Assignment 1 Report of MT5763

Student ID: 180025784

Contents

1 Introduction

2 Data

3 Analysis & Results

3 1 Read and clean the data

3 2 Exploratory analysis

3 3 Design plots to address the question

3 3 1 K_Sc

3 3 2 K_Fe

3 3 3 K_Mg

3 3 4 Sc_Fe

3 3 5 Sc_Mg

4 Conclusions

5 References

6 Appendix

6 1 Scatter plot of Fe and Mg in R

6 2 Scatter plot of K and Sc in SAS

6 3 Scatter plot of K and Fe in SAS

6 4 Scatter plot of K and Mg in SAS

6 5 Scatter plot of Sc and Fe in SAS

6 6 Scatter plot of Sc and Mg in SAS

6 7 Scatter plot of Fe and Mg in SAS

7 Programming Code

7 1 SAS Code

7 2 R Code

1 Introduction

Dion Sheppard, a scientist, studied the chemical composition of Cannabis leaves grown in different soil types. He believes that soil types can cause different chemical components in plants. In particular, he felt that there was a difference between the plants of the potted mixture purchased from the standard store and the plants in the common soil. It is considered that if the investigation is successful, it will help to prosecute the production and distribution of the drug. Therefore, in order to help study cannabis leaves of different soil types, this study will demonstrate this theory by analysing the chemical composition of cannabis leaves from different locations. The primary aim of this report is on whether the chemical composition of cannabis cultivation is the same in different soil types.

2 Data

The study used the data provided by University of St. Andrews. The variables of the data contained:

- **Sample Name**: The number of the sample.
- **Group**: Also means four soil types for growing cannabis leaves. There are four different locations, including pm (potting mix) and three other types (bhb, mb and nth) in New Zealand.
- **Elements**: Data of 38 chemistry elements from Mg to Th.

The sample name will not be the main point to be focuses in this study due to the fact that the study will do more research on the relationship between variable Group and Elements.

R Studio [5.0] and SAS [9.4] are the main softwares to be used to in this analysis. The code of both softwares used to analyse the data will be shown in Appendix. The target audience is not statistical. Thus, there will be more explanation on how to conduct the study.

3 Analysis & Results

3 1 Read and clean the data

The collected data are from three different sample set. In order to decrease discrepancy and for easy comparison. In this section, R was used to combine and clean the data. The primary step

is to clean and process the data. First, the sample sheet should be combined together for better exploration. However, while doing the combination, it can be easily found that the the group name pm in sample set two is of full name, potting mix. Also, sample set two already concluded the mean and sd of each element within these samples. In addition, there is a whole null row in sample set three. Before the combination, correction has to be done. After the combination of data from three sheets, cleaning of null data should also be done. Finally, summary of data shows that the variable of Group are character with length of 163 and there are 38 elements overall as below in Figure 1.

Group	Th	Al	K	Ca	Sc	Ti	Mn	Fe
Length:163	Min. :0.00000	Min. : 8.60	Min. : 85	Min. : 28	Min. :0.1900	Min. : 0.230	Min. : 0.36	Min. : 35.0
Class :character	1st Qu.:0.01272	1st Qu.: 26.24	1st Qu.:16867	1st Qu.:24338	1st Qu.:0.6652	1st Qu.: 5.221	1st Qu.: 273.15	1st Qu.:395.0
Mode :character	Median :0.02446	Median : 32.22	Median :22000	Median :39252	Median :0.7800	Median : 7.600	Median : 379.97	Median :474.7
NA	Mean :0.02833	Mean : 36.85	Mean :21234	Mean :40139	Mean :0.7005	Mean : 7.058	Mean : 478.60	Mean :431.5
NA	3rd Qu.:0.03894	3rd Qu.: 42.17	3rd Qu.:30115	3rd Qu.:60000	3rd Qu.:0.8442	3rd Qu.: 9.685	3rd Qu.: 606.08	3rd Qu.:567.2
NA	Max. :0.11262	Max. :104.07	Max. :49463	Max. :89422	Max. :1.1289	Max. :15.367	Max. :1693.54	Max. :741.5
Group	Ni	Cu	Zn	Ga	Ge	Se	Br	Rb
Length:163	Min. : 0.530	Min. : 1.00	Min. : 2.6	Min. : 0.040	Min. :0.0580	Min. : 1.400	Min. : 6.40	Min. : 0.040
Class :character	1st Qu.: 1.900	1st Qu.: 37.00	1st Qu.: 300.8	1st Qu.: 1.426	1st Qu.:0.3587	1st Qu.: 4.720	1st Qu.:21.99	1st Qu.: 9.474
Mode :character	Median : 2.712	Median : 50.49	Median : 400.0	Median : 2.300	Median :0.5489	Median : 6.700	Median :30.90	Median : 37.876
NA	Mean : 2.645	Mean : 58.97	Mean : 568.6	Mean : 4.335	Mean :0.5436	Mean : 6.724	Mean :31.58	Mean : 56.497
NA	3rd Qu.: 3.351	3rd Qu.: 73.37	3rd Qu.: 596.5	3rd Qu.: 3.756	3rd Qu.:0.7388	3rd Qu.: 8.344	3rd Qu.:38.87	3rd Qu.:103.812
NA	Max. :11.079	Max. :283.57	Max. :2493.0	Max. :18.466	Max. :1.6071	Max. :16.408	Max. :81.43	Max. :159.064
Group	Sr	Y	Mo	Pd	I	Ba	La	Ce
Length:163	Min. : 0.1	Min. :0.01100	Min. : 0.0720	Min. :0.00000	Min. : 0.42	Min. : 0.37	Min. :0.00340	Min. :0.0230
Class :character	1st Qu.: 232.6	1st Qu.:0.04954	1st Qu.: 0.6985	1st Qu.:0.08287	1st Qu.: 7.20	1st Qu.: 60.50	1st Qu.:0.07928	1st Qu.:0.1000
Mode :character	Median : 390.0	Median :0.07019	Median : 2.2051	Median :0.14000	Median :10.77	Median : 94.00	Median :0.10267	Median :0.1311
NA	Mean : 415.8	Mean :0.06789	Mean : 4.0155	Mean :0.14853	Mean :10.37	Mean :174.22	Mean :0.09379	Mean :0.1344
NA	3rd Qu.: 620.3	3rd Qu.:0.09168	3rd Qu.: 6.7461	3rd Qu.:0.21872	3rd Qu.:14.68	3rd Qu.:142.86	3rd Qu.:0.12000	3rd Qu.:0.1806
NA	Max. :1104.9	Max. :0.13206	Max. :13.6328	Max. :0.37479	Max. :26.84	Max. :687.77	Max. :0.22602	Max. :0.3843
Group	Pr	Nd	Sm	Eu	Gd	Tb	Dy	
Length:163	Min. :0.00100	Min. :0.00280	Min. :0.00480	Min. :0.00120	Min. :0.00450	Min. :0.000700	Min. :0.00340	
Class :character	1st Qu.:0.01400	1st Qu.:0.04875	1st Qu.:0.03200	1st Qu.:0.01500	1st Qu.:0.03237	1st Qu.:0.004670	1st Qu.:0.02119	
Mode :character	Median :0.01846	Median :0.07198	Median :0.04800	Median :0.02213	Median :0.04654	Median :0.006830	Median :0.03229	
NA	Mean :0.01672	Mean :0.06642	Mean :0.04471	Mean :0.02516	Mean :0.04495	Mean :0.006545	Mean :0.03179	
NA	3rd Qu.:0.02139	3rd Qu.:0.08859	3rd Qu.:0.05900	3rd Qu.:0.03320	3rd Qu.:0.05979	3rd Qu.:0.008345	3rd Qu.:0.04320	
NA	Max. :0.04022	Max. :0.12868	Max. :0.09278	Max. :0.07905	Max. :0.10677	Max. :0.013435	Max. :0.06750	
Group	Ho	Er	Tm	Yb	Lu	Ta	Mg	
Length:163	Min. :0.001000	Min. :0.00300	Min. :0.000900	Min. :0.00540	Min. :0.001000	Min. :0.000000	Min. : 4	
Class :character	1st Qu.:0.006707	1st Qu.:0.01603	1st Qu.:0.005000	1st Qu.:0.02300	1st Qu.:0.005945	1st Qu.:0.003261	1st Qu.:19496	
Mode :character	Median :0.009056	Median :0.02300	Median :0.007392	Median :0.03500	Median :0.008500	Median :0.007700	Median :27378	
NA	Mean :0.008416	Mean :0.02234	Mean :0.006997	Mean :0.03509	Mean :0.008105	Mean :0.010071	Mean :25613	
NA	3rd Qu.:0.010891	3rd Qu.:0.02900	3rd Qu.:0.009349	3rd Qu.:0.04742	3rd Qu.:0.010241	3rd Qu.:0.012830	3rd Qu.:33787	
NA	Max. :0.018005	Max. :0.05308	Max. :0.015288	Max. :0.09706	Max. :0.020134	Max. :0.072162	Max. :61945	

Figure 1 Summary of data

Also, compared with the summary of client, the summary of all elements seem to be more explicit. In the summary of client work, there are only five elements in each sheet, which makes the analysis become complex. However, the summary of data from all sheets can clearly show the results. It is hard to find out where is wrong in the client work due to the lack of data of all the elements. In the summary, min and max means the minimum and maximum number of data whilst 1st Quantile and 3rd Quantile means the data on the 25% and 75% of data from the

smallest to the largest. Additionally, mean represents the average of the data while median is the data lies in the middle. Plus, sd stands for standard deviation, which means the measure to quantify the dispersion of a data set (Kobayashi and Salam, 2000). The formula of sd is $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$.

		Th	Al	K	Ca	Sc
S1	mean	0.033687	39.0458	22226.65	41648.3	0.727991
	sd	0.021748	17.93301	12615.96	26490.59	0.233017
	min	0.001581	13.95571	91.43711	48.61847	0.209773
	max	0.112624	104.0717	45444.83	88380.13	1.109849
	n	56	56	56	56	56
		Mg	Al	K	Ca	Sc
S2	mean	23428.35	32.3	18637.46	35651.29	0.649231
	sd	13871.87	14.71686	10728.08	23887.95	0.214035
	min	4	8.6	85	28	0.19
	max	56000	83	36000	81000	1
	n	52	52	52	52	52
		Mg	Al	K	Ca	Sc
S3	mean	26672.4	38.92607	22677.23	42845.37	0.720984
	sd	16010.29	16.71595	13451.96	26864.62	0.234278
	min	14.84923	18.39458	91.15588	42.74571	0.200489
	max	61944.99	87.36592	49463.34	89422.25	1.128851
	n	55	55	55	55	55

Figure 2 Summary of data from client's work

3 2 Exploratory analysis

In order to choose suitable element for the research, do more exploratory analysis of data will be conducted in this step. On the basis of summary of all the data by the 'Group', mean and standard deviation (sd) of elements from Th to Mg. Then calculate the coefficient of variance (CV), which is the ratio of sd to mean. CV is a measure of dispersion of a probability distribution (Bluman, 2009). In the Figure 3 below, the collection of values demonstrated that K, Sc, Fe and Mg are observed as the most different between the groups. Hence, these elements are chosen to be paired and plot to address the question.

	Group	Th	Al	K	Ca	Sc	Ti	Mn	Fe
1	bhb	1.61908724	3.71790109	3.90113282	5.8090376	6.42350708	3.94834545	3.52061084	6.87663171
2	mb	1.50235374	2.92766337	4.30716179	1.97653816	5.52659268	1.13780262	1.49859722	3.14417436
3	nth	1.64322605	5.77297829	16.043106	5.22375243	9.17398683	4.31428614	5.23473082	11.4121074
4	pm	1.70439331	5.43757301	3.71386223	3.60725818	8.23055205	5.62417184	2.06909028	5.60730932
	Group	Ni	Cu	Zn	Ga	Ge	Se	Br	Rb
1	bhb	5.82416838	6.7126033	3.94432753	5.45943268	2.7815758	3.91105588	3.52610601	3.58891676
2	mb	4.26047042	2.2351021	3.2970789	1.10408576	2.81737387	2.74800515	7.03552355	1.37493125
3	nth	6.42038957	1.7814037	7.08679751	6.98620252	4.3721576	2.97964332	3.0440403	2.92319534
4	pm	1.91670966	1.63330861	5.34217054	3.08372491	2.67167732	2.79812417	2.34313768	4.90350607
	Group	Sr	Y	Mo	Pd	I	Ba	La	Ce
1	bhb	5.70017161	3.30186953	2.66816401	4.21769761	6.00037101	6.91636973	2.95375646	2.49549688
2	mb	1.67835545	4.04343422	2.8895303	1.59169773	3.02175228	2.53094174	3.12154819	5.02562951
3	nth	5.69929638	4.9240703	6.40233847	2.45659979	6.35811501	7.16865877	4.08207879	2.39416954
4	pm	3.46121308	3.97038954	2.79750987	2.06963013	2.71788604	4.19687609	5.92304113	3.28290311
	Group	Pr	Nd	Sm	Eu	Gd	Tb	Dy	
1	bhb	2.37732827	3.08606693	3.45123282	4.77220749	3.31597189	3.3007388	2.45957933	
2	mb	2.41278231	2.59327495	2.93200599	2.59721451	2.68284611	1.98939967	2.88136048	
3	nth	6.18939828	6.10647563	4.06153675	4.09187358	2.70456928	3.66802716	4.59471773	
4	pm	6.20511074	4.66277596	3.57985658	3.64572137	4.3368585	3.25323915	3.67310776	
	Group	Ho	Er	Tm	Yb	Lu	Ta	Mg	
1	bhb	3.81926082	3.12061891	2.77658921	2.50715897	2.55481179	1.30199093	5.19914757	
2	mb	1.85078664	3.54941273	2.37181624	4.50596052	2.77763475	1.46635615	1.68527255	
3	nth	4.91901883	5.86662376	3.68580469	2.64138928	4.64240685	1.0339227	17.97946	
4	pm	3.87799664	2.98601454	3.67661166	3.72510362	3.84401475	1.70884354	7.2333053	

Figure 3 CV of data

3 3 Design plots to address the question

This section is the main step of the research. By mapping the aesthetics in the plots to variables in the dataset and also the colours of the points to the Group variable, the information can vividly reveal the different soil type. The number of selections of pairs are five. Due to the limited space for more pairs to be analysed, the plot of Fe and Mg can also be found in Appendix with the plots ran by SAS.

3 3 1 K_Sc

Figure 4 is the plot of elements Ga and Mn. The plot demonstrates that the plants grown in bnb and pm contain more K than the other two. And in the mb area, K in plant is almost 0, which is too little compared with other areas. Overall, the content of Sc in all four areas is low, but K in bhb, nth, pm are all well.

Figure 4 Scatter plot for K and Sc

3 3 2 K_Fe

The Figure 5 represents the scatter plot of K and Fe. There are many similarities with the map and the points in Figure 4. For example, the elements of plants grown in mb area are close to 0 while the K and Fe contents grown in other three sites are relatively large. However, it can be observed that the unit number of Fe coordinates is more than 1000 times the number of units in Sc in Figure 4.

Figure 5 Scatter plot for K and Fe

3 3 3 K_Mg

It can be clearly seen in Figure 6 that there is more Mg in the place containing more K. Since the coordinates of Mg and K are both ten thousands of numbers, they are also scattered to the upper right corner. Except for the number of mb areas is too small, which is difficult to see its distribution, the distributions of other areas, especially the pm area, can clearly confirm this

idea.

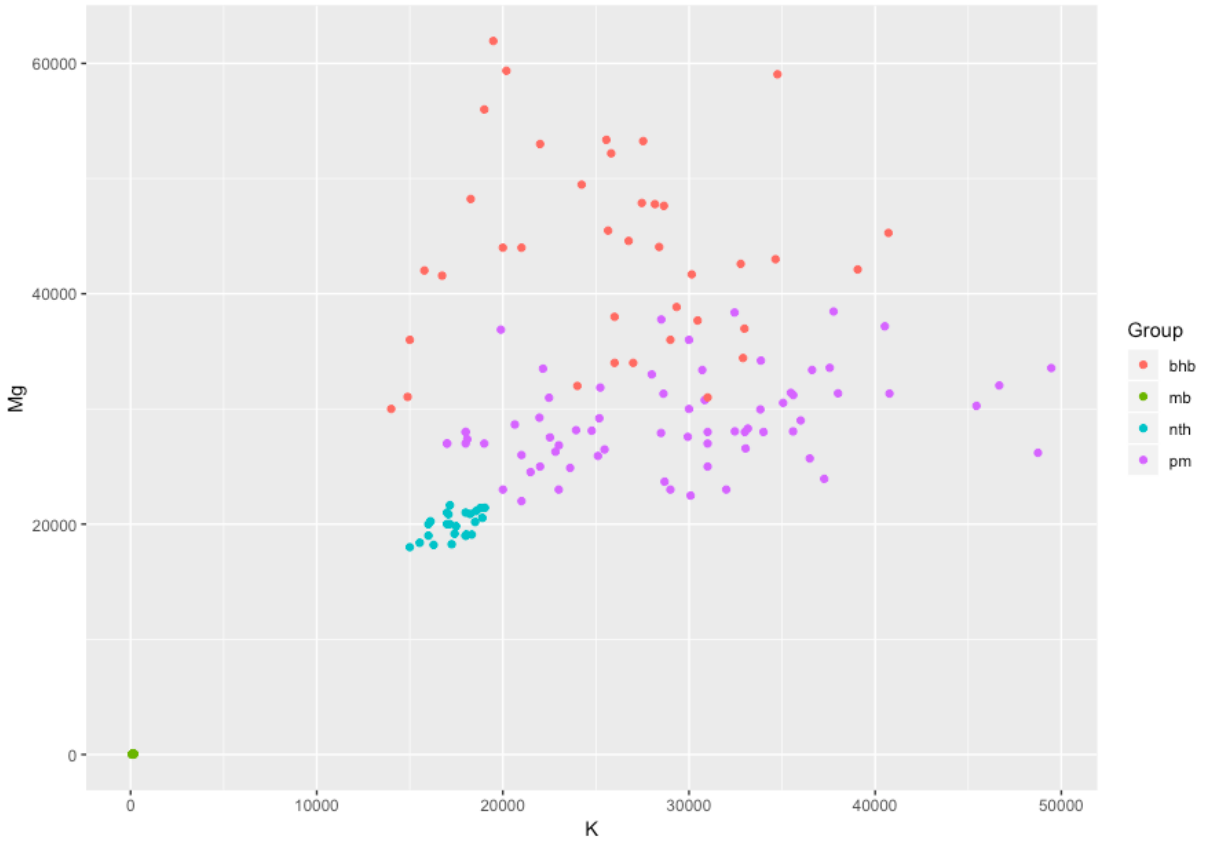


Figure 6 Scatter plot for K and Mg

3 3 4 Sc_Fe

Figure 7 also shows a certain degree of in the slightly more Sc content area, the Fe content is also more, which can be seen from the scatter distribution in the mb area. The other three regions are more obvious. However, the content of Sc is very small of all soil types, and this idea might be completely determined.

Figure 7 Scatter plot for Fe and Sc

3 3 5 Sc_Mg

Finally, a comparison of the Sc element and the Mg element is indicated in Figure 8. By observing the unit quantity of Sc coordinates, it can be found that the content of Sc element in each place is relatively small, and the Mg element is relatively large, especially for the bhb

area.

Figure 8 Scatter plot for Mg and Sc

4 Conclusion

The data and graphs do show the difference in chemical elements of cannabis plants grown in different soil types.

Some elements seem to be related to each other, for example, cannabis leaves containing more K also contain more Mg. Further research should be conducted more on is there really a linear relationship between two elements.

From my perspective, the results of this experiment are not able to determine the correlation between the soil in which the plant grows and the constituent elements of the plant leaf. There are three main reasons for this. First of all, the subject of this experiment is cannabis leaf, and there is no wider range of leaves to study. Second, the samples in this experiment were relatively small and were only sampled in New Zealand. Third, this experiment did not design a comparative experiment. Therefore, more experiments and research need to be conducted to confirm the idea of plants chemical composition is specific to the soil type.

5 References

Bluman, A. (2009) *Elementary statistics: a step by step approach*. McGraw-Hill Education.

Kobayashi, K. and Salam, M.U. (2000) 'Comparing simulated and measured values using mean squared deviation and its components', *Agronomy Journal*, 92(2), pp. 345-352. *Science Societies* [Online]. Available at: https://dl.sciencesocieties.org/publications/aj/abstracts/92/2/345/ (Accessed: 9 October 2018).

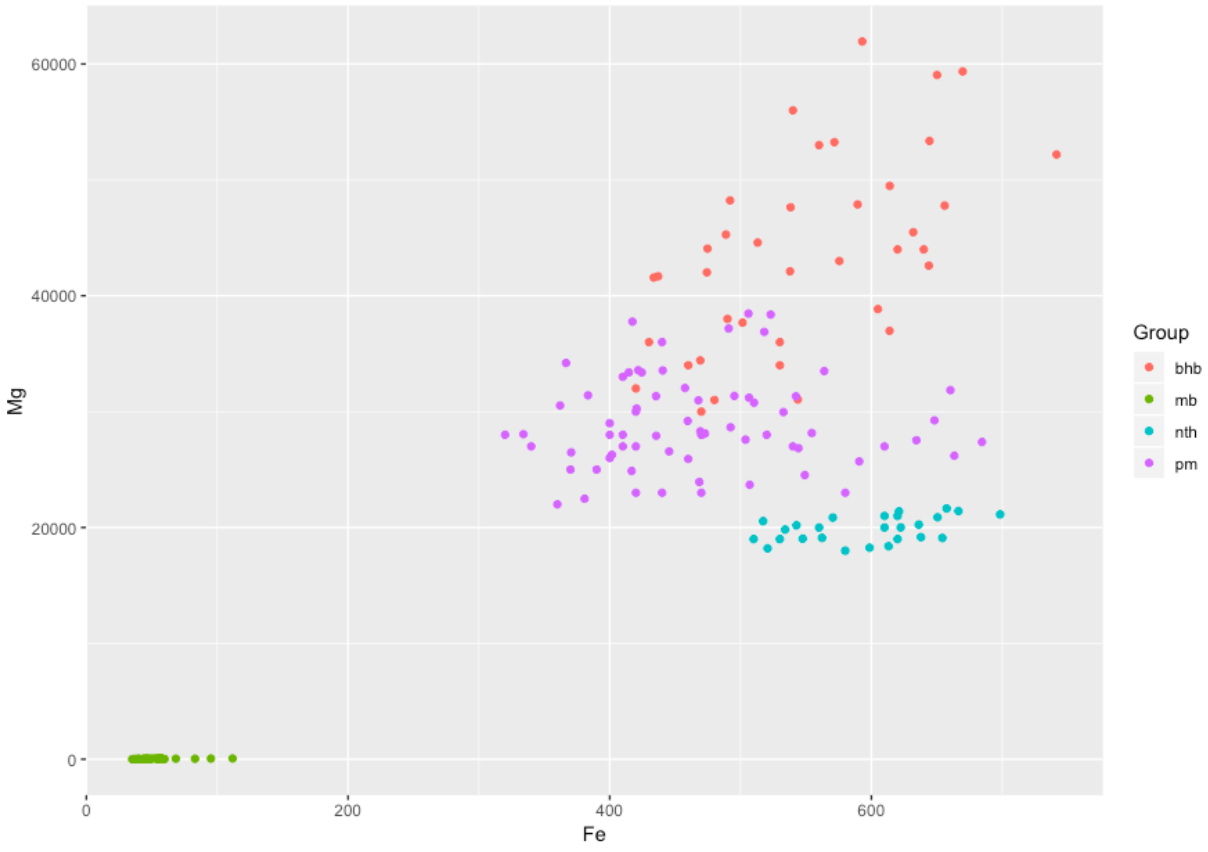
RStudio Inc. (2018) RStudio Version 1.1.456.

SAS Institute Inc. (2011) Base SAS® 9.3 Procedures Guide.

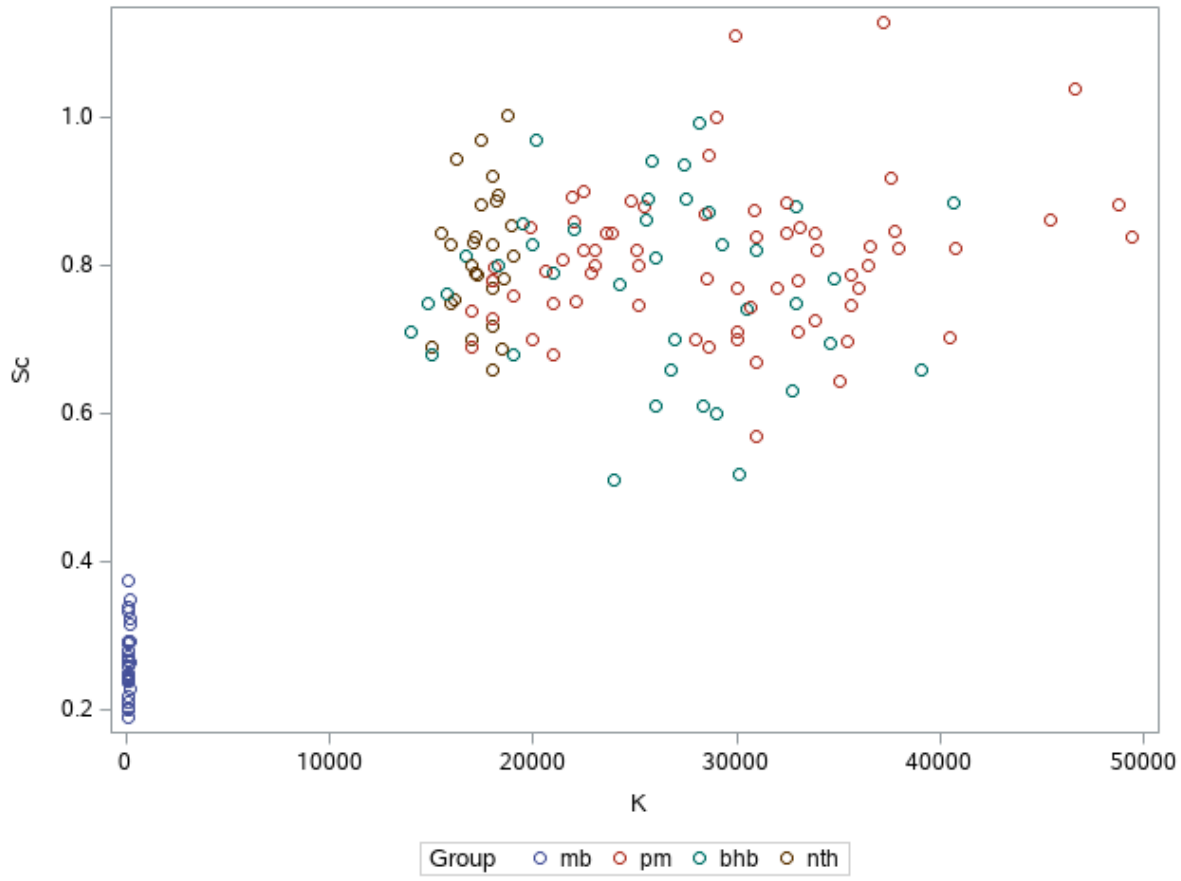
Smith, N. (2000) Criminal may rue pot from this plot. *New Zealand Herald*.

6 Appendix

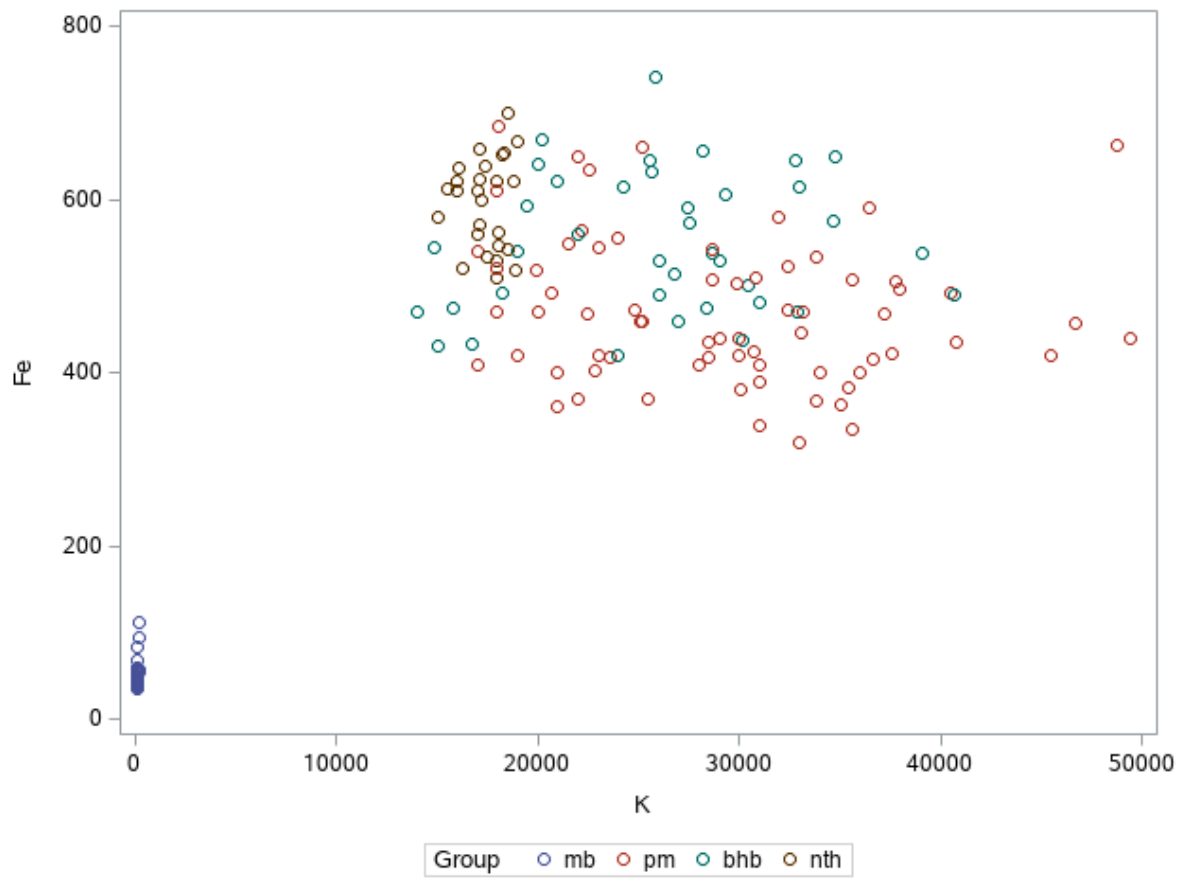
6 1 Scatter points of Fe and Mg in R



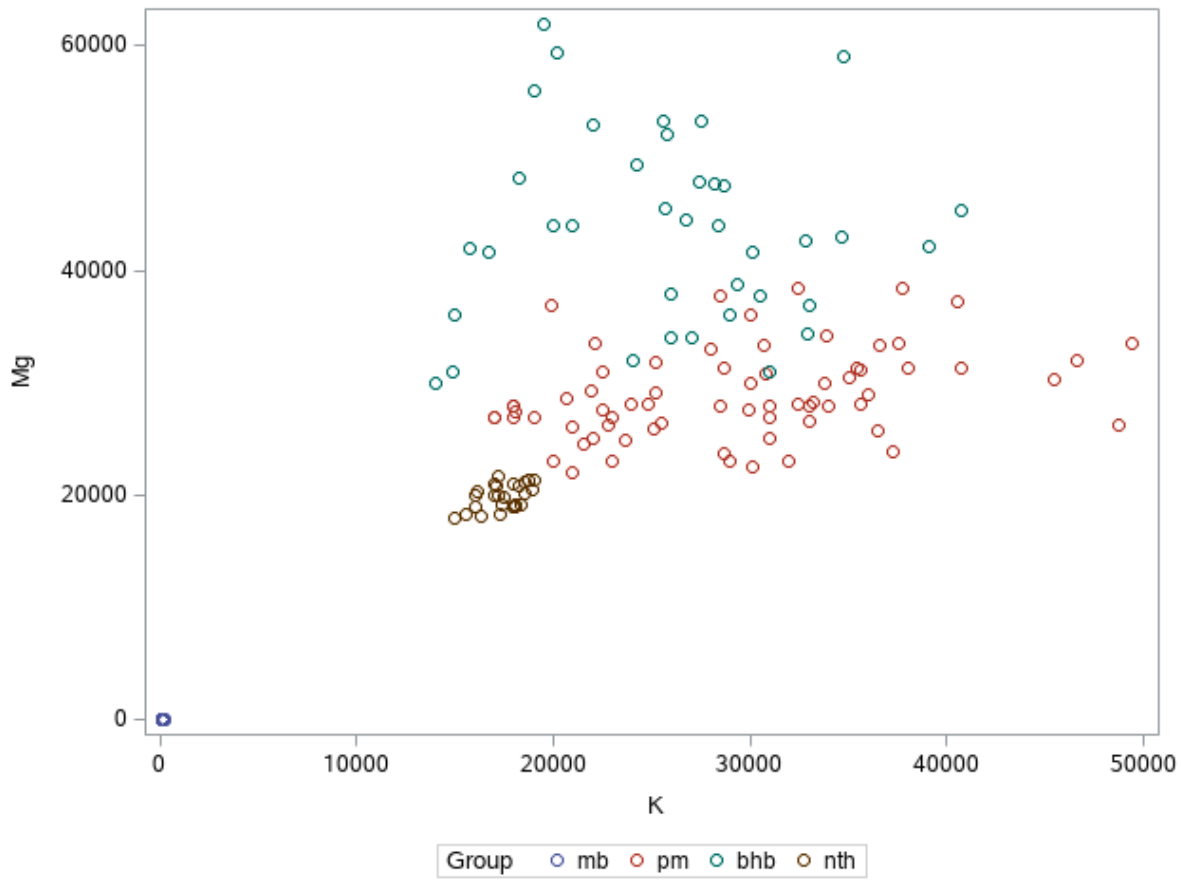
6 2 Scatter points of K and Sc in SAS



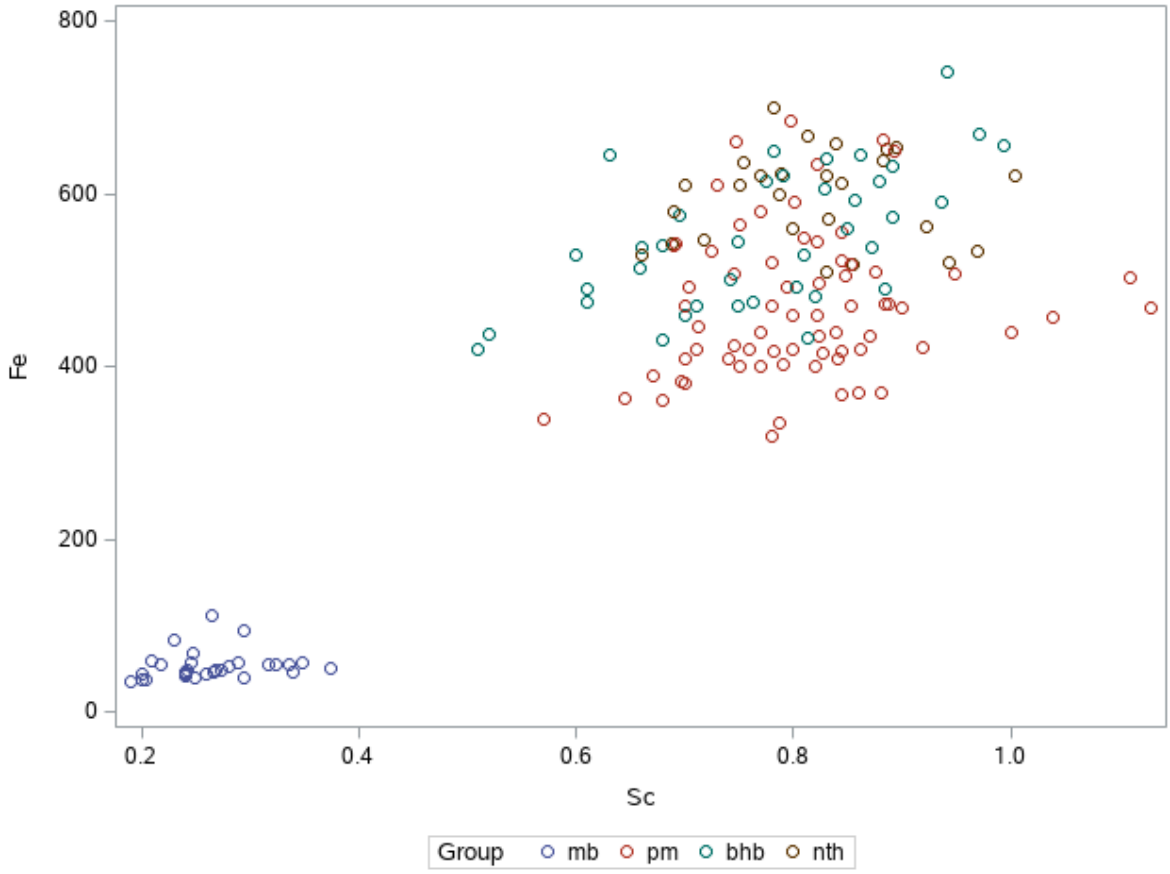
6 3 Scatter points of K and Fe in SAS



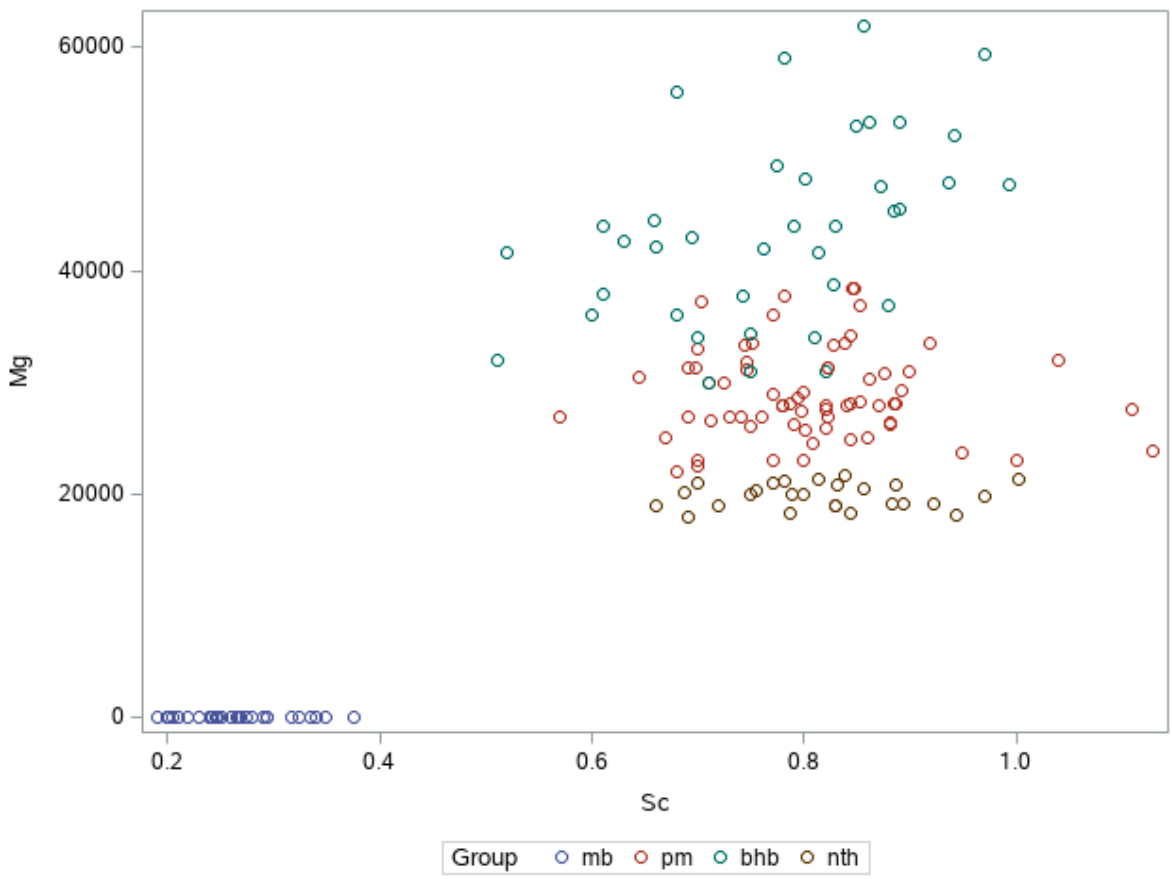
6 4 Scatter points of K and Mg in SAS



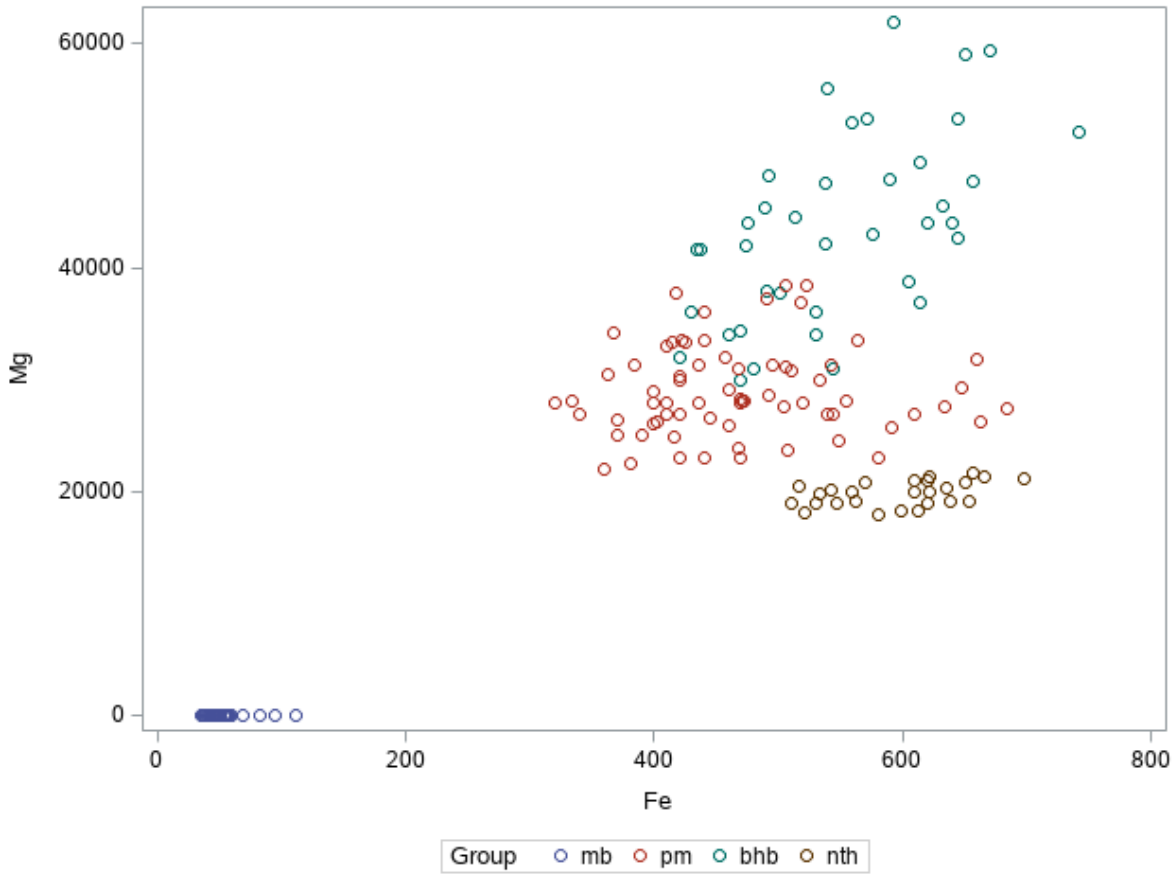
6 5 Scatter points of Fe and Sc in SAS



6 6 Scatter points of Mg and Sc in SAS



6 7 Scatter points of Fe and Mg in SAS



7 Programming Code

7 1 SAS Code

```
/*read sheet1*/
proc import out=potplants1
  datafile="D:\MT5763\Assignment 1\PotPlants_18.xlsx";
  sheet="Sample Set One";
  getnames=yes;
run;
/*read sheet2*/
proc import out=potplants2
  datafile="D:\MT5763\Assignment 1\PotPlants_18.xlsx";
  sheet="Sample Set Two";
  getnames=yes;
run;
/*read sheet3*/
```

```
proc import out=potplants3
  datafile="D:\MT5763\Assignment 1\PotPlants_18.xlsx";
  sheet="Sample Set Three";
  getnames=yes;
run;
data potplants4;
set potplants1 potplants2 potplants3;
run;
/*check Null*/
data missing(drop=i);
set potplants4;
array a_numeric_;
do i=1 to dim(a);
if missing(a) then output;
end;
array b_charater_;
do i=1 to dim(b);
if missing(b) then output;
end;
/*drop null values*/
data potplants;
set potplants4;
array x_all_;
do i=1 to dim(potplants4);
if x(i)=. then delete;
end;
run;
/*summary the data*/
proc summary data = potplants mean std n max min ;/*statistical quantities*/

var _numeric_;
class Group;
output out = aa;
```

```
proc print data=aa;

run;

/*tabulate the data*/
proc tabulate data=potplants ;
class Group ;
var _numeric_ ;
table Group ,(mean,std \)
proc print ;
run;

/*plot the K and Sc*/
proc sgplot data=potplants;
scatter x=K  y=Sc  /group=Group;
run;

/*plot the K and Fe*/
proc sgplot data=potplants;
scatter x=K  y=Fe  /group=Group;
run;

/*plot the K and Mg*/
proc sgplot data=potplants;
scatter x=K  y=Mg  /group=Group;
run;

/*plot the Sc and Fe*/
proc sgplot data=potplants;
scatter x=Sc  y=Fe  /group=Group;
run;

/*plot the Sc and Mg*/
proc sgplot data=potplants;
```

```
scatter x=Sc  y=Mg  /group=Group;
run;

/*plot the Fe and Mg*/
proc sgplot data=potplants;
scatter x=Fe  y=Mg  /group=Group;
run;
```

7 2 R Code

```
#If no "readxl" package, install it
#install.packages("readxl")
library(readxl)
#read in the data in the three sheets
potplants1<-read_xlsx("/Users/apple/Desktop/MT5763/Assignment 1/PotPlants_18.xlsx",
sheet = 1)
head(potplants1)
potplants2<-read_xlsx("/Users/apple/Desktop/MT5763/Assignment 1/PotPlants_18.xlsx",
sheet = 2)
head(potplants2)
potplants3<-read_xlsx("/Users/apple/Desktop/MT5763/Assignment 1/PotPlants_18.xlsx",
sheet = 3, col_types = c("text", "text", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric"))
head(potplants3)
#format arrangement
potplants2$Group[potplants2$Group == "potting mix" ] = "pm"
potplants2<-potplants2[-57:-59,]
potplants3<-potplants3[-49,]
#combine the data
library(dplyr)
```

```
potplants<-bind_rows(potplants1,potplants2,potplants3)
potplants<-potplants[,-1]
potplants<-potplants[,-40]
#drop NA in the data
potplants<-na.omit(potplants)
dim(potplants)
#summary of data
potplants_sum<-summary(potplants)
write.csv(potplants_sum, file = "Desktop/MT5763/Assignment 1/potplants_sum.csv")

#explore the data
potplants_mean<-summarise_all(group_by(potplants,Group),mean)
potplants_sd<-summarise_all(group_by(potplants,Group),sd)
cv<-summarise_all(group_by(potplants,Group),funs(mean(.)/sd(.)))
write.csv(cv, file = "Desktop/MT5763/Assignment 1/cv.csv")

#plot the data
library(ggplot2)
K_Sc<-ggplot(data = potplants) + geom_point(mapping = aes(x = K, y = Sc, color = Group))
ggsave("potplants_K_Sc", K_Sc, "pdf")
K_Sc
K_Fe<-ggplot(data = potplants) + geom_point(mapping = aes(x = K, y = Fe, color = Group))
ggsave("potplants_K_Fe", K_Fe, "pdf")
K_Fe
K_Mg<-ggplot(data = potplants) + geom_point(mapping = aes(x = K, y = Mg, color =
Group))
ggsave("potplants_K_Mg", K_Mg, "pdf")
K_Mg
Sc_Fe<-ggplot(data = potplants) + geom_point(mapping = aes(x = Sc, y = Fe, color =
Group))
ggsave("potplants_Sc_Fe", Sc_Fe, "pdf")
Sc_Fe
Sc_Mg<-ggplot(data = potplants) + geom_point(mapping = aes(x = Sc, y = Mg, color =
Group))
```

```
ggsave("potplants_Sc_Mg", Sc_Mg, "pdf")
Sc_Mg
Fe_Mg<-ggplot(data = potplants) + geom_point(mapping = aes(x = Fe, y = Mg, color =
Group))
ggsave("potplants_Fe_Mg", Fe_Mg, "pdf")
Fe_Mg
```