Report

MT5763 assignment1 Student ID:190019859

1 Introduction

1.1 Simple Introduction

A forensic scientist selected samples from four soils and analysed the the level of elements in the cannabis leaves. He wanted to find whether cannabis leaves grown in the same soil had the same elements. Additionally, there are some researchers who try to track people who grow cannabis by identifying elements in its leaves, and help to track people who grow and sell cannabis illegally.

1.2 The purpose of the assignment

The data used in this experiment is from the plants grown in standard store-bought potting mix, and those that are grown in general outdoor soil. And the main purposes of this assignment are as below:

- Identify the differences of elements of cannabis leaves among 4 different types of soils.
- Identify if there is a correlation in the mean level of elements in cannabis leaves.

1.3 Exploratory analysis of data

During the process that is to familiarise myself with data, descriptive analysis is performed for each element, finally, 5 elements are selected to be used in the following analysis. According to the researches written by Shibuya, E.K. (2007)and Kuras, M.J. (2011), I chose the elements showed in their study, which are the parts of the main elements in cannabis leaves. As a result, Mg, AL, K, Y, La are selected to do exploratory data analysis. In addition, the other reason I choose them is that their content in cannabis leaves represents the range from the main elements to microelement to some extent.

(Because the dimensionality of each data was not known, the data were standardized here.)

	Mg	Al	K	Y	La
Min.	0.0001449	0.2398516	0.0039036	0.1652518	0.0359252
1st Qu.	0.6883466	0.6972429	0.7348038	0.6459844	0.7819004
Median.	0.9600623	0.8088018	0.8955421	0.9163965	0.9932249
Mean.	0.8619044	0.9092048	0.8677803	0.9062024	0.8811663
3rd Qu.	1.0959202	1.0458644	1.2514627	1.2656788	1.0830378
Max.	2.0288109	2.3148466	1.6533085	1.6525183	2.3245688

Observe the data distribution of the above five elements in four different soils through data visualization.
The result is showed in figures below:

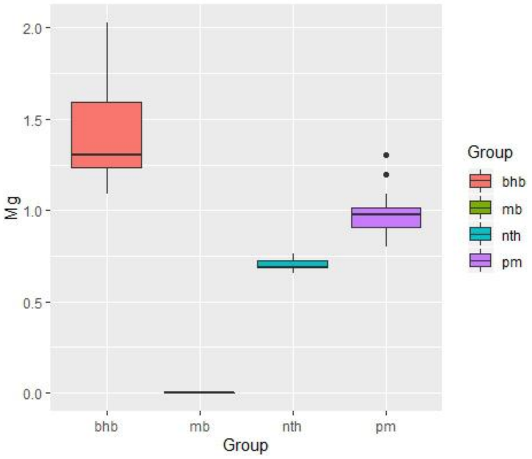


Figure 1

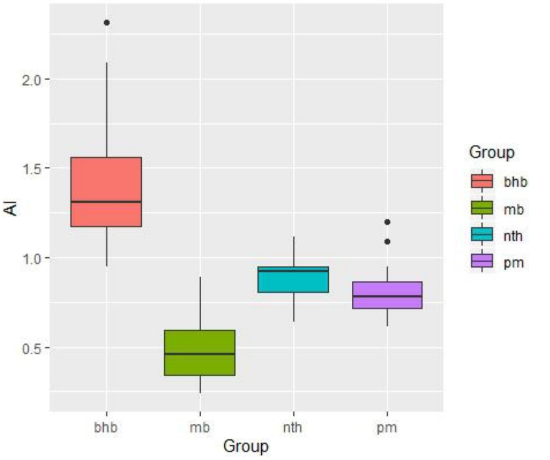


Figure 2

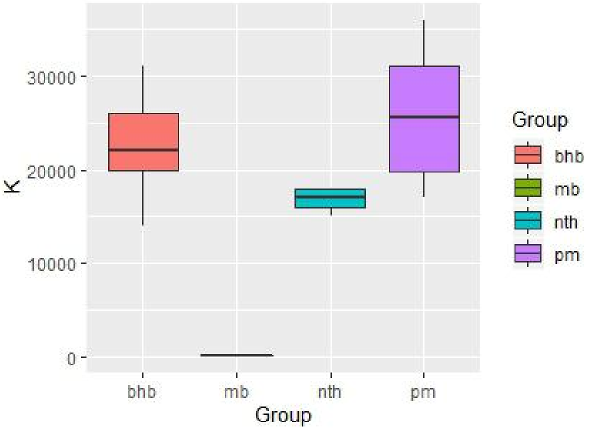


Figure 3

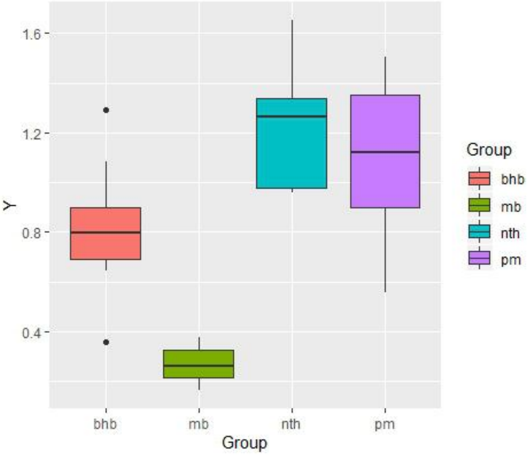


Figure 4

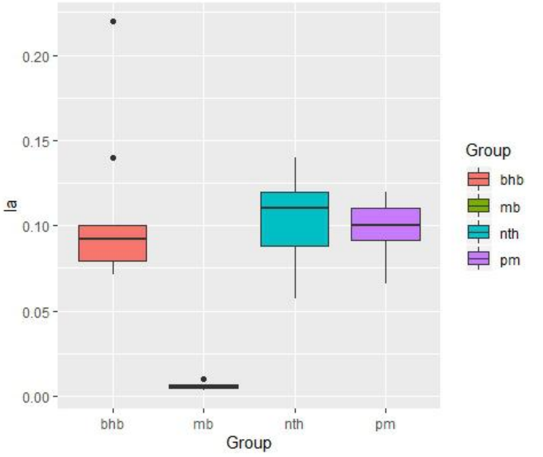


Figure 5

As the purpose of this study is whether the research data shows that in different soil types the elemental composition of Cannabis leaves are different, so when I do hypothesis test I put hypothesis "the data shows that the elemental composition of cannabis leaves in different soils does not have differences" in the null hypothesis , then the alternative hypothesis is "in different soil types cannabis leaves contain different elements". During the whole process, I use the average level to represent each element content.

It can be seen from Figure 1 and Figure 3 that the Mg content in cannabis leaves is very close in Northland (bth) and Potting mix (pm) and the content of K is very close in Blockhouse Bay(bhb) and Potting mix (pm). The specific values of the grouping descriptive statistics by group in table below are also reflected, so the t-test of the two population will be conducted in the following research.

	bhb	mb	nth	pm
Mg (mean)	39230.77	27.41	19444.44	26958.33
K (mean)	22615.38	114.80	16777.78	25500.00

According to Figure2, Figure4 and Firgue5, we can see that the contents of Al, Y and La in cannabis leaves are very similar in Blockhouse Bay(bhb), Northland (nth) and Potting mix (pm). The specific values can be seen in table below. Because we need to compare the mean values of these samples from three kinds of soils, ANOVA method will be used in later studies.

	bhb	mb	nth	pm
Al (mean)	51.92	17.66	31.78	28.67
Y (mean)	0.05	0.02	0.08	0.07
La (mean)	0.10	0.01	0.10	0.10

3 Hypothesis testing (Two samples)

3.1 The Content of Mg (t-test)

Hypothesis:

Because of the result in figure above, we want to know whether the content of Mg is different in two soils (nth and pm).

Here $\mu_1 \mu_2$ are used to represent the mean value of Mg contained in cannabis leaves grown in nth and pm respectively. With these population parameters, we conduct a t-test to with the following alternative hypothesis or null hypothesis as:

H_0： $\mu_1 - \mu_2 = 0$

H_1： $\mu_1 - \mu_2 \neq 0$

Result

The population mean hypothesis test results of Mg content in nth and pm soils are shown in the following table:

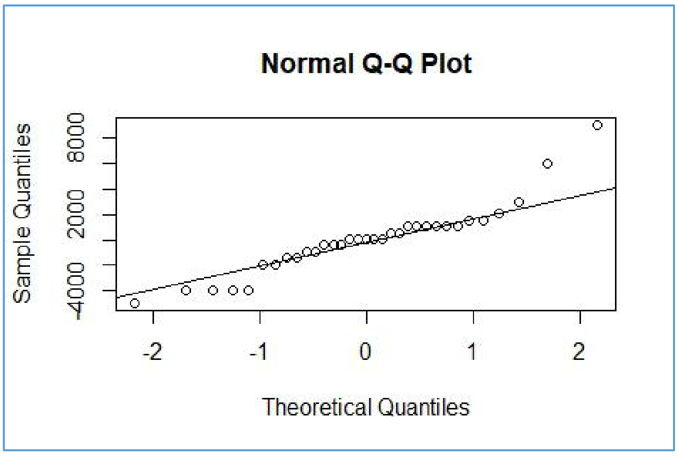
Two sample t-test						
t	df	p-value	95 percent condfidence interval		mean in nth	mean in pm
-9.9732	30.99	3.429e-11	-9050.496	-5977.281	19444.44	26958.33

According to this result, we can see that p-value is very small, so I may conclude there is a significant difference between the groups in terms of underlying mean content of Mg in cannabis leaves

Assumption Check (Normality, independence, Homogeneity)

- Normality

It could be seen from QQ-norm plot that there are data at both ends of the straight line deviating from the line. At the same time, p-value<0.05 in Shapiro's test, so we have evidence to suspect that noise is may not normal.



Shapiro-wilk normality test	
w	p-value
0.90173	0.005909

- Independence

Although in most cases we'll rely on sensible data collection and assume this is OK, but there may be some problems about the independence of noise, because the data are come from researchers who cultivate these cannabis.

- Homogeneity

I use F-test to compare the two variances to check homogeneity of noises. The result is shown in the table below, which fails to reject the H0(in this case, ratio of variances is equal to 1), with a p-value < 0.05.

F test to compare two variances				
F	p-value	95 percent confidence interval		estimate: ratio of variances
0.12508	0.004975	0.04454955	0.49515375	0.1250813

To conclude, t-test that has assumptions badly violated, so next part I will use a non-parametric equivalent.

3.2 The Content of K (t-test)

Hypothesis

According to the result in figure above, we want to know whether the content of K is different in two soils (bhb and pm). Here $\mu_1 \mu_2$ are used to represent the mean value of K contained in cannabis leaves grown in bhb and pm respectively. With these population parameters, we conduct a t-test to with the following alternative hypothesis or null hypothesis as:

H_0： $\mu_1 - \mu_2 = 0$

H_1： $\mu_1 - \mu_2 \neq 0$

Result

The population mean hypothesis test results of K content in bhb and pm soils are shown in the following table:

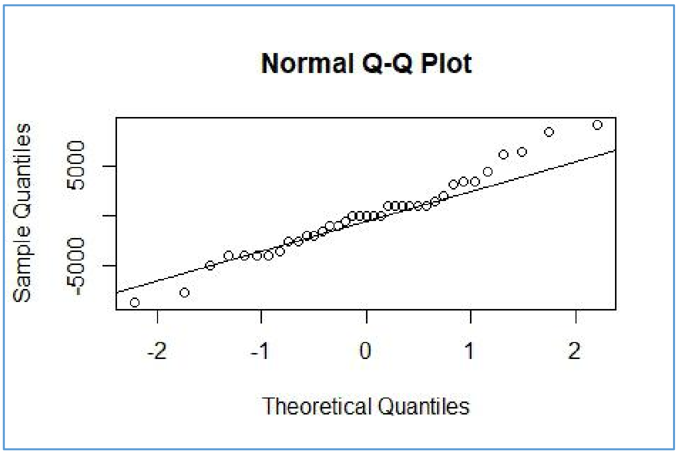
			Two sample t-test			
t	df	p-value	95 percent condfidence interval		mean in nth	mean in pm
-1.4846	29.923	0.1481	-6853.206	1083.975	22615.38	25500.00

According to this result, we can see that p-value>0.05, so we could not reject H0, which means there has no evidence to conclude there is a significant difference between the groups in terms of underlying mean content of K in cannabis leaves

Assumption Check (Normality, independence, Homogeneity)

- Normality

It could be seen from QQ-norm plot that there are data at both ends of the straight line deviating from the line. At the same time, p-value<0.05 in Shapiro's test, so we have evidence to suspect that noise is may not normal.



Shapiro-wilk normality test	
w	p-value
0.97561	0.5798

- Independence

Although in most cases we'll rely on sensible data collection and assume this is OK, but there may be some problems about the independence of noise, because the data are come from researchers who cultivate these cannabis.

- Homogeneity

I use F-test to compare the two variances to check homogeneity of noises. The result is shown in the table below, which could reject the H0, with a p-value < 0.05. So their variances are different.

F test to compare two variances				
F	p-value	95 percent confidence interval		estimate: ratio of variances
2.5865	0.04848	1.006461	7.838728	2.586547

From my perspective, except for the differences between variances, another reason is that we cannot guarantee that the sample set we used or the population data to be tested is normally distributed, so it is not reasonable for us to use t-test.

So next part I will use a non-parametric equivalent.

3.3 Non-parametric (Wilcoxon tests)

Hypothesis

The hypothesis for these two elements are same as:

H_0 : the content of Mg(K) in two types of soils share the same level

H_1 : the content of Mg(K) in two types of soils are different

The following two table show the results of non-parametric tests on these two elements respectively. According to p-value, it can be seen that the results of non-parametric tests are consistent with the conclusions of the t-test we did before. However, in fact, this result is not easy to be verified. For example, after test on Mg elements, we rejected the null hypothesis, but the confidence interval is negative, which is not very convincing.

Wilcoxon rank sum test with continuity correction					
	w	p-value	95 percent cofidence interval		estimate: difference in location
Mg	0	6.14e-06	-Inf	-6000	-7000
K	118	0.1159	-Inf	1000	-3000

4 One-way ANOVA (Multiple samples)

4.1 ANOVA

In the exploratory analysis of data, we have a simple insight of data through the data visualization. In box plot, we can see that the content of Al in mb, nth and pm is very similar, so I choose Al for ANOVA analysis in this part.

Hypothesis

Set the mean level of Al in cannabis grown in mb, nth and pm are μ_1, μ_2, μ_3 respectively.

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: In μ_1, μ_2, μ_3 , at least one mean is different from one of the others.

Result

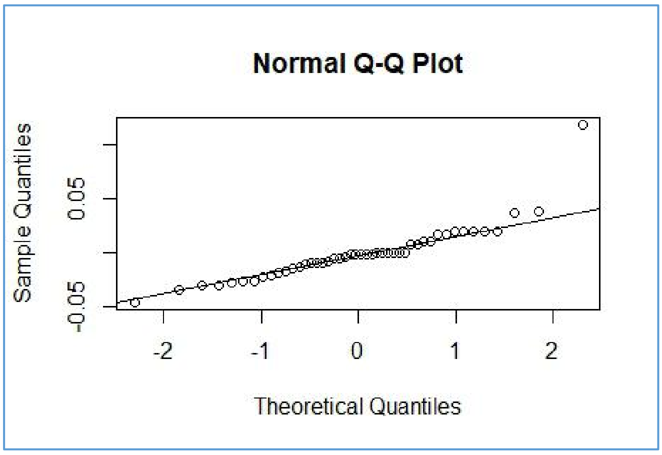
The table below shows that we have evidence to reject the H0, so there has obvious differences about the mean of Al in cannabis among three types soil.

	Df Sum	Mean	Sq	F value	Pr (>F)
Group	2	1142	571.0	18.36	2.21e-06
Residuals	40	1244	31.1		

Assumption Check (Normality, independence, Homogeneity)

- Normality

It could be seen from QQ-norm plot that there are data at both ends of the straight line deviating from the line. At the same time, p-value<0.05 in Shapiro's test, so we have evidence to suspect that noise is may not normal



Shapiro-wilk normality test	
w	p-value
0.94538	0.0405

- Independence

Although in most cases we'll rely on sensible data collection and assume this is OK, but there may be some problems about the independence of noise, because the data are come from researchers who cultivate these cannabis.

- Homogeneity

Because we have 3 populations so we can not use F-test, we can observe the SD of Al among three different soil. As the table shows below their SD are different.

Group	SD
mb	7.33
nth	5.14
pm	4.89

4.2 Non-parametric (Kruskal-Wallis test)

In this assignment I use Kruskal-Wallis test as an alternative to ANOVA. The table shows the same result of the test above.

Kruskal-Wallis	
Kruskal-Wallis chi-squared	p-value
16.792	0.0002257

The Kruskal-Wallis test told us that the content of elements in cannabis leaves grown in three different soils was different, but did not tell us which soil had different elements from other soils.In ANOVA, we often use Tukey HSD tests to perform pairwise comparisons between groups based on ANOVA results. So next part I will do Tukey test.

4.3 Post Hoc Tests (Tukey)

Tukey multiple comparisons of means 95% family-wise confidence level				
Group	diff	lwr	upr	p adj
nth-mb	14.117778	7.881031	20.354525	0.0000068
pm-mb	11.006667	5.897661	16.115673	0.0000160
pm-nth	-3.111111	-8.416689	2.194467	0.3368237

Summary:

1. The content of Al in cannabis grown in nth and mb has differences.
2. The content of Al in cannabis grown in pm and mb has differences.
3. There has no evidence show that the content of Al in cannabis grown in pm and nth has differences.

5 linear regression

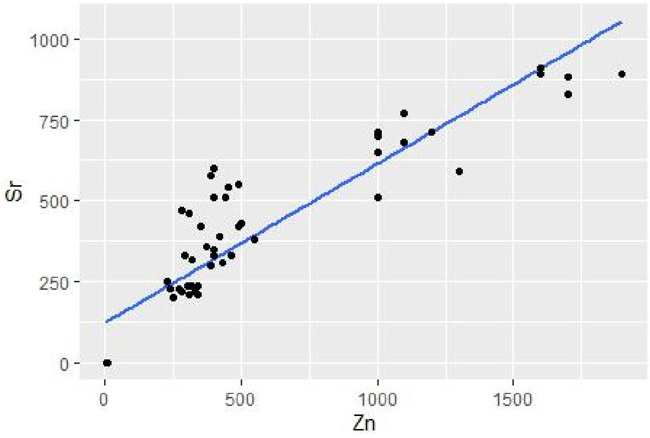
5.1 linear regression

The first pair: Zn ~ Sr

Residuals				
Min	1Q	Median	3Q	Max
-490.31	-99.61	25.92	116.71	524.20

Coefficients				
	estimate	std.error	t value	Pr(>|T|)
(Intercept)	-114.1604	49.6662	-2.299	0.0254*
Sr	1.6741	0.1072	15.613	<2e-16

Residual standard error	Multiple R--squared	Adjusted R-squared	F-statistic	P-value
211.7 on 54 degress of freedom	0.8186	0.8153	243.8 on 1 and 54 DF	2.2e-16

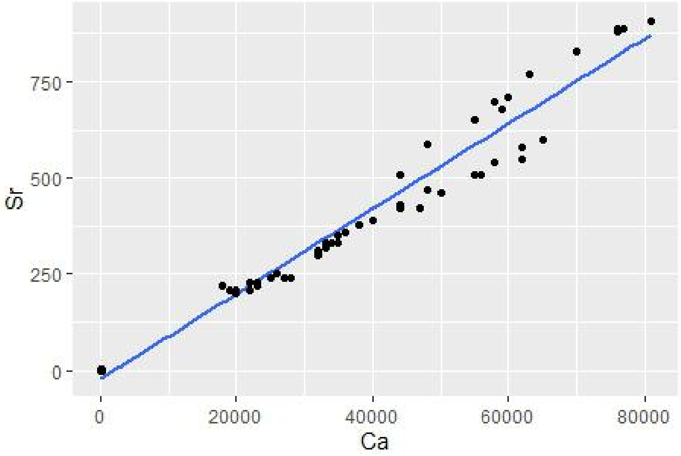


The first pair: Ca ~ Sr

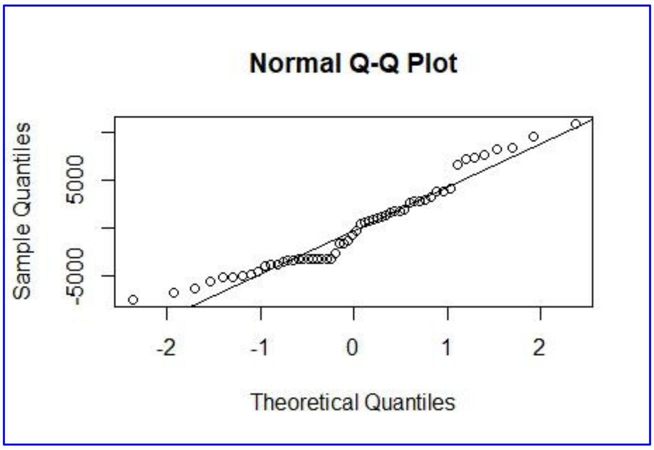
Residuals				
Min	1Q	Median	3Q	Max
-7354.8	-3273.3	-487.9	2779.3	10817.7

Coefficients				
	estimate	std.error	t value	Pr(>|T|)
(Intercept)	2151.116	49.6662	3.042	0.00362**
Sr	87.148	2.307	37.771	<2e-16***

Residual standard error	Multiple R--squared	Adjusted R-squared	F-statistic	P-value
4556 on 54 degress of freedom	0.9635	0.9629	1427 on 1 and 54 DF	2.2e-16



5.2 Assumption check



Shapiro-wilk normality test Ca~Sr	
w	p-value
0.94484	0.01257

From the QQ-norm plot and the result of shapiro test, we do not obtain a very valid result of the assumption check about the residuals.

6 Conclusion

Question1:

In this experiment, we analysis the differences of the content of three elements in various soils. The results of two hypothesis tests show that the content of K in cannabis leaves is not different between bhb and pm, while the content of Mg is different between nth and pm. As for Al, its content in the samples we selected was different.

From my perspective, the data indicates differences in the elemental composition of Cannabis leaves grown in different soil types. But in reality, this conclusion is not so convening, because there exists some elements share the same level in a few type of soils.

Question 2:

I choose two pairs with interesting relationships as:
The content of Zn and Sr are linear correlation.
The content of Ca and Sr are linear correlation.

Question 3:
However, I do not think the results of this experiment is able to decide which soils the cannabis are belong to, especially depend on the elemental composition of the leaves.

Firstly, as for the population, we want to draw a conclusion which can be apply to the world wherever the cannabis grow in. However, the data in this experiment just comes from 4 types of soils and the cannabis are just those cultivated by researcher. So the result of this experiment is not practical enough. Secondly, the sample size is small and the variances has huge differences. Thirdly, if cannabis is not cultivated by researcher, there exist more factors which can have huge impacts on their growth.

As for the solution, I think bigger sample size will make the result more ideal. Additionally, we should read more materials and find other impact factors so that the result can be more practical and reliable.

Reference：

Smith,N., Criminals May Rue Pot from This Plot, New Zealand Herald, 2000.

ShibuyaI, E.K., SarkisI, J.E.S., Negrini-NetoIII, N.O., Ometto and J.P.H.B., *Journal of the Brazilian Chemical Society*, Multivariate classification based on chemical and stable isotopic profiles in sourcing the origin of marijuana samples seized in Brazil, 18(1), 2007.

Kuras,M.J. and Marek Jan Wachowicz M.S.Eng., *Journal of Forensic Sciences, Cannabis Profiling Based on Its Elemental Composition—Is It Possible?, 56(5), 1250-1255, 2011.*

Method: Lecture slides

Software: Rstudio

Code

```
##MT5762 project 1
#-------------------------------------------------------------------------------------------



##-----------Part1 exploratory--------------------------------------------------------------
library(ggplot2)
library(tidyverse)
library(doBy)
library(psych)
library(knitr)

#read data
data <- read.csv("potplants_MT5762.csv")

#convert characters into numberic variences
data[,c(1,2)] <- lapply(data[,c(1,2)], as.numeric)
#data standardization
data$Mg <- scale(data$Mg,center=F,scale=T)
data$Al <- scale(data$Al,center=F,scale=T)
data$K <- scale(data$K,center=F,scale=T)
data$Y <- scale(data$Y,center=F,scale=T)
data$La <- scale(data$La,center=F,scale=T)

#do descriptive analysis and select 5 elements
elements <- summary(select(data,'Mg', 'Al','K','Y', 'La'))
kable(elements)
#-------------boxplot data exploratory-----------------------------------
ggplot(data,aes(x=Group, y=Mg)) + geom_boxplot(aes(Group, Mg, fill=Group))
ggplot(data,aes(x=Group, y=Al)) + geom_boxplot(aes(Group, Al, fill=Group))
ggplot(data,aes(x=Group, y=K)) + geom_boxplot(aes(Group, K, fill=Group))
ggplot(data,aes(x=Group, y=Y)) + geom_boxplot(aes(Group, Y, fill=Group))
ggplot(data,aes(x=Group, y=la)) + geom_boxplot(aes(Group, la, fill=Group))


#-------------descriptive analysis --------------------------------------------
mg_k <- c('Mg','K')
describeBy(data[mg_k], list(Group=data$Group))
al_y_la <- c('Al','Y','La')
describeBy(data[al_y_la], list(Group=data$Group))

###-----------------------------Mg--------------------------------------------
##-----------Part2 t test-----------------------------------------------------
mg_test<- data %>% filter(Group == 'nth' | Group == 'pm')
data %>% group_by(Group) %>%
   summarise(mean = mean(Mg), SD = sd(Mg), n = n())
t <- t.test(Mg ~ Group, data = mg_test)
##----------check assumption---------------------------------------------
#QQ-norm plot and shapiro.test
nth <- data %>% filter(Group == 'nth') %>% mutate(noise = Mg - mean(Mg))
pm <- data %>% filter(Group == 'pm') %>% mutate(noise = Mg - mean(Mg))
allNoise <- c(nth$noise, pm$noise)
qqnorm(allNoise)
qqline(allNoise)
```

```
shapiro.test(allNoise)
#Homogeneity
data %>% select(Mg, Group) %>% group_by(Group) %>% summarise(sd = sd(Mg))

###------------------------------K-----------------------------------------------
##------------t test-------------------------------------------------------------
k_test<- data %>% filter(Group == 'bhb' | Group == 'pm')
data %>% group_by(Group) %>%
   summarise(mean = mean(K), SD = sd(K), n = n())
t <- t.test(K ~ Group, data = k_test)
##-----------check assumption--------------------------------------------------
#QQ-norm plot and shapiro.test
bhb <- data %>% filter(Group == 'bhb') %>% mutate(noise = K - mean(K))
pm <- data %>% filter(Group == 'pm') %>% mutate(noise = Mg - mean(Mg))
allNoise <- c(bhb$noise, pm$noise)
qqnorm(allNoise)
qqline(allNoise)
shapiro.test(allNoise)
#Homogeneity
data %>% select(K, Group) %>% group_by(Group) %>% summarise(sd = sd(K))
var.test(bhb$noise, pm$noise)

#-----------Non-parametric Wilcoxon tests-----------------------------------------------
wilcox.test(Mg ~ Group, data = mg_test, alternative = 'l', conf.int = TRUE)
wilcox.test(K ~ Group, data = k_test, alternative = 'l', conf.int = TRUE)

##------------Part3 ANOVA----------------------------------------------------------------
#ANOVA analysis
al_test <- data %>% filter(Group == 'mb'|Group == 'nth'|Group == 'pm')
al_ANOVA <- aov(Al ~ Group, data = al_test)
summary(al_ANOVA)
##-----------check assumption--------------------------------------------------
#QQ-norm plot and shapiro.test
qqnorm(al_ANOVA$residuals)
qqline(al_ANOVA$residuals)
shapiro.test(al_ANOVA$residuals)
al_test %>% group_by(Group) %>% summarise(SD = sd(Al))
#-----------Non-parametric ------------------------------------------------------
al_non_test <- data %>% filter(Group == 'mb'|Group == 'nth'|Group == 'pm')
kruskal.test(Al~Group, al_non_test)

#-----------Post Hoc Tests Turkey------------------------------------------------------
turkey <- TukeyHSD(al_ANOVA)
turkey
tktable <- map_df(list(tk), tidy)
kable(tktable, digits = c(2,30), caption = "Tukey Honest Significant Differences")

#-----------linear regression------------------------------------------------------
#Ca~Zn
CaZn<-lm(Ca~Zn,data=data)
summary(CaZn)
ggplot(data=data,aes(x = Ca,y = Zn))+geom_smooth(se=FALSE,method='lm')+geom_point()

#Zn~Sr
znsr<-lm(Zn~Sr,data=data)
summary(znsr)
ggplot(data=data,aes(x = Zn,y = Sr))+geom_smooth(se=FALSE,method='lm')+geom_point()
```

```
#Ca~Sr
CaSr<-lm(Ca~Sr,data=data)
summary(CaSr)
ggplot(data=data,aes(x = Ca,y = Sr))+geom_smooth(se=FALSE,method='lm')+geom_point()

#------------linear regression test-------------------------------------
Ca_Sr <- lm(Ca ~ Sr, data, x = TRUE, y = TRUE)
shapiro.test(Ca_Sr$residuals)
qqnorm(Ca_Sr$residuals)
qqline(Ca_Sr$residuals)
```