Factors Impact Student Performance on Course Assessment under Pandemic

Hanqi Zeng August 21, 2022

Introduction

Summer of 2022, beyond everyone's expectation, the world is still under a global pandemic caused by the novel coronavirus called COVID-19. Students are ordered to stay home if any symptoms are detected, traveling to campus is restricted, and despite in-person exams, lectures are recorded to accommodate students' needs.

The topic that we are addressing is *What are the factors that predict student performance on the course assessment*?. Besides, unlike most existing studies focusing on individual variables, this study intends to involve covariates effect and explore multiple covariates simultaneously to assess their collective effect on test grades.

Our purpose of developing this model is to provide guidance for professors and students. With the help of this study, teaching stream professors can gain more insights, know students' learning habits under pandemic and improve their teaching. Students taking this course can know significant factors influencing course grades and adjust their learning strategies for better performance in the future.

Experiment information

Data Preparation

The cohort of this dataset were STA302 students enrolled in the summer 2022 (July - August) semester. The data we are using is collected through the weekly Quercus quiz. The weekly assessment quiz is held at the end of each Monday lecture. Here are the variables provided in the data set (see Table1).

Variables	Meaning	Type
ID	id for each student	ordinary numeric
Studying	time (in hours) that the student put into studying STA302 each week	continuous numeric
Studying2		
Studying3		
Studying4		
COVID	time (in hours) that the student thinks about COVID-19 each week	continuous numeric
COVID2		
COVID3		
COVID4		
Miscellaneous	time (in hours) that the student spends on miscellaneous activities each week	continuous numeric
Miscellaneous2		
Miscellaneous3		
Misceallenous4		
OH	office hour attendance frequency	categorical
Famiiliar	self-evaluation on familarity of course materials	categorical
Term.Test	term test grade out of total score 55	ordinary numeric

TABLE 1

Summary of Important Variables

The score of term test (Term.Test) is the response variable. Full mark is 55. In order to eliminate potential influence of the value of total marks, we choose to transform the grades into a percentage format (See Figure 1 and Figure 2). The distribution of term test grades shows a bimodal trend generally and a 60% average score.

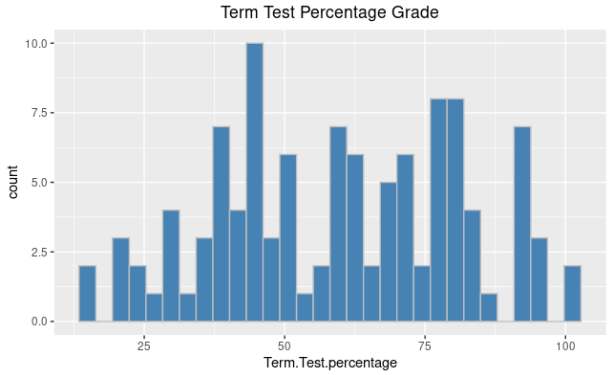


Figure 2

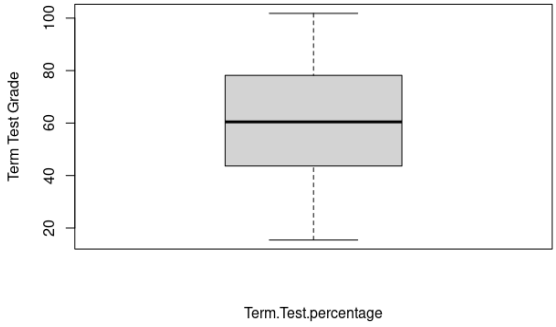


Figure 1

Table 4 shows a summary of term test percentage scores. (Note: The appearance of over 100 is due to bonus marks.)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15.45	43.64	60.45	59.93	78.18	101.82

TABLE 2

Two categorical variables, familiarity on course materials and office hour attendance are summarized in bar plots in Figure 3 and Figure 4 respectively.

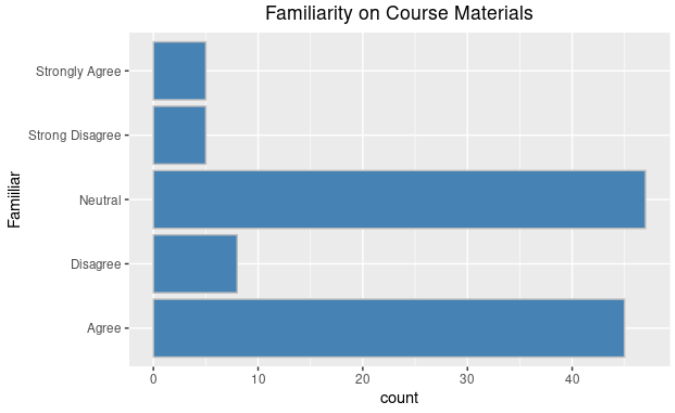


Figure 4

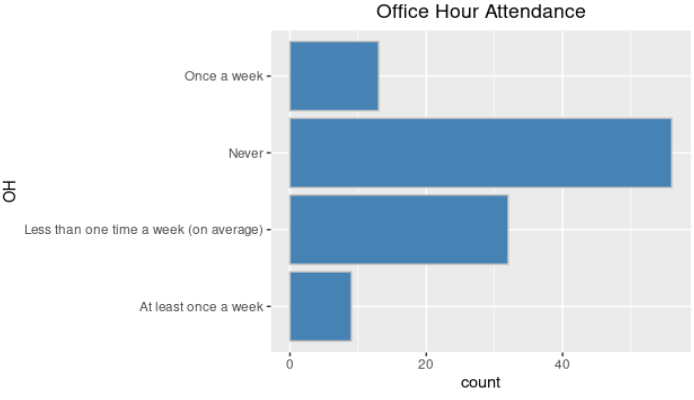


Figure 3

At least once a week	Less than one time a week (on average)
9	32
Never	Once a week
56	13

TABLE 3

Table 3 shows a summary of office hour attendance numbers.

A majority of students show positive attitudes towards their familiarity on course materials. 5% students make use of office hours to ask questions while the remaining 95% in this course do not go to office hours frequently.

Statistics of studying hours over 4 weeks is summarized as below (Table 2 and Figure 5).

Histogram	peak	# of mode	symmetry
Studying	48	unimodal	right skewed
Studying2	27	unimodal	symmetric
Studying3	40	unimodal	right skewed
Studying4	63	unimodal	right skewed

TABLE 4

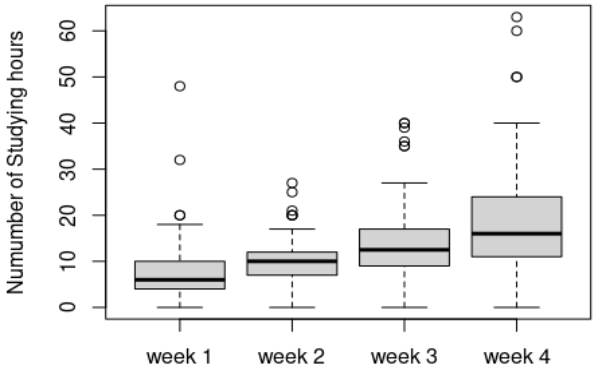


Figure 5

Students' studying hours, particularly the median, stably increase through 4 weeks. For each week, the distribution is unimodal. A few outliers are detected as well, nevertheless, they are acceptable in this stage.

However, different from studying time, miscellaneous time shows no obvious patterns in general. This pheonena is partially because everyone has different standards for miscellaneous, in other words, misconception errors are highly possible to exist in our collected data.

Likewise, distribution of covid concern time does not have dramatic changes, left-skewed each, most clustered around 1 hour, and there are only under 10% people concern more than 2 hours on a weekly basis.

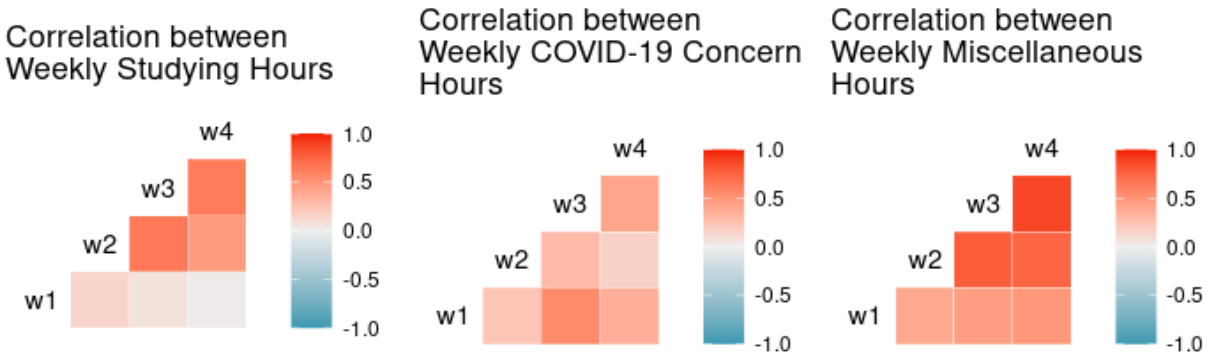


Figure 6

As we can see, from the above correlation maps (See Figure 6), correlation across any of weekly studying hours, weekly COVID-19 hours, and weekly miscellaneous hours is super high. Therefore, it might be a bad idea to include weekly raw data into the model. We need to find a solution to avoid multicollinearity. Thus, we consider taking the average of different activity hours as new variables in our next step.

Exploratory Data Analysis (EDA)

There are 110 observations and 16 variables in this data set (See previous Table 1). From these variables, 14 of them could be used as potential predictors to predict the grade a student receives on term test.

Firstly, a series of scatter plots (See Figure 7) reveal that there is an extremely weak negative relation between miscellaneous time and test grade. Thus, variables related to miscellaneous time may not significantly contribute to students' performance in the term test. Determined in previous section, and by common mathematical sense, we choose to aggregate similar variables into mean-averages. Denote mean study time before term test for each student as a new "study_avg" variable. Take the mean of the first four miscellaneous time for each student and denote it as a new "miscel_avg" variable. Denote the mean time a student thinks about covid as a new "COVID_avg" variable. The scatter plot (See Figure 8) shows there is a weak negative correlation between test

grade and miscellaneous hours. Likewise, based on the scatter plot, there is also a weak negative correlation between test grade and covid concern hours (See Figure 9).

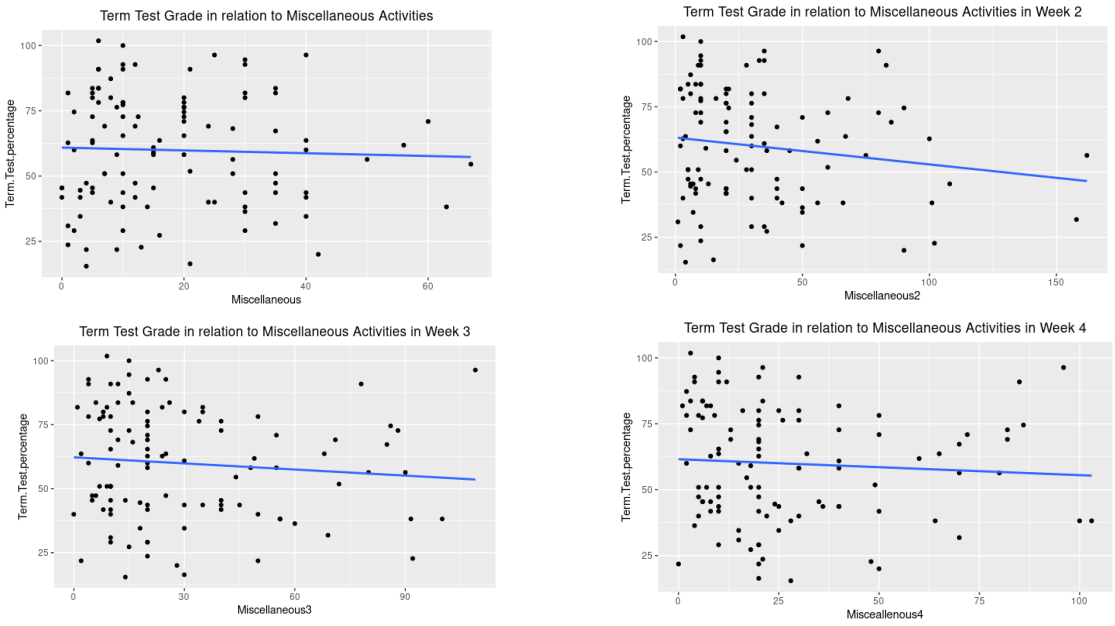


Figure 9

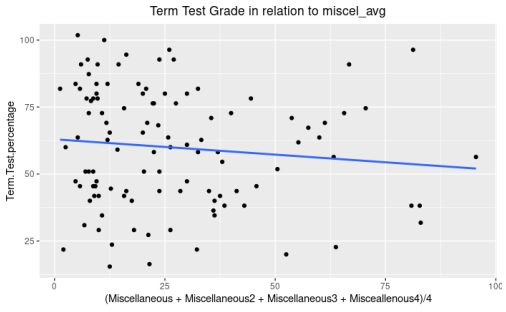


Figure 7

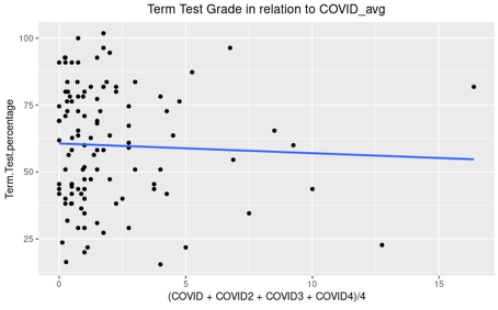


Figure 8

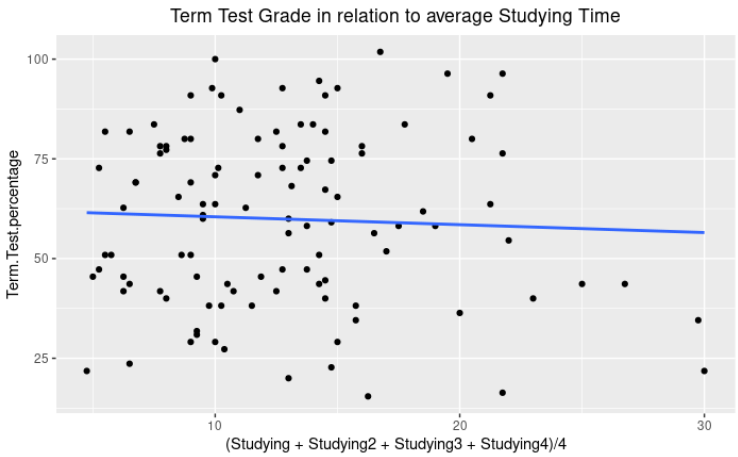


Figure 10

Vertical line means there is no relation. Figures show there are no relations between miscellaneous spending hours and term test grades, covid concern hours and test grade. However, there might be one or two leverage points which need to be checked in the next step.

Average studying hours seem not a significant factor as well (See Figure 10). This is against our common sense. In our common sense, studying hours should at least be a positive factor to test grade. Influential points need to be checked in the next step.

Next step, we are going to develop our candidate models and use F-test to verify our guess. If F-test indicates that any variables have no relationship with test grade, then we can remove them when developing the models, otherwise, we keep it.

Use side-by-side boxplots for a combination on numerical and categorical variables.

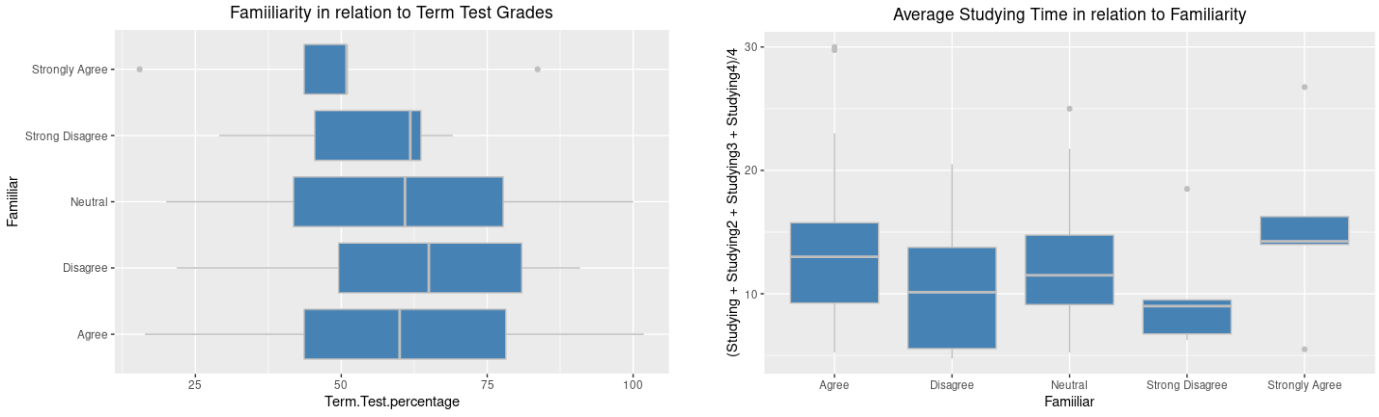


Figure 11

We can find no pattern in relation between familiarity and test grades except that the group of students who are comfortable with course materials tend to have longer studying hours and not

bad test grade (See Figure 11). However, when it comes to office hour attendance, test grade median shows large differences among groups of students. Most students who attend office hour at least once a week obviously have higher term test grades and spend more time studying compared to students who rarely go to office hour (See Figure 12). This is consistent with our common sense.

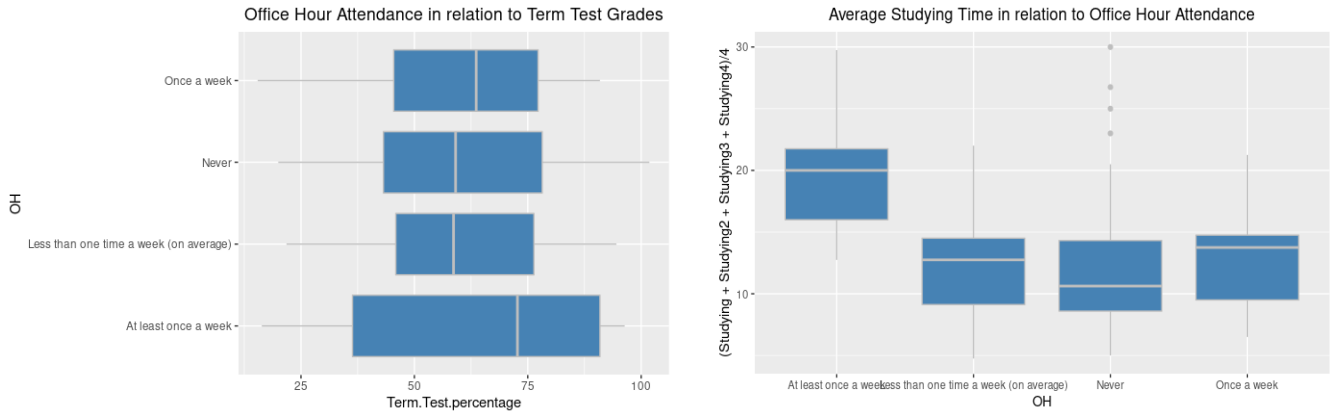


Figure 12

Model Development

Variable Selection

Step1: Start with an initial model

In this study, we start with our initial model. By our common sense and previous data exploration, miscellaneous time should be eliminated from our model. Notice that we cannot put the aggregate variables in the full model as it will get aliased coefficients and create perfect collinearity. Check VIF to see whether multicollinearity exists. Mean VIF is under 3. From this model, we notice some key factors by t-value, such as studying in first week and covid concern hours in week 3.

Step2: Do stepwise AIC backward elimination

Stepwise AIC backward elimination shows office hour attendance, studying time in week 2, week3, and week4, covid concern time in week1 and week2 should be eliminated. The remaining variables are studying time in week1, covid concern time in week3 and 4. The summary output shows p-value is 0.01937, which is extremely small.

Step3: Choose a few candidate models

After doing AIC backward Elimination, based on previous EDA results and the summary output, a few candidate models are listed as follows.

Candidate model 1:

The idea of the model is that as time goes by, pandemic is recovering, and covid concern time variation among students is small. Continuous efforts in studying, in other words, studying hours should be the biggest factor to determine students' performance while covid factor can nearly be omitted. From initial model, we know studying time in first week may matter more than those in other weeks, and high correlation between weekly studying hours is detected in previous part. Studying time in the first week should be a significant variable to build the model. It is a good signal that F-test shows p-value is under 0.057, much smaller compared with initial model, which means we are moving in the right direction.

Candidate model 2:

The idea of the model is keeping variables in candidate model 1 and involving what we find from initial model, the interaction factors of familiarity with studying hours. This model seems also not bad with p-value around 0.058.

Candidate model 3:

The idea of this model is replacing one variable in candidate model 1 with another similar variable for comparison.

Model Selection

	K	RSS	R squared	adjusted R squared	AICc	AICcWt
full model	17	12.53	0.2386	0.05134	634.34	0
AIC_reduced_model	17	12.28	0.1261	0.09021	611.42	0.53
candidate_model 1	4	12.54	0.0749	0.04991	613.52	0.19
candidate_model 2	8	12.32	0.1554	0.08303	616.07	0.05
candidate_model 3	4	12.51	0.08056	0.05571	613.04	0.23

TABLE 5

Table 5 gives the information of model selection. K is the number of parameters in the model. AICc is the information score of the model calculated from the AIC test corrected for small sample sizes. The smaller the AIC value, the better the model fit. AICcWt is the proportion of the total amount of predictive power provided by the full set of models contained in the model being assessed. Therefore, the best model is AIC_reduced_model.

Model Diagnostic

Figure 13 is a summary of diagnostic plots for our selected model.

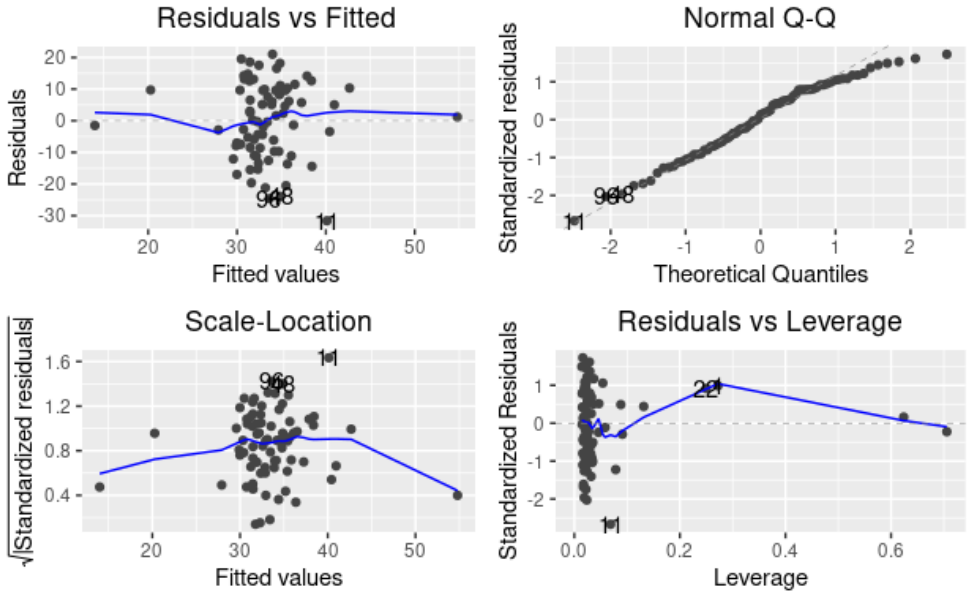


Figure 13

In this step, we use residual plots and Q-Q plots to check any assumption violations.

Step1: Independence assumption

Our data is independently collected, therefore, we can assume the independence assumption hold.

Step2: Linearity Verification

There are no obvious non-random patterns in the Residual vs Fitted value plot, which suggests that the linearity assumption is satisfied. There are some strange diagonal lines in the graph. These diagonal lines are possibly due to test grade being not truly continuous, but most predictors are continuous. See the Limitations of Model section for more discussion on this.

Step 3: Outliers, Leverages, Influential Points

Leverage value for outlying observations and the cook's distance for influential observation are checked. Since there are no observations that are both outlying and influential, we do not need to remove any observation (See Table 6).

row	Y	Y_hat	t	Cook's distance	isOutlier	isInfluential
1	53	42.67	0.9866049	9.10E-02	FALSE	FALSE
2	25	36.11	-0.919596	7.32E-03	FALSE	FALSE
3	45	30.77	1.177733	9.06E-03	FALSE	FALSE
4	28	34.12	-0.50517	2.34E-03	FALSE	FALSE
5	32	31.76	0.019556	1.83E-06	FALSE	FALSE
6	52	37.88	1.1753604	1.31E-02	FALSE	FALSE

TABLE 6

Step4: Normality Assumption

Before checking for homoscedasticity, we should make sure the normality assumption is satisfied. Looking at the Normal Q-Q plot, the lower portion of the plot is over the Q-Q line, meaning we have a right-skewed distribution. This result matches up with the conclusion from the summary of test score in the previous section (See Table 4).

Step5: Homoscedasticity Assumption

The Scale-Location plot for the refitted model shows that except the right portion fits not so well, the line is approximately straight in the left portion of the plot, and the point appears random, suggesting that the constant variance assumption is mostly satisfied.

Step6: Transformation

Box-Cox tells that the 95% confidence interval for λ contains 1 (See Figure 14), which means it is unnecessary to do transformation. No transformation is better. Therefore, we keep as original.

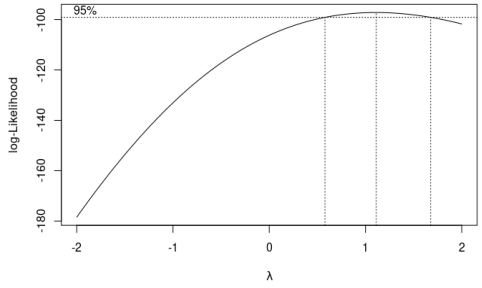


Figure 14

Step7: Multicollinearity

None of the VIF of the selected model is greater than 1.1. The mean VIF is 1. Based on all above, candidate model 1without transformation is the final model.

Final Model:

Term.Test = 28.4971 + 0.4968*Studying + 1.5220 * COVID3 - 0.5892 * COVID4

Where Term.Test is test grade out of 55, COVID3 is the number of hours a student thinks about Covid-19 in week 3, Studying is the number of hours a student studies in week 1 (See Table 1 for complete data description).

Model Validation

Step1: check whether the model coefficients are similar when fitting on the training and testing data

We validate our model against the testing data set. The model coefficients on the testing data are close but p-value 0.15 is much larger than the p-value on training data. This indicates our model may suffer on predictive ability.

Step2: evaluate whether assumption violation is similar when fitting on the training data and testing data

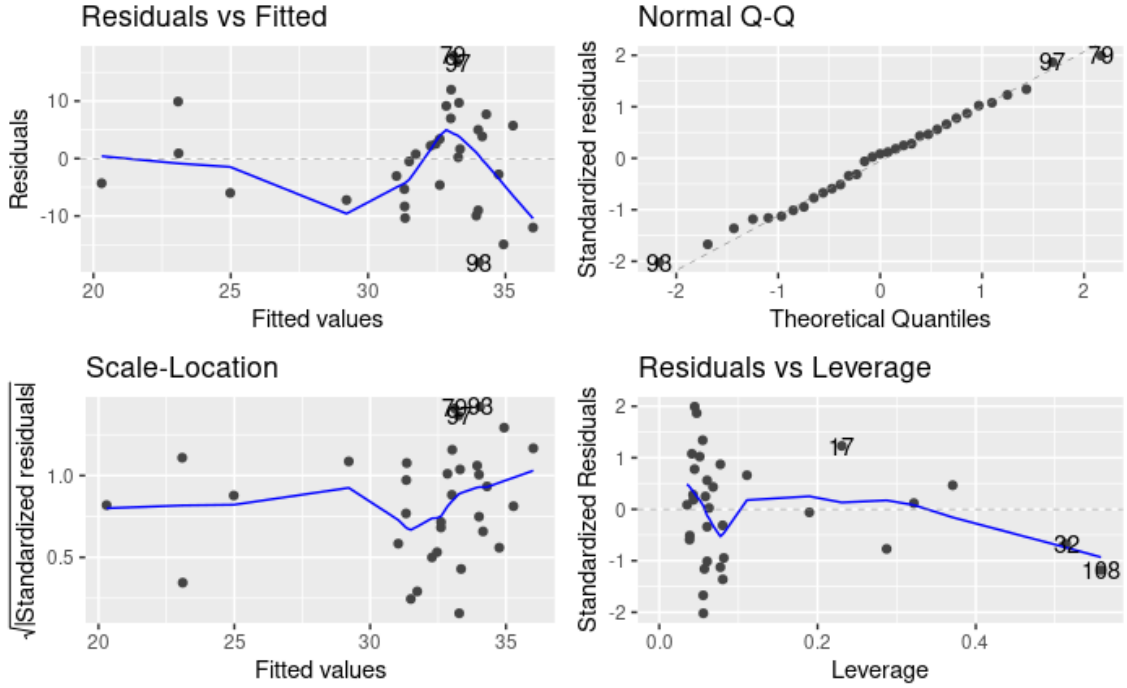


Figure 15

Plots are generally similar as before except that the lower portion of the plot is under the Q-Q line this time, meaning a left-skewed distribution (See Figure 15).

Partial F test of this model also pass with small p-value shown in ANOVA.

Conclusion & Discussion

Model Interpretation

The proposed final model is in the form:

Term Test Grade = $\beta 0 + \beta 1 \times$ Studying + $\beta 2 \times$ COVID3 + $\beta 3 \times$ COVID4

Where $\beta 0$, $\beta 1$, $\beta 2$, $\beta 3$ are coefficients calculated using least square method for linear models.

$\beta 0 = 2.6279$ is the coefficient of the intercept term. It means that the true mean of Term Test Grade of students who does not spend any time studying and concerning about COVID is 2.6279.

$\beta 1 = 0.4968$ is the coefficient of studying hours in the first week. It means that when other variables are held constant, students' mean grades would improve 0.4968 per hour spending in studying in the first week.

$\beta 2 = 1.5220$ is the coefficient of covid concern hours in week 3. It means that when other variables are held constant, students' mean grades would improve 1.5220 per hour thinking about Covid-19 in the third week.

$\beta 3 = -0.5892$ is the coefficient of covid concern hours in week 4. It means that when other variables are held constant, students' mean grades would lose 10.5892 per hour thinking about Covid-19 in the fourth week.

Advantage of Model

The final model is very simple and straightforward. It only contains three variables and computationally simple. Each of the variables is intuitive to understand.

Disadvantage of Model

Data limitations:

The size of the data set is tiny. The data set only has 110 observations, which is already quite small to start with. Furthermore, response bias exists during data collection. Some data points are clearly out of range.

Survey questions are somewhat unclear. Just as the analysis in Experiment information and EDA part, it is very hard for respondents to measure variables such as miscellaneous activity hours fairly.

Model limitations:

Since the marking of the quiz is done by summing the partial marks of each question, therefore t term test grade may not be continuous in real situations. There will always be some score that can't be reached.

Taking average in EDA might not be a perfect solution for all students since their studying habits vary. For example, some students may not start to review course materials until the final second. Sometimes for this group of students, even though their sum of studying hours is not large, their efficiency can maximize as exam date approaches. Taking the equal weight to the studying hours of each week might underestimate the significance of studying hours in the last week before exam.

The model we are proposing is possibly not the best. As mentioned in validation part, our model may not generalize well into other samples from the population and suffer from weak predictive competency.

Conclusion

In this study, a linear regression model is proposed to predict students' performance in the term test. Despite the association is very low, there are many factors that influenced the poor association, including the hybrid method of class lectures. Recording and having in person lectures can affect the hours studied per week. The summer term accelerated teaching and learning pace is also a big challenge for students.

Proposed Future Work

Based on the existing limitations, one way to improve the model is to try to collect as many composite variables as possible. These variables had better have real-life meanings regardless of the pandemic background. For example, factors like sleeping habits, sports frequency, comfort of joining in study groups and so on can enrich our dataset and provide more insights into our current work. Moreover, another improvement is to introduce learning pattern of courses as a variable since different from calculation-oriented science courses, writing-heavy social science and humanities courses have unique styles of learning. For some courses, grades cannot improve rapidly even adequate studying hours have been paid into them. Last but not least, the difficulty level of the exam varies from instructor to instructor as well. In the future, we can think about how to minimize this influence and make our model more generalized in order to benefit more professors and students.

Appendix

Reference

Fadhli, M. (2022). Optimize Learning Process during the Covid-19 Pandemic through IT-Based Learning Media at SMPN 10 Bengkulu. *Engagement: Jurnal Pengabdian Kepada Masyarakat*, *6*(1), 243-251.

Gonzalez, T., De La Rubia, M. A., Hincz, K. P., Comas-Lopez, M., Subirats, L., Fort, S., & Sacha, G. M. (2020). Influence of COVID-19 confinement on students' performance in higher education. *PloS one*, *15*(10), e0239490.

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., ... & Chen, P. K. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, *12*(1), 1-15.

Weisburd, D., Wilson, D. B., Wooditch, A., & Britt, C. (2022). Multiple regression. In *Advanced Statistics in Criminology and Criminal Justice* (pp. 15-72). Springer, Cham.

Code

(See next page)

RegProject

Hanqi Zeng

Contents

```
data <- read.csv("302data.csv",header = TRUE)
```

```
glimpse(data)
```

```
## Rows: 110
## Columns: 16
## $ ID              <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
## $ Studying        <dbl> 5, 9, 2, 5, 4, 8, 1, 9, 11, 10, 14, 10, 8, 5, 3, 11, 6,~
## $ COVID           <dbl> 6.0, 3.0, 1.0, 5.0, 0.0, 2.0, 1.0, 2.0, 3.0, 0.5, 3.0, ~
## $ Miscellaneous   <dbl> 40, 5, 5, 10, 15, 3, 20, 30, 28, 20, 4, 20, 15, 6, 1, 5~
## $ Studying2       <dbl> 27.0, 8.0, 4.0, 5.0, 9.0, 15.0, 6.0, 10.0, 12.5, 14.0, ~
## $ COVID2          <dbl> 5.00, 3.00, 1.00, 2.00, 2.00, 6.00, 1.00, 0.00, 1.50, 0~
## $ Miscellaneous2 <dbl> 80, 7, 2, 5, 45, 7, 30, 10, 30, 20, 4, 90, 12, 28, 10, ~
## $ Studying3       <int> 24, 8, 8, 6, 21, 20, 12, 18, 14, 20, 14, 16, 14, 12, 7,~
## $ COVID3          <dbl> 10.00, 4.00, 2.00, 4.00, 2.00, 12.00, 0.00, 4.00, 1.00,~
## $ Miscellaneous3 <dbl> 109, 5, 9, 10, 48, 18, 20, 15, 16, 40, 14, 86, 12, 12, ~
## $ Studying4       <int> 31, 12, 8, 6, 36, 20, 12, 20, 15, 20, 25, 16, 24, 12, 7~
## $ COVID4          <dbl> 6.00, 5.00, 3.00, 5.00, 3.00, 10.00, 0.00, 2.00, 0.50, ~
## $ Miseallenous4 <int> 96, 6, 7, 10, 40, 15, 20, 10, 20, 30, 28, 86, 18, 12, 2~
## $ OH              <chr> "At least once a week", "Once a week", "Less than one t~
## $ Famiiliar       <chr> "Agree", "Agree", "Neutral", "Strongly Agree", "Neutral~
## $ Term.Test       <dbl> 53.0, 25.0, 45.0, 28.0, 32.0, 19.0, 42.0, 52.0, 37.5, 4~
```

Creating training and test set

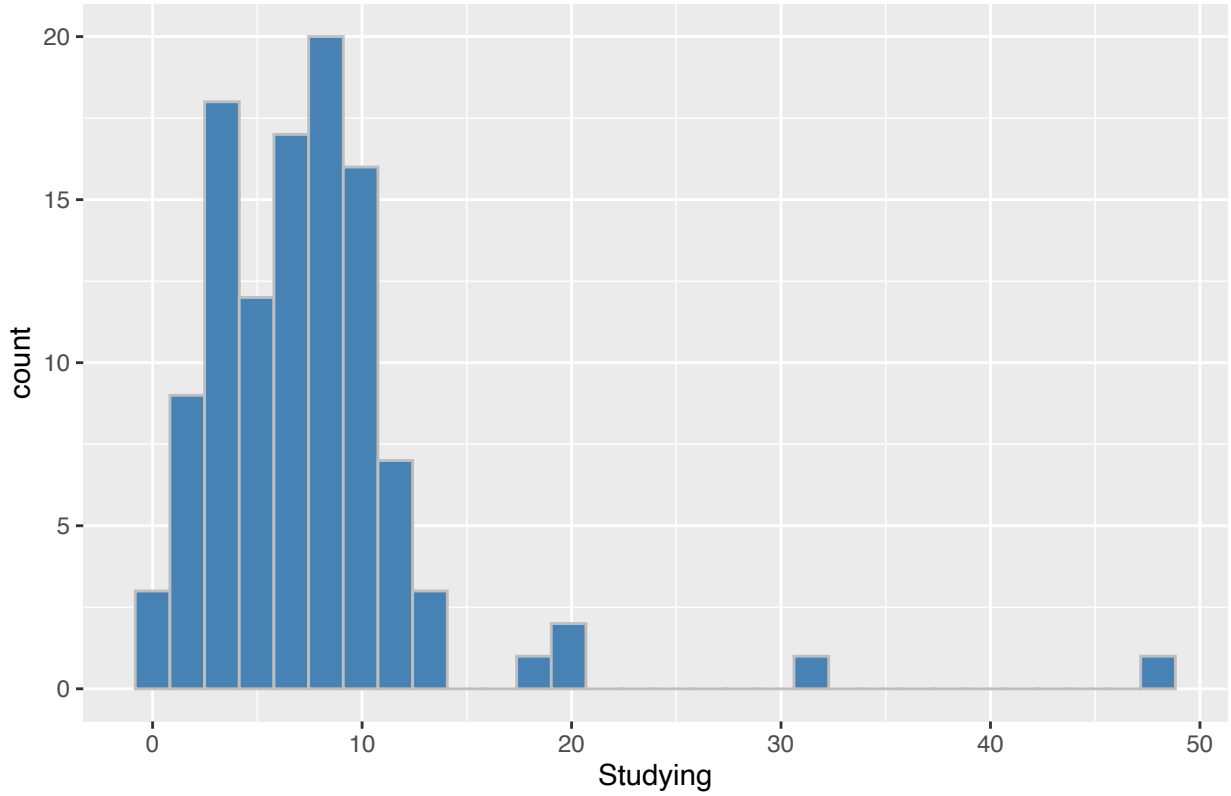
```
set.seed(1008124245)
```

```
    # add aggregate variables
    data <- data %>%
      mutate(study_avg = rowMeans(cbind(Studying, Studying2, Studying3, Studying4))) %>%
      mutate(COVID_avg = rowMeans(cbind(COVID, COVID2, COVID3, COVID4)))
```

```
# random split train 70% test 30%
dt = sort(sample(nrow(data), nrow(data)*.7))
train<-data[dt,]
test<-data[-dt,]
```

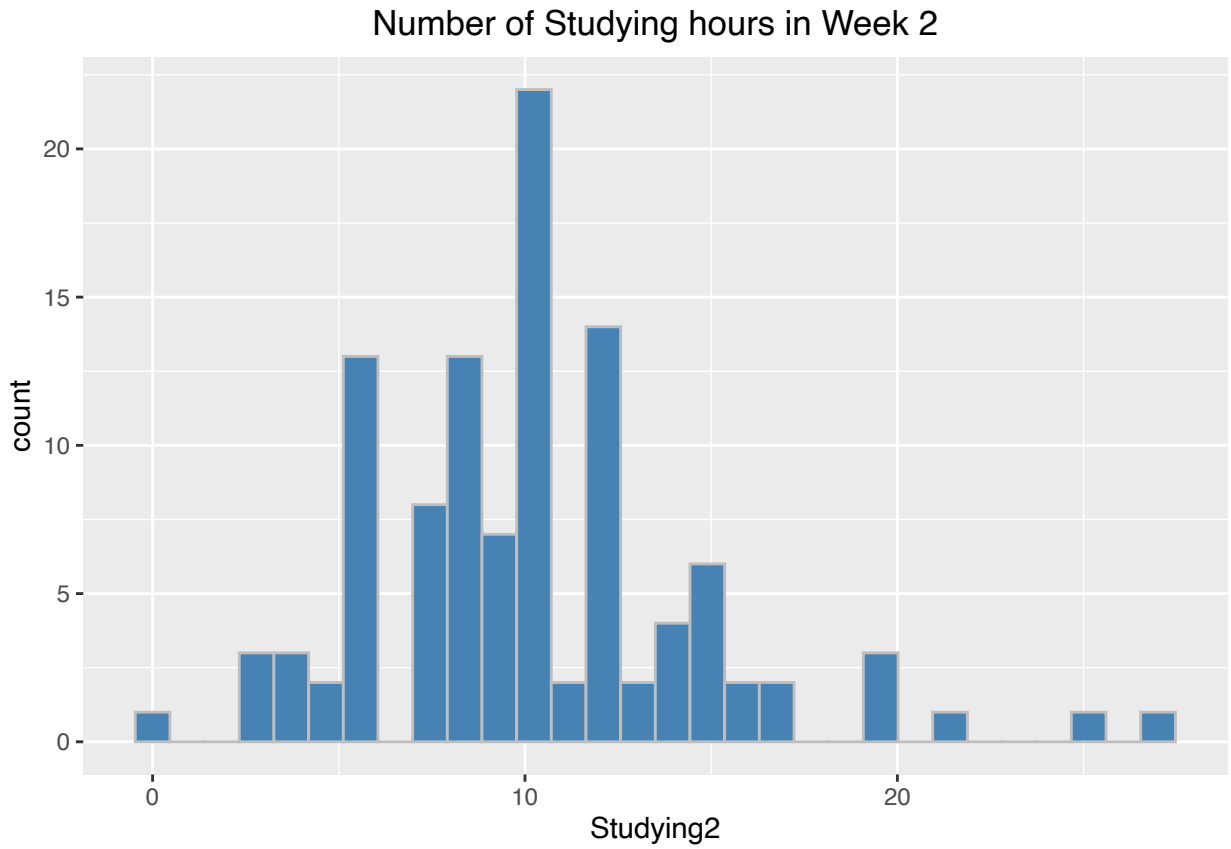
```
data %>%
  ggplot(aes(x=Studying))+geom_histogram(
  color="gray",
  fill="steelblue")+
  labs(title="Number of Studying hours in Week 1")+
  theme(plot.title=element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



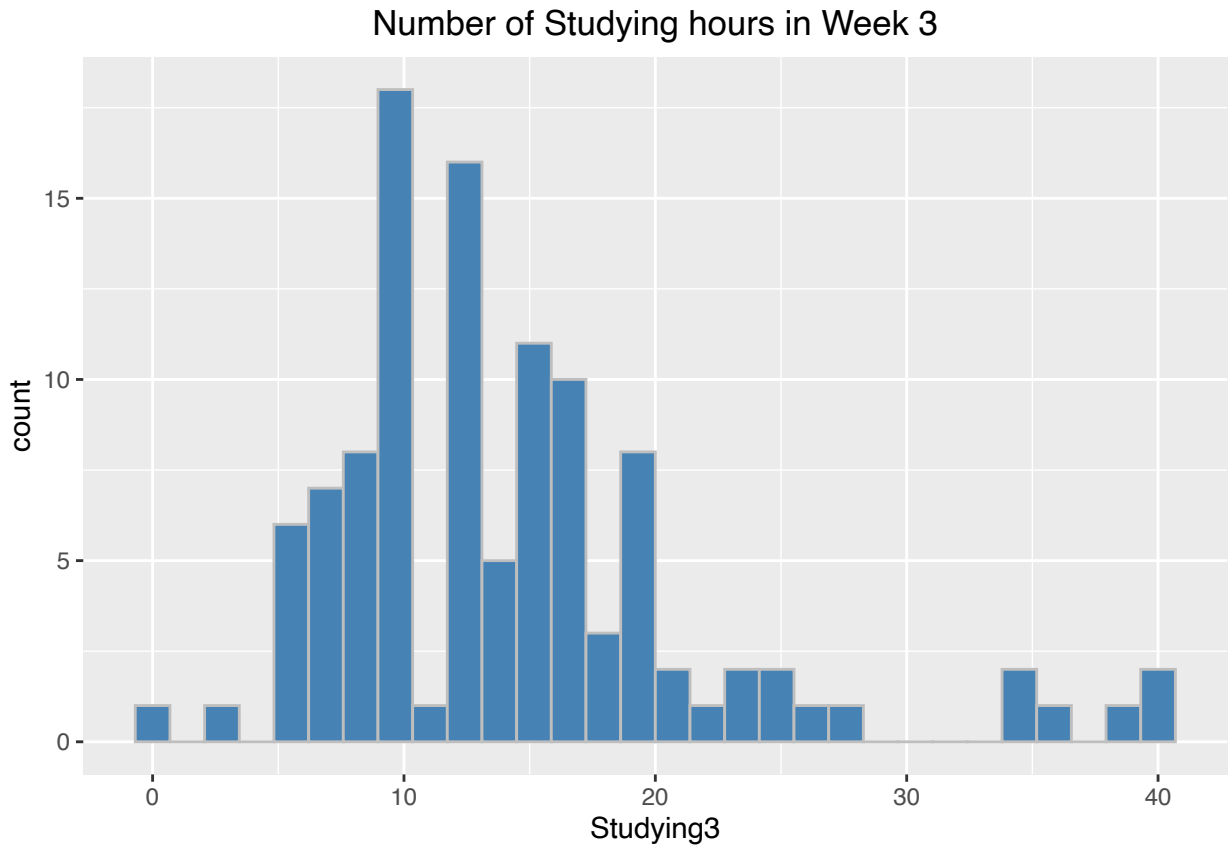

```
data %>%
  # by default bins=30
  ggplot(aes(x=Studying2))+geom_histogram(
  color="gray",
  fill="steelblue")+
  labs(title="Number of Studying hours in Week 2")+
  theme(plot.title=element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



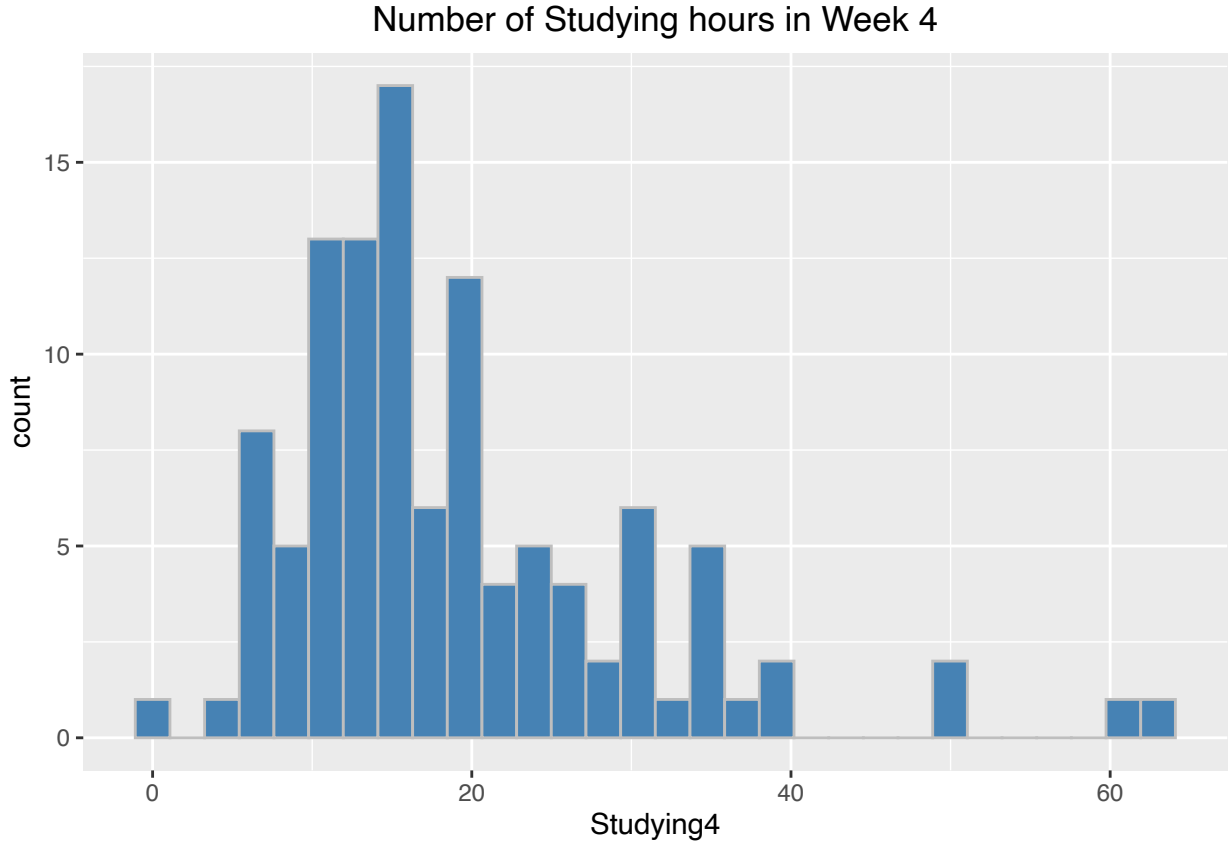

```
data %>%
  # by default bins=30
  ggplot(aes(x=Studying3))+geom_histogram(
  color="gray",
  fill="steelblue")+
  labs(title="Number of Studying hours in Week 3")+
  theme(plot.title=element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

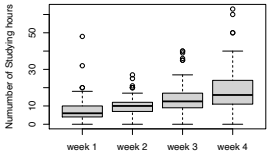



```
data %>%
  # by default bins=30
  ggplot(aes(x=Studying4))+geom_histogram(
  color="gray",
  fill="steelblue")+
  labs(title="Number of Studying hours in Week 4")+
  theme(plot.title=element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

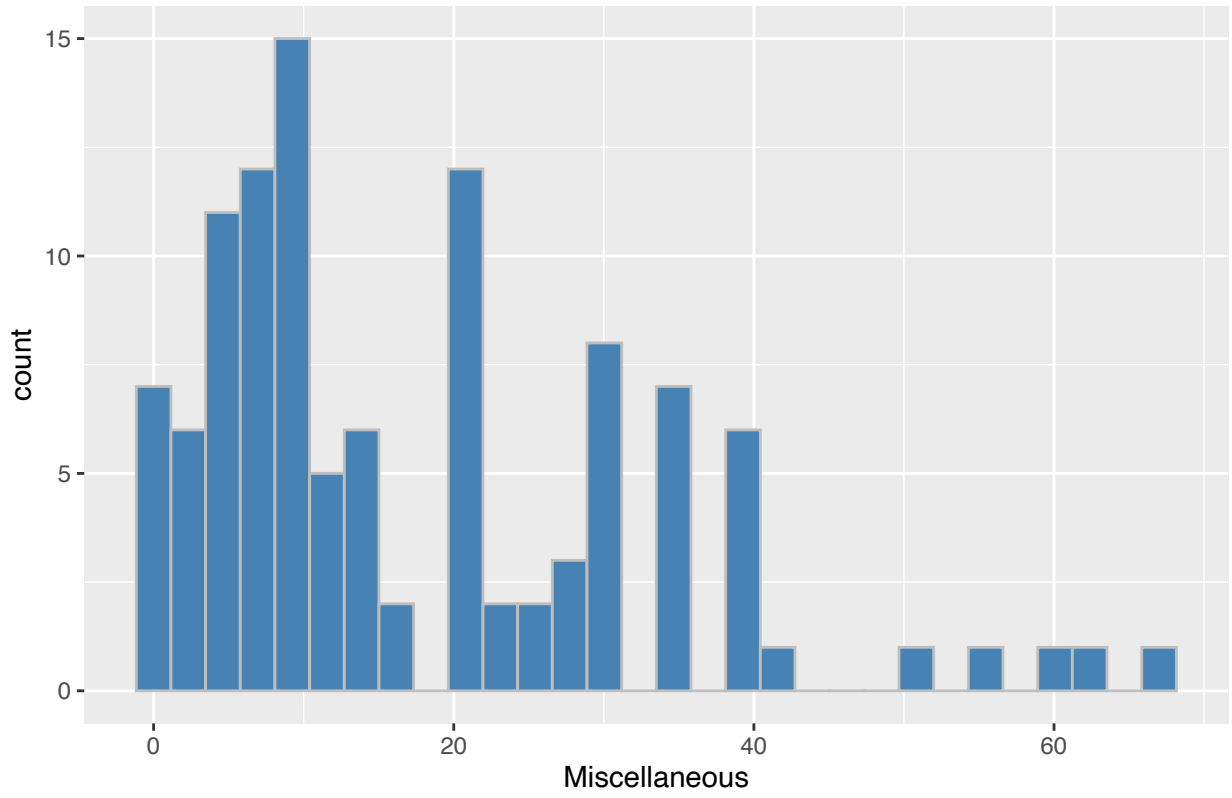



```
boxplot(data$Studying, data$Studying2, data$Studying3, data$Studying4,
names=c("week 1", "week 2", "week 3", "week 4"),
ylab="Numumber of Studying hours")
```



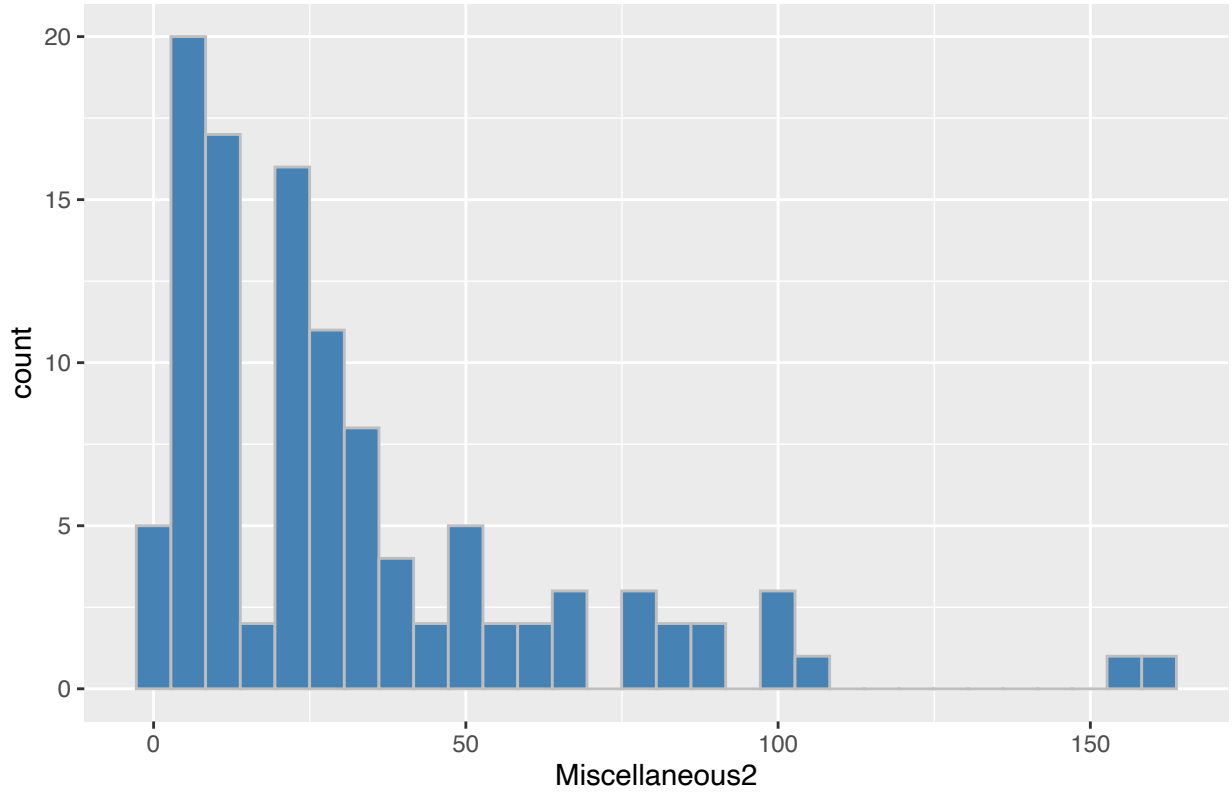

```
data %>%
  # by default bins=30
  ggplot(aes(x=Miscellaneous))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of Miscellaneous hours in Week 1")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



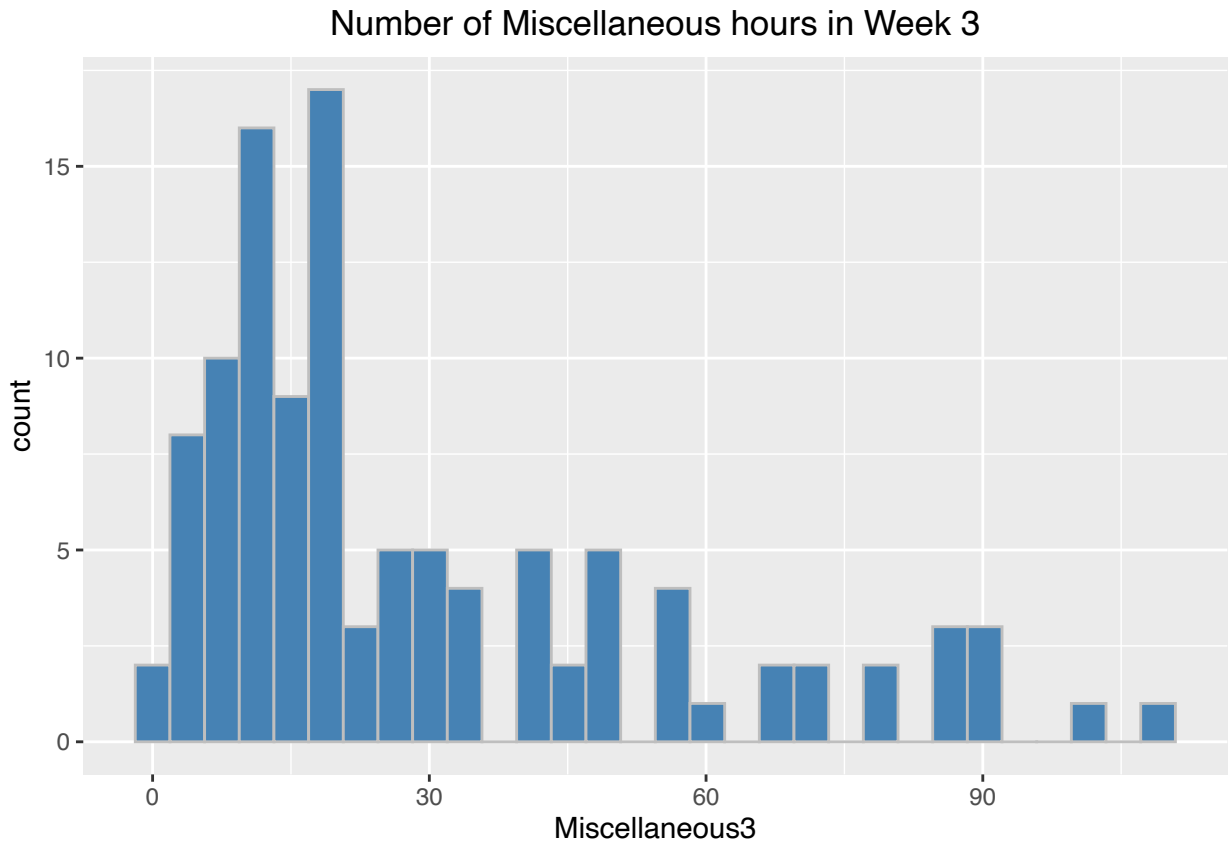

```
data %>%
  # by default bins=30
  ggplot(aes(x=Miscellaneous2))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of Miscellaneous hours in Week 2")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



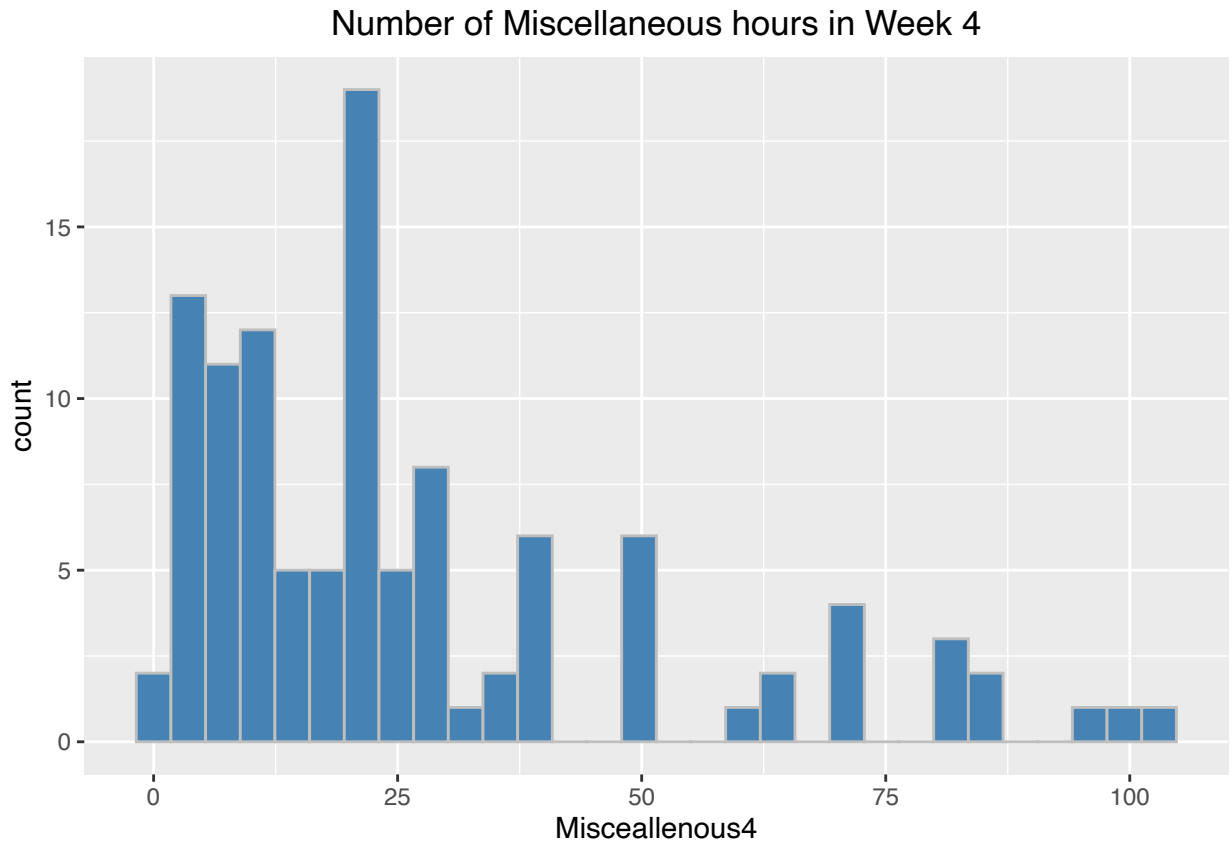

```
data %>%
  # by default bins=30
  ggplot(aes(x=Miscellaneous3))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of Miscellaneous hours in Week 3")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

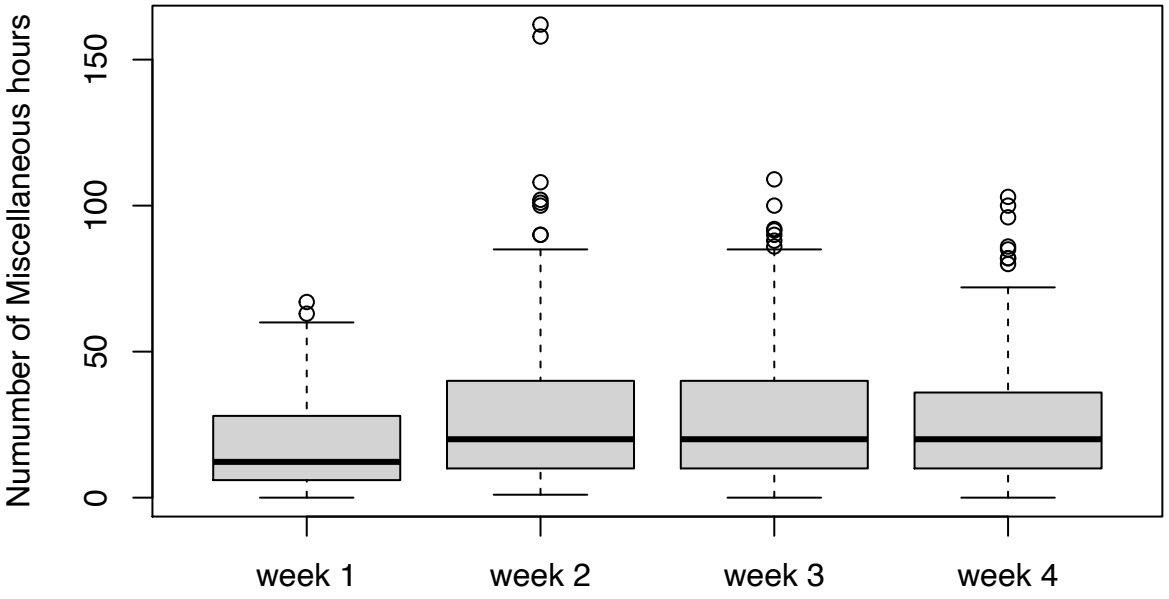



```
data %>%
  # by default bins=30
  ggplot(aes(x=Miseallenous4))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of Miscellaneous hours in Week 4")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

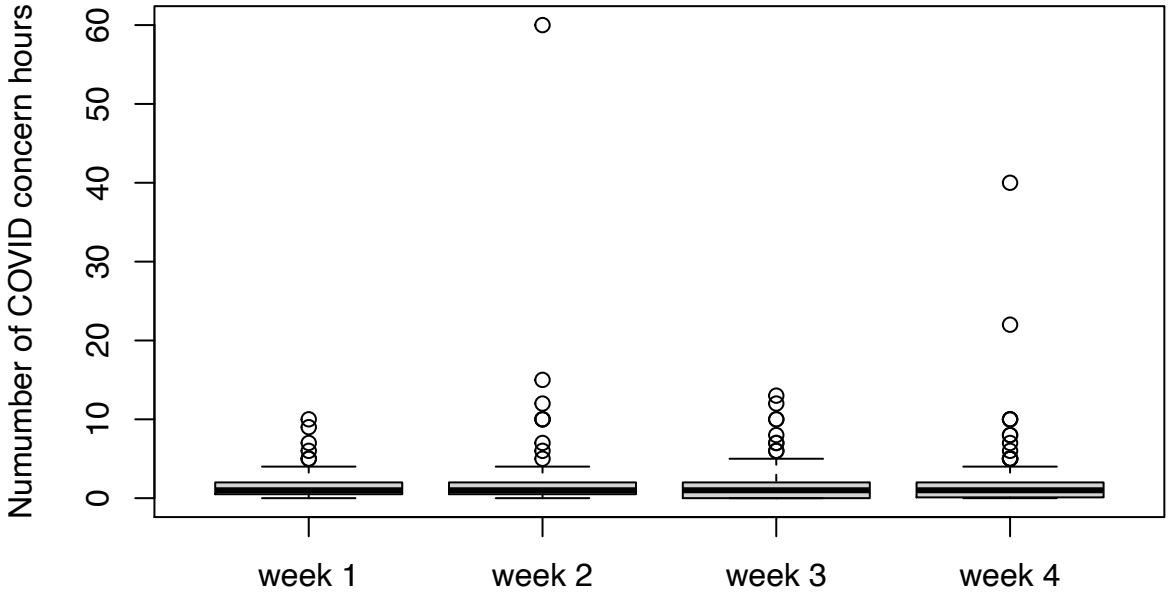



```
boxplot(data$Miscellaneous, data$Miscellaneous2, data$Miscellaneous3, data$Misceallenous4,
names=c("week 1", "week 2", "week 3", "week 4"),
ylab="Numumber of Miscellaneous hours")
```



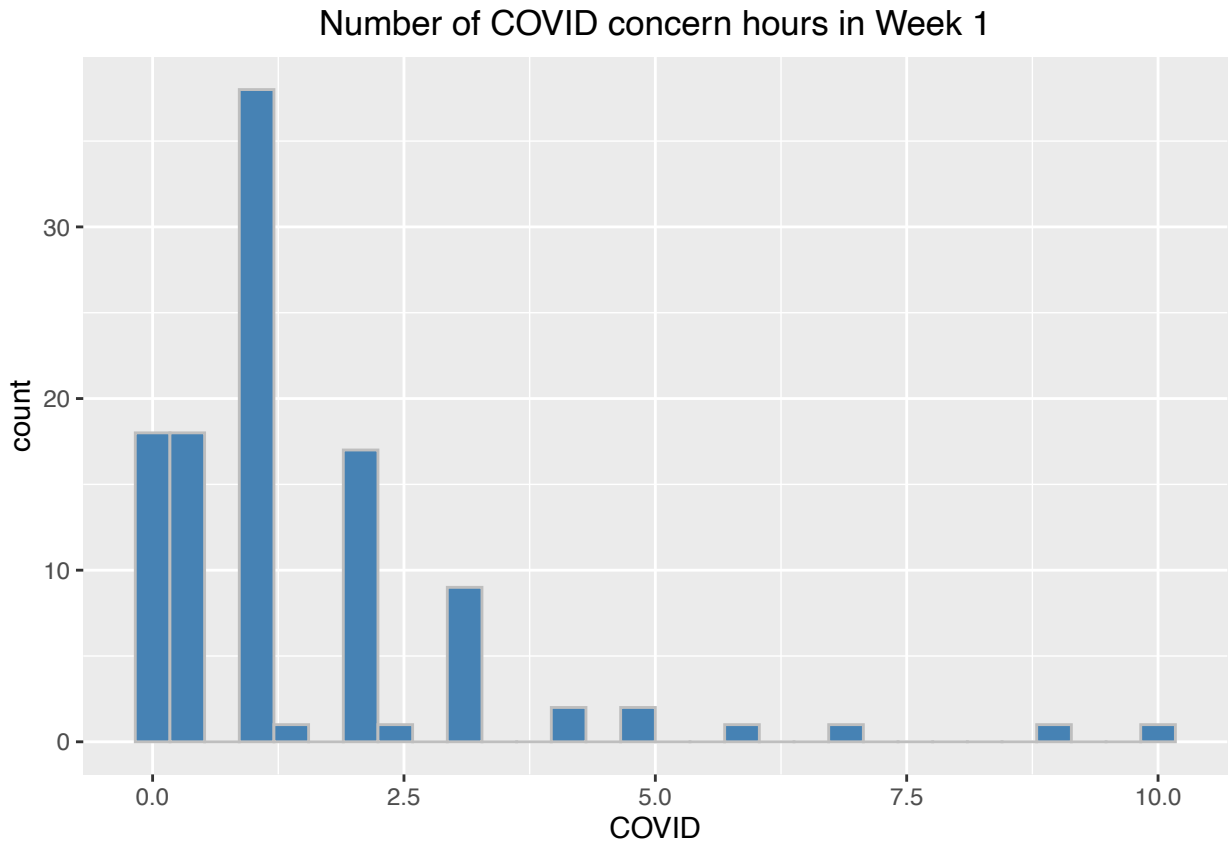

```
boxplot(data$COVID, data$COVID2, data$COVID3, data$COVID4,
```

```
names=c("week 1", "week 2", "week 3", "week 4"),
ylab="Numumber of COVID concern hours")
```



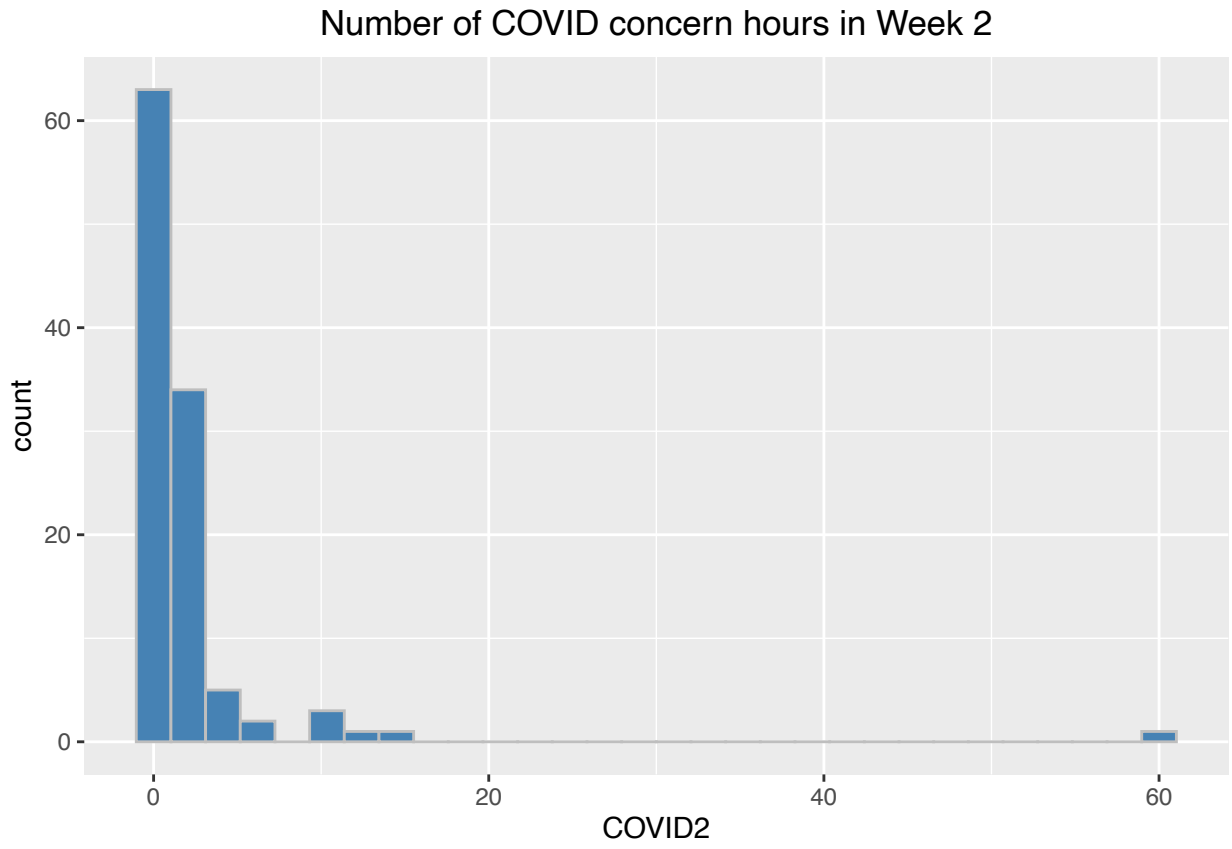

```
data %>%
  # by default bins=30
  ggplot(aes(x=COVID))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of COVID concern hours in Week 1")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



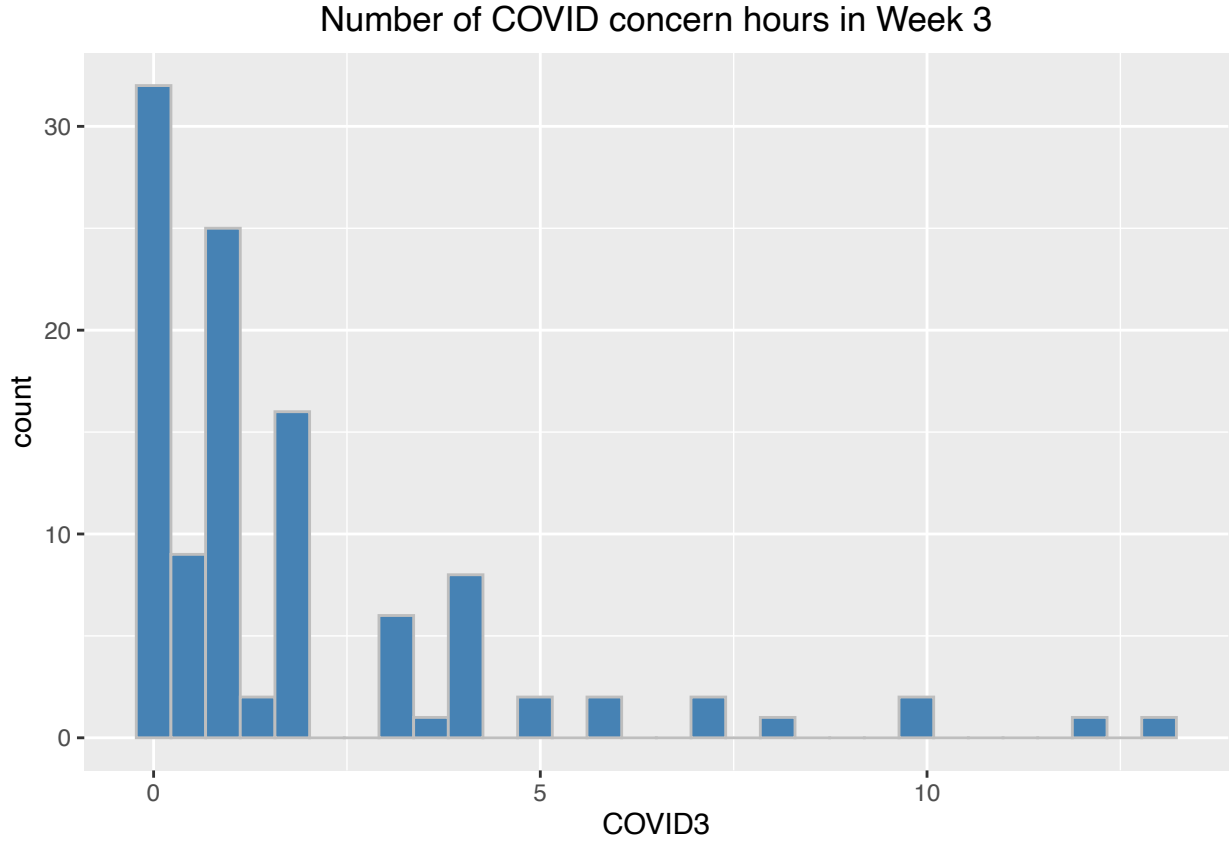

```
data %>%
  # by default bins=30
  ggplot(aes(x=COVID2))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of COVID concern hours in Week 2")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



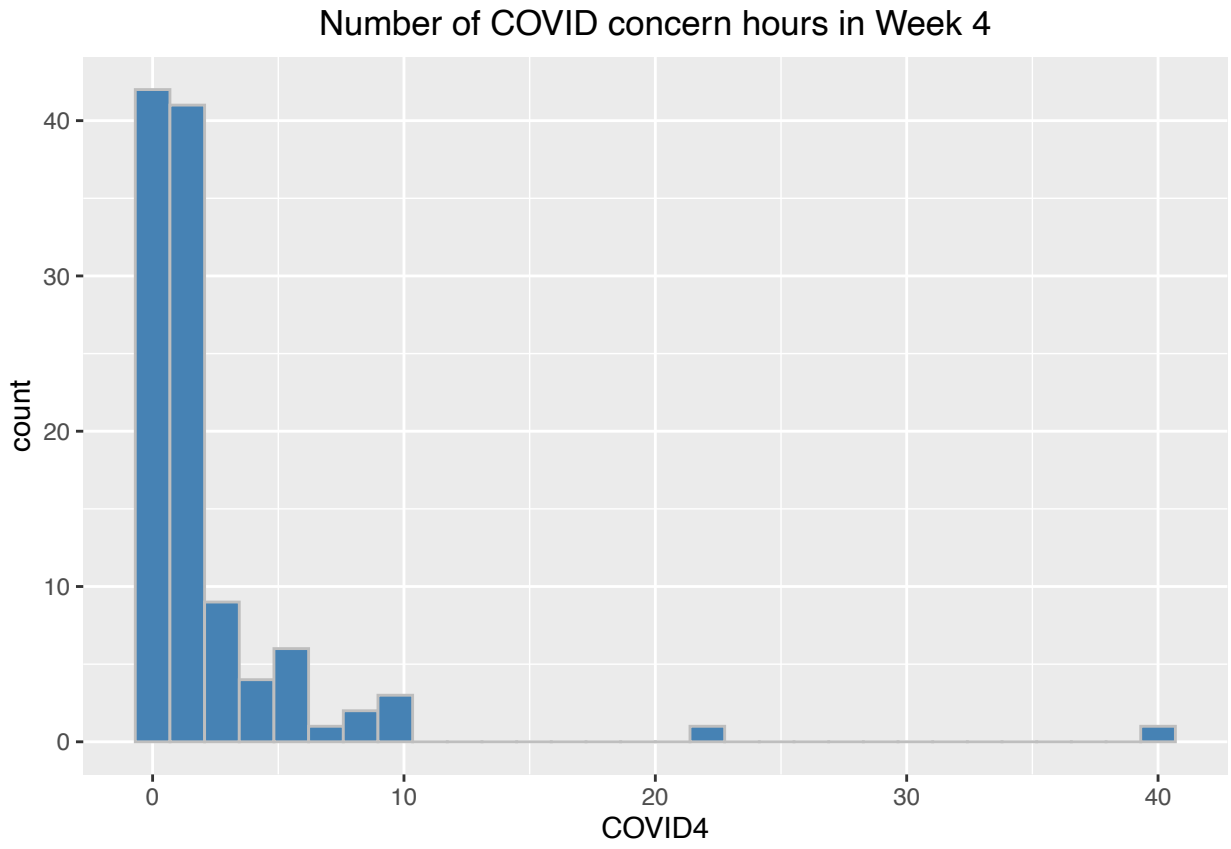

```
data %>%
  # by default bins=30
  ggplot(aes(x=COVID3))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of COVID concern hours in Week 3")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



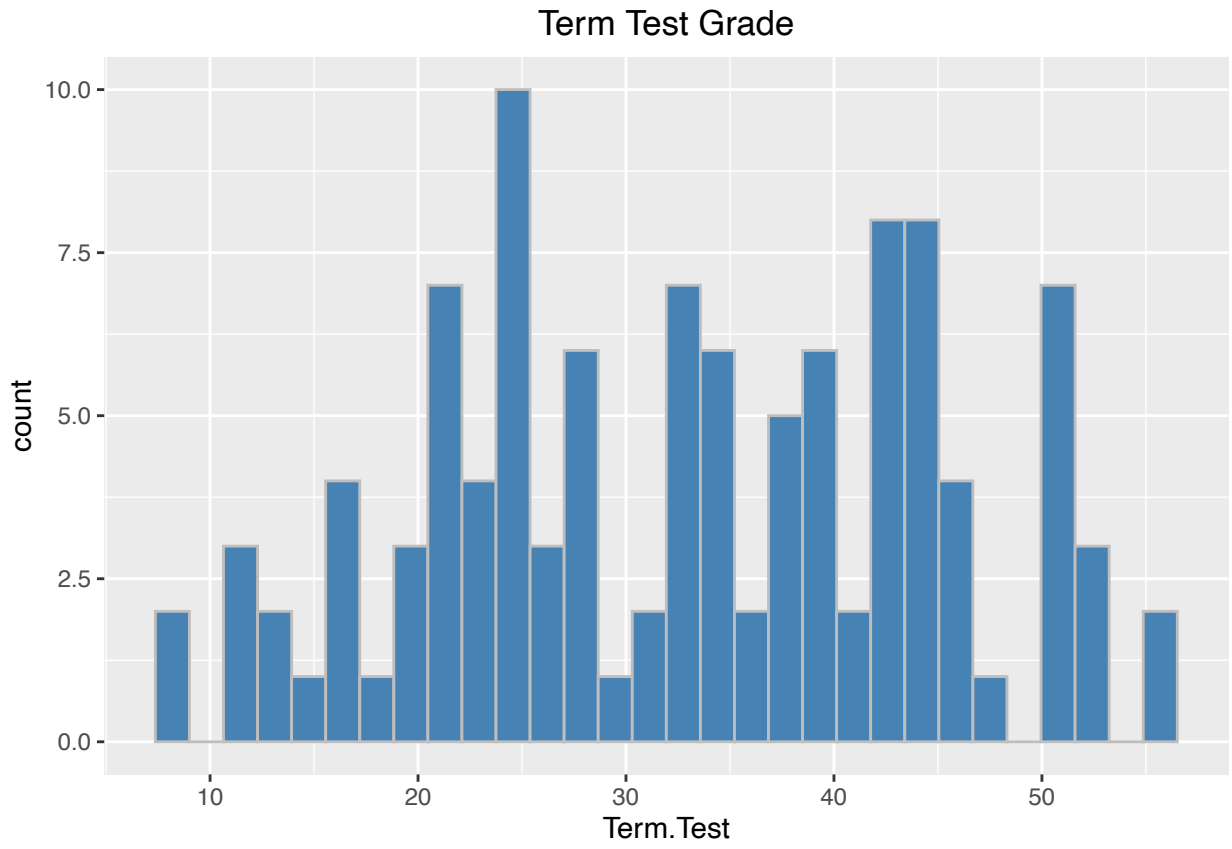

```
data %>%
  # by default bins=30
  ggplot(aes(x=COVID4))+
  geom_histogram(color="gray", fill="steelblue")+
  labs(title="Number of COVID concern hours in Week 4")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

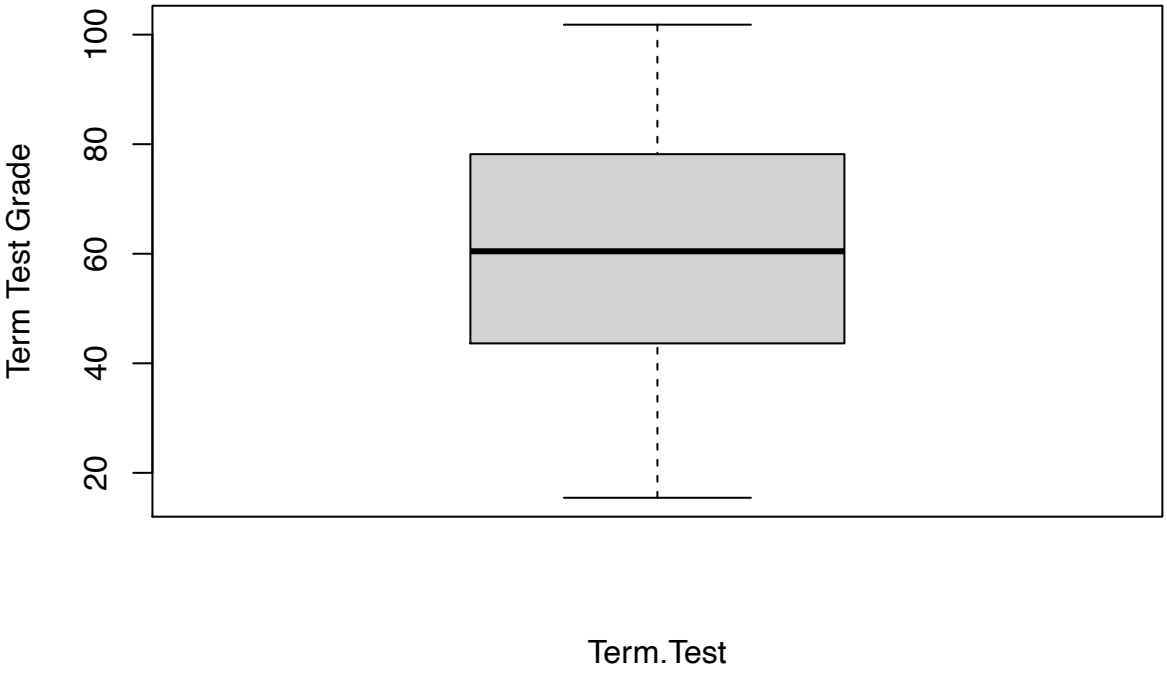



```
data %>%
  # by default bins=30
  ggplot(aes(x=Term.Test))+geom_histogram(color="gray", fill="steelblue")+
  labs(title="Term Test Grade")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



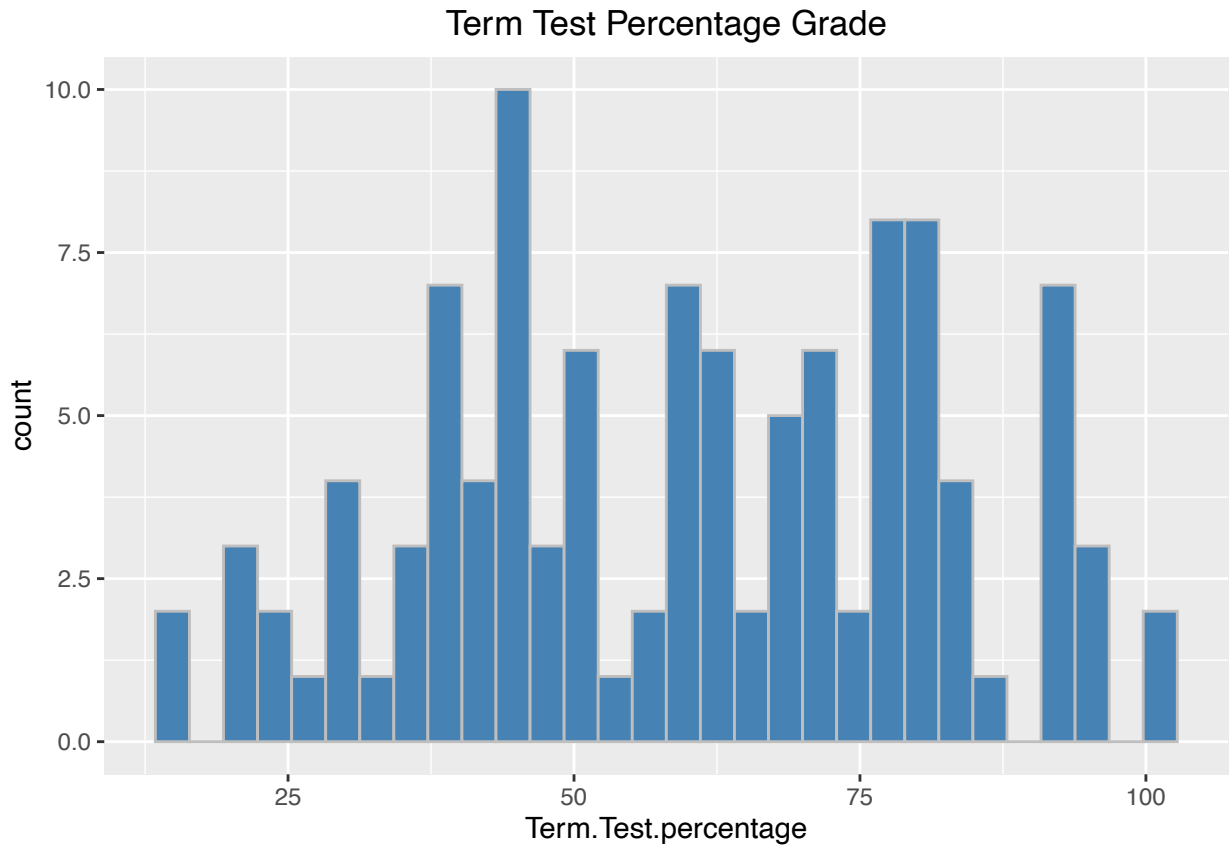

```
data$Term.Test.percentage <- data$Term.Test/55 * 100
boxplot(data$Term.Test.percentage, xlab="Term.Test", ylab="Term Test Grade")
```



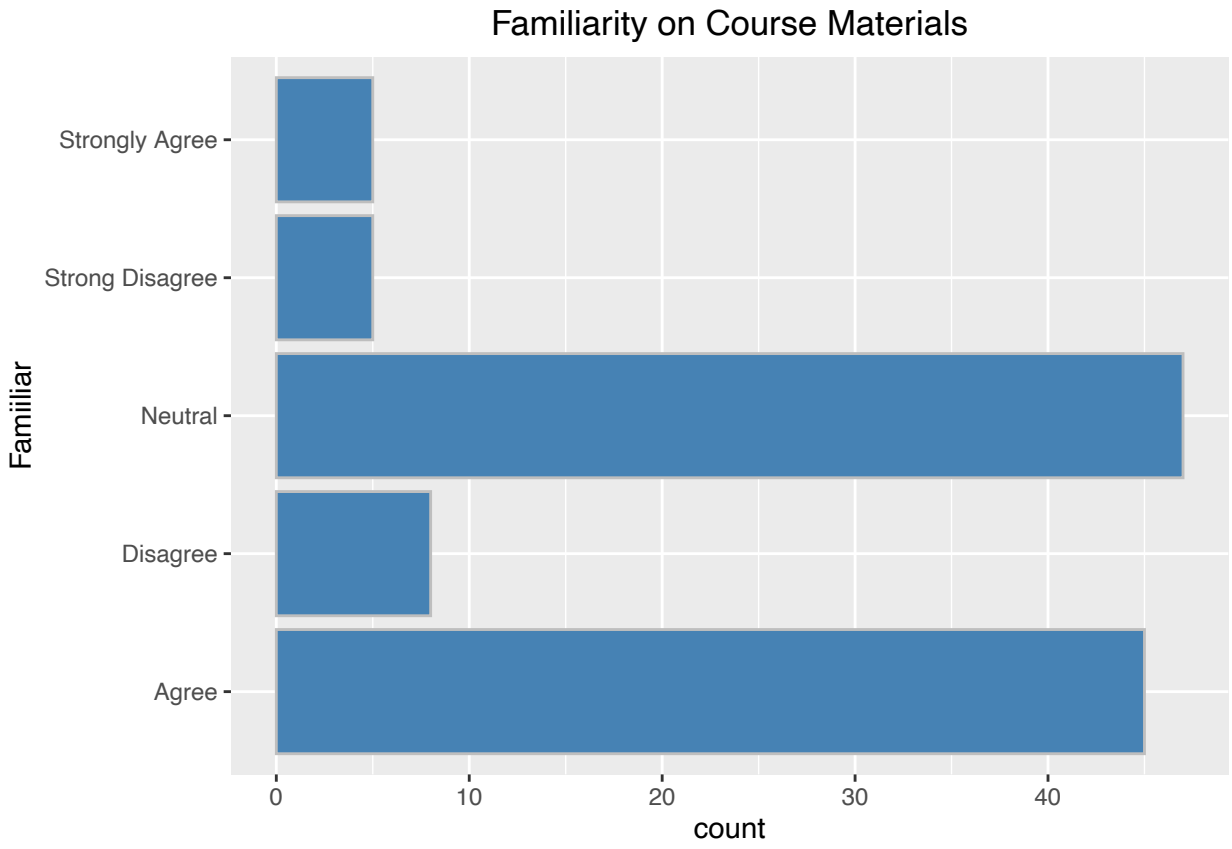

```
data %>%
  # by default bins=30
```

```
  ggplot(aes(x=Term.Test.percentage))+geom_histogram(color="gray", fill="steelblue")+
  labs(title="Term Test Percentage Grade")+
  theme(plot.title = element_text(hjust = 0.5))
```

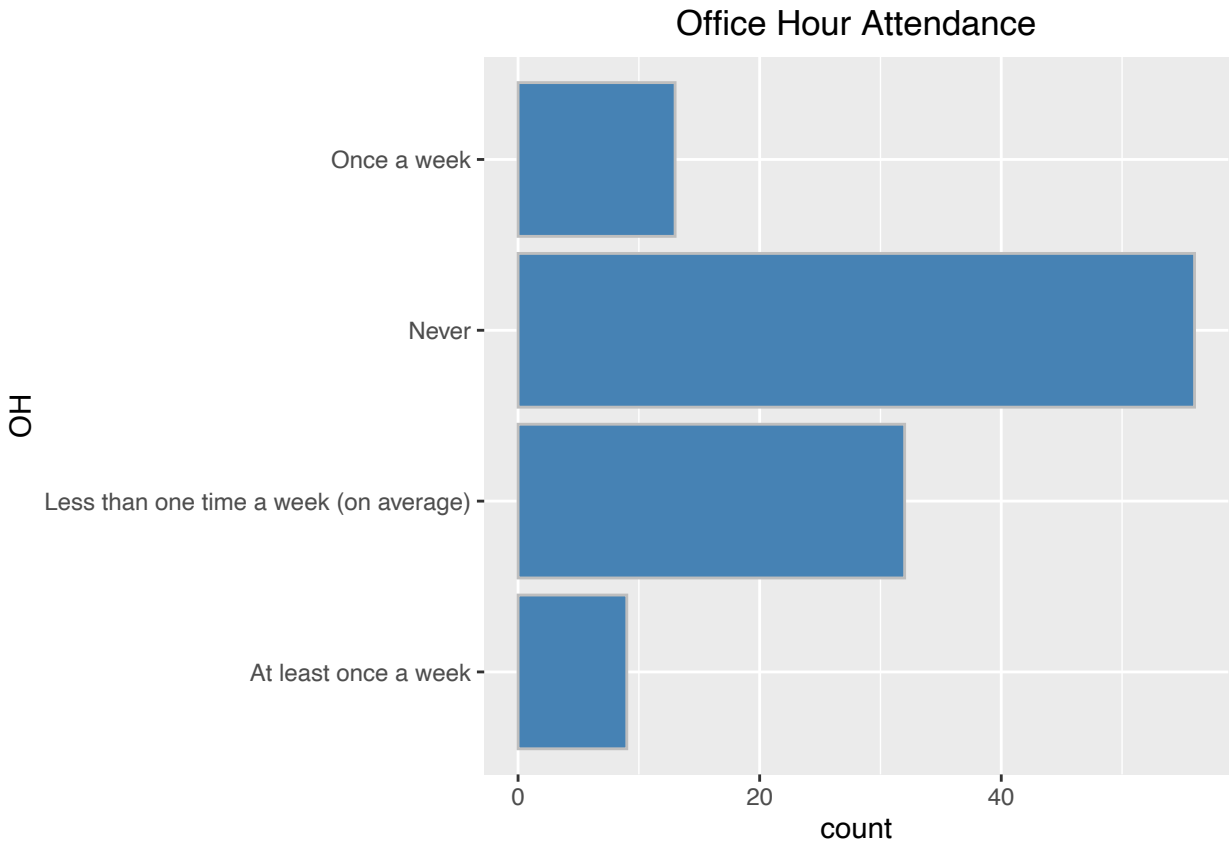
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



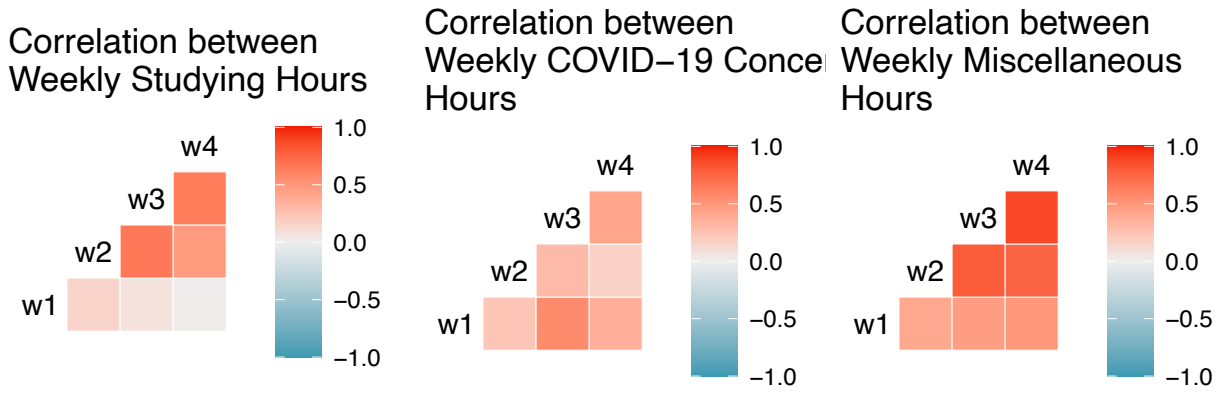

```
data %>%
  ggplot(aes(x=Famiiliar))+
  geom_bar(color="gray", fill="steelblue")+
  labs(title="Familiarity on Course Materials")+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
```




```
data %>%
  ggplot(aes(x=OH))+
  geom_bar(color="gray",fill="steelblue")+
  labs(title="Office Hour Attendance")+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
```

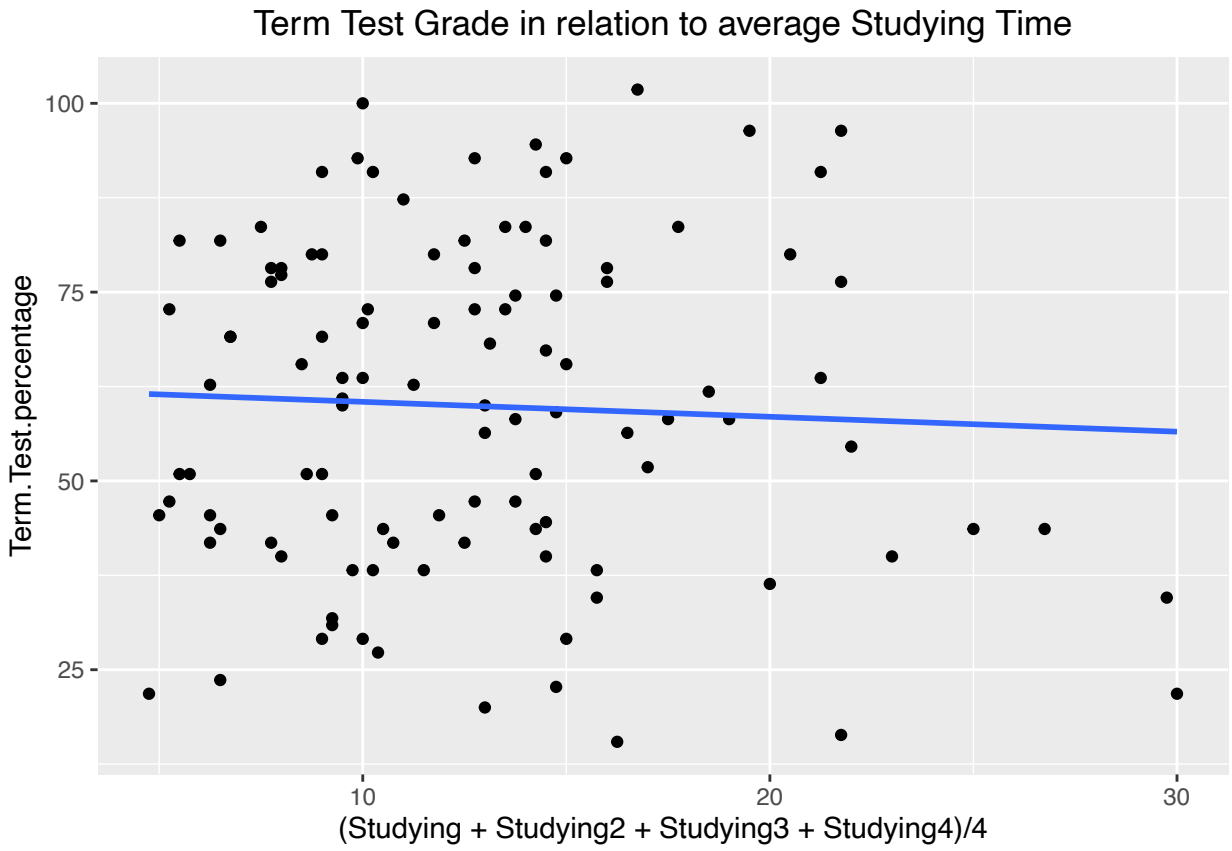



```
    # corrolation maps
    corr_study <- ggcorr(cbind(
      w1 = data$Studying,
      w2 = data$Studying2,
      w3 = data$Studying3,
      w4 = data$Studying4
    )) + ggtitle("Correlation between
Weekly Studying Hours")
    corr_covid <- ggcorr(cbind(
        w1 = data$COVID,
        w2 = data$COVID2,
        w3 = data$COVID3,
        w4 = data$COVID4
    )) + ggtitle("Correlation between
Weekly COVID-19 Concern
Hours")
    corr_miscel <- ggcorr(cbind(
        w1 = data$Miscellaneous,
        w2 = data$Miscellaneous2,
        w3 = data$Miscellaneous3,
        w4 = data$Miseallenous4
    )) + ggtitle("Correlation between
Weekly Miscellaneous
Hours")
    grid.arrange(corr_study, corr_covid, corr_miscel, nrow = 1)
```



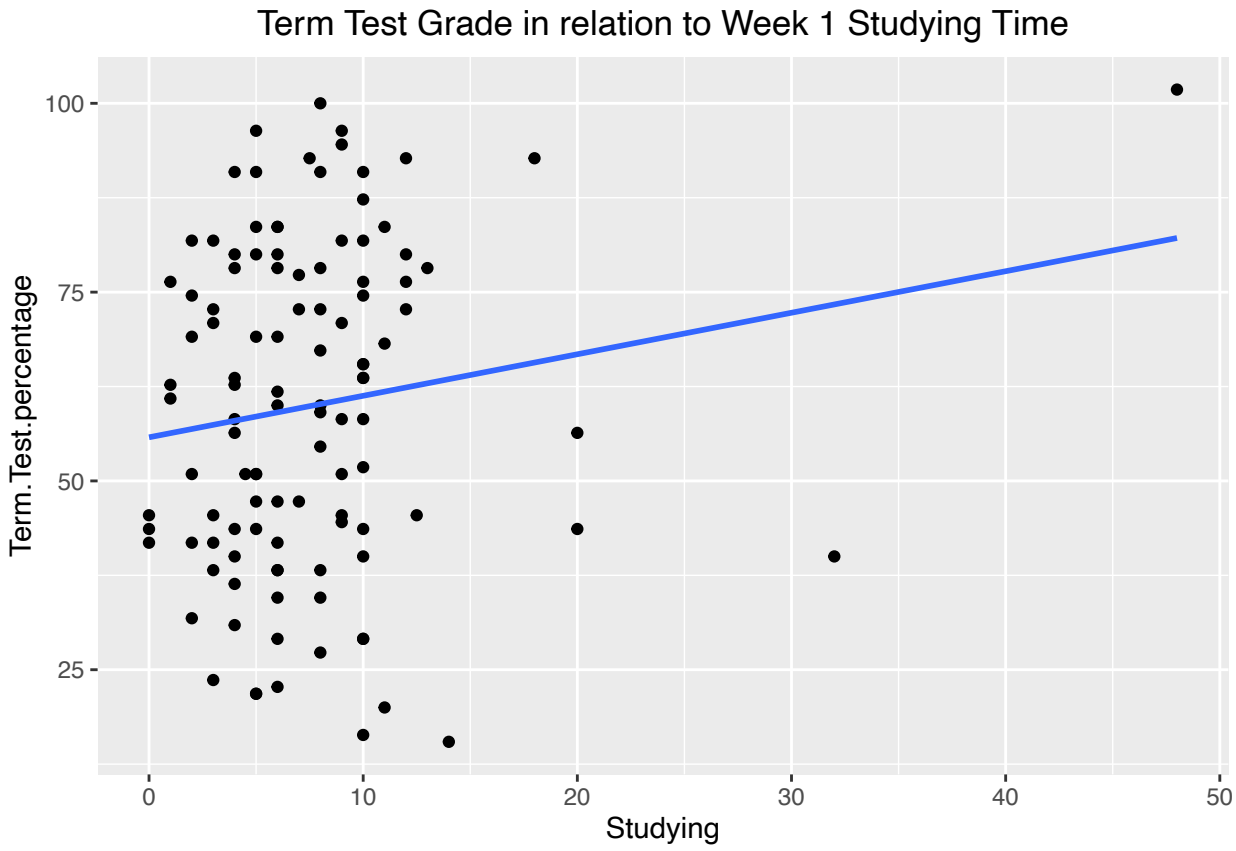

```
data %>%
  ggplot(aes(x=(Studying+Studying2+Studying3+Studying4)/4,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to average Studying Time")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



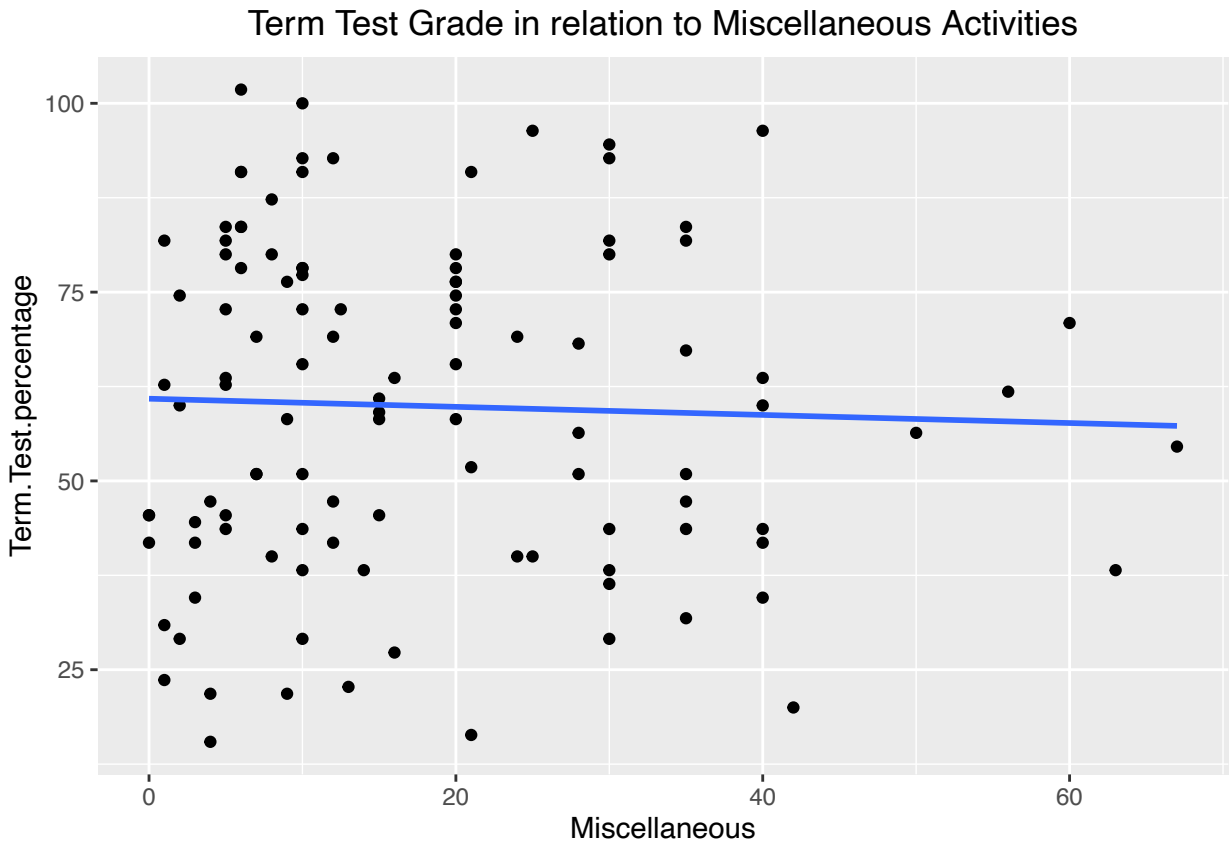

```
data %>%
  ggplot(aes(x=Studying,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to Week 1 Studying Time")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



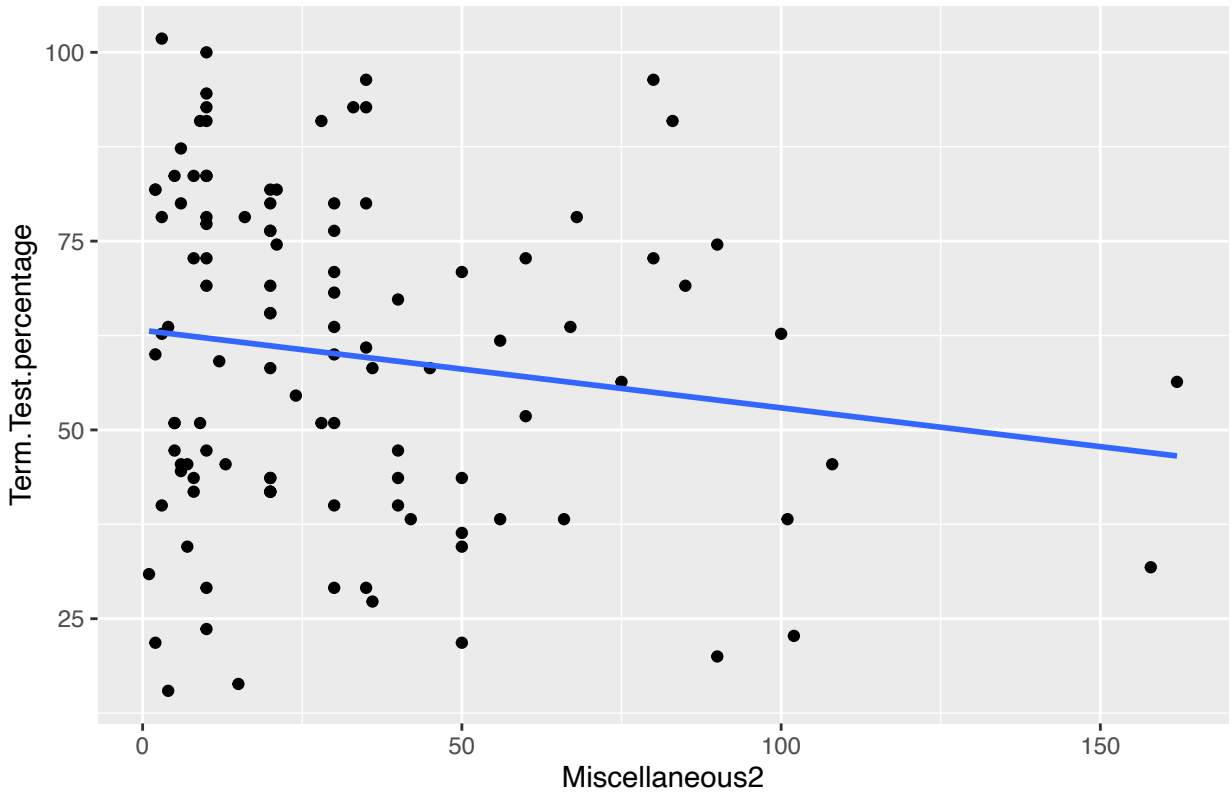

```
data %>%
  ggplot(aes(x=Miscellaneous,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to Miscellaneous Activities")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



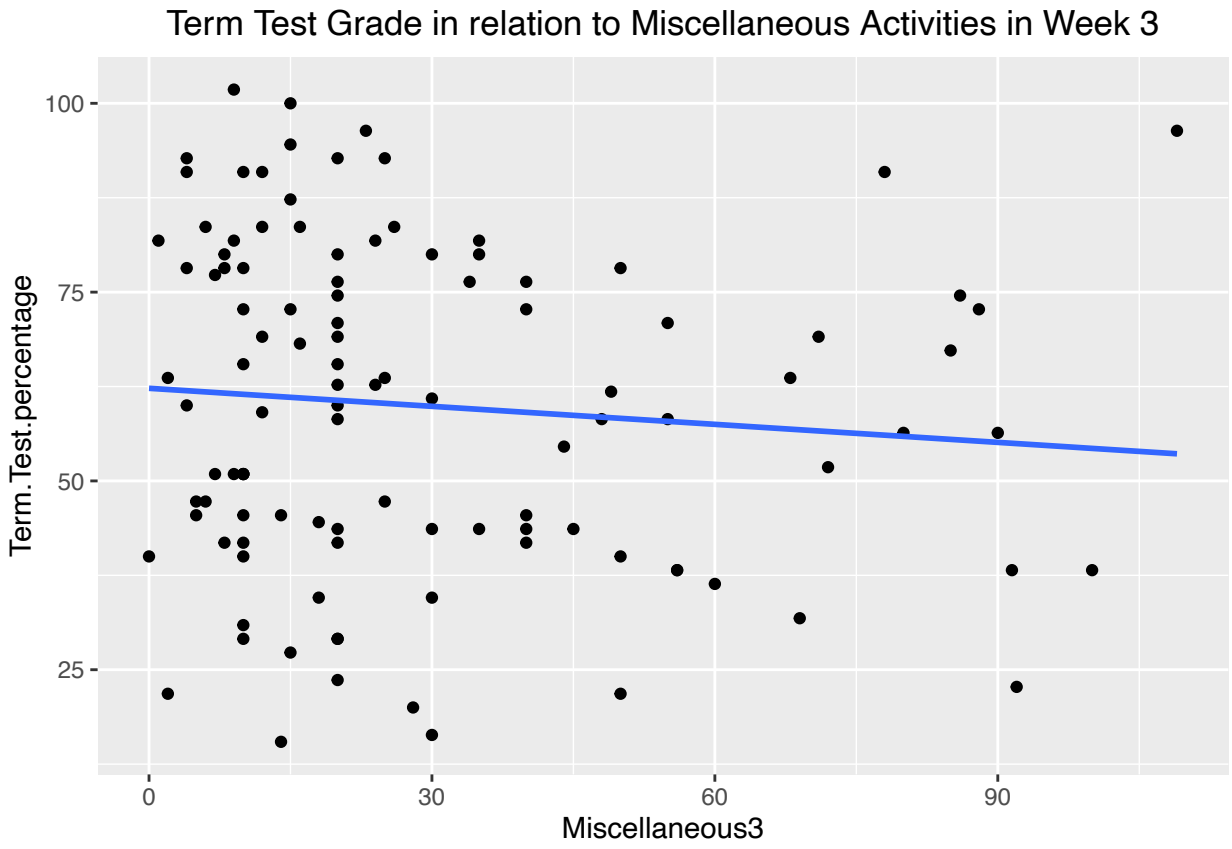

```
data %>%
  ggplot(aes(x=Miscellaneous2,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to Miscellaneous Activities in Week 2")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



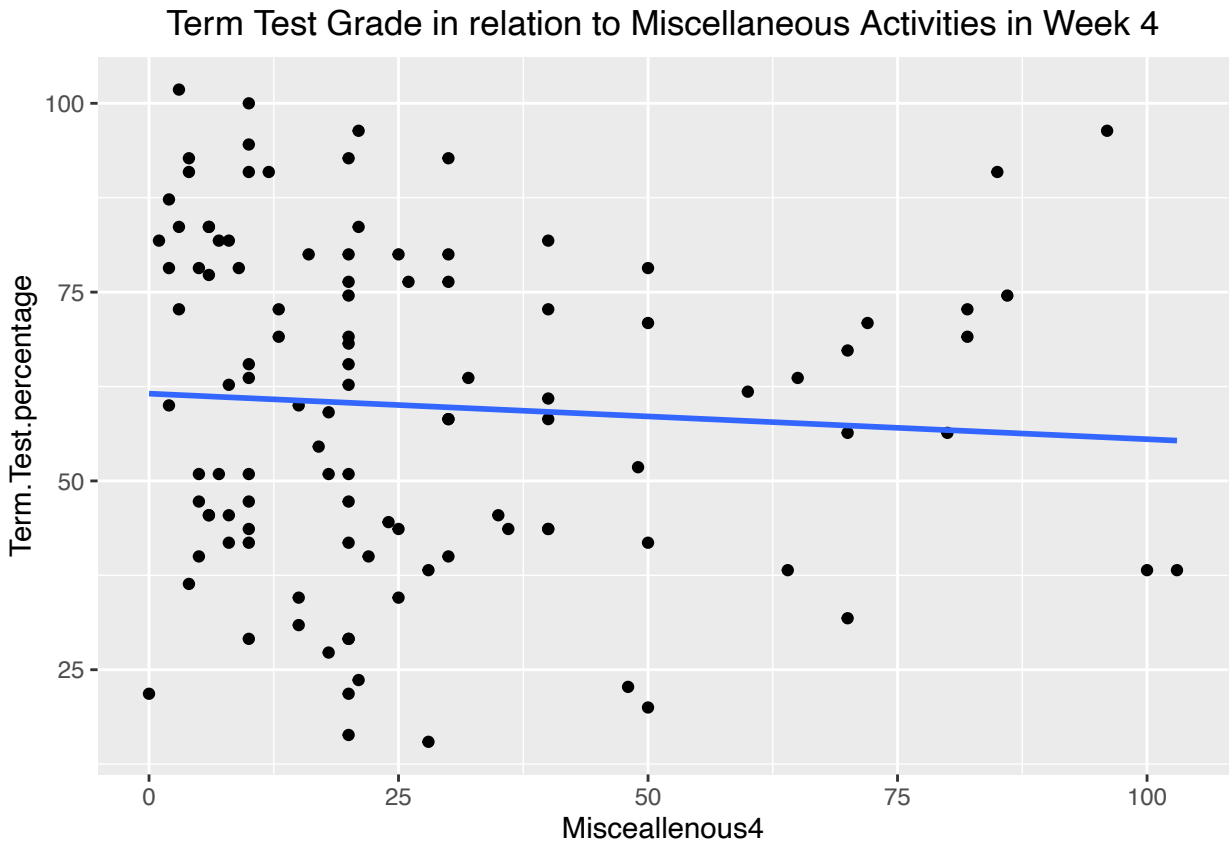

```
data %>%
  ggplot(aes(x=Miscellaneous3,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to Miscellaneous Activities in Week 3")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



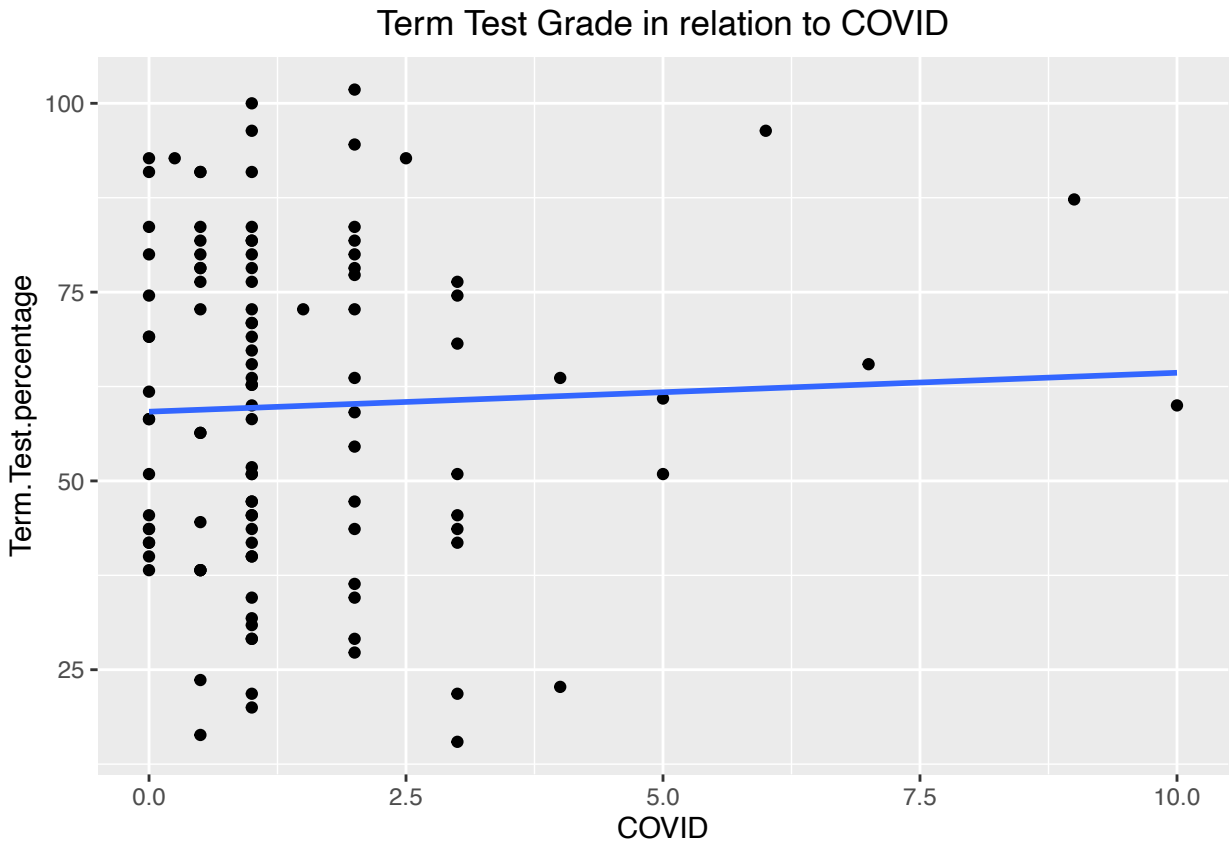

```
data %>%
  ggplot(aes(x=Misceallenous4,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to Miscellaneous Activities in Week 4")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



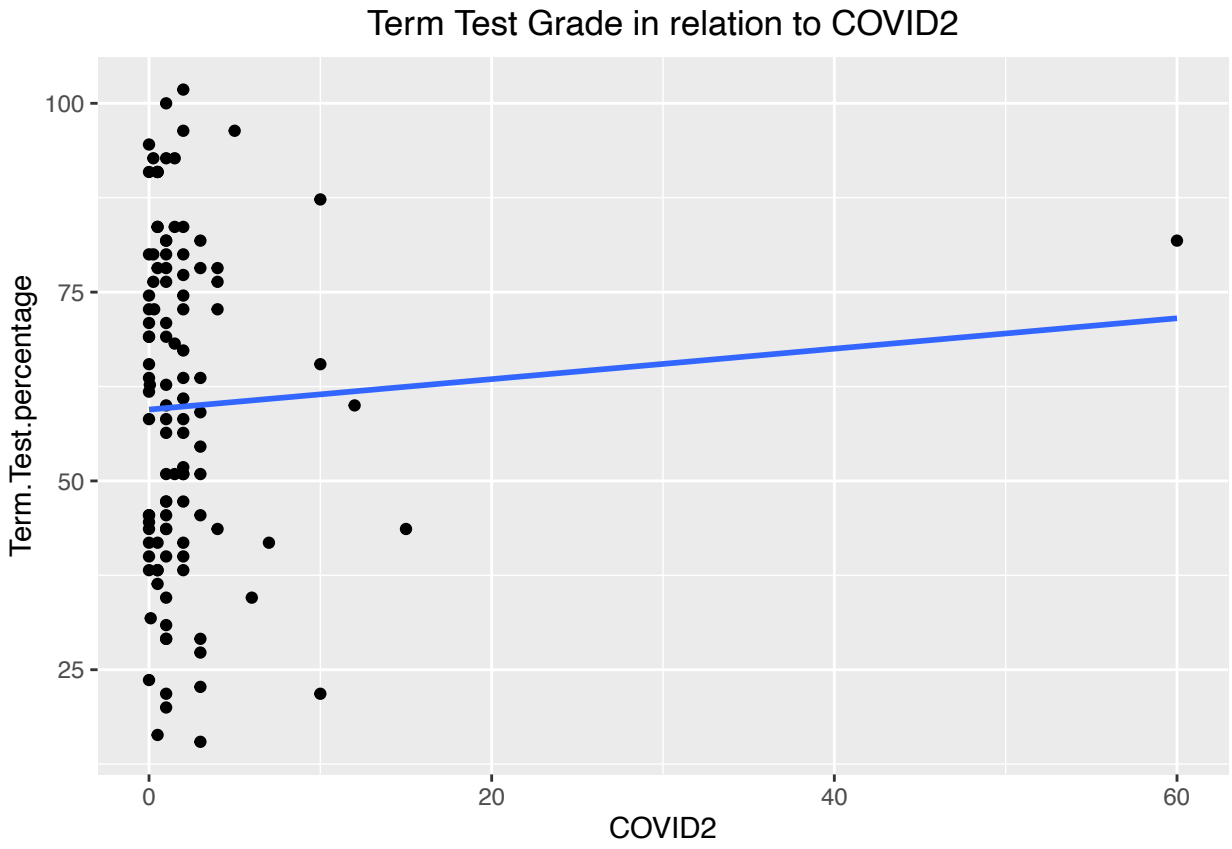

```
data %>%
  ggplot(aes(x=COVID,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to COVID")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



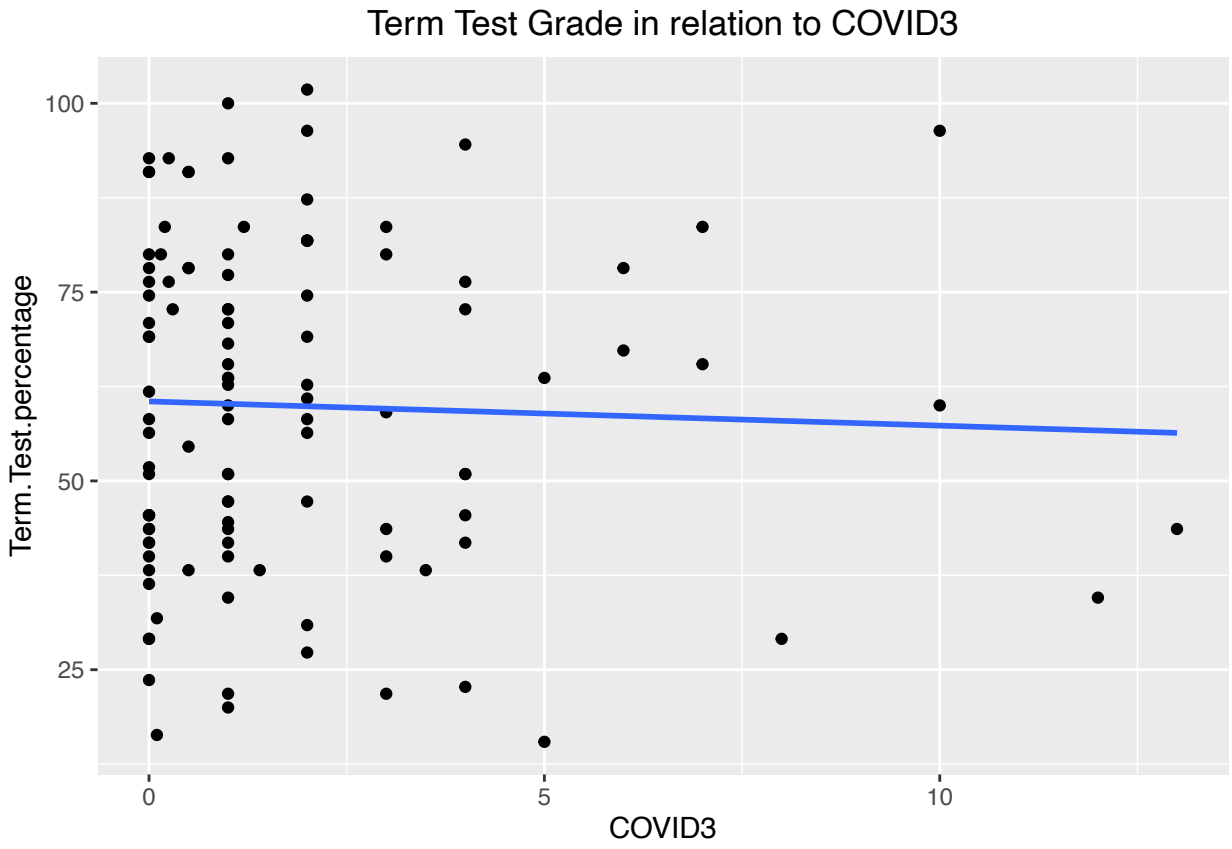

```
data %>%
  ggplot(aes(x=COVID2,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to COVID2")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



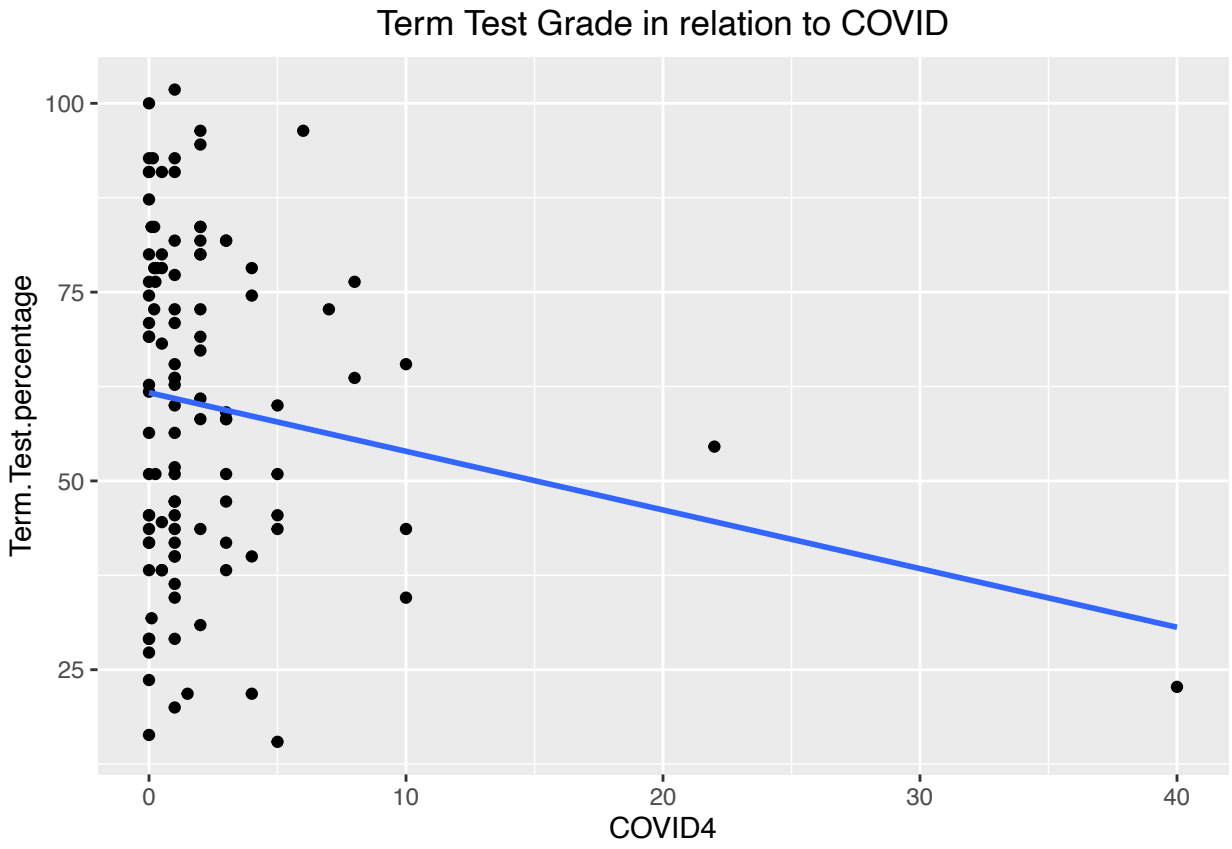

```
data %>%
  ggplot(aes(x=COVID3,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to COVID3")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



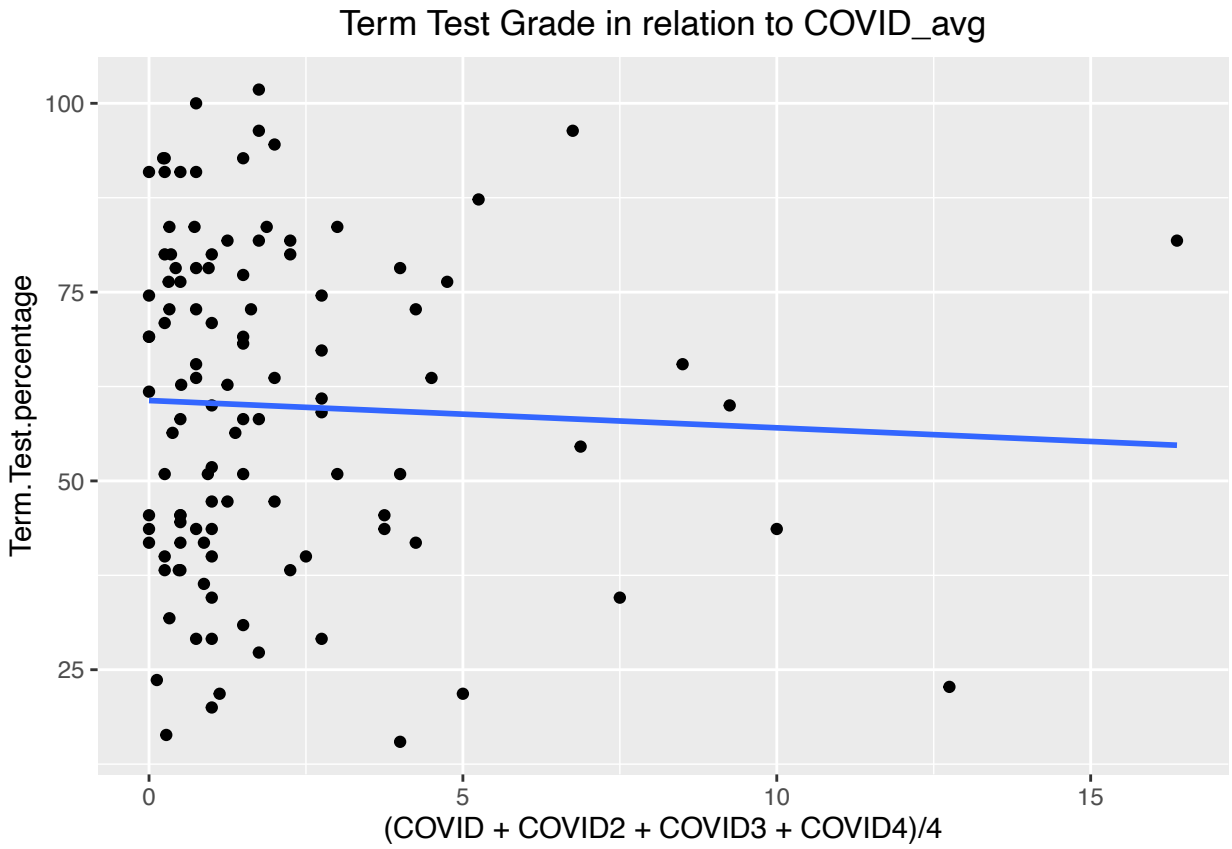

```
data %>%
  ggplot(aes(x=COVID4,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to COVID")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



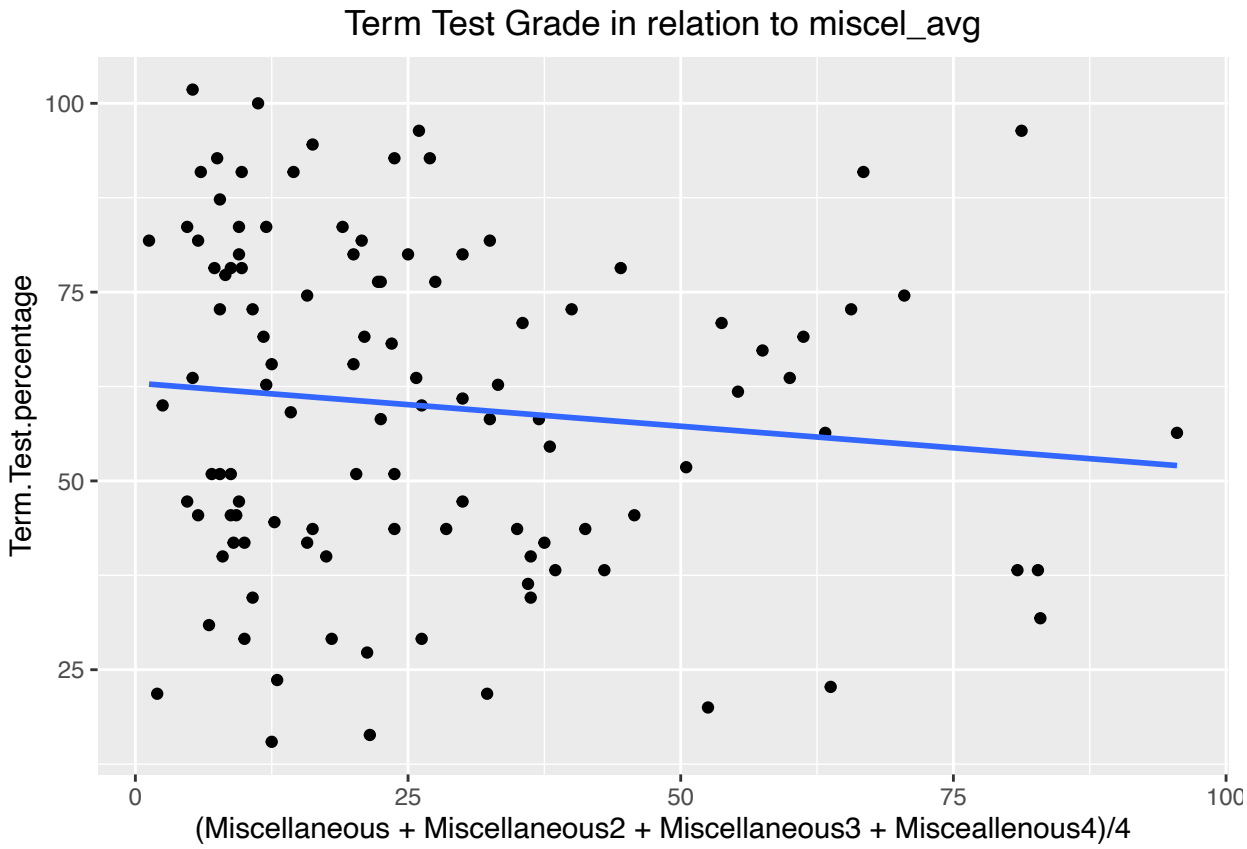

```
data %>%
  ggplot(aes(x=(COVID+COVID2+COVID3+COVID4)/4,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to COVID_avg")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

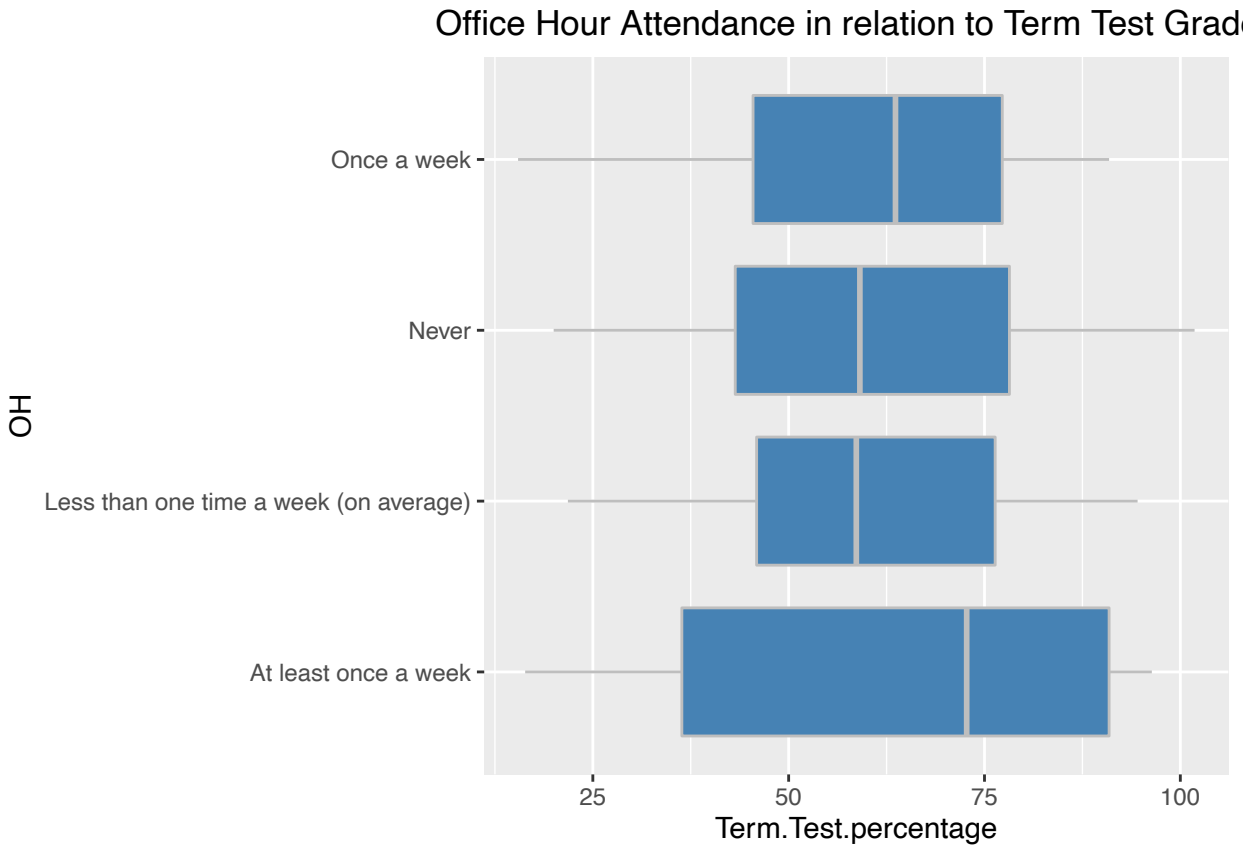



```
data %>%
  ggplot(aes(x=(Miscellaneous+Miscellaneous2+Miscellaneous3+Misceallenous4)/4,y=Term.Test.percentage))+
  geom_point()+
  geom_smooth(se=FALSE,method="lm")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Term Test Grade in relation to miscel_avg")
```

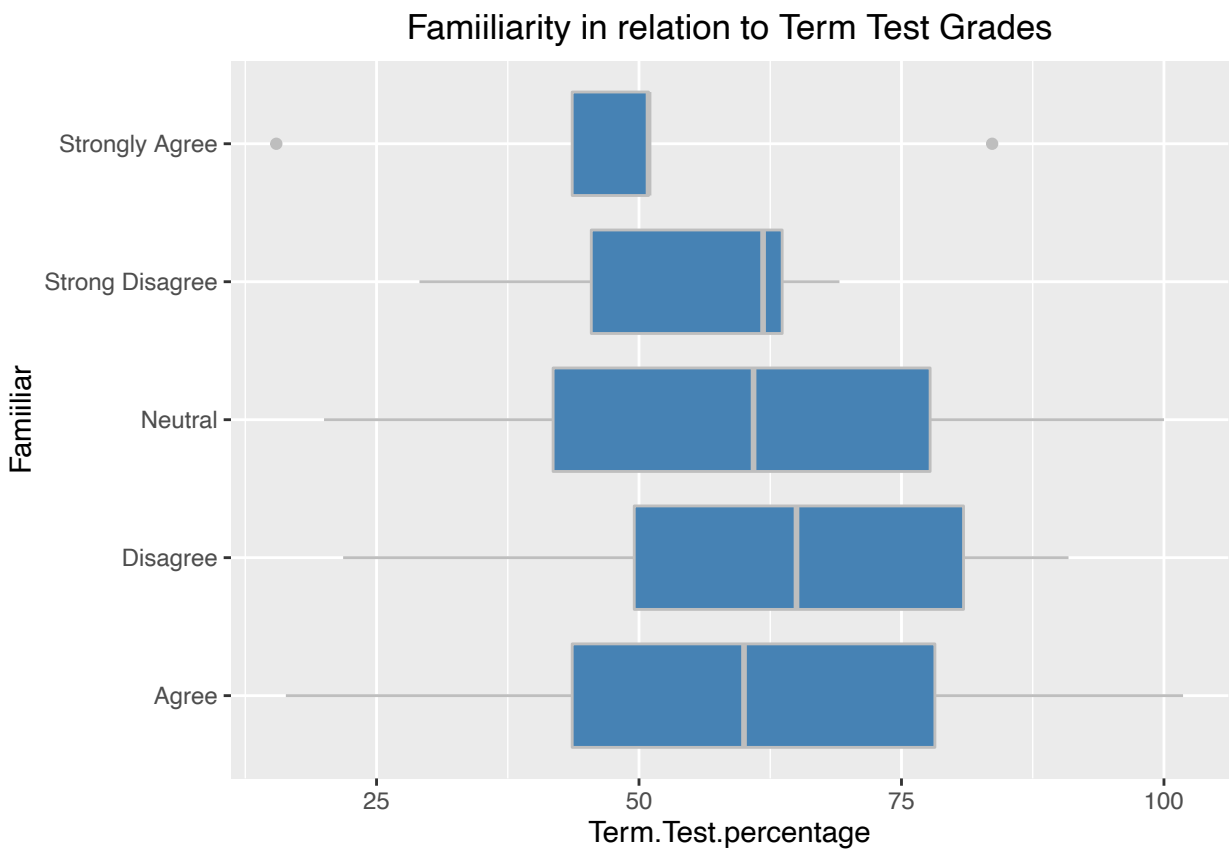
```
## `geom_smooth()` using formula 'y ~ x'
```



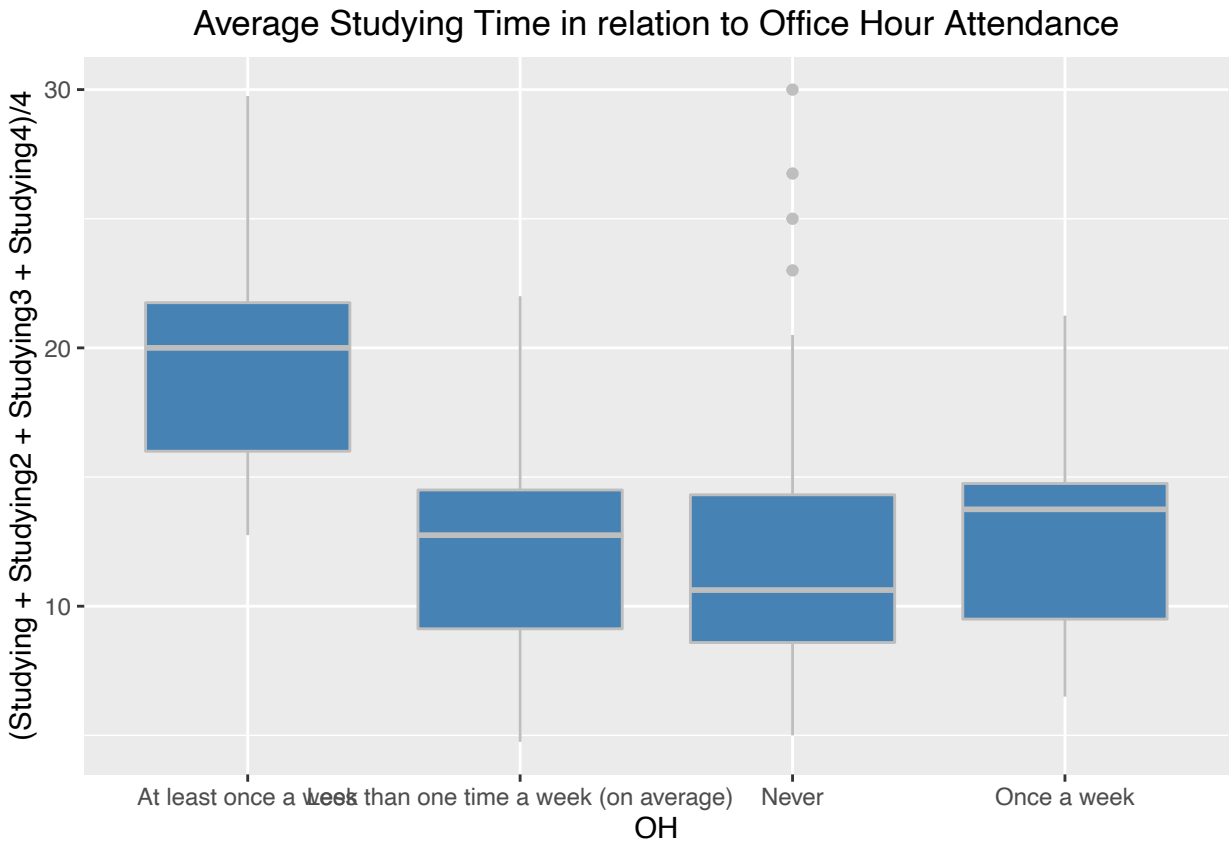

```
data %>%
  ggplot(aes(x=OH,y=Term.Test.percentage))+
  geom_boxplot(color="gray",fill="steelblue")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Office Hour Attendance in relation to Term Test Grades")+
  coord_flip()
```



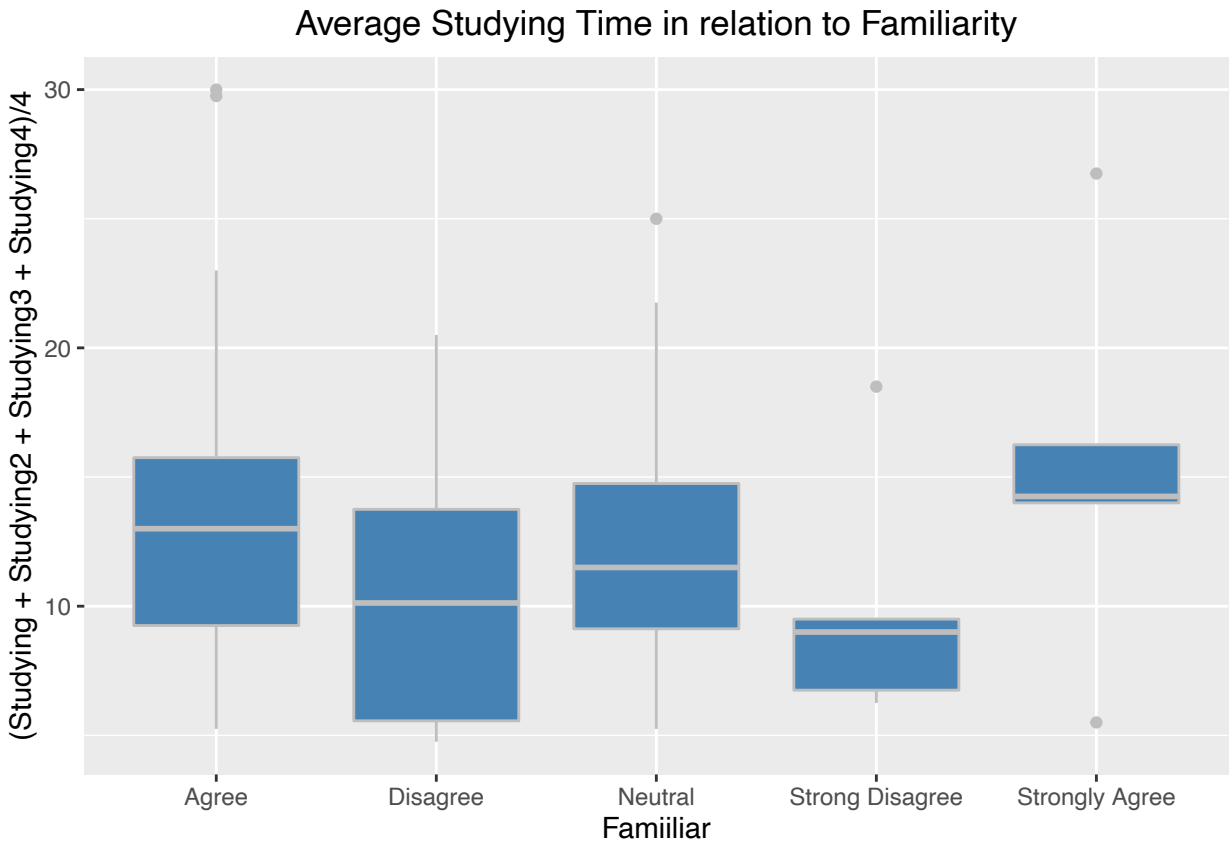

```
data %>%
  ggplot(aes(x=Famiiliar,y=Term.Test.percentage))+
  geom_boxplot(color="gray",fill="steelblue")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Famiiliarity in relation to Term Test Grades")+
  coord_flip()
```




```
data %>%
  ggplot(aes(x=(Studying+Studying2+Studying3+Studying4)/4,y=OH))+
  geom_boxplot(color="gray",fill="steelblue")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Average Studying Time in relation to Office Hour Attendance")+
  coord_flip()
```




```
data %>%
  ggplot(aes(x=(Studying+Studying2+Studying3+Studying4)/4,y=Famiiliar))+
  geom_boxplot(color="gray",fill="steelblue")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Average Studying Time in relation to Familiarity")+
  coord_flip()
```



Summary

```
table(data$OH)
```

```
##
##                   At least once a week Less than one time a week (on average)
##                                      9                                       32
##                                  Never                                Once a week
##                                     56                                       13
```

```
summary(data$Term.Test.percentage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.45   43.64   60.45   59.93   78.18  101.82
```

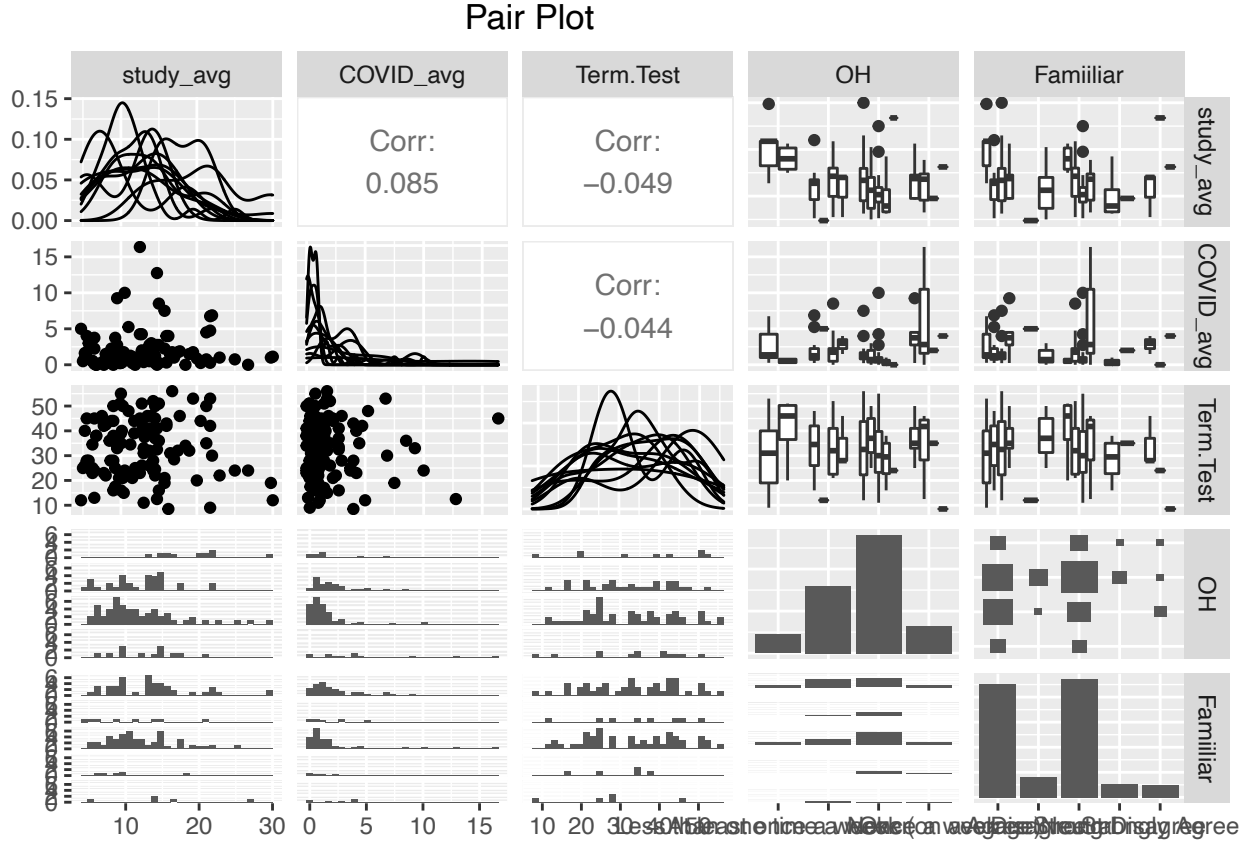
```
data %>%
ggpairs(data, columns = c("study_avg","COVID_avg","Term.Test","OH","Famiiliar"), title = "Pair Plot") + theme(
```

```
## Warning: Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

```
## Warning: Ignoring unknown aesthetics: ID, Studying, COVID, Miscellaneous, Studying2, COVID2, Miscellaneous2
```

```
## Ignoring unknown aesthetics: ID, Studying, COVID, Miscellaneous, Studying2, COVID2, Miscellaneous2, Studyin
## Warning: Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning: Ignoring unknown aesthetics: ID, Studying, COVID, Miscellaneous, Studying2, COVID2, Miscellaneous2
## Ignoring unknown aesthetics: ID, Studying, COVID, Miscellaneous, Studying2, COVID2, Miscellaneous2, Studyin
## Warning: Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
## Warning: Ignoring unknown aesthetics: ID, Studying, COVID, Miscellaneous, Studying2, COVID2, Miscellaneous2
## Ignoring unknown aesthetics: ID, Studying, COVID, Miscellaneous, Studying2, COVID2, Miscellaneous2, Studyin
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Model Development

```
full_model <- lm(Term.Test~.-ID-study_avg-COVID_avg-Miscellaneous-Miscellaneous2-Miscellaneous3-Miscellenous4
summary(full_model)
```

```
##
## Call:
## lm(formula = Term.Test ~ . - ID - study_avg - COVID_avg - Miscellaneous -
##     Miscellaneous2 - Miscellaneous3 - Miscellenous4, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1814  -8.6325   0.4426   8.3147  19.9686
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                29.28041    8.45564   3.463 0.000983
## Studying                                    0.52179    0.25950   2.011 0.048780
## COVID                                       1.79467    1.65881   1.082 0.283555
## Studying2                                   0.14974    0.40792   0.367 0.714825
## COVID2                                     -2.01024    1.26614  -1.588 0.117525
## Studying3                                   0.12376    0.31114   0.398 0.692196
## COVID3                                      2.35782    1.02626   2.297 0.025035
## Studying4                                  -0.23525    0.19684  -1.195 0.236664
## COVID4                                     -0.58202    0.32233  -1.806 0.075903
## OHLess than one time a week (on average)    1.22691    6.47024   0.190 0.850233
## OHNever                                     0.12545    6.39846   0.020 0.984421
## OHOnce a week                              -0.07528    7.36044  -0.010 0.991873
## FamiiliarDisagree                           3.67100    5.80384   0.633 0.529416
## FamiiliarNeutral                           -0.55101    3.55588  -0.155 0.877366
## FamiiliarStrong Disagree                   -4.02377    9.59007  -0.420 0.676269
## FamiiliarStrongly Agree                   -16.38222    7.26388  -2.255 0.027714
##
```

```
## (Intercept)                              ***
## Studying                                 *
## COVID
## Studying2
## COVID2
## Studying3
## COVID3                                   *
## Studying4
## COVID4                                   .
## OHLess than one time a week (on average)
## OHNever
## OHOnce a week
## FamiiliarDisagree
## FamiiliarNeutral
## FamiiliarStrong Disagree
## FamiiliarStrongly Agree                  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.53 on 61 degrees of freedom
## Multiple R-squared:  0.2386, Adjusted R-squared:  0.05134
## F-statistic: 1.274 on 15 and 61 DF,  p-value: 0.2462
```

```
vif(full_model)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## Studying  1.141462  1        1.068392
## COVID     2.853908  1        1.689351
## Studying2 1.757852  1        1.325840
## COVID2    2.637259  1        1.623964
## Studying3 2.876500  1        1.696025
## COVID3    1.893544  1        1.376061
## Studying4 2.579148  1        1.605973
## COVID4    1.402834  1        1.184413
## OH        2.004063  3        1.122842
## Famiiliar 2.022299  4        1.092020
```

```
mean(vif(full_model))
```

```
## [1] 1.665125
```

```
reduced_model = lm(full_model, data=train)
ols_step_backward_aic(reduced_model, progress = FALSE, details = FALSE)
```

```
##
##
##                        Backward Elimination Summary
## ------------------------------------------------------------------------
## Variable        AIC         RSS        Sum Sq      R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------
## Full Model    623.969     9584.118    3002.973    0.23858      0.05134
## OH            618.114     9602.169    2984.922    0.23714      0.09411
## Studying2     616.261     9620.568    2966.523    0.23568      0.10633
## Studying3     614.838     9692.897    2894.194    0.22993      0.11326
## Famiiliar     613.767    10605.624    1981.466    0.15742      0.08520
## COVID         612.482    10704.560    1882.531    0.14956      0.08967
## COVID2        611.221    10807.759    1779.332    0.14136      0.09366
## Studying4     610.576    10999.588    1587.503    0.12612      0.09021
## ------------------------------------------------------------------------
```

```
stepwise_AIC_model <- stepAIC(full_model, trace = FALSE)
summary(stepwise_AIC_model)
```

```
##
```

```
## Call:
## lm(formula = Term.Test ~ Studying + COVID3 + COVID4, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.616  -8.601   1.203   9.669  21.007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.4971     2.6387  10.800   <2e-16 ***
## Studying      0.4968     0.2381   2.087   0.0404 *
## COVID3        1.5220     0.7801   1.951   0.0549 .
## COVID4       -0.5892     0.2848  -2.069   0.0421 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.28 on 73 degrees of freedom
## Multiple R-squared:  0.1261, Adjusted R-squared:  0.09021
## F-statistic: 3.512 on 3 and 73 DF,  p-value: 0.01937
```

```
candidate_model_1 <- lm(Term.Test~COVID3+Studying, data = train)
summary(candidate_model_1)
```

```
##
## Call:
## lm(formula = Term.Test ~ COVID3 + Studying, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.3852  -9.3350   0.1227  10.5003  22.0224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.9095     2.6808  10.411  3.8e-16 ***
## COVID3        0.9558     0.7465   1.280   0.2044
## Studying      0.5140     0.2431   2.114   0.0379 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.54 on 74 degrees of freedom
## Multiple R-squared:  0.0749, Adjusted R-squared:  0.0499
## F-statistic: 2.996 on 2 and 74 DF,  p-value: 0.0561
```

```
candidate_model_2 <- lm(Term.Test~COVID3+Studying+Famiiliar:Studying, data=train)
summary(candidate_model_2)
```

```
##
## Call:
## lm(formula = Term.Test ~ COVID3 + Studying + Famiiliar:Studying,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.7856  -9.0000  -0.0688   9.7330  21.5243
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     27.319453   3.029081   9.019 2.44e-13 ***
## COVID3                           1.114126   0.762155   1.462    0.148
## Studying                         0.635478   0.249416   2.548    0.013 *
## Studying:FamiiliarDisagree       0.489260   0.680492   0.719    0.475
## Studying:FamiiliarNeutral       -0.005209   0.373123  -0.014    0.989
```

```
## Studying:FamiiliarStrong Disagree -1.022053   1.511437  -0.676    0.501
## Studying:FamiiliarStrongly Agree  -1.185615   0.521384  -2.274    0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.32 on 70 degrees of freedom
## Multiple R-squared:  0.1554, Adjusted R-squared:  0.08303
## F-statistic: 2.147 on 6 and 70 DF,  p-value: 0.05862
```

```
candidate_model_3 <- lm(Term.Test~COVID4+Studying, data = train)
summary(candidate_model_3)
```

```
##
## Call:
## lm(formula = Term.Test ~ COVID4 + Studying, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.011  -9.879   1.693  10.376  22.318
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.5831     2.4576  12.444   <2e-16 ***
## COVID4       -0.3942     0.2717  -1.451   0.1511
## Studying      0.4928     0.2425   2.032   0.0458 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.51 on 74 degrees of freedom
## Multiple R-squared:  0.08056,    Adjusted R-squared:  0.05571
## F-statistic: 3.242 on 2 and 74 DF,  p-value: 0.04471
```

```
install.packages("AICcmodavg")
```

```
## Installing package into '/opt/r'
## (as 'lib' is unspecified)
```

```
library(AICcmodavg)

#define list of models
models <- list(full_model, reduced_model, candidate_model_1,candidate_model_2,candidate_model_3)

#specify model names
mod.names <- c('full model', 'AIC reduced model', 'candidate model 1','candidate model 2','candidate model 3')

#calculate AIC of each model
aictab(cand.set = models, modnames = mod.names)
```

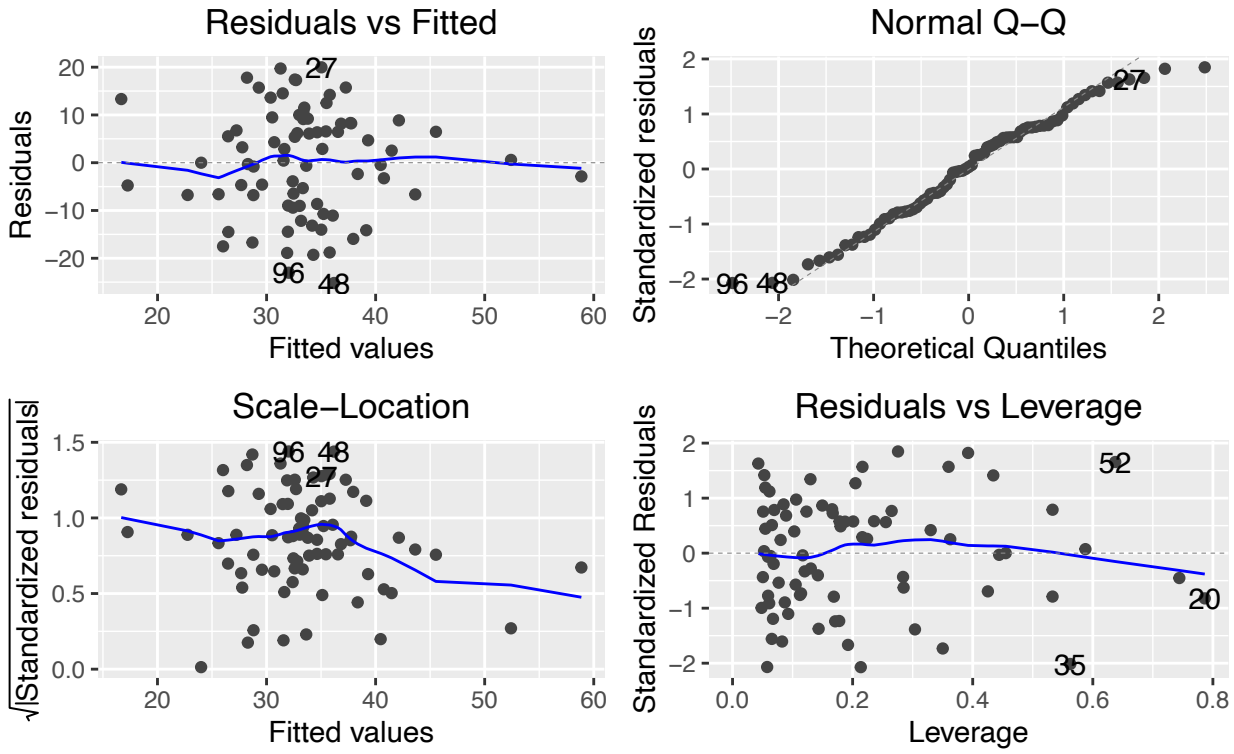
```
## Warning in aictab.AIClm(cand.set = models, modnames = mod.names):
## Check model structure carefully as some models may be redundant
```

```
##
## Model selection based on AICc:
##
##                     K   AICc Delta_AICc AICcWt Cum.Wt      LL
## candidate model 3   4 613.04       0.00   0.50   0.50 -302.24
## candidate model 1   4 613.52       0.47   0.39   0.89 -302.48
## candidate model 2   8 616.07       3.02   0.11   1.00 -298.97
## full model         17 634.34      21.30   0.00   1.00 -294.98
## AIC reduced model  17 634.34      21.30   0.00   1.00 -294.98
```

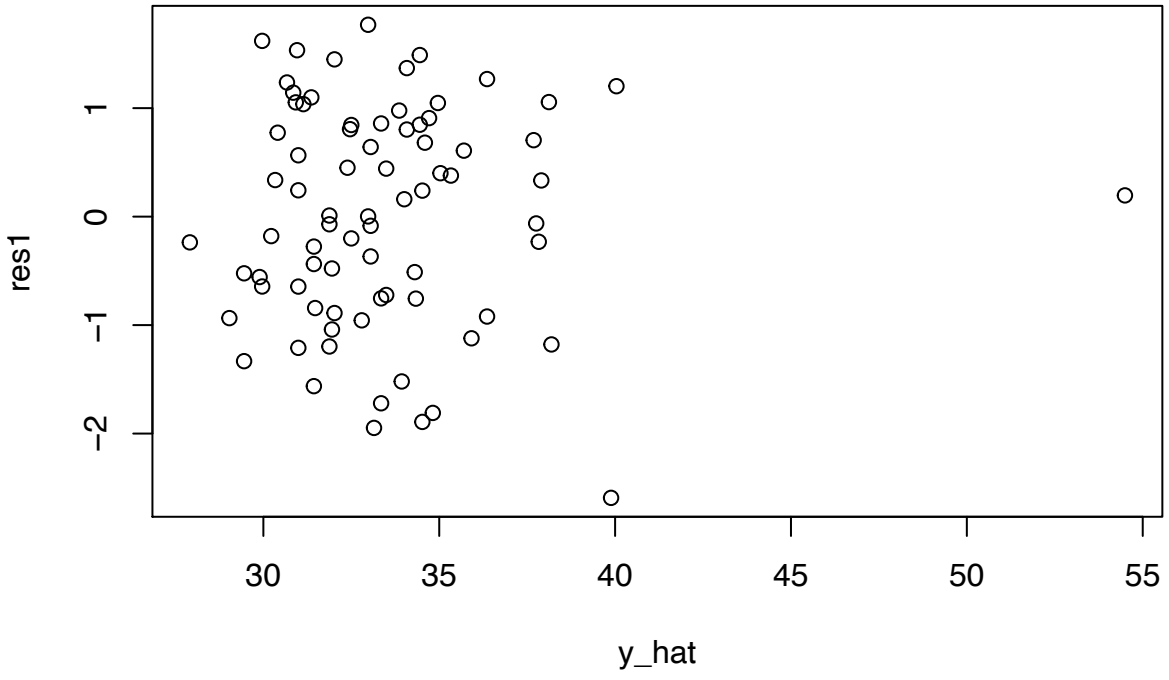
Diagnostic Plots

```
# diagnostic plots
```

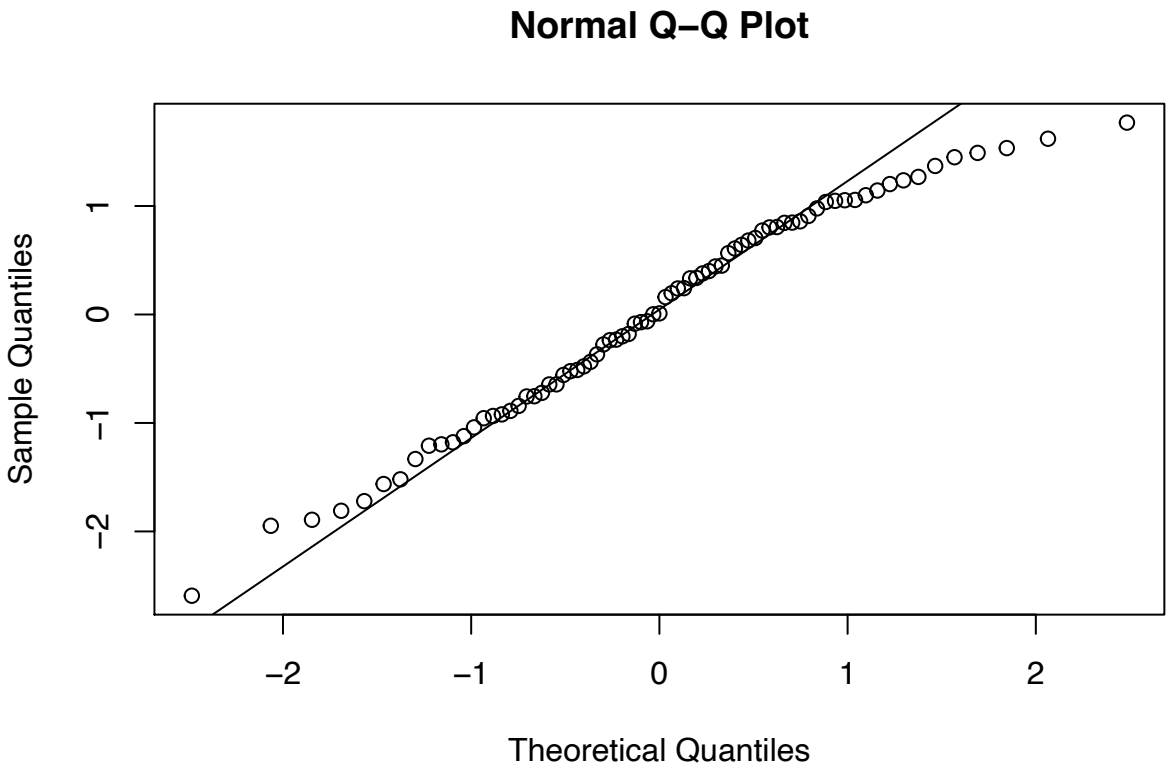
```
autoplot(reduced_model)+theme(plot.title = element_text(hjust = 0.5))
```



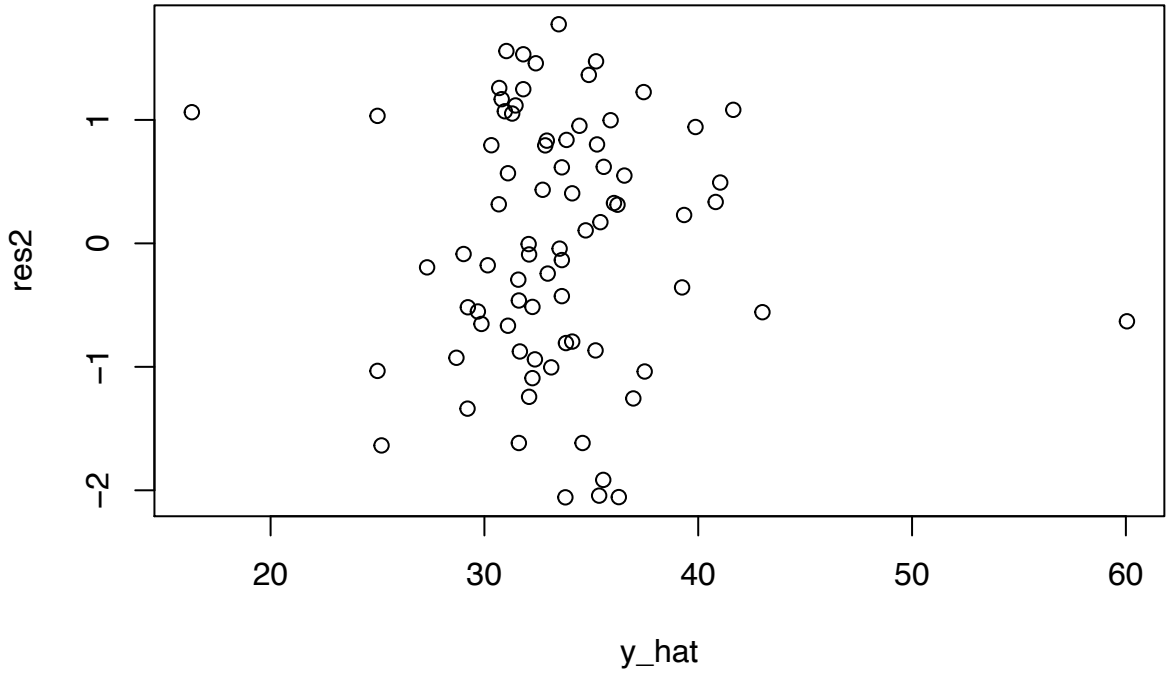

```
res1<-rstandard(candidate_model_1)
y_hat<-fitted(candidate_model_1)
plot(y_hat, res1)
```




```
qqnorm(res1)
qqline(res1)
```

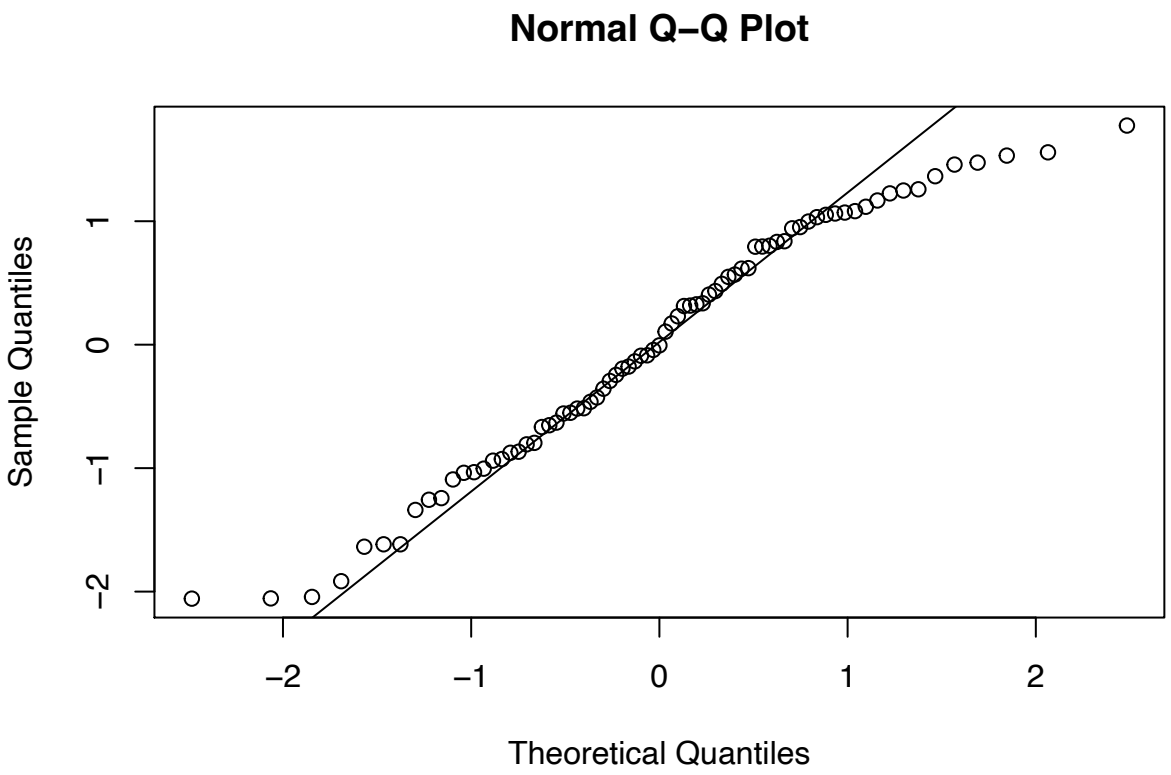



```
res2<-rstandard(candidate_model_2)
y_hat<-fitted(candidate_model_2)
plot(y_hat, res2)
```




```
qqnorm(res2)
```

```
qqline(res2)
```




```
# Leverage point
h <- hatvalues(candidate_model_1)
threshold <- 2*(length(candidate_model_1$coefficients)/nrow(train))
length(which(h>threshold))
```

```
## [1] 6
```

```
# Outlier
std_res <- rstandard(candidate_model_1)
length(which(abs(std_res)>2))
```

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

```
# Cook's distance
D <-  cooks.distance(candidate_model_1)
cutoff_D <- 4/(nrow(train)-2)
length(which(D>cutoff_D))
```

```
## [1] 2
```

```
vif(candidate_model_1)
```

```
##   COVID3 Studying
## 1.000498 1.000498
```

```
AIC(candidate_model_1)
```

```
## [1] 612.9617
```

```
BIC(candidate_model_1)
```

```
## [1] 622.3369
```

```
summary(candidate_model_1)$adj.r.squared
```

```
## [1] 0.04989768
```

‘

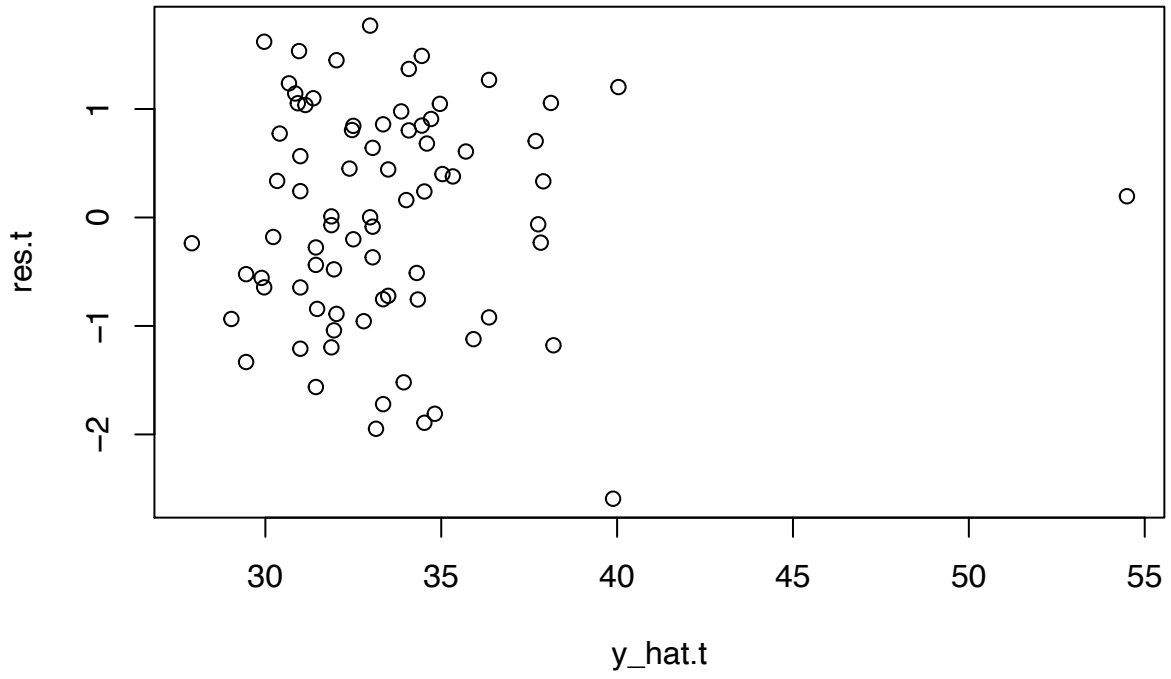
```r
train_model <- lm(Term.Test~COVID3+Studying,data=train)
summary(train_model)
```

```
##
## Call:
## lm(formula = Term.Test ~ COVID3 + Studying, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.3852  -9.3350   0.1227  10.5003  22.0224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.9095     2.6808  10.411  3.8e-16 ***
## COVID3        0.9558     0.7465   1.280   0.2044
## Studying      0.5140     0.2431   2.114   0.0379 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.54 on 74 degrees of freedom
## Multiple R-squared:  0.0749, Adjusted R-squared:  0.0499
## F-statistic: 2.996 on 2 and 74 DF,  p-value: 0.0561
```

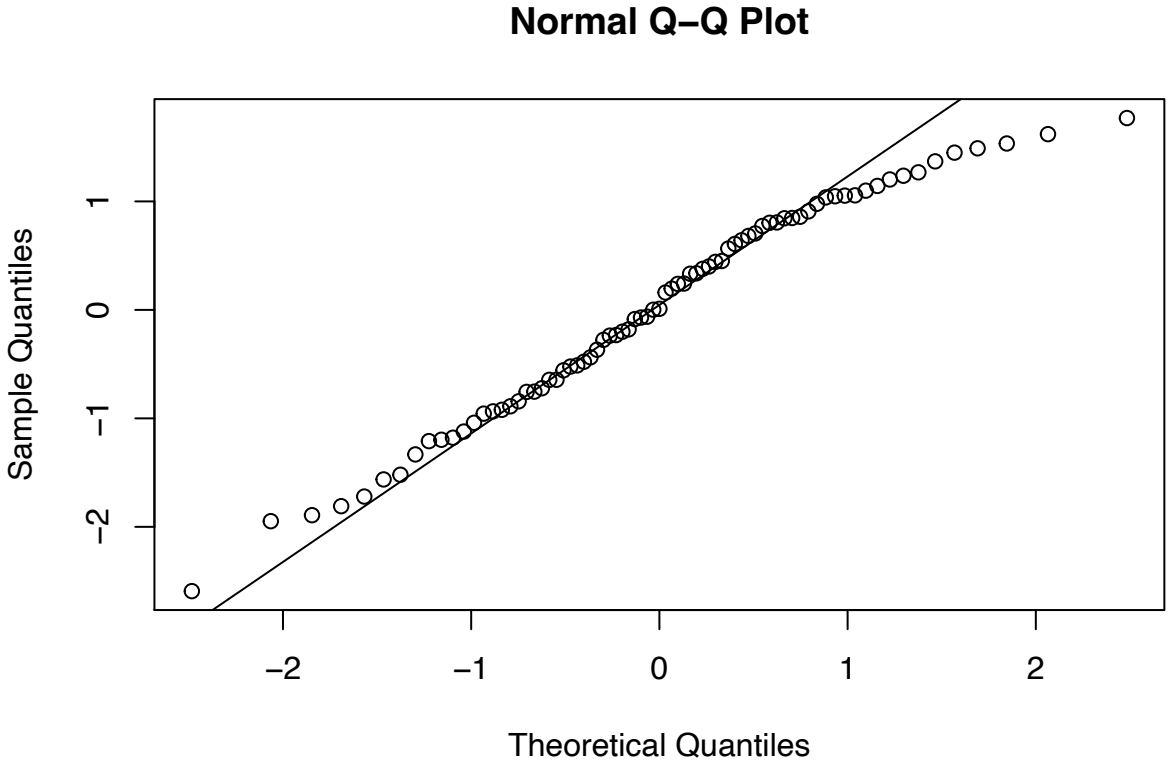
```r
test_model <- lm(Term.Test~COVID3+Studying,data=test)
summary(test_model)
```

```
##
## Call:
## lm(formula = Term.Test ~ COVID3 + Studying, data = test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.471  -7.355   1.013   7.200  17.377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.9471     2.6944  12.970  7.8e-14 ***
## COVID3       -0.8693     0.4596  -1.891   0.0683 .
## Studying     -0.1476     0.2671  -0.553   0.5846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.213 on 30 degrees of freedom
## Multiple R-squared:  0.1173, Adjusted R-squared:  0.05845
## F-statistic: 1.993 on 2 and 30 DF,  p-value: 0.1539
```

```r
res.t <- rstandard(train_model)
y_hat.t <- fitted(train_model)
plot(y_hat.t, res.t)
```




```
qqnorm(res.t)
qqline(res.t)
```



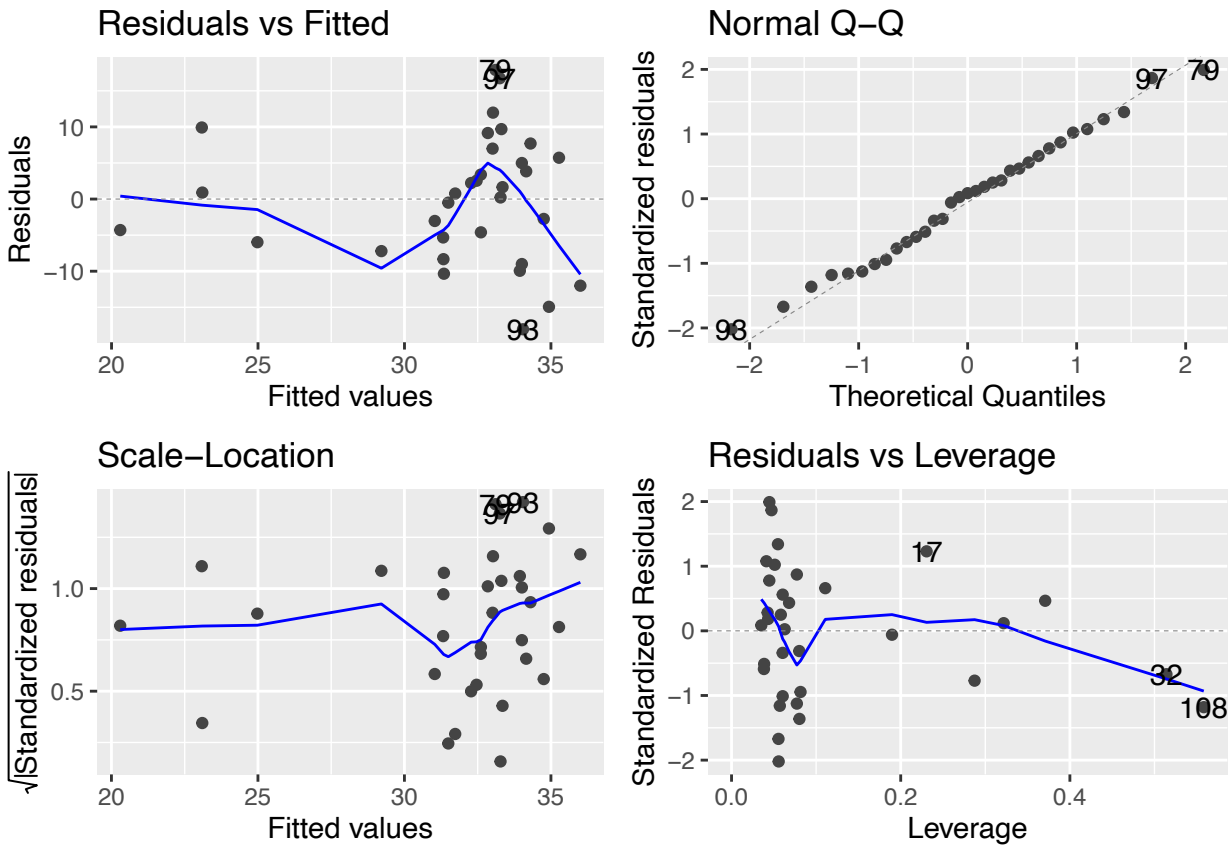

```
summary(candidate_model_2)
```

```
##
## Call:
## lm(formula = Term.Test ~ COVID3 + Studying + Famiiliar:Studying,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.7856  -9.0000  -0.0688   9.7330  21.5243
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      27.319453   3.029081   9.019 2.44e-13 ***
## COVID3                            1.114126   0.762155   1.462    0.148
## Studying                          0.635478   0.249416   2.548    0.013 *
## Studying:FamiiliarDisagree        0.489260   0.680492   0.719    0.475
## Studying:FamiiliarNeutral        -0.005209   0.373123  -0.014    0.989
## Studying:FamiiliarStrong Disagree -1.022053   1.511437  -0.676    0.501
## Studying:FamiiliarStrongly Agree  -1.185615   0.521384  -2.274    0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.32 on 70 degrees of freedom
## Multiple R-squared:  0.1554, Adjusted R-squared:  0.08303
## F-statistic: 2.147 on 6 and 70 DF,  p-value: 0.05862
```

```
test_model <- lm(Term.Test~COVID3+COVID4+Studying,data=test)
summary(test_model)
```

```
##
## Call:
## lm(formula = Term.Test ~ COVID3 + COVID4 + Studying, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0466  -5.9801   0.7666   5.7236  17.8918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.4511     2.7279  12.629 2.59e-13 ***
## COVID3       -1.5844     0.8114  -1.953   0.0606 .
## COVID4        1.0724     1.0039   1.068   0.2943
## Studying     -0.1477     0.2664  -0.554   0.5836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.191 on 29 degrees of freedom
## Multiple R-squared:  0.1507, Adjusted R-squared:  0.06285
## F-statistic: 1.715 on 3 and 29 DF,  p-value: 0.1857
```

```
autoplot(test_model)
```



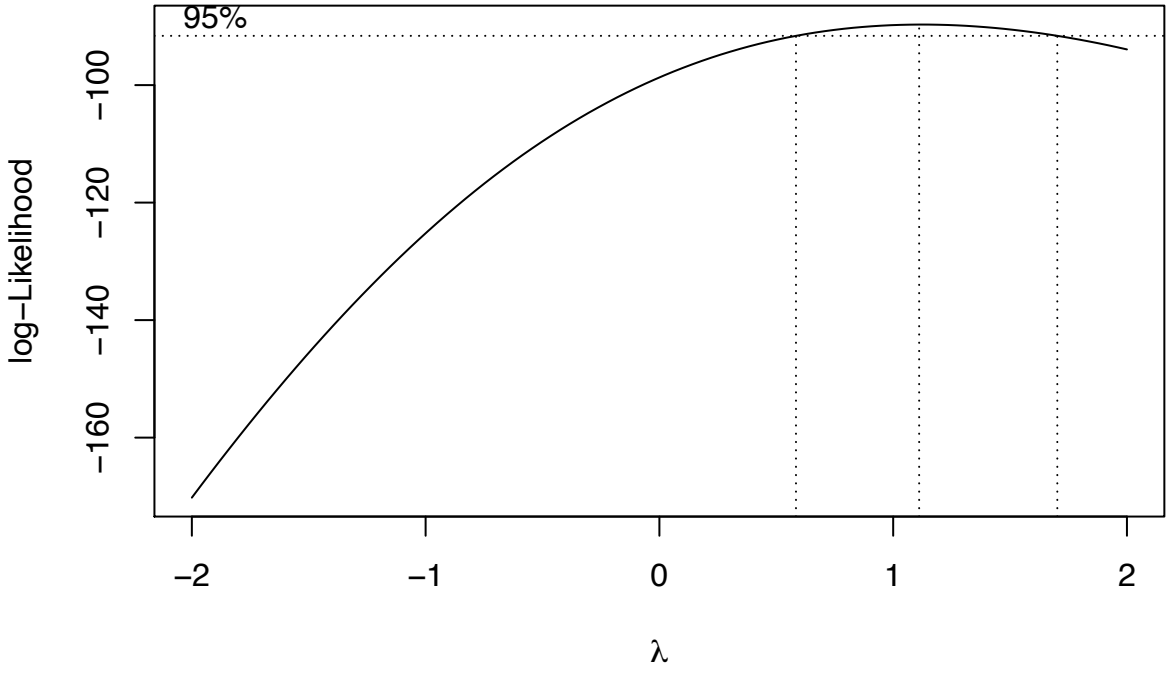

```r
    # Outlying and influential for Y
    n <- nrow(train)
    p_prime <- length(coef(reduced_model))
    # calculate studentized deleted residual
    model_studentized_deleted_residual <- rstudent(reduced_model)
    # alpha value
    alpha <- 0.05
    # construct report
    model_outlying_influential_y <- cbind(
        row_index = seq(1:n),
        Y = train$Term.Test,
        Y_hat = fitted(reduced_model),
        t = model_studentized_deleted_residual,
        cooks_distance = cooks.distance(reduced_model)
      ) %>%
      as_tibble() %>%
      mutate(is_outlier = abs(t) > qt(1 - alpha/(2*n), n - p_prime - 1)) %>%
      mutate(is_influential = cooks_distance > qf(0.5, p_prime, n - p_prime))
    head(model_outlying_influential_y)
```

```
## # A tibble: 6 x 7
##   row_index     Y Y_hat       t cooks_distance is_outlier is_influential
##       <dbl> <dbl> <dbl>   <dbl>          <dbl> <lgl>      <lgl>
## 1         1    53  52.4  0.0724       0.000475  FALSE      FALSE
## 2         2    25  39.1 -1.24         0.0198    FALSE      FALSE
## 3         3    45  33.5  0.973        0.00702   FALSE      FALSE
## 4         4    28  28.3 -0.0305       0.0000474 FALSE      FALSE
## 5         5    32  26.5  0.484        0.00325   FALSE      FALSE
## 6         6    52  45.5  0.570        0.00473   FALSE      FALSE
```

```r
model_outlying_influential_y %>% filter(is_outlier & is_influential)
```

```
## # A tibble: 0 x 7
## # ... with 7 variables: row_index <dbl>, Y <dbl>, Y_hat <dbl>, t <dbl>,
## #   cooks_distance <dbl>, is_outlier <lgl>, is_influential <lgl>
```

```
bc <- boxcox(reduced_model)
```




```
(lambda <- bc$x[which.max(bc$Term.Test)])
```

```
## numeric(0)
```

```
# Partial F Test
final_model <- lm(Term.Test~Studying+COVID4+COVID3,data=train)
anova(final_model)
```

```
## Analysis of Variance Table
##
## Response: Term.Test
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Studying   1   684.8  684.83  4.5450 0.03638 *
## COVID4     1   329.2  329.16  2.1845 0.14371
## COVID3     1   573.5  573.51  3.8062 0.05490 .
## Residuals 73 10999.6  150.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```