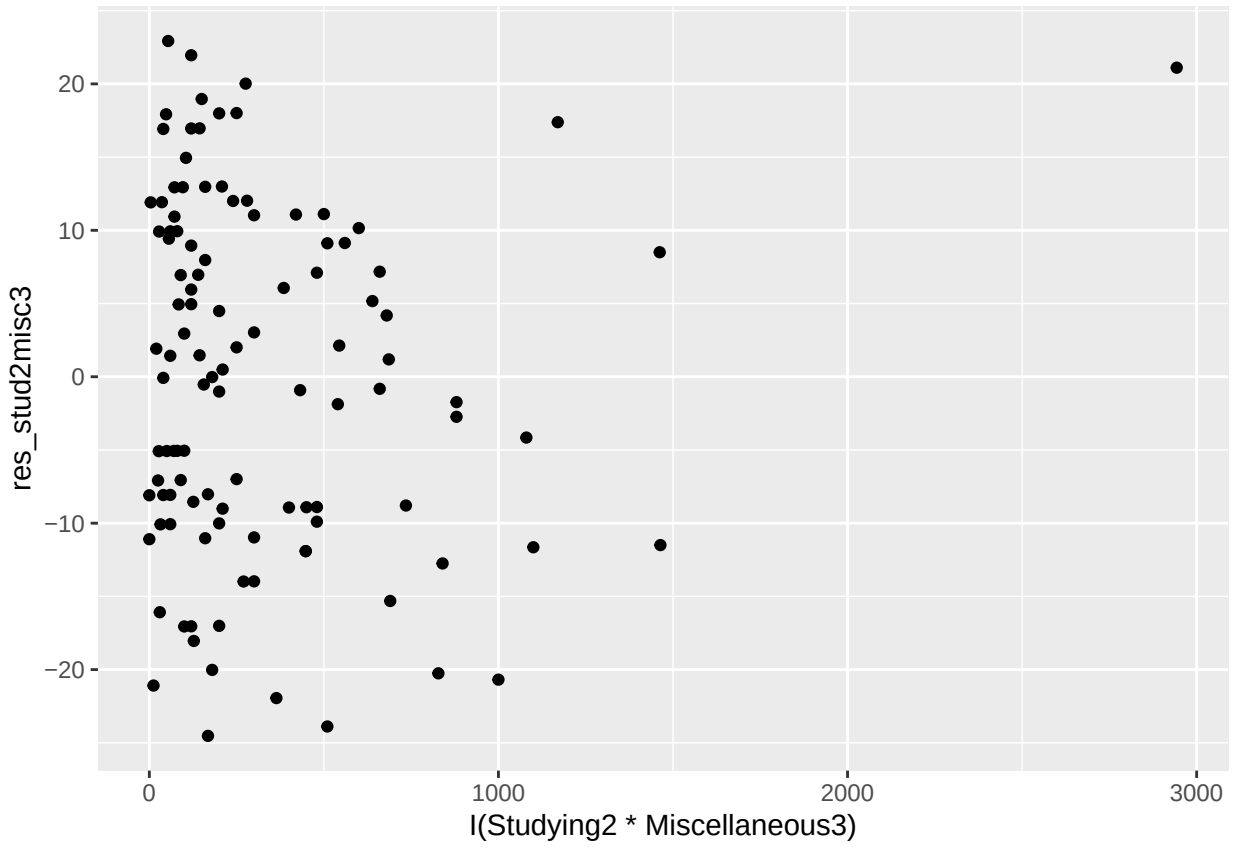

res_misc2misc4
20
10
0
−10
−20
0
5000
10000
I(Miscellaneous2 * Misceallenous4)

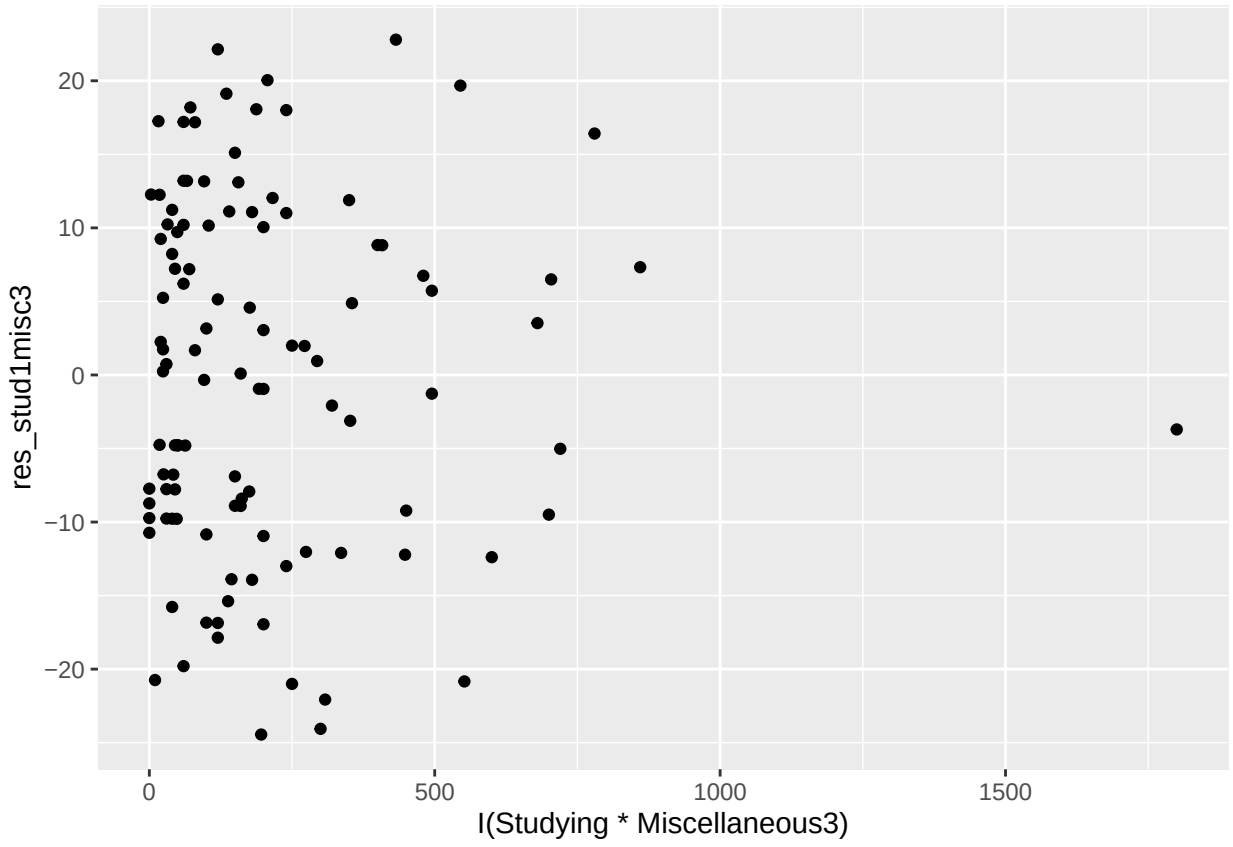

res_misc1misc4
20
10
0
−10
−20
0
2000
4000
6000
I(Miscellaneous * Misceallenous4)

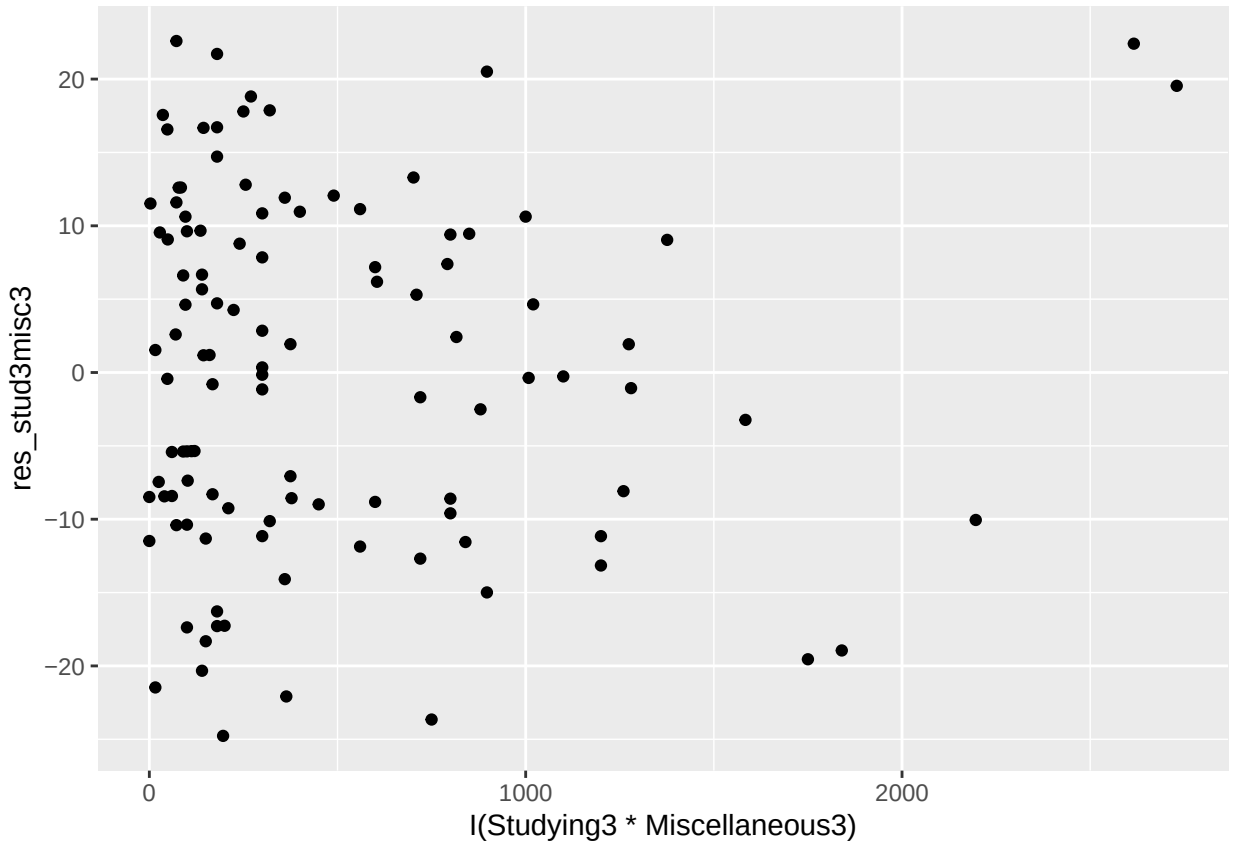

res_misc3misc3
20
10
0
−10
−20
0
2500
5000
7500
10000
I(Miscellaneous3 * Miscellaneous3)

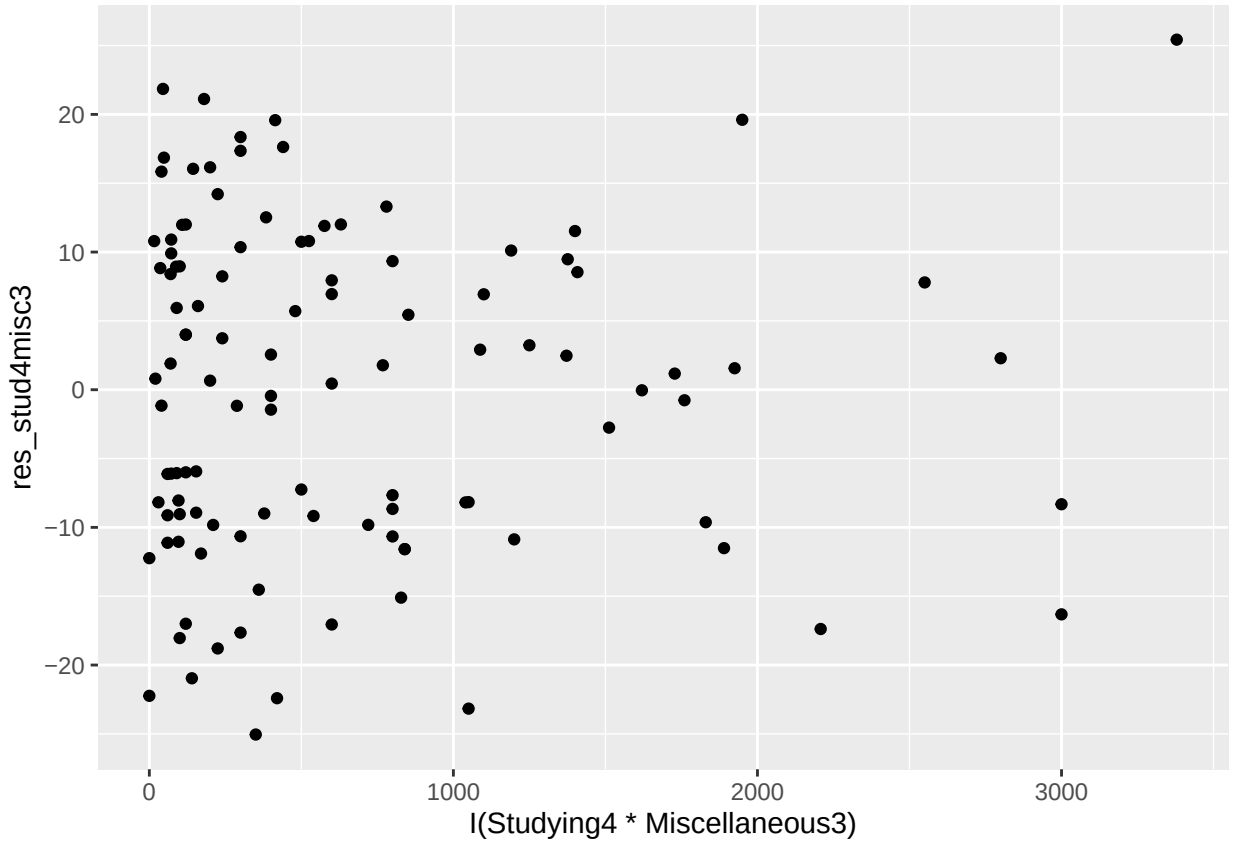

20
10
0
−10
−20
res_studto2
0
500
1000
1500
2000
I(Studying^2)

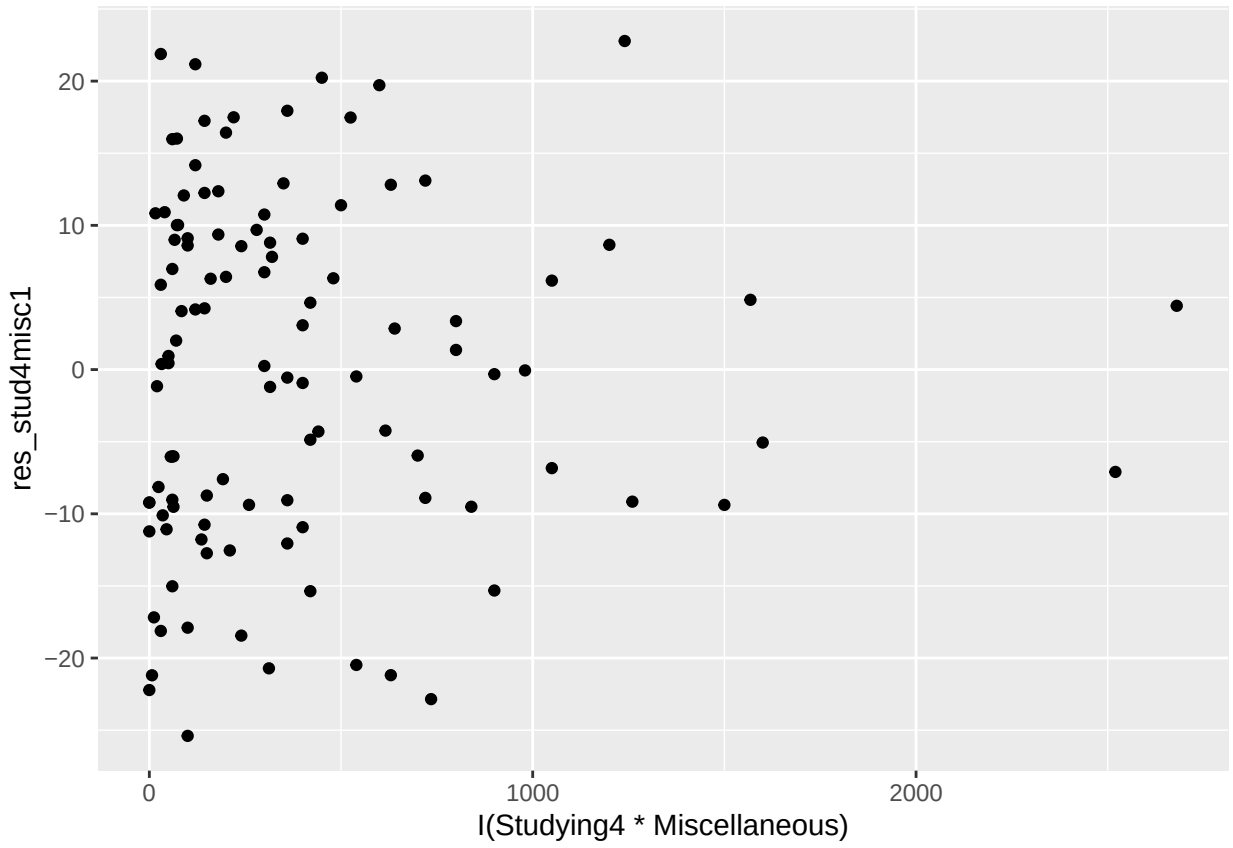

20
10
0
−10
−20
res_stud2to2
0
200
400
600
I(Studying2^2)

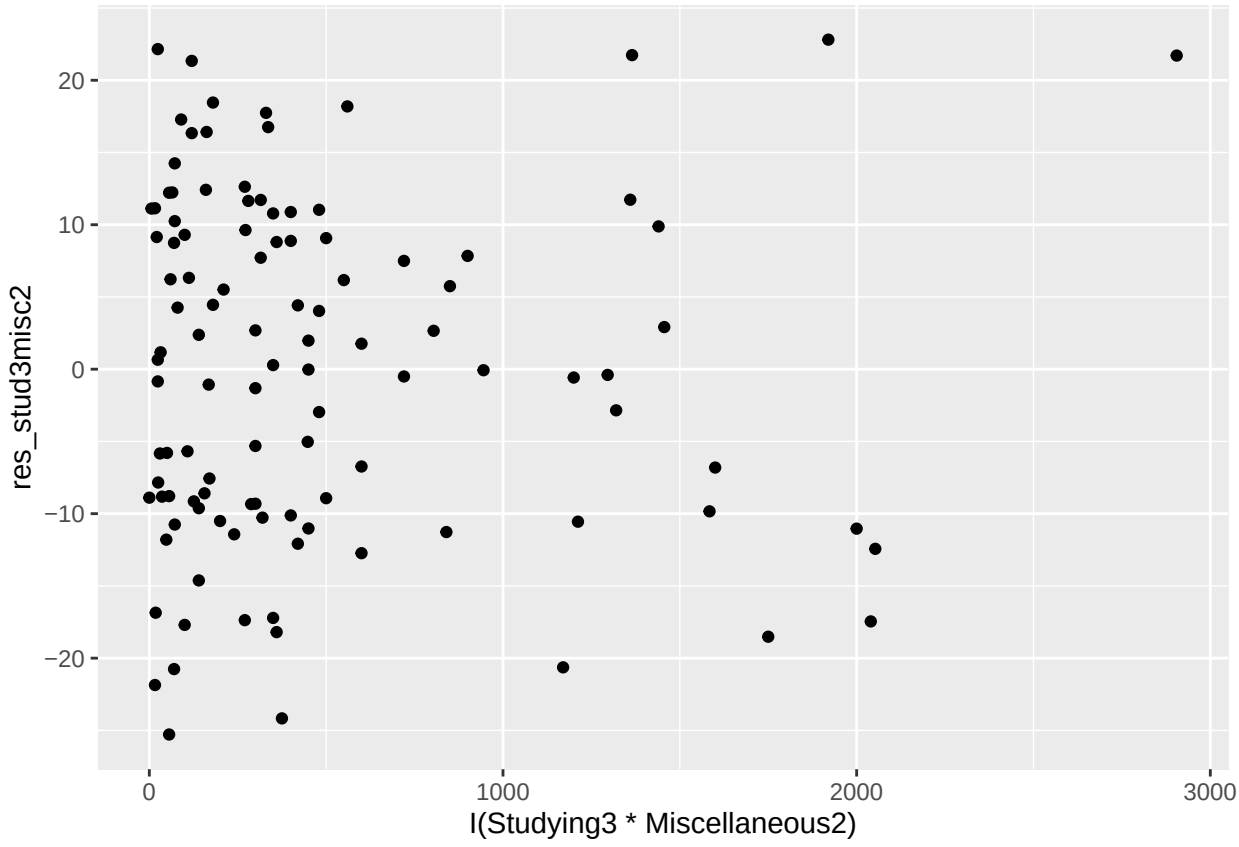

res_stud2misc3
20
10
0
−10
−20
0
1000
2000
3000
I(Studying2 * Miscellaneous3)

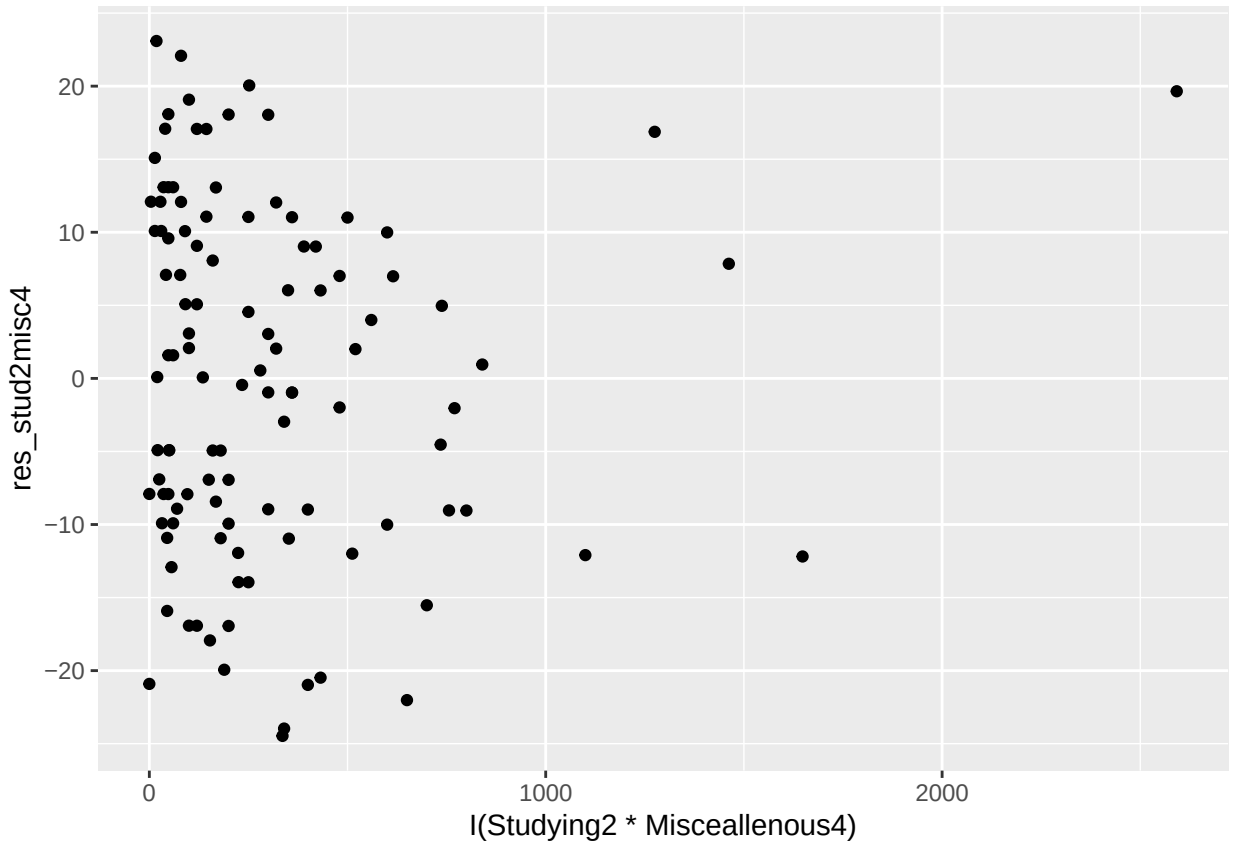

res_stud1misc3
20
10
0
−10
−20
0
500
1000
1500
I(Studying * Miscellaneous3)


res_stud3misc3
20
10
0
−10
−20
0
1000
2000
I(Studying3 * Miscellaneous3)


res_stud4misc3
20
10
0
−10
−20
0
1000
2000
3000
I(Studying4 * Miscellaneous3)


res_stud4misc1
20
10
0
−10
−20
0
1000
2000
I(Studying4 * Miscellaneous)


res_stud3misc2
20
10
0
−10
−20
0
1000
2000
3000
I(Studying3 * Miscellaneous2)



Model Validity

In order to check the validity of the final model, the 4 model assumptions will be used to clarify and check on the residuals (or errors) that correspond to the statistics of the model. A t-test will be performed on the residuals as well.

The 4 Assumptions

Assumption 1 (Linearity of the Relationship between the Residuals)

In order to check this, the fitted values are plotted against the residuals in order to see the relationship between the different residuals of different estimated term test values. Then, it could be clearly seen that there is hardly any relationship in between the residuals. The same could be said for the residual plots that are compared against the predictor variables (and interactions) in the final model. This suggests that no non-linearity exists in the selected final model.

Assumption 2 (Independence of Errors)

There is a possibility that the independence of errors between the residuals exist because the data was not collected from a random sample. More specifically, the purpose of discovering the final model in this project is to see if there are any significant predictors that could help to determine the influence on term test marks. However, this data could be collected from everyone who has studied STA302 but during the data collection process, only the students that study STA302 in 2022 summer are selected for the data collection. Nevertheless, as mentioned in assumption 1, from the plots of the residuals against all the predictor variables, there is an insignificant relationship shown between them if the outliers are excluded. Hence, the availability of the independence of errors is achieved.

Assumption 3 (Homoscedasticity)

From the fitted term mark plot on page 26, there is no significant pattern shown between the variance of the residuals (blue line that crosses through the plot). Despite a bit of curve shown in the center of the box plot, the residuals are spreaded around an approximately horizontal straight line; neither spread out nor spread is shown from the increase of the term test marks. Therefore, the homoscedasticity is not violated.

Assumption 4 (Normality of the Errors)

Referring to the histogram of the residual on page 25 and the normal QQ plot of the residuals on page 24, the residuals almost show a perfect normal distribution which the normality is also achieved. The possibility of non-normal errors exists is tiny, which insists that sufficient samples are collected for the analysis of this research.

50/50 Training/Testing Split

The data is split into two halves. The first half of the model has been used to test the model that we have trained to ensure that the model is reproducible for different datasets.

On running the final model, it is found that the mean of the residuals is -6.77173×10^{-17}. This is arbitrarily close to 0 which shows that the model has a high predictive power.

T-Test Statistics of the Residuals (Joshua)

$$H_0 : \mu_{residuals} = 0$$

$$H_a : \mu_{residuals} \neq 0$$

$$\mu_{residuals}$$

here means the average of the residuals.

Our 95% confidence interval is (-1.745,1.745). Since our mean of the residuals from our final model using the entire data set gives us -1.26966e-16 , which is arbitrarily close to 0 and since 0 lies within our confidence interval, we can say with 95% confidence that the average of our residuals is equal to 0.

Summary from the Assumptions

All four assumptions checked on the residuals are all valid, which means that the final model that comes out from this research is significant and valid as a reference for the discovery that was aimed for the purpose (to predict the term test marks by the chosen predictors). No variable transformation is needed for the final model because the error terms are linear, independent, homoscedastic, and approximately normal.

Conclusion

Purpose of the Research

Referring back to the purpose of this whole research, the final model is made to see if any of the predictor variables (i.e. studying time in different weeks, miscellaneous time in different weeks, time spent on thinking about COVID, frequency of attending office hours, familiarity of the course) could help to predict the term test marks that students get in STA302H1.

Interpretation of the Final Model

According to the final model, there is a significant relationship that exists between studying time spent on week 1, miscellaneous time spent on week 1 and week 2 and term test marks in the course STA302. In context, these relations are also strong as studying 1 more hour in week 4 would on average and with all variables held constant, net an increase of 3% on the term test mark. A note that needs to be emphasized is the fact that when one predictor variable is compared with the term test mark, the others would remain constant. There is also a significant relationship that exists in between the interaction of studying time week 1 & 2, week 2 & week 4, miscellaneous time week 1 & week 2, week 2 & 4, week 1 & 4, etc. and more could be determined from the final model. Additionally, the quadratic variables of miscellaneous week 3 and studying time week 1 and week 2 also strongly correlates to the term test mark. The interactions and quadratic relationship should all be taken into the considerations while modeling out the multi-linear relationship between the predictor variables and response in the final model.

Remaining Limitations and Problems with the Final Model

First of all, although a significant final model is produced to predict the relationship that is aimed to discover the purpose, the model is not generalizable to predict the term test scores of different students that enrol into STA302H1 across different terms. The reason is the sample data was only collected from the students who study in STA302H1 in summer 2022. Across different sections taught at different terms by different professors, the predictor variables could vary and change significantly, and the hardness of the term test could change as well.

Second, there are more variables that could be taken into considerations that affect the term test marks as well. For example, the final marks of the Statistics course that students take in UofT before STA302H1 (level of understanding of base Statistics knowledge) or the level of stress of students that take the course. Overall, there are missing predictors which could heavily influence the response that is aimed to look at.

Third, it is possible for some students to mis-input the data of studying time during the data collection, as the professor mentions about how people reporting only 1 hour of studying in one week but if the lectures are attended, then the base hours of studying time would be 6 hours each week (Unless those students actually do not go to lectures). In addition, the miscellaneous hours per week are incorrectly recorded for some students as well, since some people consider sports or social media as miscellaneous but some do not. Therefore, human errors exist during the data collection of this research.

Fourth, which was mentioned in the second assumptions in the diagnostic section as well, is that the sample is not randomly selected for the aim of the research. The whole data is collected and processed only in the STA302H1 summer 2022 class. Therefore, bias does exist in the final model since the sampling model.

Improvements for the Final Model

In order to improve and prevent the limitations and problems that are listed above, more sufficient (and intuitive) predictor variables could be added to consider in the initial model. Then, the research that is done for the same purpose later could now collect data from the students who study in the same course during different terms. These steps could help to solve the problem of random selection of the data as well. Then, the definition of studying time and miscellaneous could be given clearer in order for more accurate collection of the data. These changes might not solve all the problems that exist, but at least remove some of the bias that already exists, especially from what should have been easy, the sampling model. Some further research could also be done to solve the existing problems as well.

Student Contributions

Joshua: Model Development Residual Plots Pairwise Scatter plots and correlation R markdown formatting

Jiyun (Lyla): Purpose statement Variables description Histogram description + R visualization Boxplot description + R visualization Scatterplots exploratory explanation Pairwise exploratory explanation R

markdown formatting

Janhavi: Introduction R Code for Visualization Making scatterplots for exploratory data visualization Residual Plots Clean up and R Markdown (Formatting)

Jercy: Model Development (along with Joshua's help in coding) 4 Assumptions in Model Diagnostics Conclusion

```
summary(poly.fit) #initial model
```

```
##
## Call:
## lm(formula = Term.Test ~ Studying + Studying2 + Studying3 + Studying4 +
##     Miscellaneous + Miscellaneous2 + Miscellaneous3 + Miseallenous4 +
##     I(Studying * Studying3) + I(Studying * Studying4) + I(Studying *
##     Studying2) + I(Studying2 * Studying3) + I(Studying3 * Studying4) +
##     I(Studying2 * Studying4) + I(Miscellaneous * Miscellaneous2) +
##     I(Miscellaneous2 * Miscellaneous3) + I(Miscellaneous3 * Miseallenous4) +
##     I(Miscellaneous2 * Miseallenous4) + I(Miscellaneous * Miseallenous4) +
##     I(Miscellaneous * Miscellaneous3) + I(Miscellaneous2^2) +
##     I(Miscellaneous3^2) + I(Miseallenous4^2) + I(Studying^2) +
##     I(Studying2^2) + I(Studying3^2) + I(Studying4^2) + I(Studying *
##     Miscellaneous) + I(Studying * Miscellaneous2) + I(Studying *
##     Miscellaneous3) + I(Studying * Miseallenous4) + I(Studying2 *
##     Miscellaneous) + I(Studying2 * Miscellaneous2) + I(Studying2 *
##     Miscellaneous3) + I(Studying2 * Miseallenous4) + I(Studying3 *
##     Miscellaneous) + I(Studying3 * Miscellaneous2) + I(Studying3 *
##     Miscellaneous3) + I(Studying3 * Miseallenous4) + I(Studying4 *
##     Miscellaneous) + I(Studying4 * Miscellaneous2) + I(Studying4 *
##     Miscellaneous3) + I(Studying4 * Miseallenous4), data = midterm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5953  -5.4819  -0.5932   5.7204  20.1114
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       21.2175678  8.3255637   2.548   0.0132 *
## Studying                          -0.3277725  1.3765733  -0.238   0.8125
## Studying2                          0.6683734  1.5515912   0.431   0.6680
## Studying3                         -0.1936831  1.0960986  -0.177   0.8603
## Studying4                          1.9279914  0.7670116   2.514   0.0144 *
## Miscellaneous                      0.7622154  0.3968022   1.921   0.0591 .
## Miscellaneous2                    -0.6381537  0.4223126  -1.511   0.1355
## Miscellaneous3                    -0.1957699  0.5827338  -0.336   0.7380
## Miseallenous4                     -0.2518103  0.6102616  -0.413   0.6812
## I(Studying * Studying3)            0.0453730  0.0947433   0.479   0.6336
## I(Studying * Studying4)            0.0040306  0.0419138   0.096   0.9237
## I(Studying * Studying2)           -0.0972645  0.1468782  -0.662   0.5101
## I(Studying2 * Studying3)          -0.0198340  0.1469581  -0.135   0.8931
## I(Studying3 * Studying4)           0.0282820  0.0288358   0.981   0.3303
## I(Studying2 * Studying4)          -0.1572484  0.0972935  -1.616   0.1108
## I(Miscellaneous * Miscellaneous2) -0.0216698  0.0086808  -2.496   0.0151 *
## I(Miscellaneous2 * Miscellaneous3) -0.0025643  0.0103430  -0.248   0.8050
## I(Miscellaneous3 * Miseallenous4) -0.0034368  0.0126784  -0.271   0.7872
## I(Miscellaneous2 * Miseallenous4)  0.0152575  0.0117590   1.298   0.1990
```

```
## I(Miscellaneous * Miscellenous4)  -0.0049021  0.0150217  -0.326   0.7452
## I(Miscellaneous * Miscellaneous3)   0.0218219  0.0147603   1.478   0.1441
## I(Miscellaneous2^2)                -0.0009945  0.0023284  -0.427   0.6707
## I(Miscellaneous3^2)                -0.0069431  0.0114576  -0.606   0.5466
## I(Miscellenous4^2)                  0.0032855  0.0059883   0.549   0.5851
## I(Studying^2)                       0.0099156  0.0206707   0.480   0.6330
## I(Studying2^2)                      0.1171740  0.0806883   1.452   0.1512
## I(Studying3^2)                     -0.0050315  0.0337740  -0.149   0.8820
## I(Studying4^2)                     -0.0156643  0.0141991  -1.103   0.2740
## I(Studying * Miscellaneous)         0.0060101  0.0377179   0.159   0.8739
## I(Studying * Miscellaneous2)       -0.0045536  0.0205570  -0.222   0.8254
## I(Studying * Miscellaneous3)        0.0696826  0.0543907   1.281   0.2046
## I(Studying * Miscellenous4)        -0.0318735  0.0577268  -0.552   0.5827
## I(Studying2 * Miscellaneous)        0.0383017  0.0473362   0.809   0.4213
## I(Studying2 * Miscellaneous2)      -0.0029605  0.0421675  -0.070   0.9442
## I(Studying2 * Miscellaneous3)       0.0757139  0.0564997   1.340   0.1848
## I(Studying2 * Miscellenous4)       -0.0663752  0.0676352  -0.981   0.3300
## I(Studying3 * Miscellaneous)        0.0151807  0.0358798   0.423   0.6736
## I(Studying3 * Miscellaneous2)       0.0497629  0.0309398   1.608   0.1125
## I(Studying3 * Miscellaneous3)      -0.0779713  0.0376106  -2.073   0.0421 *
## I(Studying3 * Miscellenous4)        0.0100388  0.0266551   0.377   0.7077
## I(Studying4 * Miscellaneous)       -0.0568242  0.0230801  -2.462   0.0164 *
## I(Studying4 * Miscellaneous2)       0.0072800  0.0158491   0.459   0.6475
## I(Studying4 * Miscellaneous3)       0.0178756  0.0185510   0.964   0.3388
## I(Studying4 * Miscellenous4)        0.0097630  0.0158384   0.616   0.5397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.72 on 66 degrees of freedom
## Multiple R-squared:  0.5122, Adjusted R-squared:  0.1943
## F-statistic: 1.611 on 43 and 66 DF,  p-value: 0.0397
```

```
summary(final_model)
```

```
##
## Call:
## lm(formula = Term.Test ~ Studying4 + Miscellaneous + Miscellaneous2 +
##     I(Studying * Studying2) + I(Studying2 * Studying4) + I(Miscellaneous *
##     Miscellaneous2) + I(Miscellaneous2 * Miscellenous4) + I(Miscellaneous *
##     Miscellenous4) + I(Miscellaneous3^2) + I(Studying^2) + I(Studying2^2) +
##     I(Studying * Miscellaneous3) + I(Studying2 * Miscellaneous3) +
##     I(Studying2 * Miscellenous4) + I(Studying3 * Miscellaneous2) +
##     I(Studying3 * Miscellaneous3) + I(Studying4 * Miscellaneous) +
##     I(Studying4 * Miscellaneous3), data = midterm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.1438  -6.5233  -0.8763   7.5220  19.2220
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   19.917103   4.134334   4.817 5.78e-06 ***
## Studying4                      1.663136   0.379718   4.380 3.17e-05 ***
## Miscellaneous                  0.800071   0.217471   3.679 0.000396 ***
## Miscellaneous2                -0.611032   0.123671  -4.941 3.52e-06 ***
```

```
## I(Studying * Studying2)            -0.084651   0.034590  -2.447 0.016315 *
## I(Studying2 * Studying4)           -0.173602   0.031839  -5.453 4.23e-07 ***
## I(Miscellaneous * Miscellaneous2)  -0.021017   0.004710  -4.462 2.32e-05 ***
## I(Miscellaneous2 * Misceallenous4)  0.011038   0.002480   4.452 2.41e-05 ***
## I(Miscellaneous * Misceallenous4)   0.013797   0.005434   2.539 0.012816 *
## I(Miscellaneous3^2)                -0.009222   0.002282  -4.042 0.000111 ***
## I(Studying^2)                       0.014218   0.006223   2.285 0.024658 *
## I(Studying2^2)                      0.183627   0.039721   4.623 1.24e-05 ***
## I(Studying * Miscellaneous3)        0.036260   0.009542   3.800 0.000261 ***
## I(Studying2 * Miscellaneous3)       0.077177   0.020034   3.852 0.000218 ***
## I(Studying2 * Misceallenous4)      -0.072848   0.018909  -3.852 0.000217 ***
## I(Studying3 * Miscellaneous2)       0.040709   0.009627   4.229 5.59e-05 ***
## I(Studying3 * Miscellaneous3)      -0.043003   0.012288  -3.500 0.000723 ***
## I(Studying4 * Miscellaneous)       -0.019604   0.006598  -2.971 0.003794 **
## I(Studying4 * Miscellaneous3)       0.014171   0.006473   2.189 0.031154 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 91 degrees of freedom
## Multiple R-squared:  0.4016, Adjusted R-squared:  0.2832
## F-statistic: 3.393 on 18 and 91 DF,  p-value: 6.146e-05
```

```
summary(midterm_develop_half)
```

```
##        ID          Studying          COVID       Miscellaneous
##  Min.   : 1.0   Min.   : 0.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:14.5   1st Qu.: 4.500   1st Qu.: 0.75   1st Qu.: 5.50
##  Median :28.0   Median : 7.000   Median : 1.00   Median :12.00
##  Mean   :28.0   Mean   : 7.709   Mean   : 2.00   Mean   :17.58
##  3rd Qu.:41.5   3rd Qu.:10.000   3rd Qu.: 3.00   3rd Qu.:26.00
##  Max.   :55.0   Max.   :48.000   Max.   :10.00   Max.   :67.00
##    Studying2         COVID2       Miscellaneous2     Studying3
##  Min.   : 3.00   Min.   : 0.000   Min.   :  1.00   Min.   : 6.00
##  1st Qu.: 7.00   1st Qu.: 0.750   1st Qu.:  7.00   1st Qu.:10.00
##  Median :10.00   Median : 2.000   Median : 20.00   Median :12.00
##  Mean   :10.55   Mean   : 2.301   Mean   : 27.87   Mean   :14.65
##  3rd Qu.:13.50   3rd Qu.: 3.000   3rd Qu.: 40.00   3rd Qu.:18.00
##  Max.   :27.00   Max.   :12.000   Max.   :102.00   Max.   :40.00
##      COVID3       Miscellaneous3     Studying4        COVID4
##  Min.   : 0.000   Min.   :  2.00   Min.   : 0.0   Min.   : 0.000
##  1st Qu.: 0.750   1st Qu.: 10.00   1st Qu.:10.0   1st Qu.: 0.225
##  Median : 1.400   Median : 20.00   Median :16.0   Median : 1.000
##  Mean   : 2.509   Mean   : 28.69   Mean   :18.8   Mean   : 3.254
##  3rd Qu.: 4.000   3rd Qu.: 42.00   3rd Qu.:24.0   3rd Qu.: 4.000
##  Max.   :12.000   Max.   :109.00   Max.   :60.0   Max.   :40.000
##  Misceallenous4       OH             Famiiliar           Term.Test
##  Min.   : 0.00   Length:55          Length:55          Min.   : 8.50
##  1st Qu.:10.00   Class :character   Class :character   1st Qu.:24.00
##  Median :20.00   Mode  :character   Mode  :character   Median :34.00
##  Mean   :27.24                                         Mean   :32.55
##  3rd Qu.:40.00                                         3rd Qu.:42.00
##  Max.   :96.00                                         Max.   :56.00
##   covid_total      study_total       misc_total    fam_strong_agree
##  Min.   : 0.000   Min.   : 19.00   Min.   :  8.0   Min.   :0.00000
##  1st Qu.: 2.475   1st Qu.: 36.50   1st Qu.: 39.0   1st Qu.:0.00000
```

```
##  Median : 6.500   Median : 52.00   Median : 81.0   Median :0.00000
##  Mean   :10.064   Mean   : 51.72   Mean   :101.4   Mean   :0.07273
##  3rd Qu.:15.000   3rd Qu.: 63.50   3rd Qu.:145.0   3rd Qu.:0.00000
##  Max.   :51.000   Max.   :120.00   Max.   :325.0   Max.   :1.00000
##  fam_strong_disagree   fam_agree       fam_disagree      fam_neutral
##  Min.   :0.00000     Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000     1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000     Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.03636     Mean   :0.3818   Mean   :0.07273   Mean   :0.4364
##  3rd Qu.:0.00000     3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000     Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##     oh_never         oh_less          oh_once           oh_alo
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.4182   Mean   :0.3818   Mean   :0.1636   Mean   :0.03636
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
```

```
mean(midterm_develop_half)
```

```
## Warning in mean.default(midterm_develop_half): argument is not numeric or
## logical: returning NA
```

```
## [1] NA
```