Module Report 2

Code

470518795 480050719

20/09/2019

Introduction

The survey conducted by Dr.Garth is about the different characteristics as well as social and academic activities of student enrolled in data20X2 in 2019 semester 2 at University of Sydney. The data from this survey is collected from 110 enrolled students.

Importing data

Hide

```
mydata=readr::read_csv("data/DATA2X02 class survey.csv",na=c("","  "))
```

Data cleaning

Hide

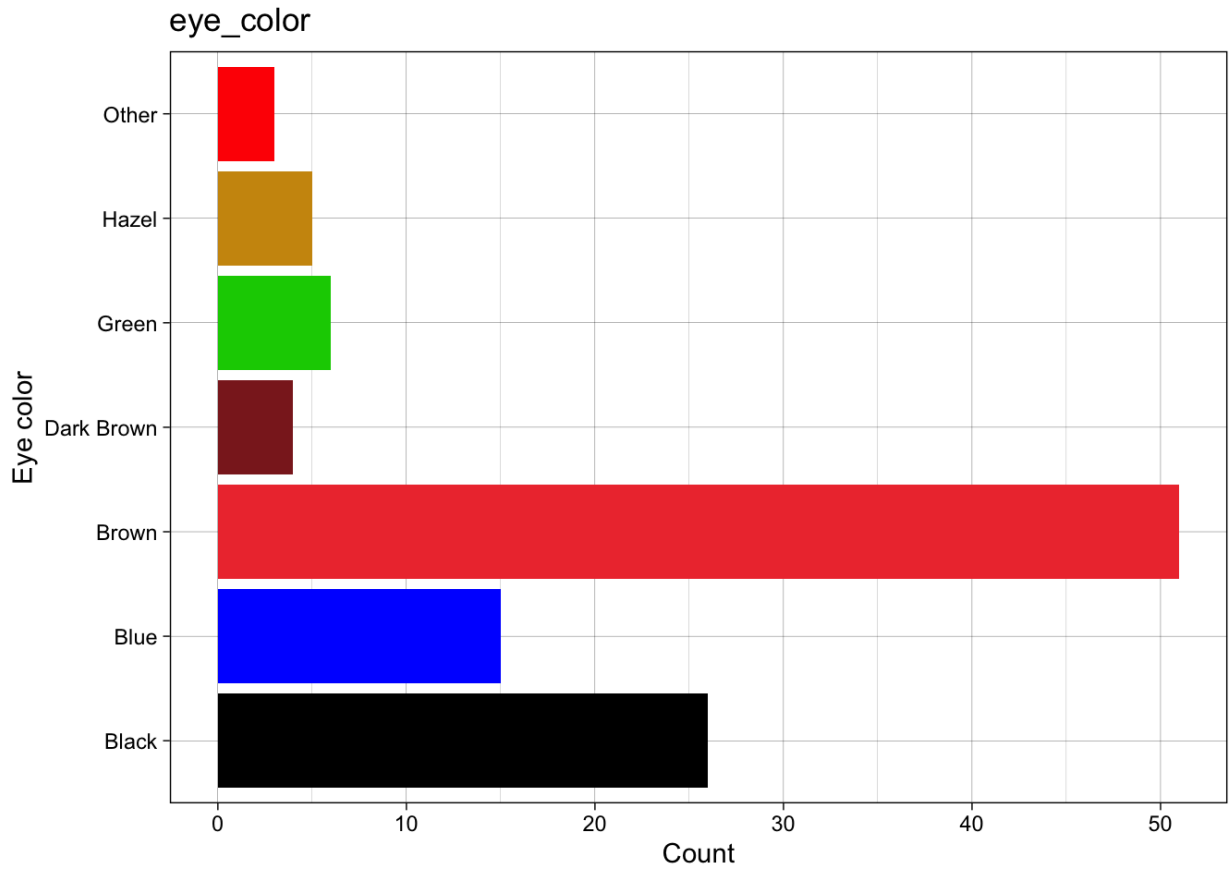
```
clean_data<-janitor::clean_names(mydata)
# Using the function to clean the column names
# however, there are some columns which are quite long, it can be clean
ed further.
new_colnames<-c("survey_time","user_name","gender","home_postcode","pre
vious_statistics_courses",
                "number_of_clubs_attended","last_time_went_to_dentist",
"study_time","social_media",
                "number_of_siblings","pet","live_with_parents","hours_o
f_exercise_per_week","eye_color","work_time","favourite_season","shoe_s
ize","height","frequency_of_floss_teeth","with_glasses","dominant_hand"
,"steak_cooked_preference")
colnames(mydata)<-new_colnames
colnames(mydata)
```

```
##  [1] "survey_time"                 "user_name"
##  [3] "gender"                      "home_postcode"
##  [5] "previous_statistics_courses" "number_of_clubs_attended"
##  [7] "last_time_went_to_dentist"   "study_time"
##  [9] "social_media"                "number_of_siblings"
## [11] "pet"                         "live_with_parents"
## [13] "hours_of_exercise_per_week"  "eye_color"
## [15] "work_time"                   "favourite_season"
## [17] "shoe_size"                   "height"
## [19] "frequency_of_floss_teeth"    "with_glasses"
## [21] "dominant_hand"               "steak_cooked_preference"
```

Eye color

Hide

```
mydata=mydata %>% mutate(eye_color=stringr::str_to_title(eye_color),
                         eye_color=case_when(eye_color=="Balack"~"Blac
k",eye_color=="Hazelnut"~"Hazel",TRUE~eye_color))
mydata = mydata %>%
  mutate(
    eye_color = forcats::fct_lump_min(eye_color, min = 2)
  )
mydata %>% ggplot(aes(x = eye_color)) + geom_bar(fill = c("black",
    "blue", "brown2", "brown4", "green3", "darkgoldenrod3", "red")) +
    labs(title = "eye_color", y = "Count",
        x = "Eye color") + theme_linedraw() + coord_flip()
```



Hide

```
mydata %>% tabyl(eye_color) %>% adorn_pct_formatting() %>% kable()
```

eye_color	n	percent
Black	26	23.6%
Blue	15	13.6%
Brown	51	46.4%
Dark Brown	4	3.6%
Green	6	5.5%
Hazel	5	4.5%
Other	3	2.7%

Gender

Hide

```
recoded_gender=recode_gender(gender=mydata$gender)
mydata$gender=recoded_gender
mydata %>% tabyl(gender)%>% adorn_pct_formatting()%>%kable()
```

gender	n	percent	valid_percent
female	41	37.3%	38.0%
male	66	60.0%	61.1%
non-binary	1	0.9%	0.9%
NA	2	1.8%	•

Social media

Hide

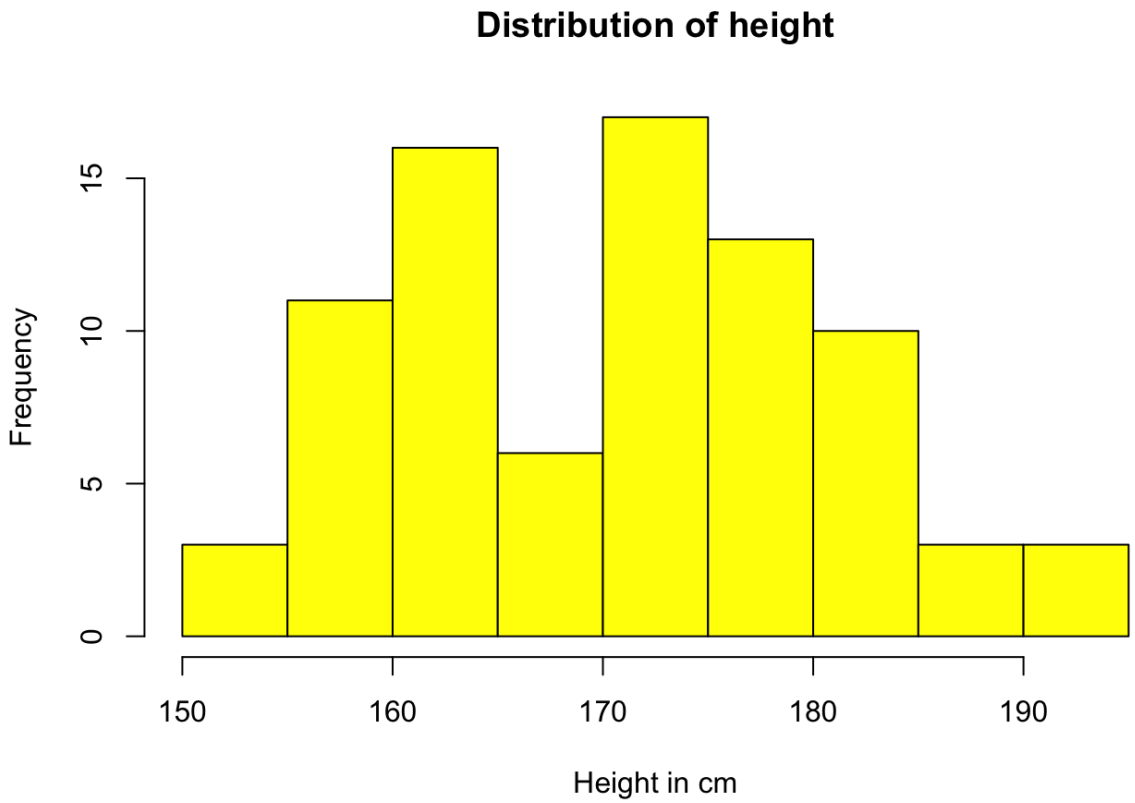
```
mydata$social_media=mydata$social_media %>% tolower()
mydata$social_media=forcats::fct_recode(mydata$social_media,
                                        "facebook"="facebook messenger"
,
                                        "facebook"="fb",
                                        "instagram"="ig",
                                        "none"="i never use social medi
a.")
mydata$social_media=forcats::fct_lump(mydata$social_media,n=7)
mydata %>% tabyl(social_media) %>% adorn_pct_formatting()%>%kable()
```

social_media	npercent	valid_percent
facebook	2926.4%	27.1%
none	65.5%	5.6%
instagram	1917.3%	17.8%
reddit	98.2%	8.4%
snapchat	76.4%	6.5%
wechat	1311.8%	12.1%
youtube	87.3%	7.5%
Other	1614.5%	15.0%
NA	32.7%	•

Height

Hide

```
mydata = mydata %>% mutate(
  height =case_when(
    height<2.2 ~ height*100,
    height<4 ~ NA_real_,
    height<8 ~height *30.48,
    height>250 ~NA_real_,
    TRUE ~ height)
) %>%drop_na()
hist(mydata$height,main="Distribution of height",xlab = "Height in cm",
col = "yellow")
```



Exercise

Hide

```
mydata$hours_of_exercise_per_week<-as.numeric(mydata$hours_of_exercise_
per_week)
mydata=mydata %>%mutate(
  hours_of_exercise_per_week=case_when(
    hours_of_exercise_per_week>60 ~ NA_real_,
    TRUE ~ hours_of_exercise_per_week
  )
)
mydata$hours_of_exercise_per_week=as.numeric(mydata$hours_of_exercise_p
er_week)
```

clubs

Hide

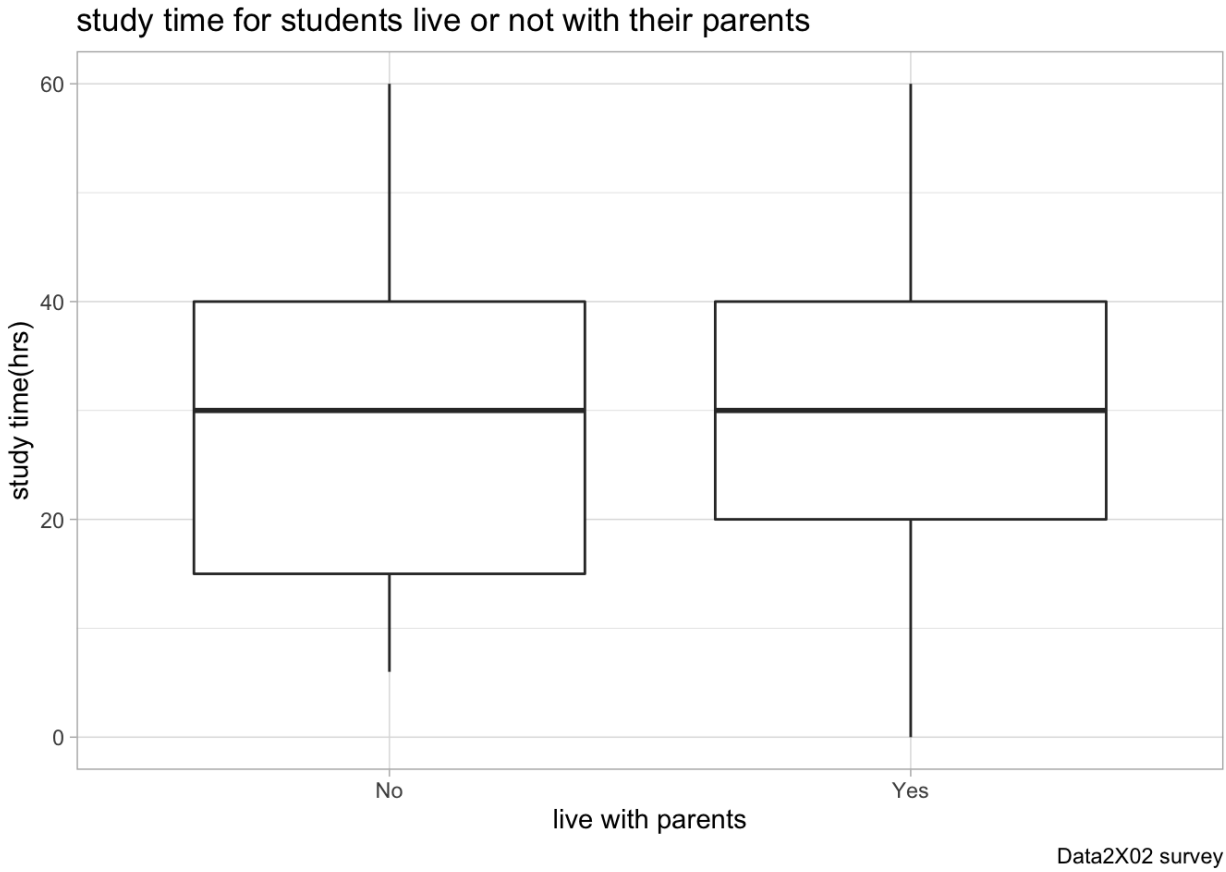
```
mydata$number_of_clubs_attended=as.double(mydata$number_of_clubs_attend
ed)
mydata = mydata %>%
  mutate(number_of_clubs_attended=case_when(
    number_of_clubs_attended<0 ~ NA_real_,
    number_of_clubs_attended>10 ~ NA_real_,
    TRUE ~ number_of_clubs_attended
  )) %>% drop_na
mydata %>% tabyl(number_of_clubs_attended) %>% adorn_pct_formatting()%
>%kable()
```

number_of_clubs_attended	n	percent
0	30	36.6%
1	15	18.3%
2	16	19.5%
3	16	19.5%
4	1	1.2%
5	2	2.4%
7	1	1.2%
10	1	1.2%

Study

Hide

```
study <-mydata %>% select(live_with_parents,study_time) %>% filter(stud
y_time<70)
with_parents<- study %>%
  filter(live_with_parents=="Yes")
without_parents <-study %>%
  filter(live_with_parents=="No")
study %>% ggplot(aes(x=live_with_parents,y=study_time)) +
  geom_boxplot()+
  theme_light()+
  labs(x="live with parents",y="study time(hrs)",title = "study time fo
r students live or not with their parents", caption = "Data2X02 survey"
)
```



Save the cleaned data

Hide

```
readr::write_csv(mydata,"mydata_cleaned.csv")
```

Data analysis

1: Is this a random sample of DATA2002 students?

This survey is not random sample. Students who completed this survey is actively interact with lecturer, however, not all response to this survey. Therefore, samples only can represent a part of active DATA2X02 students, it cannot represent the whole population. This survey suffers from selection bias. Those who do not respond survey is inactive comparing with those who complete it, certain group(inactive students) are under-represented. The probability of completing survey is not the same across the students, therefore the sample is not random.

2: What are the potential biases in this data generation?

There might be selection bias, people answering this survey is likely to be more active. Some inactive students do not answer the survey. Therefore, the sample does not acurately represents the population.

There also likely to exists measurement bias. For example, if it is hard to identify whether student studying 30 hours per week or 30 min if he just type 30 in the survey. Also , there are some people who study more than 144 hours one week, which is unrealistic. This is because some people does not take this survey seriously. Therefore, there exists measurement bias in the dataset.

Another measurement bias would be on study time, student may tend to overestimate their study time since they may feel inappropriate to study less than certain amount of time. Since we can not detect the truth and there is no penalty or award for eliciting the real time of study. Therefore, this is not incentive compatible.

3: Which variables are most likely to be subjected to this bias?

In this survey, there are several biases, and the free-response variables are most likely to be subjected to this bias because the answers of these questions often have unrelated or even impossible answers, such as exercise time, study time. For example, it is impossible for a person to study more than 144 hours per week. These kind of response of those variables could even be outliers there, which would definitely lead to bias in this survey.

Besides, the variable that did not clarify the standard measurement units are also most likely to be subjected to this bias, such as height(it could be measured in cm or m), study hours(it could be measured in minutes or hours), shoe size (it is different in different country). For example, the height, someone use cm as measurement unit (for example, 178) and someone use m as measurement unit here (for example, 1.89).

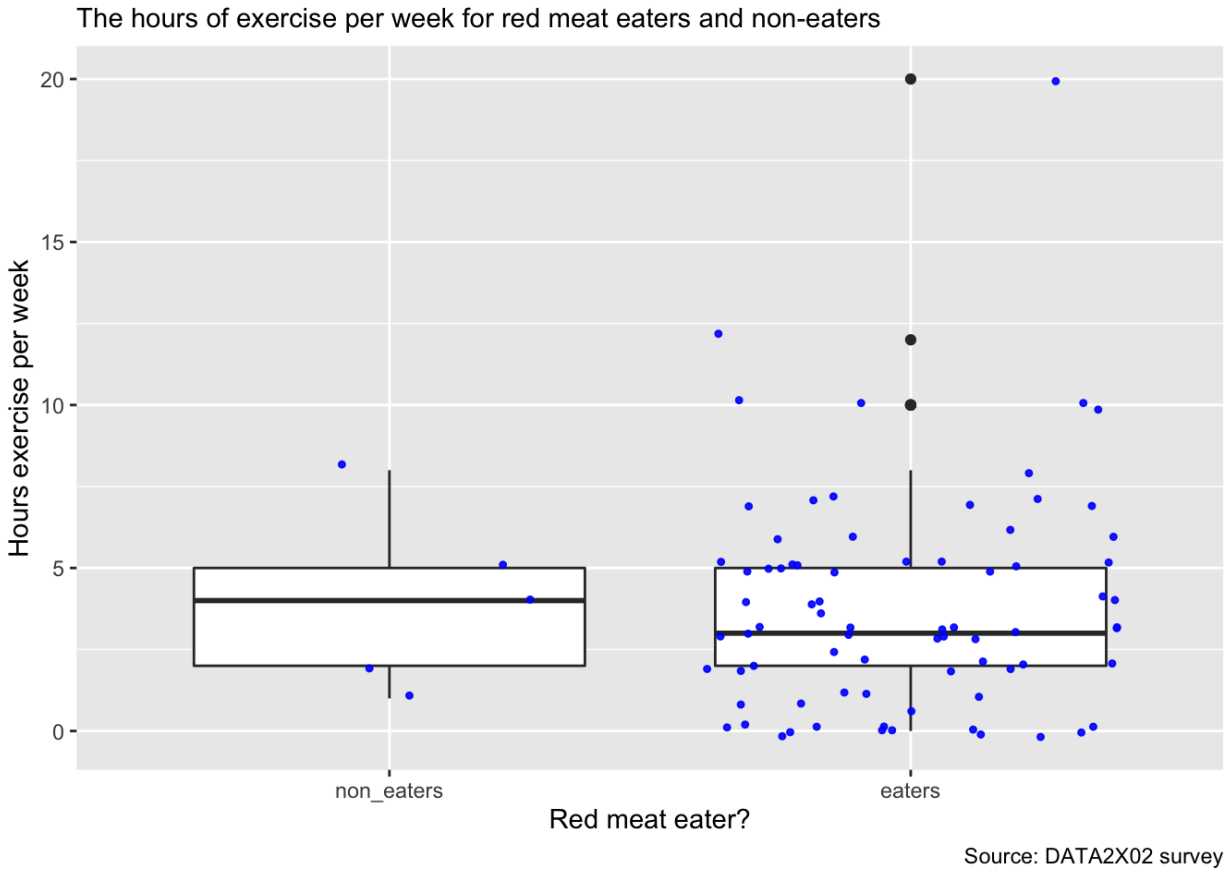
4: Is there any evidence to suggest that there's a difference in the amount of exercise done by people who eat red meat compared to people who don't eat red meat?

Visualisation

Hide

```
mydata$steak_cooked_preference=fct_collapse(mydata$steak_cooked_prefere
nce,
                                            eaters=c("Blue","Rare","Med
ium-rare","Medium","Medium-well done","Well done"),
                                            non_eaters=c("I don't eat r
ed meat"))
exercise = mydata %>% filter(steak_cooked_preference=="eaters"||steak_c
ooked_preference=="non-eaters")%>% select(steak_cooked_preference,hours
_of_exercise_per_week)

ggplot(subset(exercise,!is.na(steak_cooked_preference)),aes(x=steak_coo
ked_preference,y=hours_of_exercise_per_week))+
  geom_boxplot()+
  theme(
    legend.position = "none",
    plot.title = element_text(size = 11)
  )+
  geom_jitter(color="blue",size=0.9,alpha=0.9)+
  labs(title = "The hours of exercise per week for red meat eaters and
 non-eaters",
       x="Red meat eater?",
       y="Hours exercise per week",
       caption = "Source: DATA2X02 survey")
```



Two sample test

This is a two sample test.

Check for assumption: Equal Variance

Firstly, looking at the pooled two sample t-test, the assumption here is there have equal variance. By calculation, it can be seen that the standard deviation differences among the sample is approximately 1. By test of equal variance assumption, p-value is much more than 0.05. Therefore, we have sufficient evidence to assume equal variance.

Hide

```
eaters=exercise %>% filter(steak_cooked_preference=="eaters",as.numeric
(hours_of_exercise_per_week)) %>% drop_na
sd_eaters=sd(eaters$hours_of_exercise_per_week)
non_eaters=exercise %>% filter(steak_cooked_preference=="non_eaters",a
s.numeric(hours_of_exercise_per_week)) %>% drop_na
sd_noneaters=sd(non_eaters$hours_of_exercise_per_week)
c((sd_eaters),(sd_noneaters))
```

```
## [1] 3.206120 2.738613
```

Hide

```
var.test(eaters$hours_of_exercise_per_week,non_eaters$hours_of_exercise
_per_week)
```

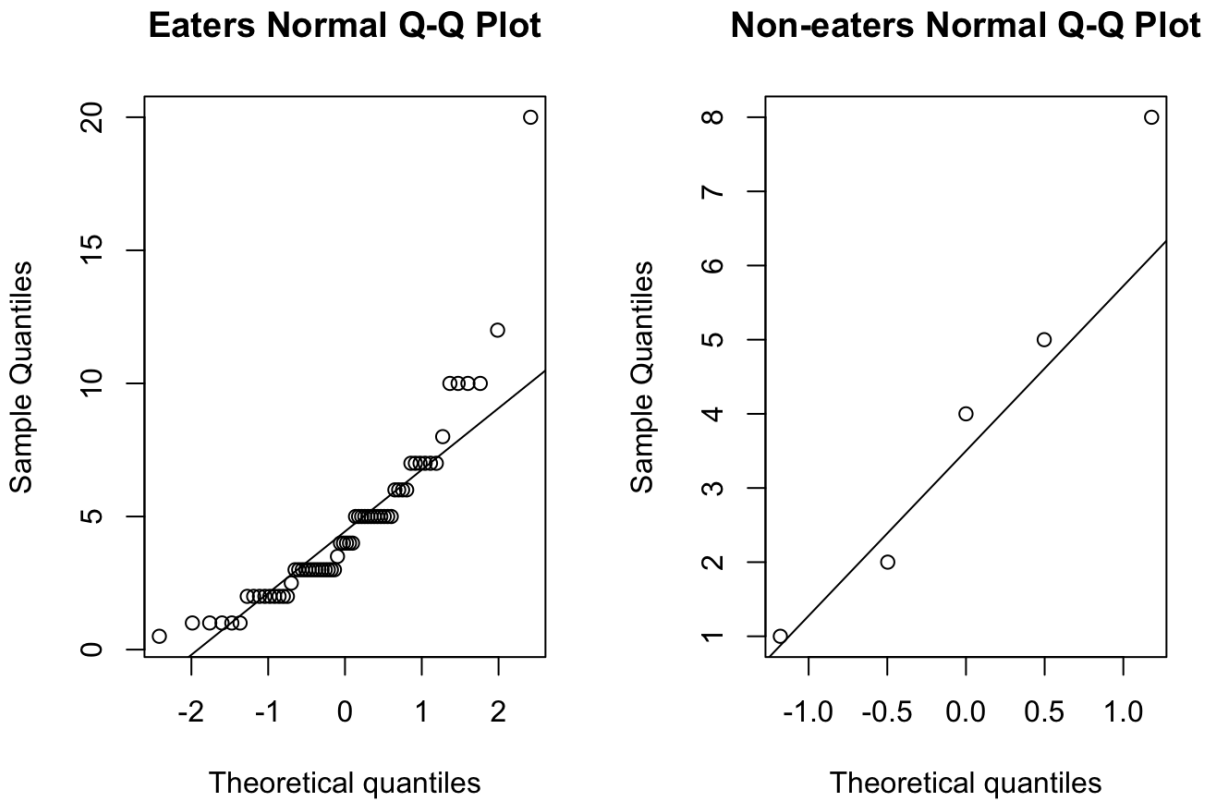
```
##
##  F test to compare two variances
##
## data:  eaters$hours_of_exercise_per_week and non_eaters$hours_of_exe
rcise_per_week
## F = 1.3706, num df = 63, denom df = 4, p-value = 0.8498
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1640294 4.1070874
## sample estimates:
## ratio of variances
##           1.370561
```

Check for assumption: Normality

Secondly, it is assumed that the distribution of these data is normal. This assumption cen be checked by drawing QQ-plot to see wehther it is approximate straight line. By looking at two diagram, it is obvious that the sample is quite fit with theoretical line. The problem is that we have only few samples of non-eaters. Therefore, it is reasonable to assume hours of exercise is normally distributed.

Hide

```
par(mfrow=c(1,2))
qqnorm(eaters$hours_of_exercise_per_week,main="Eaters Normal Q-Q Plot",
xlab="Theoretical quantiles",ylab="Sample Quantiles")
qqline(eaters$hours_of_exercise_per_week)
qqnorm(non_eaters$hours_of_exercise_per_week,main="Non-eaters Normal Q-
Q Plot",xlab="Theoretical quantiles",ylab="Sample Quantiles")
qqline(non_eaters$hours_of_exercise_per_week)
```



It is suggested that the pooled two sample t-test should be used in this analysis since the equal variance assumption is statisfied.

Pooled two sample t-test

Hypothesis $H_0 : u_{eaters} = p_{non-eaters}$ vs $H_1 : u_{eaters} \neq u_{non-eaters}$ Null hypothesis: There is no difference in hours of exercise per week between red meat eaters and non-eaters Alternative: The hours of exercise between these two groups are different.

Assumption $X_1, \ldots, X_{n_x}$ are iid with $N(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_{n_x}$ are iid with $N(\mu_Y, \sigma^2)$ and X_i are independent of Y_i

Test statistics $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}$, where $S_p^2 = \frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2}$. Under H_0, $T \sim t_{n_x+n_y-2}$

Observed test statistics $t_0 = 0.4073664$

Hide

```
n1=nrow(eaters)
n2=nrow(non_eaters)
Sp=sqrt(((n1-1)*(sd_eaters)^2+(n2-1)*(sd_noneaters)^2)/(n1+n2-2))
t_0=(mean(eaters$hours_of_exercise_per_week)-mean(non_eaters$hours_of_e
xercise_per_week))/(Sp*sqrt(1/n1+1/n2))
t_0
```

```
## [1] 0.4073664
```

P-value $p = 2P(T \geq t_0) = 0.6850$

Hide

```
(p=2*pt(t_0,df=n1+n2-2,lower.tail = FALSE))
```

```
## [1] 0.6850376
```

Decision Since the p-value is approximately 0.68, this is not significant at 5% significance level, therefore, we do not reject null hypthothesis that there is no difference on hours of exercise per week between red meat eaters and non-eaters.

Hide

```
t.test(eaters$hours_of_exercise_per_week,non_eaters$hours_of_exercise_p
er_week,alternative="two.sided",var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  eaters$hours_of_exercise_per_week and non_eaters$hours_of_exe
rcise_per_week
## t = 0.40737, df = 67, p-value = 0.685
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.345966  3.549091
## sample estimates:
## mean of x mean of y
##  4.601562  4.000000
```

As mentioned, the only concern is that for non-eaters, there are only few samples. It is doubt that the normality assumption is hold. To make our analysis more robust, we can use Wilcoxon rank-sum test which relaxes on normality assumption.

Wilcoxon rank-sum test

Hypothesis $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x \neq \mu_y$

Assumption $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_x}$ are iid and follow the same distribution but differ by a shift.

Test statistics $W = R_1 + R_2 + \ldots + R_{n_x}$. Under H_0, $W \sim WRS(100, 7)$ distribution

Observed test statistics w=r_1+r_2+...+r_n_x=175

P-value $E(W) = \frac{100(107+1)}{2} = 5400$

$Var(W) = \frac{100*7}{107(107-1)}(\sum_{i=1}^{N}(r_i^2) - \frac{107(107+1)^2}{4})$

$p \approx 2P(Z \geq |\frac{W-E(W)}{\sqrt{Var(W)}}|) = 0.726$

Decision As the p-value is much more than 0.05, there is insufficient evidence to reject H_0.

Hide

```
wilcox.test(eaters$hours_of_exercise_per_week,non_eaters$hours_of_exerc
ise_per_week,correct = FALSE,alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test
##
## data:  eaters$hours_of_exercise_per_week and non_eaters$hours_of_exe
rcise_per_week
## W = 175, p-value = 0.726
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion:

In these two test, we do not have sufficient evidence to reject H0: There is no difference in weekly hours of exercise between eaters and non-eaters. Therefore, we do not reject null hypothesis and find evidence for difference in weekly exercise hours between red meat eaters and non-eaters.

5: Is there any evidence that the weekly exercise time of students who participate in more than 3 univeristy clubs is different to those who don't?

Hide

```
data = data.frame(mydata$number_of_clubs_attended,mydata$hours_of_exerc
ise_per_week)
moreclub= data %>%
  filter(mydata.number_of_clubs_attended>3)
lessclub=data %>%
  filter(mydata.number_of_clubs_attended<=3)
```

Two sample test

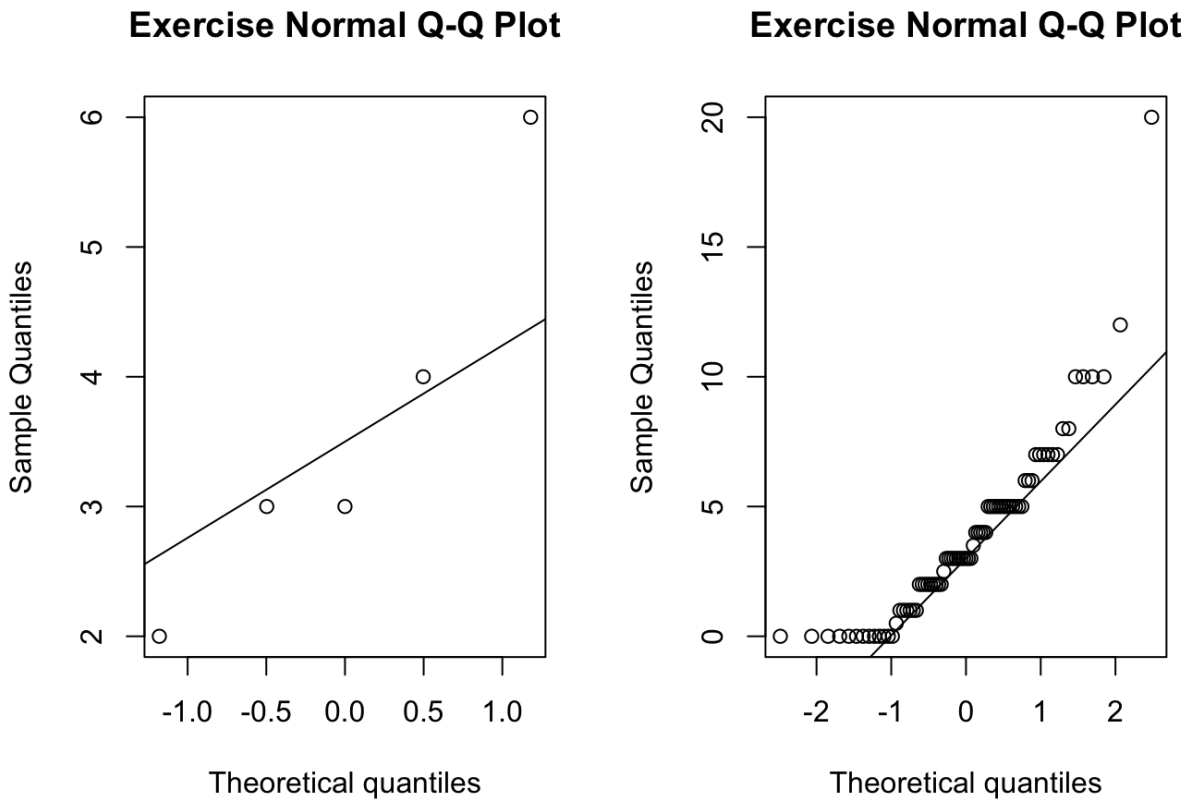
This is a two sample test.

Check for assumption: Normality

By checking the qqplot, it can be seen that the qqplot isn't fit well as well. Therefore, it is suggested that we should relax the normality assumption and use Wilcoxon rank-sum test.

Hide

```
par(mfrow=c(1,2))
qqnorm(moreclub$mydata.hours_of_exercise_per_week,main="Exercise Normal
Q-Q Plot",xlab="Theoretical quantiles",ylab="Sample Quantiles")
qqline(moreclub$mydata.hours_of_exercise_per_week)
qqnorm(lessclub$mydata.hours_of_exercise_per_week,main="Exercise Normal
Q-Q Plot",xlab="Theoretical quantiles",ylab="Sample Quantiles")
qqline(lessclub$mydata.hours_of_exercise_per_week)
```

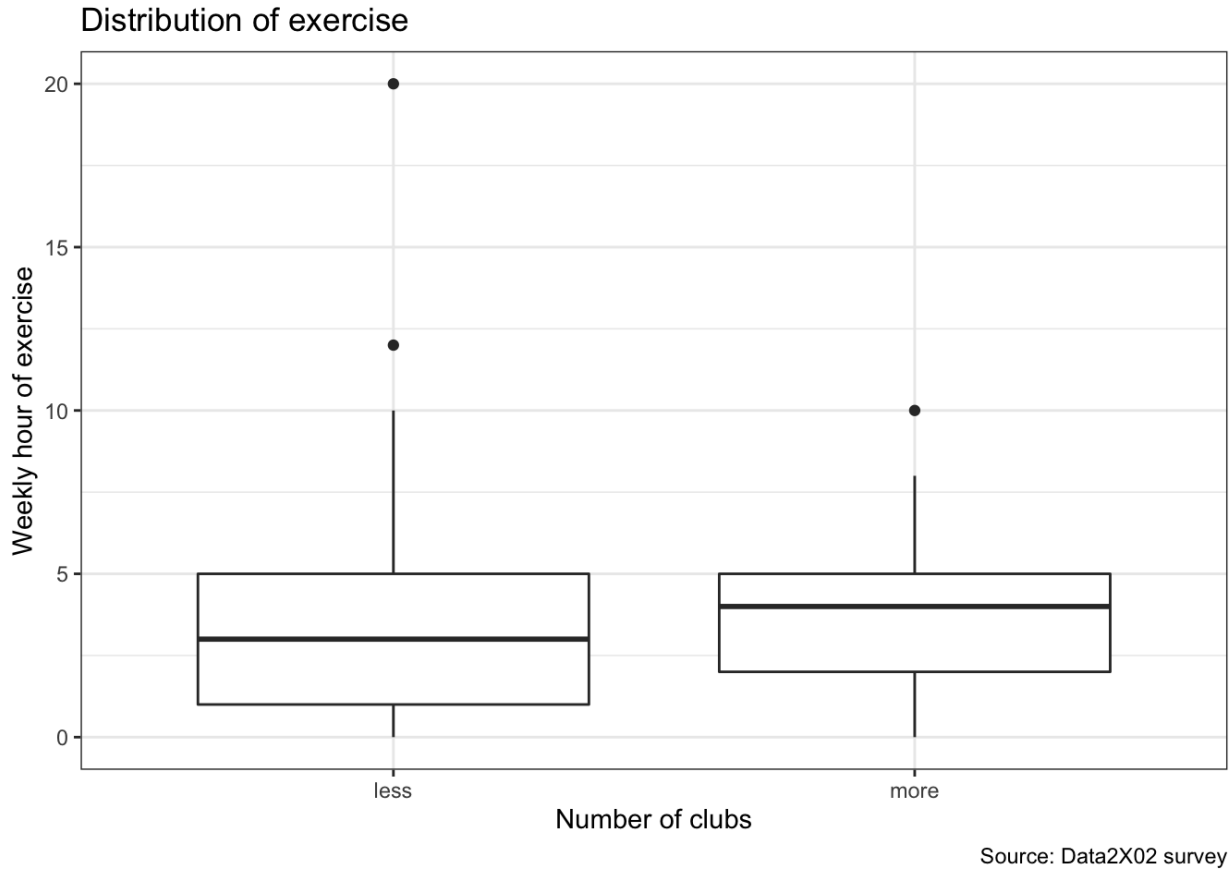


Wilcoxon Ranked-sum test

We need to do is check the assumption of Wilcoxon Ranked sum test which is the 2 samples follow the same distribution. By looking at the boxplot, these two are roughly symmetric. Therefore, we can use Wilcoxon Ranked-sum test.

Hide

```
data = data %>%
  mutate(number_of_clubs_attended =
           case_when(
             mydata.number_of_clubs_attended < 3 ~ "less",
             mydata.number_of_clubs_attended >= 3 ~ "more"
           ))
ggplot(subset(data, !is.na(mydata.number_of_clubs_attended),!is.na(myda
ta.hours_of_exercise_per_week)), aes(x = number_of_clubs_attended, y =
 mydata$hours_of_exercise_per_week)) +
geom_boxplot() +
theme_bw()+
  labs(title="Distribution of exercise",
       x="Number of clubs",
       y="Weekly hour of exercise",
       caption = "Source: Data2X02 survey")
```



Hypothesis $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x \neq \mu_y$

Assumption $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_x}$ are iid and follow the same distribution but differ by a shift.

Test statistics $W = R_1 + R_2 + \ldots + R_{n_x}$. Under H_0, $W \sim WRS(n_x, n_y)$ distribution

Observed test statistics w=r_1+r_2+...+r_n_x=183.5

P-value

$$p \approx 2P(Z \geq |\frac{W - E(W)}{\sqrt{Var(W)}}|) = 0.8605$$

Decision As the p-value is much more than 0.05, there is insufficient evidence to reject H_0.

Hide

```
wilcox.test(lessclub$mydata.hours_of_exercise_per_week,moreclub$mydata.
hours_of_exercise_per_week,correct=FALSE,alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test
##
## data:  lessclub$mydata.hours_of_exercise_per_week and moreclub$mydat
a.hours_of_exercise_per_week
## W = 183.5, p-value = 0.8605
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion

In the test, we do not have sufficient evidence to reject H0: There is no difference in weekly hours of exercise between people who attend less or more club. Therefore, we do not reject null hypothesis.

6: Is there evidence that students who live with their parents study more hours per week than students who don't live with their parents?

Checking for assumption: Equal variance

Hide

```
sd_with_parents=sd(with_parents$study_time)
sd_noparents=sd(without_parents$study_time)
c((sd_with_parents),(sd_noparents))
```

```
## [1] 13.10353 14.25833
```

Hide

```
var.test(with_parents$study_time,without_parents$study_time)
```

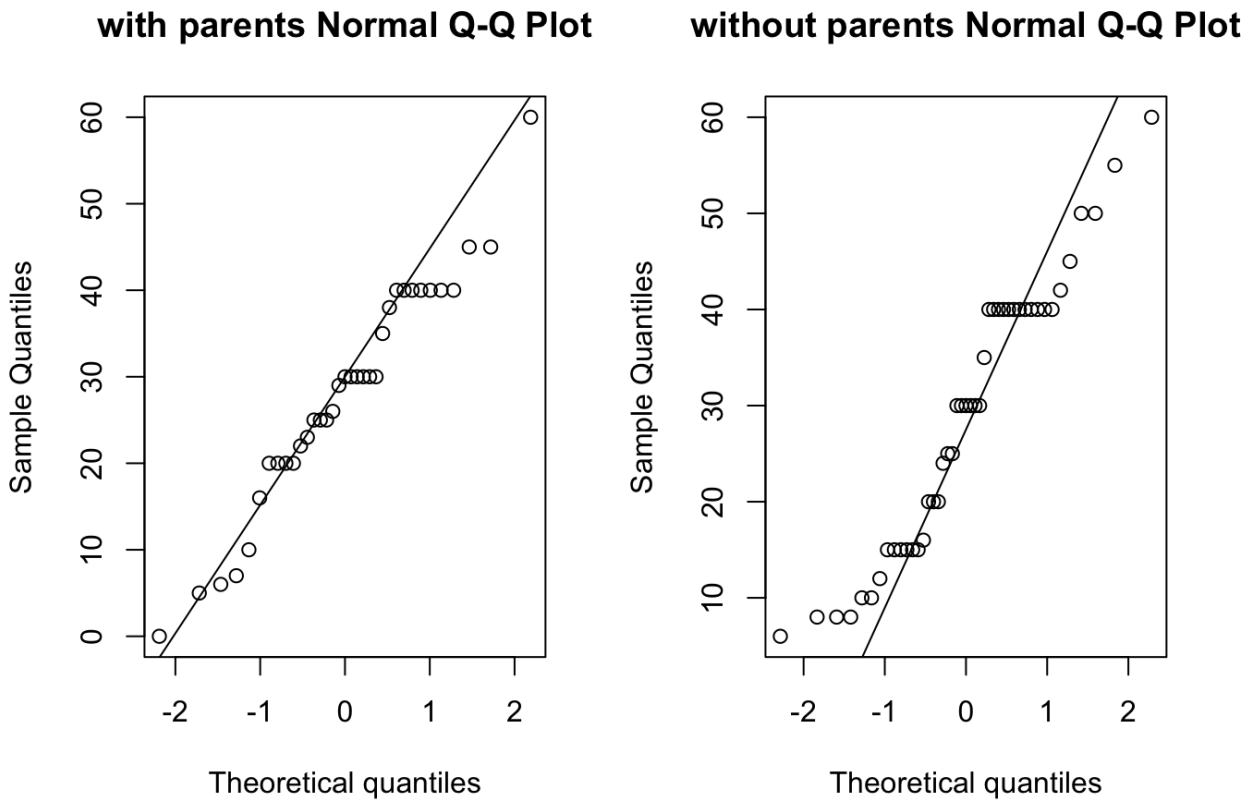
```
##
##  F test to compare two variances
##
## data:  with_parents$study_time and without_parents$study_time
## F = 0.84458, num df = 34, denom df = 44, p-value = 0.6145
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4501835 1.6273622
## sample estimates:
## ratio of variances
##          0.8445771
```

It can be seen that our variance test is not being rejected, therefore, we can say that the equal variance assumption is not violated. Hence, we can use pooled two sample t-test.

Checking for assumption: Normality

Hide

```
par(mfrow=c(1,2))
qqnorm(with_parents$study_time,main="with parents Normal Q-Q Plot",xlab
="Theoretical quantiles",ylab="Sample Quantiles")
qqline(with_parents$study_time)
qqnorm(without_parents$study_time,main="without parents Normal Q-Q Plo
t",xlab="Theoretical quantiles",ylab="Sample Quantiles")
qqline(without_parents$study_time)
```



By looking at the QQ plot, it can be seen that the line is quite fitting to our sample quantiles. Therefore, the normality assumption is also hold. We can use pooled sample t-test and Wilcoxon test.

Pooled two sample t-test

Hypothesis $H_0 : u_{eaters} = p_{non-eaters}$ vs $H_1 : u_{eaters} \neq u_{non-eaters}$ Null hypothesis: There is no difference in hours of study per week between those who living or not living with their parents Alternative: The hours of study between these two groups are different.

Assumption $X_1, \ldots, X_{n_x}$ are iid with $N(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_{n_x}$ are iid with $N(\mu_Y, \sigma^2)$ and X_i are independent of Y_i

Test statistics $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}$, where $S_p^2 = \frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2}$. Under $H_0, T \sim t_{n_x+n_y-2}$

Observed test statistics $t_0 = -0.26091$

Hide

```
t.test(with_parents$study_time,without_parents$study_time,var.equal = T
RUE,alternative = "two.sided")
```

```
##
##  Two Sample t-test
##
## data:  with_parents$study_time and without_parents$study_time
## t = -0.26091, df = 78, p-value = 0.7949
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.986527  5.367480
## sample estimates:
## mean of x mean of y
##  28.05714  28.86667
```

P-value $p = 2P(T \geq t_0) = 0.7949$

Decision Since the p-value is approximately 0.79, this is not significant at 5% significance level, therefore, we do not reject null hypthothesis that there is no difference on hours of study per week between these two groups.

As mentioned, the only concern is that for non-eaters, there are only few samples. It is doubt that the normality assumption is hold. To make our analysis more robust, we can use Wilcoxon rank-sum test which relaxes on normality assumption.

Wilcoxon rank-sum test

Hypothesis $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x \neq \mu_y$

Assumption $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_x}$ are iid and follow the same distribution but differ by a shift.

Test statistics $W = R_1 + R_2 + \ldots + R_{n_x}$. Under H_0, $W \sim WRS(n_x, n_y)$ distribution

Observed test statistics w=r_1+r_2+...+r_n_x=761

P-value

$p \approx 2P(Z \geq |\frac{W - E(W)}{\sqrt{Var(W)}}|) = 0.7953$

Decision As the p-value is much more than 0.05, there is insufficient evidence to reject H_0.

Hide

```
wilcox.test(with_parents$study_time,without_parents$study_time,correct
= FALSE,alternative = "two.sided")
```

```
## Warning in wilcox.test.default(with_parents$study_time,
## without_parents$study_time, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test
##
## data:  with_parents$study_time and without_parents$study_time
## W = 761, p-value = 0.7953
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion

Firstly, by looking at the boxplot in data cleaning part, there is no much difference between two plot. And there have similar distribution. Moreover, we assumes the equal variance and normality, pooled two sample t-test is being used. Then we relaxes assumption normality and equal variance. We use wilcoxan test. By these two test, it doesm't found any evidence of difference on study time between these two groups. Therefore, we have insufficent evidence to say there is a difference on study time between these two groups.

7: What other questions could you ask?

Is handedness independent of wearing glasses?

Assumption: expected cell counts>=5

It can be seen that the assumption needed for chi-squared test does not hold, therefore, it is recommended to use Fisher exact test

Hide

```
handedglasseschi <- mydata %>%
  select(with_glasses, dominant_hand) %>%
  drop_na() %>%
  group_by(with_glasses, dominant_hand) %>%
  count() %>%
  spread(key = dominant_hand, value = n) %>%
  column_to_rownames(var = "with_glasses")
handedglasseschi
```

```
##     Left handed Right handed
## No            4           29
## Yes           3           46
```

Hide

```
n <- sum(handedglasseschi)
r <- rowSums(handedglasseschi)
c <- colSums(handedglasseschi)
e <- r %*% t(c) / n
paste("The expected cell count assumption is", valid <- all(e >= 5))
```

```
## [1] "The expected cell count assumption is FALSE"
```

Hide

```
fisher.test(handedglasseschi)
```

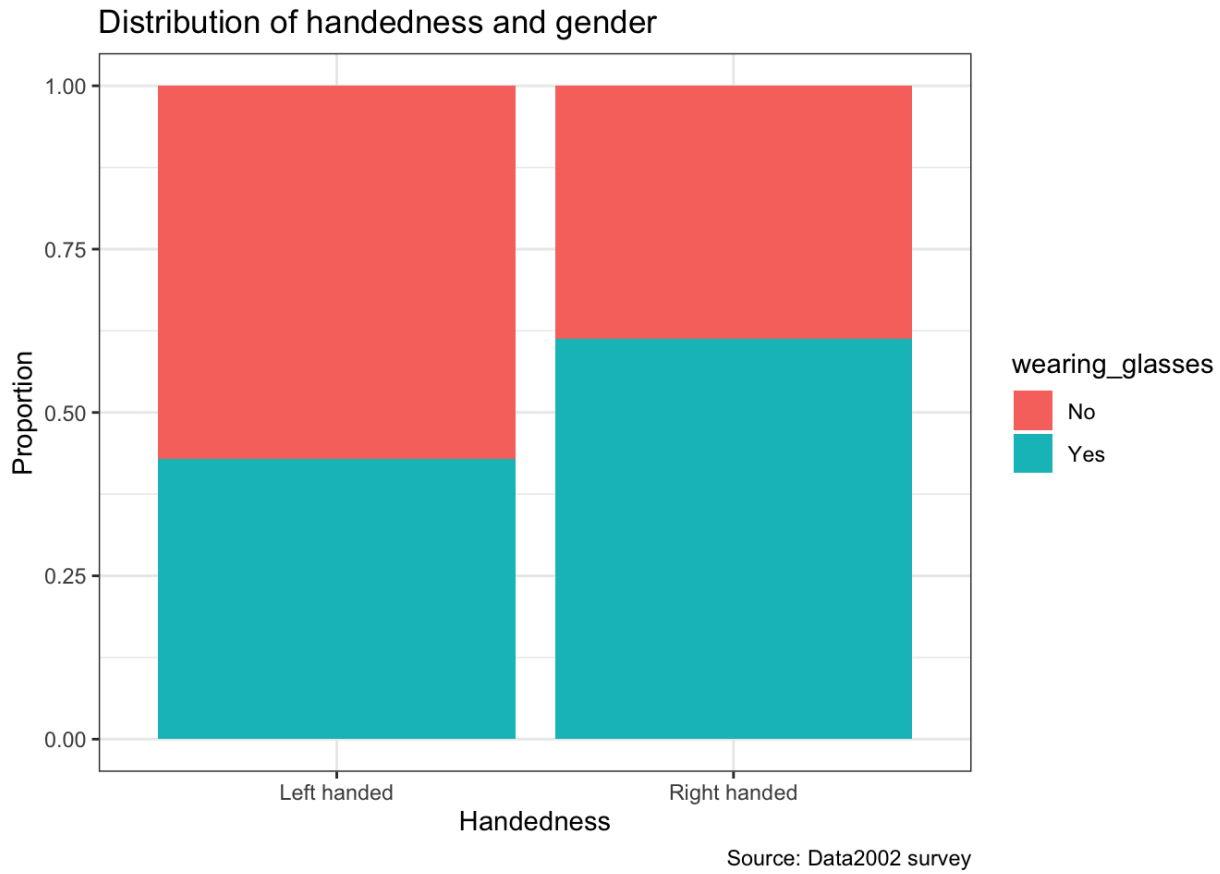
```
##
##  Fisher's Exact Test for Count Data
##
## data:  handedglasseschi
## p-value = 0.4311
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.3286176 15.3534302
## sample estimates:
## odds ratio
##   2.095038
```

The p-value is 0.7818 which is quite large comparing to 0.05. Therefore, we have sufficient evidence to claim that wearing glasses is independent of handedness.

Data visualisation

Hide

```
ggplot(data = mydata %>%drop_na() ,
       aes(x = dominant_hand, fill = with_glasses) )+
  geom_bar(position = "fill") +
  labs(x = "Handedness",
       y = "Proportion",
       fill = "wearing_glasses",
       title = "Distribution of handedness and gender",
       caption = "Source: Data2002 survey") +
  theme_bw()
```



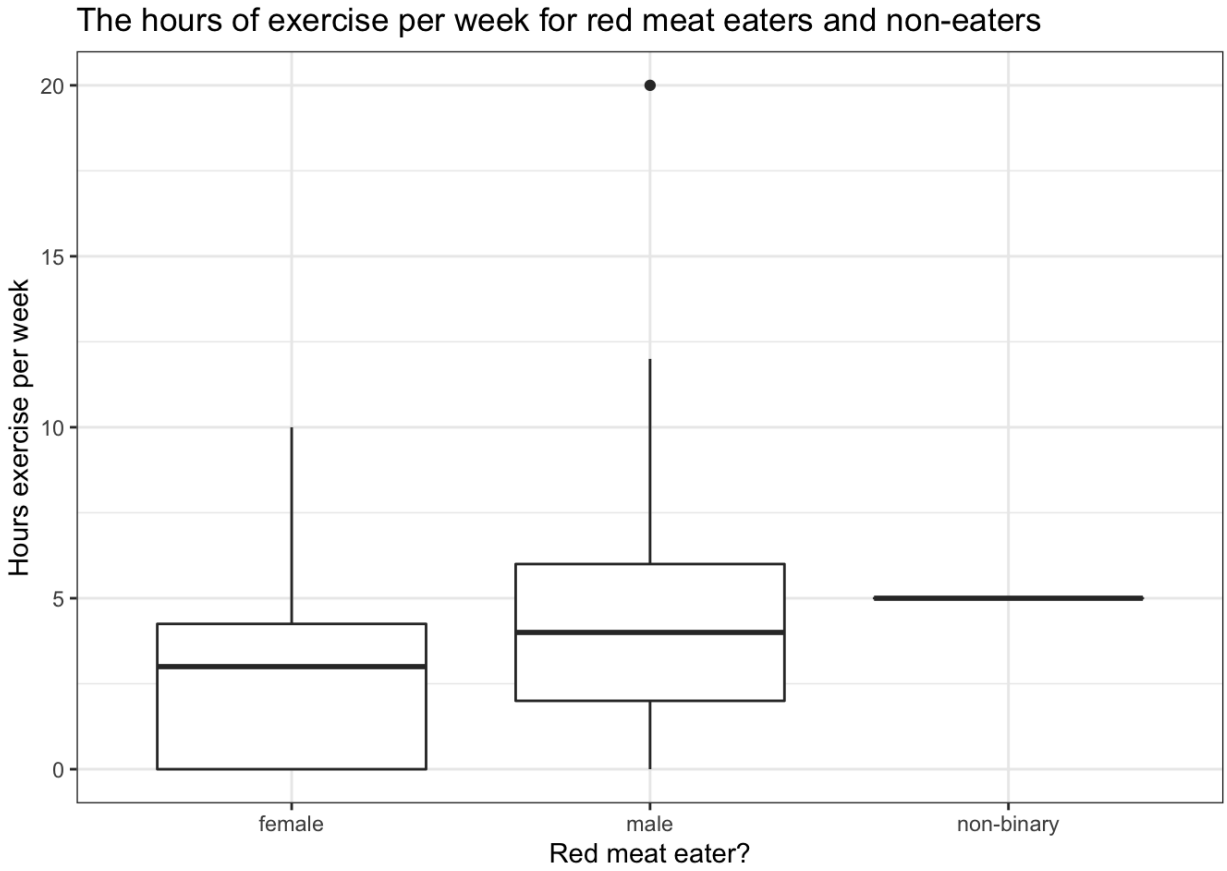
It can be seen that while there were roughly the same amount of left-handed wearing glasses and without glasses, the right-handed subpopulation tended to be skewed towards wearing glasses and the left-handed subpopulation tended to be skewed towards without glasses.

Is there any evidence that there is an difference on hours of exercise per week between male and female?

Visualisation

Hide

```
gender = mydata %>% filter(mydata$gender=="male"||mydata$gender=="femal
e")%>% select(gender,hours_of_exercise_per_week)
ggplot(subset(gender,!is.na(hours_of_exercise_per_week)),aes(x=gender,y
=hours_of_exercise_per_week))+
  geom_boxplot()+
  theme_bw()+
  labs(title = "The hours of exercise per week for red meat eaters and
 non-eaters",
       x="Red meat eater?",
       y="Hours exercise per week")
```



This is a two sample t-test.

Checking assumption: Equal Variance

Firstly, looking at the pooled two sample t-test, the assumption here is there have equal variance. By test of equal variance assumption, p-value is less than 0.05. Therefore, we have insufficient evidence to assume equal variance.

Hide

```
male=gender %>% filter(gender=="male",as.numeric(hours_of_exercise_per_
week)) %>% drop_na

sd_male=sd(male$hours_of_exercise_per_week)
female=gender %>% filter(gender=="female",as.numeric(hours_of_exercise_
per_week)) %>% drop_na
sd_female=sd(female$hours_of_exercise_per_week)
c((sd_male),(sd_female))
```

```
## [1] 3.606377 2.213942
```

Hide

```
var.test(male$hours_of_exercise_per_week,female$hours_of_exercise_per_w
eek)
```

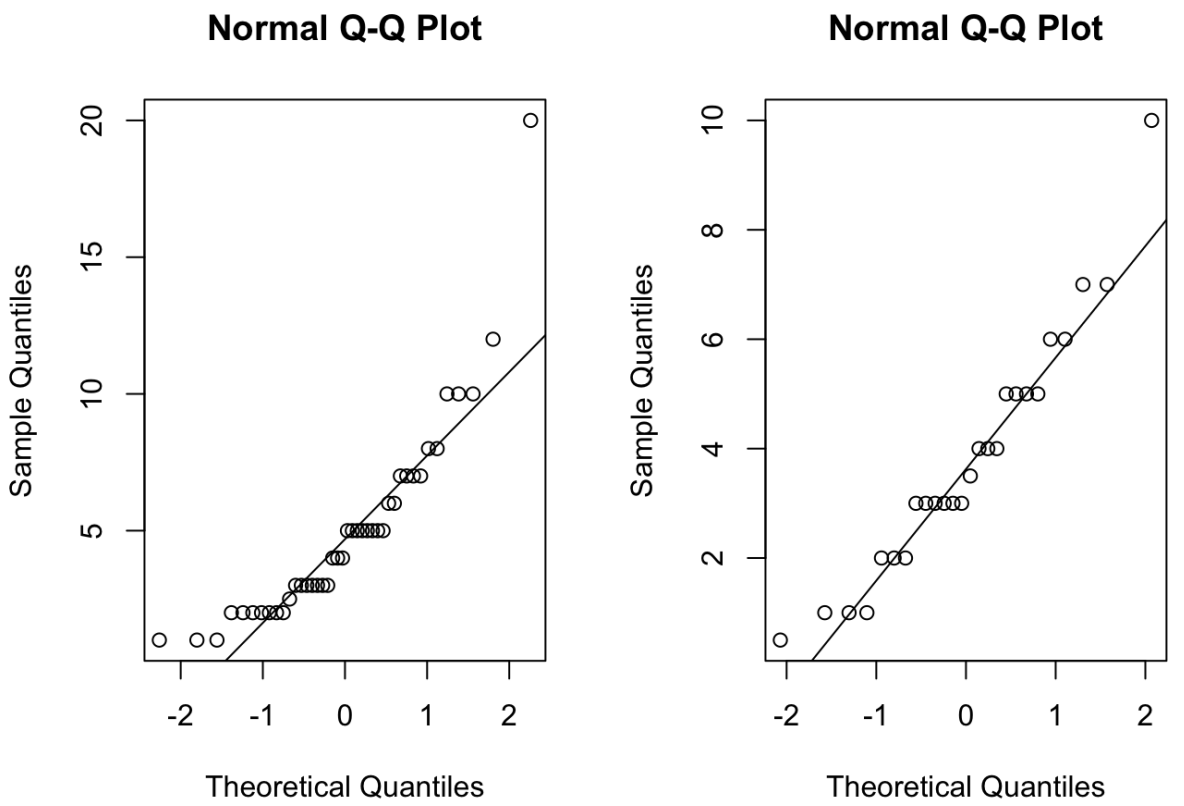
```
##
##  F test to compare two variances
##
## data:  male$hours_of_exercise_per_week and female$hours_of_exercise_
per_week
## F = 2.6534, num df = 41, denom df = 25, p-value = 0.01154
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.255471 5.267171
## sample estimates:
## ratio of variances
##           2.653443
```

Checking assumption: Normality

Therefore, cosidering the second assumption normality, it is assumed that the distribution of these data is normal. This assumption cen be checked by drawing QQ-plot to see wehther it is approximate straight line. By looking at two diagram, it is obvious that the sample is quite fit with theoretical line. Therefore, it is reasonable to assume hours of exercise is normally distributed. There are also sufficient samples.

Hide

```
par(mfrow=c(1,2))
qqnorm(male$hours_of_exercise_per_week)
qqline(male$hours_of_exercise_per_week)
qqnorm(female$hours_of_exercise_per_week)
qqline(female$hours_of_exercise_per_week)
```



Welch test

Hide

```
t.test(male$hours_of_exercise_per_week,female$hours_of_exercise_per_wee
k,alternative = "two.sided",correct=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  male$hours_of_exercise_per_week and female$hours_of_exercise_
per_week
## t = 1.7061, df = 66, p-value = 0.09269
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2050088  2.6134337
## sample estimates:
## mean of x mean of y
##  5.011905  3.807692
```

Since the equal variance does not hold in our case,but normality is still hold, therefore, we use Welch test which assumes normality but relaxes the assumption of equal variance.

Conclusion

In conclusion, we found the following results for our sample of DATA2X02 from the survey: 1: This is not an random sample. More active students have more probabilities to fill out the survey. 2: The data is likely to suffers from selection bias and measurement bias 3: The study time, exercise hours and height are likely to suffer from these bias 4: We con't find evidence that read meat eaters do exercise more than non-eaters 5: There is no evidence that people who join more club tend to do exercise different amount of time than those who join less club 6: There is no evidence that people who live with parents tend to study different amount of time than those who live without parents 7: Handedness is independent of wearing glasses 8: We do not found evidence that male spend less study time than female.

Reference

Data source:

https://docs.google.com/spreadsheets/d/1hNeBWmVXTLyUwl6b9yTnfPn2jfbQuBdPJuhJu-FZJSo/export?gid=690048681&format=csv (https://docs.google.com/spreadsheets/d/1hNeBWmVXTLyUwl6b9yTnfPn2jfbQuBdPJuhJu-FZJSo/export?gid=690048681&format=csv)

Package used:

readr: https://github.com/tidyverse/readr (https://github.com/tidyverse/readr)

Tidyverse: https://www.tidyverse.org (https://www.tidyverse.org)

Dplyr: https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html (https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html)

Janitor: https://cran.r-project.org/web/packages/janitor/index.html (https://cran.r-project.org/web/packages/janitor/index.html)

Ggplot2: https://ggplot2.tidyverse.org/ (https://ggplot2.tidyverse.org/)

Lubridate: https://www.rdocumentation.org/packages/lubridate/versions/1.7.4 (https://www.rdocumentation.org/packages/lubridate/versions/1.7.4)

kableExtra: https://cran.r-project.org/web/packages/kableExtra/index.html (https://cran.r-project.org/web/packages/kableExtra/index.html)

Forcats: https://cran.r-project.org/web/packages/forcats/index.html (https://cran.r-project.org/web/packages/forcats/index.html)

GenderRcode: https://github.com/ropenscilabs/gendercoder (https://github.com/ropenscilabs/gendercoder)