Report for homework#7

12032924 李熹成

Problem#1

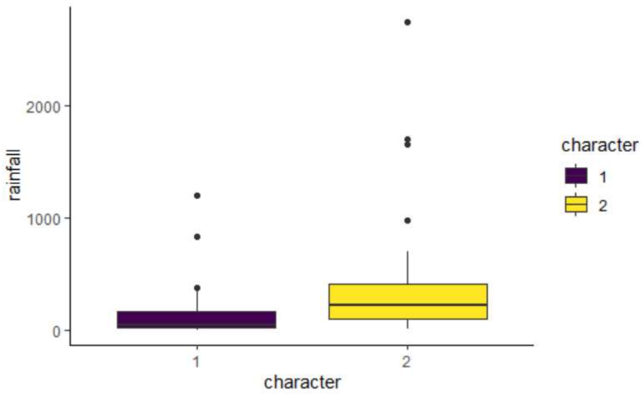
```r
library(tidyr)
library(dplyr)
library(ggplot2)

#1 Cloud Seeding
##1.1
seedday<-read.csv(file = 'ese5023hw3_1.csv',header = T)
class(seedday)

seedday_tbl<-as_tibble(seedday) %>%
  mutate(character=factor(character,ordered = T))

seedday_tbl %>%
  group_by(character) %>%
  summarise(
    count = n(),
    mean_rainfall = mean(rainfall, na.rm = TRUE),
    sd_rainfall = sd(rainfall, na.rm = TRUE)
  )

seedday_tbl%>%
  ggplot(aes(x=character,y=rainfall,fill=character))+
  geom_boxplot() +
  theme_classic()#Seems there is difference between two dataset.
```

```r
##1.2
#Normality
uns<-seedday_tbl %>%
  filter(character=='1') %>%
  pull(rainfall)
s<-seedday_tbl %>%
  filter(character=='2') %>%
  pull(rainfall)

shapiro.test(uns)#p-value = 3.134e-07
shapiro.test(s)#p-value = 1.411e-06
#The two datasets do not obey normal distribution

#Homogeneity of variance
bartlett.test(rainfall~character,data=seedday_tbl)#p-value = 6.754e-05
#Reject the hypothesis

anova_one_way<-aov(rainfall ~ character,data = seedday_tbl)
summary(anova_one_way)
#The Pr() is 0.0511, which can just reject the hypothesis and have
#little significance.But from the boxplot, it seems there are differences
#between the two sets.
```

The results are as below:

```
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 4
  character count mean_rainfall sd_rainfall
  <ord>     <int>         <dbl>       <dbl>
1 1            26          165.        278.
2 2            26          442.        651.
```

From the simple statistics, the mean and SD are distinct.



It seems to have differences from boxplot.

```
> shapiro.test(uns)#p-value = 3.134e-07

        Shapiro-Wilk normality test

data:  uns
W = 0.60219, p-value = 3.134e-07

> shapiro.test(s)#p-value = 1.411e-06

        Shapiro-Wilk normality test

data:  s
W = 0.65626, p-value = 1.411e-06

> #The two datasets do not obey normal distribution
>
> #Homogeneity of variance
> bartlett.test(rainfall~character,data=seedday_tbl)#p-value = 6.754e-05

        Bartlett test of homogeneity of variances

data:  rainfall by character
Bartlett's K-squared = 15.879, df = 1, p-value = 6.754e-05
```

After doing normality and heterogeneity test the data seems not conform to the condition to do ANOVA.

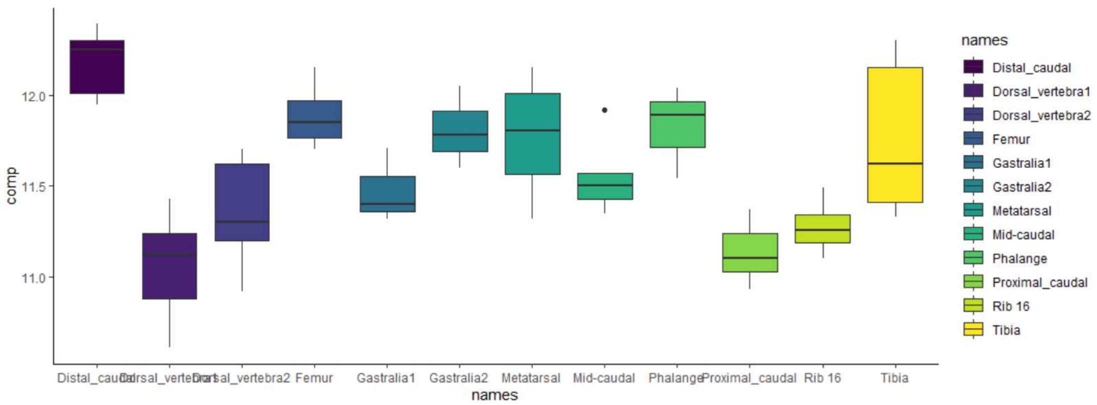
```
> anova_one_way<-aov(rainfall ~ character,data = seedday_tbl)
> summary(anova_one_way)
            Df   Sum Sq Mean Sq F value Pr(>F)
character    1  1000360 1000360   3.993 0.0511 .
Residuals   50 12525457  250509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can just reject the hypothesis, which is there is a difference between the variances while there is little significancy. However, from the simple statistics and boxplot, there seems to have differences. **This may be caused by the data itself is not fit the requirement to do ANOVA.**

Problem#2

```r
#2. Was Tyrannosaurus Rex Warm-Blooded?
bone<-read.csv(file = 'ese5023hw3_2.csv',header = T)

bone_tbl<-as_tibble(bone) %>%
  mutate(names=factor(names,ordered = T))
glimpse(bone_tbl)

bone_tbl%>%
  ggplot(aes(x=names,y=comp,fill=names))+
  geom_boxplot() +
  theme_classic()

anova_one_way2<-aov(comp ~ names,data = bone_tbl)
summary(anova_one_way2)#The Pr(>F)=9.73e-07 ***, which means the variances are
#significantly differences.
#Thus the conclusion is that Tyrannosaurus Rex is not warm-blooded.
TukeyHSD(anova_one_way2)
```

The results are as below:



It seems there are distinct differences among the different bone groups.

```
> anova_one_way2<-aov(comp ~ names,data = bone_tbl)
> summary(anova_one_way2)
            Df Sum Sq Mean Sq F value   Pr(>F)    
names       11  6.067  0.5516   7.427 9.73e-07 ***
Residuals   40  2.971  0.0743                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

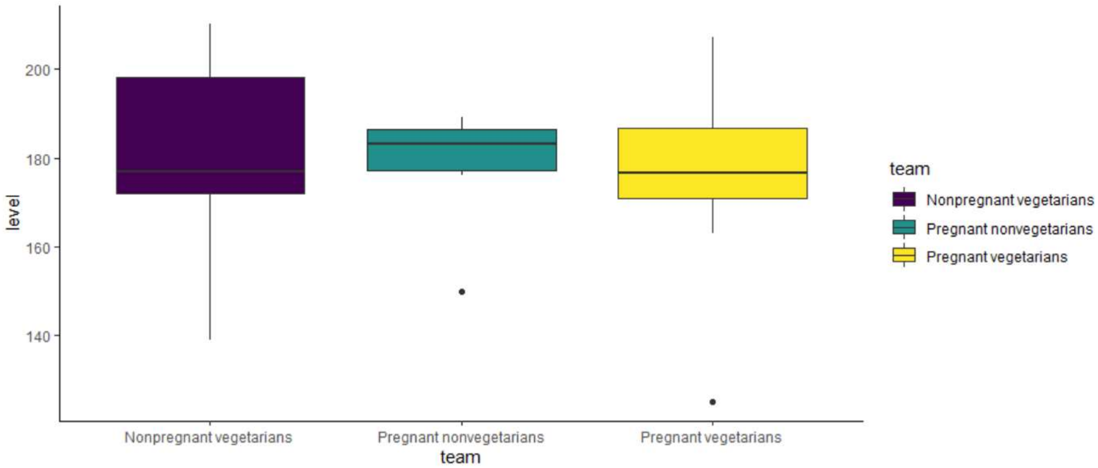
The P-value shows that the difference in variances are significant and thus the conclusion is that Tyrannosaurus Rex is not warm-blooded.

Problem#3

```r
#3. Vegetarians and Zinc
veg<-read.csv(file = 'ese5023hw3_3.csv',header = T)

veg_tbl<-as_tibble(veg) %>%
  mutate(team=factor(team,ordered = T))
glimpse(veg_tbl)

veg_tbl%>%
  ggplot(aes(x=team,y=level,fill=team))+
  geom_boxplot() +
  theme_classic()

veg_tbl %>%
  group_by(team) %>%
  summarise(
    count = n(),
    mean_level = mean(level, na.rm = TRUE),
    sd_level = sd(level, na.rm = TRUE)
  )#On simple mean aspect, The zinc level in Pregnant Vegetarians is
#slightly less than that in nonpregnant vegetarians.
```

```r
#Normality
name3<-unique(veg$team)
p_nv<-veg_tbl %>%
  filter(team==name3[1]) %>%
  pull(level)
p_v<-veg_tbl %>%
  filter(team==name3[2]) %>%
  pull(level)
np_v<-veg_tbl %>%
  filter(team==name3[3]) %>%
  pull(level)

shapiro.test(p_nv)#p-value = 0.03533 reject
shapiro.test(p_v) #p-value = 0.1418
shapiro.test(np_v)#p-value = 0.8142

#Homogeneity of variance
bartlett.test(level~team,data=veg_tbl)#p-value = 0.445

anova_one_way3<-aov(level~team,data=veg_tbl)
summary(anova_one_way3)#Pr(>F)=0.982, accept the hypothesis.
#There is no distinct differences in variance among the three different groups.
#Although the means have slight differences.
TukeyHSD(anova_one_way3)
```

The results are as below:



From the boxplot, it seems there is little difference among the three groups.

```
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 3 x 4
  team                     count mean_level sd_level
  <ord>                    <int>      <dbl>    <dbl>
1 Nonpregnant vegetarians      5       179.     27.3
2 Pregnant nonvegetarians      6       178      14.5
3 Pregnant vegetarians        12       177.     20.9
```

On simple mean aspect, the zinc level in Pregnant Vegetarians is slightly less than that in nonpregnant vegetarians.

```
> shapiro.test(p_nv)#p-value = 0.03533 reject

        Shapiro-Wilk normality test

data:  p_nv
W = 0.77596, p-value = 0.03533

> shapiro.test(p_v) #p-value = 0.1418

        Shapiro-Wilk normality test

data:  p_v
W = 0.89624, p-value = 0.1418

> shapiro.test(np_v)#p-value = 0.8142

        Shapiro-Wilk normality test

data:  np_v
W = 0.9609, p-value = 0.8142

>
> #Homogeneity of variance
> bartlett.test(level~team,data=veg_tbl)#p-value = 0.445

        Bartlett test of homogeneity of variances

data:  level by team
Bartlett's K-squared = 1.6192, df = 2, p-value = 0.445
```

The data can pass the normality and heterogeneity test.

```
> summary(anova_one_way3)
            Df Sum Sq Mean Sq F value Pr(>F)
team         2     16     8.1   0.018  0.982
Residuals   20   8816   440.8
>
```

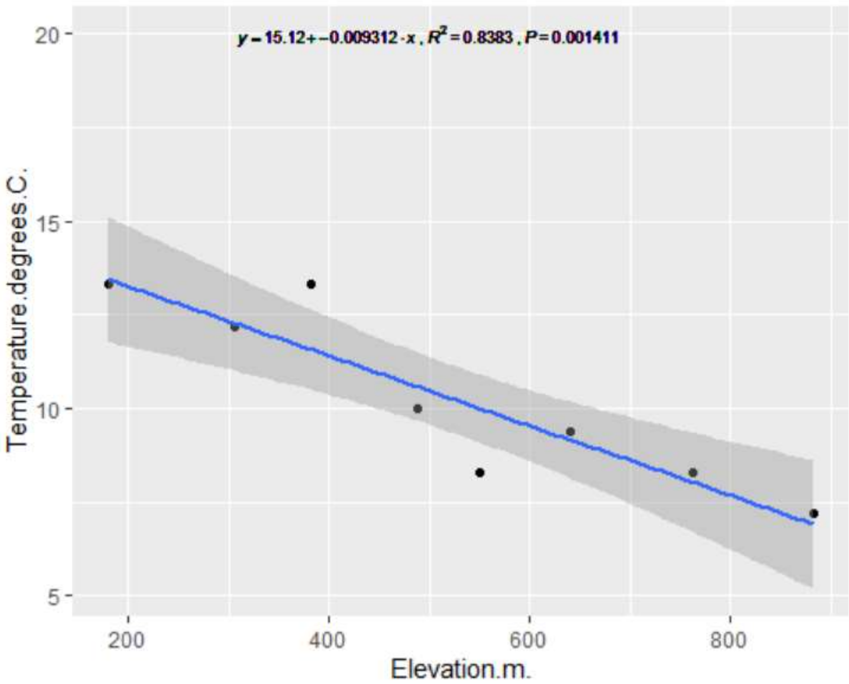
However, the P-value shows that, there is no distinct differences in variance among the three different groups. **Although the means have slight differences.**

```
> TukeyHSD(anova_one_way3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = level ~ team, data = veg_tbl)

$team
                                                     diff       lwr      upr     p adj
Pregnant nonvegetarians-Nonpregnant vegetarians -1.2000000 -33.36377 30.96377 0.9951012
Pregnant vegetarians-Nonpregnant vegetarians    -2.1166667 -30.39020 26.15686 0.9804367
Pregnant vegetarians-Pregnant nonvegetarians    -0.9166667 -27.47502 25.64169 0.9958057
```

Problem#4

```
#4. Atmospheric Lapse Rate
Huawei<-read.csv(file = 'ese5023hw3_4.csv',header = T)

linearr4 <- lm(Temperature.degrees.C.~Elevation.m., data = Huawei)
summary(linearr4)
coef(linearr4)

plot4<-ggplot(Huawei, aes(x=Elevation.m., y=Temperature.degrees.C.))+
  geom_point()+
  geom_smooth(method = "lm")

l<- list(a = as.numeric(format(coef(linearr4)[1], digits = 4)),
         b = as.numeric(format(coef(linearr4)[2], digits = 4)),
         r2 = format(summary(linearr4)$r.squared, digits = 4),
         p = format(summary(linearr4)$coefficients[2,4], digits = 4))
eq <- substitute(italic(y) == a + b %.% italic(x)~","~
                   italic(R)^2~"="~r2~","~italic(P)~"="~p, l)
#Methods of adding text on the plot
#refer to:https://blog.csdn.net/weixin_43948357/article/details/105336901
plot4 + geom_text(aes(x = 500, y = 20,
                  label = as.character(as.expression(eq))),
                  parse = TRUE,size = 2.5)
#The lapse rate here is 9.312 degrees C km-1.
```

The results are as below:

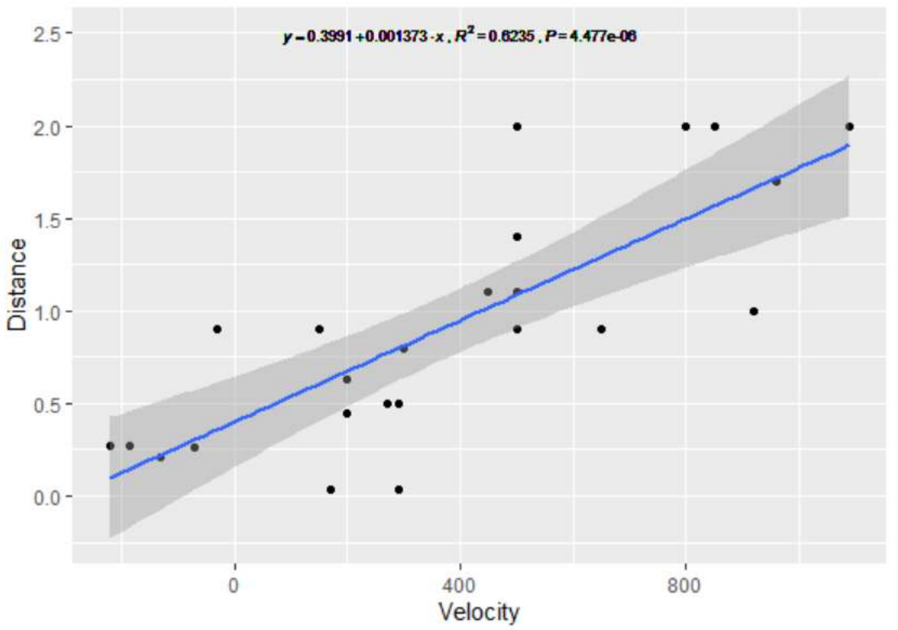


From the expression, the lapse rate here is **9.312 degrees C km-1.**

Problem#5

```r
#5. The Big Bang Theory
bbt<-read.csv(file = 'ese5023hw3_5.csv',header = T)
##5.1&5.2
plot5<-ggplot(bbt, aes(x=Velocity, y=Distance))+
  geom_point()+
  geom_smooth(method = "lm")
#It seems there is a trend in point distribution but more scattered.
linearr5 <- lm(Distance~Velocity, data = bbt)
summary(linearr5)
coef(linearr5)
l5<- list(a5 = as.numeric(format(coef(linearr5)[1], digits = 4)),
          b5 = as.numeric(format(coef(linearr5)[2], digits = 4)),
          r25 = format(summary(linearr5)$r.squared, digits = 4),
          p5 = format(summary(linearr5)$coefficients[2,4], digits = 4))
eq5 <- substitute(italic(y) == a5 + b5 %.% italic(x)~","~
                    italic(R)^2~"="~r25~","~italic(P)~"="~p5, l5)
plot5 + geom_text(aes(x = 400, y = 2.5,
                      label = as.character(as.expression(eq5))),
                  parse = TRUE,size = 2.5)
```

The results and calculating processes are as below:



$$Age = \frac{Distance}{Velocity} = \frac{1\times10^{6}\times 30.9\times10^{12}\,km}{1\,km/s} \times 1.373\times10^{-3} = 4.24257\times10^{15}\,s \Big/ 3600\times24\times365\,s/yr$$

$$= \frac{4.24257\times10^{15}}{3.1536\times10^{7}}\,yr$$

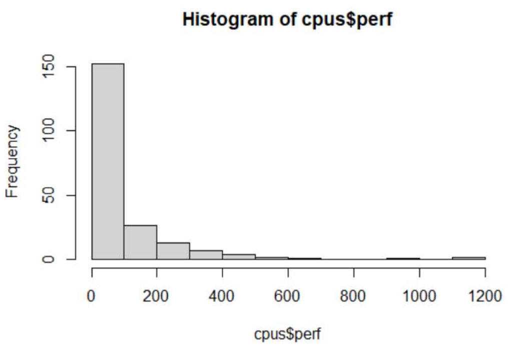
$$= 1.345\times10^{9}\,yr$$

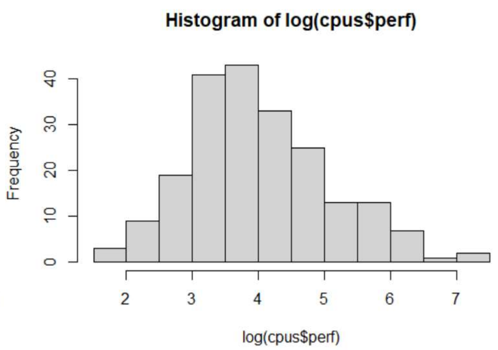
```
##5.3
###The explanation to the two assumptions:
###a.The intercept should be zero:
###If the BBT is correct, the fact is that the outer celestial body
###has a larger velocity compared to the inner celestial bodies derided from
###red shift. Thus, celestial bodies close to each other should have similar
###velocity. This leads to if the velocity(relative velocity) is 0,and thus
###the distance should be very close to 0.
###b.The slope is the age of the universe:
###The slope is Distance/Velocity and the unit should be time. If the Theory
###is correct, the universe continues to expand since singular point.
###The expand time is the age of the universe.

###The intercept is 0.3991 because the detection is not particularly accurate.
###From the handwriting calculation, the age of universe
###is only 1.345*10^(9) years.


##5.4
###As explained in 5.3,the improvement of distance measurement will probably
###enhance the R^2 as well as the fitting slope and intercept.
```

Problem#6

```r
#6. CPU Performance
library(MASS)
data(cpus)
##6.1
hist(test$perf)
# hist(log(test$perf))#The log data is better normally but I did not get
#the idea of whether should be logarithmetics because other parameters are
#all skewed.
# hist(test$syct)
# cpus<-cpus %>%
#   mutate(perf_log = log(perf))

sample_index <- sample(nrow(cpus),nrow(cpus)*0.80)
train <- cpus[sample_index,]
test  <- cpus[-sample_index,]
model6 <- lm(perf~syct+mmin+mmax+cach+chmin+chmax, data=train)
summary(model6)
##6.2
coef(model6)
perf_predict <- predict(model6,test)
plot(test$perf, perf_predict)
cor(test$perf, perf_predict)
mean(perf_predict)
mean(test$perf)
(mean(perf_predict) - mean(test$perf))/
  mean(test$perf)*100
```

The results are as below:

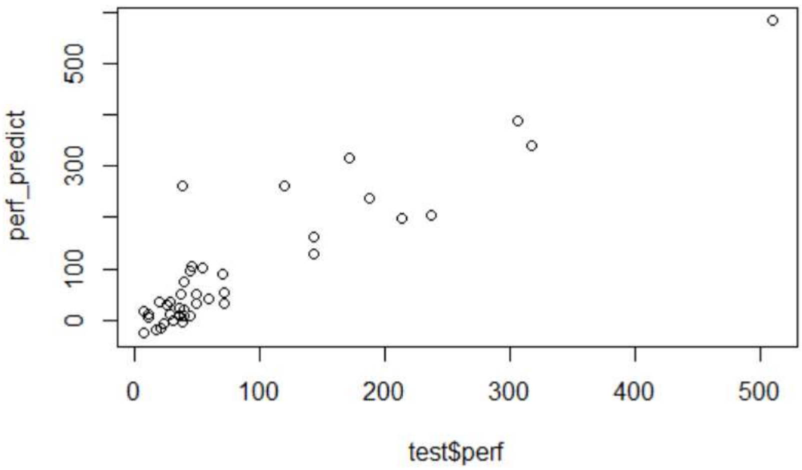




The log data is better normally but I did not get the idea of whether data should be logarithmetics because other parameters are all skewed.

```
Call:
lm(formula = perf ~ syct + mmin + mmax + cach + chmin + chmax,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-185.03  -25.42    3.60   27.25  338.59

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.076e+01  9.093e+00  -6.682 3.71e-10 ***
syct         5.390e-02  2.058e-02   2.619  0.00966 **
mmin         1.591e-02  2.052e-03   7.753 9.83e-13 ***
mmax         5.405e-03  7.177e-04   7.531 3.48e-12 ***
cach         7.033e-01  1.665e-01   4.224 4.01e-05 ***
chmin       -4.149e-01  1.082e+00  -0.384  0.70179
chmax        1.775e+00  2.514e-01   7.059 4.83e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.61 on 160 degrees of freedom
Multiple R-squared:  0.8767,    Adjusted R-squared:  0.8721
F-statistic: 189.6 on 6 and 160 DF,  p-value: < 2.2e-16
```

```
> coef(model6)
 (Intercept)         syct         mmin         mmax         cach        chmin
-60.75818894   0.05390156   0.01591050   0.00540480   0.70332906  -0.41493436
       chmax
  1.77451441
> perf_predict <- predict(model6,test)
> plot(test$perf, perf_predict)
> cor(test$perf, perf_predict)
[1] 0.913668
> mean(perf_predict)
[1] 95.55509
> mean(test$perf)
[1] 84.64286
> (mean(perf_predict) - mean(test$perf))/
+   mean(test$perf)*100
[1] 12.89209
```



The mean bias is 12.89209.

Problem#7

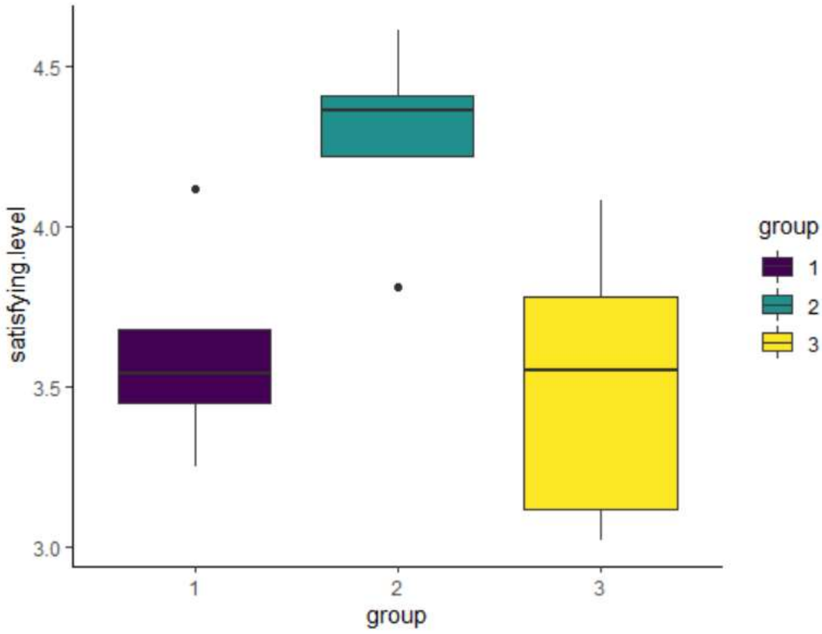
```r
#7. Analysis of Data Sets from Your Group
satisfying<-read.csv(file = 'ese5023hw3_7.csv',header = T)
satisfying_tbl<-as_tibble(satisfying) %>%
  mutate(group=factor(group,ordered = T))
glimpse(satisfying_tbl)

satisfying_tbl%>%
  ggplot(aes(x=group,y=satisfying.level,fill=group))+
  geom_boxplot() +
  theme_classic()
```

```r
##7.1
###Question:Whether the satisfaction are difference between service and delivery.
#Normality
name7<-unique(satisfying$group)
service<-satisfying_tbl %>%
  filter(group==name7[1]) %>%
  pull(satisfying.level)
deliver<-satisfying_tbl %>%
  filter(group==name7[2]) %>%
  pull(satisfying.level)
price<-satisfying_tbl %>%
  filter(group==name7[3]) %>%
  pull(satisfying.level)

shapiro.test(service)#p-value = 0.7004
shapiro.test(deliver) #p-value = 0.6326
shapiro.test(price)#p-value = 0.6845
#All be accepted.

#Homogeneity of variance
bartlett.test(satisfying.level~group,data=satisfying_tbl)#p-value = 0.7182
#The data are homogeneity

#Do t_test to service satisfaction and deliver satisfaction.
t.test(service, deliver)
#Showing that for different types, the differences of variance are significant.
```

```r
##7.2
###Question:Do the three satisfactions have significant differences?
#We have exam the normality and the Homogeneity
anova_one_way7<-aov(satisfying.level~group,data=satisfying_tbl)
summary(anova_one_way7)#Pr(>F)=0.0109, reject the hypothesis.
#The three satisfactions have significant differences among them.

##7.3
###Question:Is there some relationship between delivery satisfaction and
###service satisfaction?
data7<-data.frame(service,deliver)
plot7<-ggplot(data7,aes(x=deliver, y=service))+
  geom_point()+
  geom_smooth(method = "lm")
#It seems there is a trend in point distribution.
linearr7 <- lm(service~deliver, data = data7)
summary(linearr7)
coef(linearr7)
l7<- list(a7 = as.numeric(format(coef(linearr7)[1], digits = 4)),
          b7 = as.numeric(format(coef(linearr7)[2], digits = 4)),
          r27 = format(summary(linearr7)$r.squared, digits = 4),
          p7 = format(summary(linearr7)$coefficients[2,4], digits = 4))
eq7 <- substitute(italic(y) == a7 + b7 %.% italic(x)~","~
                    italic(R)^2~"="~r27~","~italic(P)~"="~p7, l7)
plot7 + geom_text(aes(x = 3.6, y = 5,
                      label = as.character(as.expression(eq7))),
                  parse = TRUE,size = 2.5)
##From the graph, the R_square is 0.4549.
##There is limited relationship between deliver satisfaction
##and service satisfaction.
```

The results are as below:

I choose a dataset which is about the satisfaction rate to service(1), delivery(2), and price(3).



From the boxplot the difference seems distinct.

```
> shapiro.test(service)#p-value = 0.7004

        Shapiro-Wilk normality test

data:  service
W = 0.94485, p-value = 0.7004

> shapiro.test(deliver) #p-value = 0.6326

        Shapiro-Wilk normality test

data:  deliver
W = 0.93526, p-value = 0.6326

> shapiro.test(price)#p-value = 0.6845

        Shapiro-Wilk normality test

data:  price
W = 0.94262, p-value = 0.6845

> #All be accepted.
>
> #Homogeneity of variance
> bartlett.test(satisfying.level~group,data=satisfying_tbl)

        Bartlett test of homogeneity of variances

data:  satisfying.level by group
Bartlett's K-squared = 0.66203, df = 2, p-value = 0.7182
```

The data **pass** the normality and heterogeneity test.

7.1 Question: Whether the satisfactions are difference between service and delivery?

Do t-test to service satisfaction and deliver satisfaction.

```
> t.test(service, deliver)

        Welch Two Sample t-test

data:  service and deliver
t = -3.4091, df = 7.9392, p-value = 0.009342
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.130521 -0.217479
sample estimates:
mean of x mean of y
    3.608     4.282
```

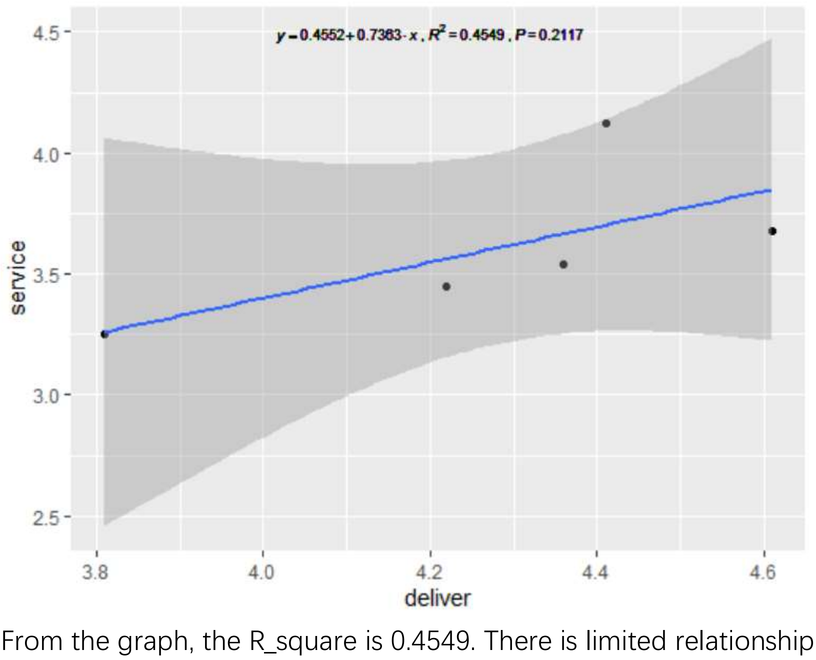
The results show that for different types, the differences of variance are significant.

7.2 Question: Do the three satisfactions have significant differences?

```
> anova_one_way7<-aov(satisfying.level~group,data=satisfying_tbl)
> summary(anova_one_way7)#Pr
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  1.766  0.8832   6.736 0.0109 *
Residuals   12  1.573  0.1311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The three satisfactions have significant differences among them.
Pr(>F)=0.0109, reject the hypothesis.

7.3 Question: Is there some relationship between delivery satisfaction and service satisfaction?



From the graph, the R_square is 0.4549. There is limited relationship between deliver satisfaction and service satisfaction.

```
> summary(linearr7)

Call:
lm(formula = service ~ deliver, data = data7)

Residuals:
       1        2        3        4        5 
-0.11235  0.41775 -0.01047 -0.12543 -0.16950 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4552     1.9965   0.228    0.834
deliver       0.7363     0.4653   1.582    0.212

Residual standard error: 0.2779 on 3 degrees of freedom
Multiple R-squared:  0.4549,	Adjusted R-squared:  0.2732 
F-statistic: 2.503 on 1 and 3 DF,  p-value: 0.2117

> coef(linearr7)
(Intercept)     deliver 
  0.4552052   0.7362902 
```