2018

Project 1- Exploratory data analysis

MT5763

180029243

Executive Summary

This report analyses a data set containing data regarding the elemental composition of cannabis leaves grow in different soils of New Zealand (Smith, 2000). This data analysis was done through data manipulation and plotting in R Studio and SAS. Some packages were used in R Studio, such as dplyr, tidyverse, ggplot, corrplot, data.table and ggpubr. The main conclusions from this report are:

- The data indicate differences in the elemental composition of cannabis leaves grown in different soil types.
- Most elements are related to one another in terms of their levels in the sampled leaves.
- The results of this experiment do not seem to ultimately allow the determination of what soil the plants were grown in, just from the elemental composition of the leaves.

Contents

Introduction

There has been a tendency for cannabis use to increase in developed countries, especially among teenagers. Studies show that over two thirds of young people up to 21 years old have tried cannabis at least once in New Zealand (Ferguson and Horwood, Poulton et al, cited in Ferguson, et al, 2003). Besides the obvious health hazards that can occur due to cannabis use, a massive illegal market arises, which the police is keen on stopping (Hall and Solowij, Wodak et al, Ferguson et al, Kander et al, McGee et al, cited in Ferguson et al, 2003). One possibility to increase efficiency when prosecuting criminals is to identify unique elements in the plant's composition and the soil type. Scientist Dion Sheppard is writing his Master's thesis based on this hypothesis. If a correlation between the soil the plant was grown in and its elements is found, then this research can be used by the police in the court of law (Smith, 2000). This report aims to analyse the observations of this experiment and answer the following questions:

- Do the data indicate differences in the elemental composition of the cannabis leaves grown in different soil types?
- Are some of the elements related to one another in terms of their levels in the samples leaves?
- Is it possible to determine what soil these plants were grown in, based solely on the elemental composition of the cannabis leaves?

All the R and SAS and complementary images for this report are in the Appendix section. The images in the body of this report are from the R code (SAS, 2001)(R Studio, 2017).

Analysis

Statistical concepts

This report presents some statistical concepts, therefore, their understanding is key to understand the content and analysis of this paper.

The mean is calculated by adding up all the values in a population distribution and dividing by the number of values.

The min and max are the minimum and the maximum values in a distribution, respectively.

The standard deviation is the average distance from the mean that values in a distribution fall. A small standard deviation means values in a distribution cluster close to the mean, and a large standard deviation means the values deviate widely from the mean (R Studio, 2017) (Randolph and Myers, 2013).

Data Set Overview

The data set used for this report records the elemental composition of cannabis leaves across different soil types. This data set was divided in three sample sets, which were clean, organised and combined prior to analysis. There are 40 variables in this data set, the first of which is Sample Name, with 153 character type entries, and refers to each individual cannabis leaf. The second variable is Group, it refers to the soil type and it is a character type data. The four types of soil in this data set were store bought potting mix (pm), Blockhouse Bay (bhb), Mission Bay (mb) and Northland (nth). Blockhouse bay and Mission Bay are geographically close to each other in the suburbs of Auckland, whereas Northland is soil from a norther region of New Zealand (Smith, 2000). The remaining 38 variables are the units of elements present in these cannabis leaves, and this is a numeric variable (Figures 1-3).

Methodology

The software used for this report is R Studio version 3.5.1 and SAS version 9.4., and Microsoft Word 2013 for word processing. Additional packages used in R Studio are ggplot2, tidyverse, dplyr, corrplot, data.table and ggpubr.

Data Set Summary

A summary of the data was made in order to get the following results: mean, standard deviation, minimum and maximum values, and number of entries. This summary was mode for five elements in each of the Sample Sets. There are no differences in these results to the results in the summary stats sheet. The results are detailed in the tables below (Tables 1-3).

	Mean	Sd	Min	Max	n
Th	0.03368741	0.0217476732	0.00158077	0.11262422	56
Al	39.0458	17.9330106	13.9557	104.0717	56
K	22226.6546	12615.9588	91.4371	45444.8317	56
Ca	41648.2958	26490.5917	48.6185	88380.1279	56
Sc	0.727991	0.233016987	0.209773	1.109849	56

Table 1- Summary, Sample Set One

	Mean	Sd	Min	Max	n
Mg	23428.3	13871.866	4.0	56000.0	52
Al	32.30	14.7168571	8.60	83.00	52
K	18637.5	10728.081	85.0	36000.0	52
Ca	35651.3	23887.9529	28.0	81000.0	52
Sc	0.649231	0.214035421	0.190000	1.000000	52

Table 2- Summary, Sample Set Two

	Mean	Sd	Min	Max	n
Mg	26672.4006	16010.2855	14.8492	61944.9859	55
Al	33.0866	16.7159458	18.3946	87.3659	55
K	22677.2295	13451.9649	91.1559	49463.3389	55
Ca	42845.3681	26864.6237	42.7457	89422.2519	55
Sc	0.720984	0.234277615	0.200489	1.128851	55

Table 3- Summary, Sample Set Three

Question 1

The first question that arises from this data set is whether there is a difference in the composition of the cannabis leaves in function of the soil type they grew in. The elements chosen to answer this question were aluminium, calcium, potassium, magnesium and titanium.

It is possible to conclude from this data set that the soil from the Blockhouse Bay shows greater deviation of elements. As the value for its standard deviation is high, measurements for Calcium values in this soil are highly variable, from a minimum of 48000 to a maximum of 89422 units. Potassium has a mean of 25900 and Magnesium of 6647. These two elements also deviate considerably less in value than Calcium. Comparatively to the other elements in this soil type, the mean values of Aluminium and Scandium are miniscule, with means of 60.1 and 0.77 respectively (Figures 1-3).

The samples taken from Mission Bay all show low composition of values comparatively to other to other soil types. The element with the highest mean for this soil type is Potassium, with 126.6 units. This drastic change of values might indicate that Mission Bay has a different soil composition than the other types for this experiment

Northland, on the other hand, has highly deviated values of Calcium, 11081.8 units, but not so much for Potassium and Magnesium with 1085.6 and 1104.6 standard deviations, respectively. The values of the mean and standard deviation for Aluminium and Scandium are also comparatively low.

Finally, samples taken from the Potting Mix show more similar values across some elements. The means for Calcium, Potassium and Magnesium differ by no more than 5500 units. Scandium and Aluminium remain low in this soil type.

As described above and shown at Figures 2 and 3, there are some noticeable changes in the elements for each different soil for these five elements, the most different of all being Mission Bay. Therefore, the data suggest differences in the elemental composition of cannabis leaves grown in different soils (R Studio, 2016).

	Element	Group	mean	sd
1	Al	bhb	6.074711e+01	1.633909e+01
2	Al	mb	2.199235e+01	7.511911e+00
3	Al	nth	3.443588e+01	5.965012e+00
4	Al	pm	3.106674e+01	5.713348e+00
5	Ca	bhb	7.044974e+04	1.212761e+04
6	Ca	mb	1.101249e+02	5.571605e+01
7	Ca	nth	5.788848e+04	1.108178e+04
8	Ca	pm	3.416192e+04	9.470328e+03
9	K	bhb	2.593031e+04	6.646867e+03
10	K	mb	1.266418e+02	2.940261e+01
11	K	nth	1.741698e+04	1.085636e+03
12	K	pm	2.926226e+04	7.879199e+03
13	Mg	bhb	4.367183e+04	8.399805e+03
14	Mg	mb	3.595889e+01	2.133714e+01
15	Mg	nth	1.986067e+04	1.104631e+03
16	Mg	pm	2.895552e+04	4.003083e+03
17	Sc	bhb	7.721968e-01	1.202142e-01
18	Sc	mb	2.642842e-01	4.782045e-02
19	Sc	nth	8.167015e-01	8.902362e-02
20	Sc	pm	8.068892e-01	9.803586e-02

Figure 1- Table, Mean and Standard Deviation of each element

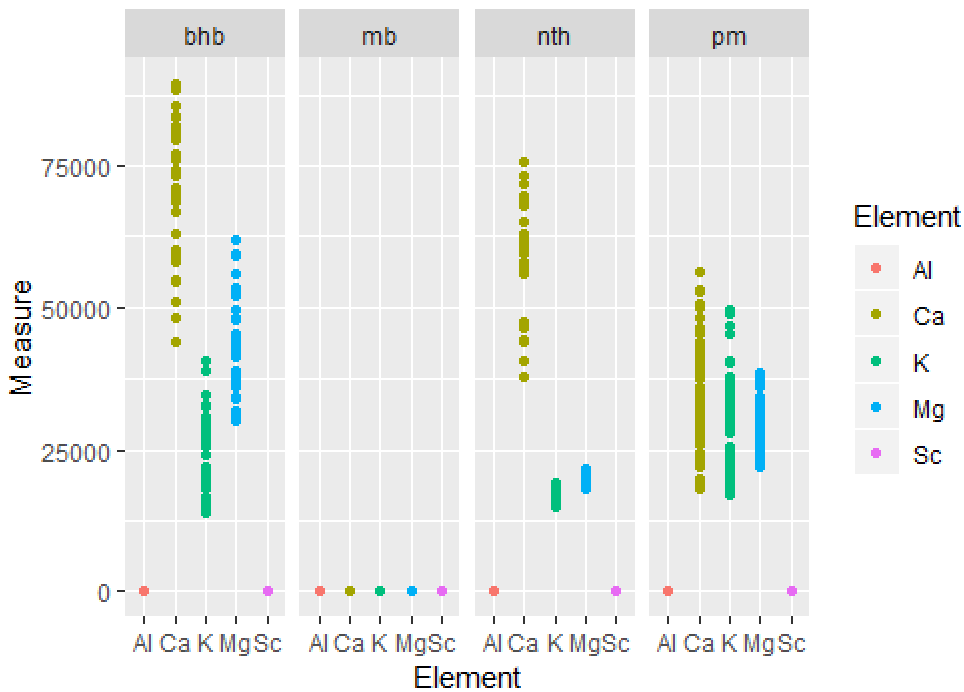

Figure 2- Scatterplot, elements in function of soil types.

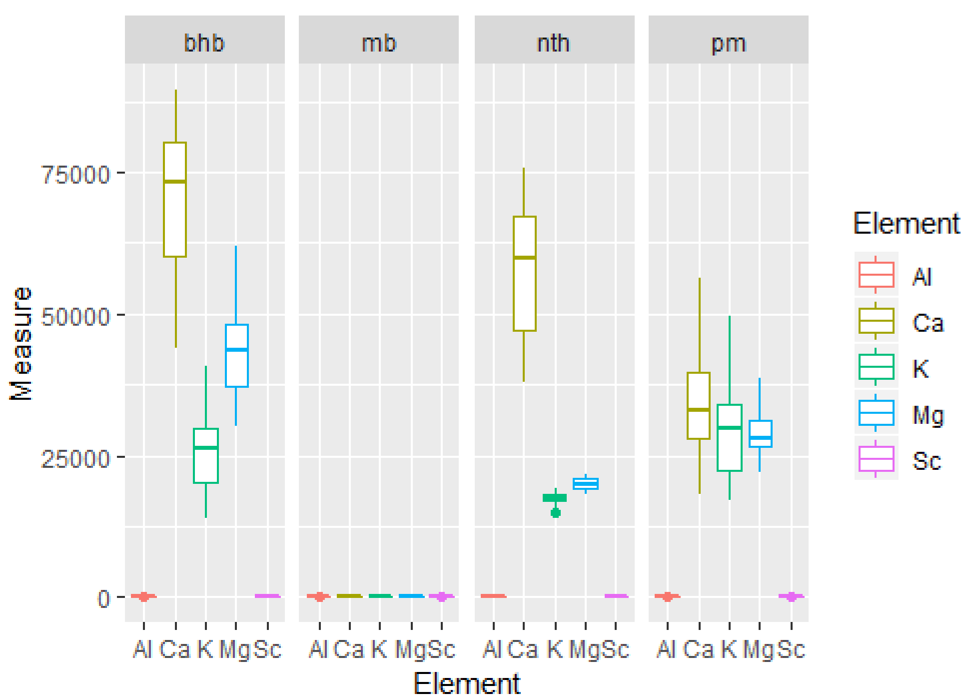

Figure 3- Boxplot, elements in function of soil types.

Question 2

This data set definitely shows interesting correlations between pairs of sampled leaves. Figure 4 is a graphical representation of these correlations. Elements such as Barium and Europium show an almost perfect correlation, 0.912. The same phenomenon is seen with the pair of elements Zinc and Strontium and their 0.890 correlation. On the other hand, Molybdenum and Gallium have a negatively correlated, as well as Molybdenum and Barium. Elements such as Molybdenum and Calcium can be considered non-correlated, as their correlation is extremely close to zero. Figures 5-9 illustrate the correlations between pairs of elements. There are in fact some element related to one another. In fact, there are 348 cases of correlation between pairs of elements, or 50.8% of all possible pairs.

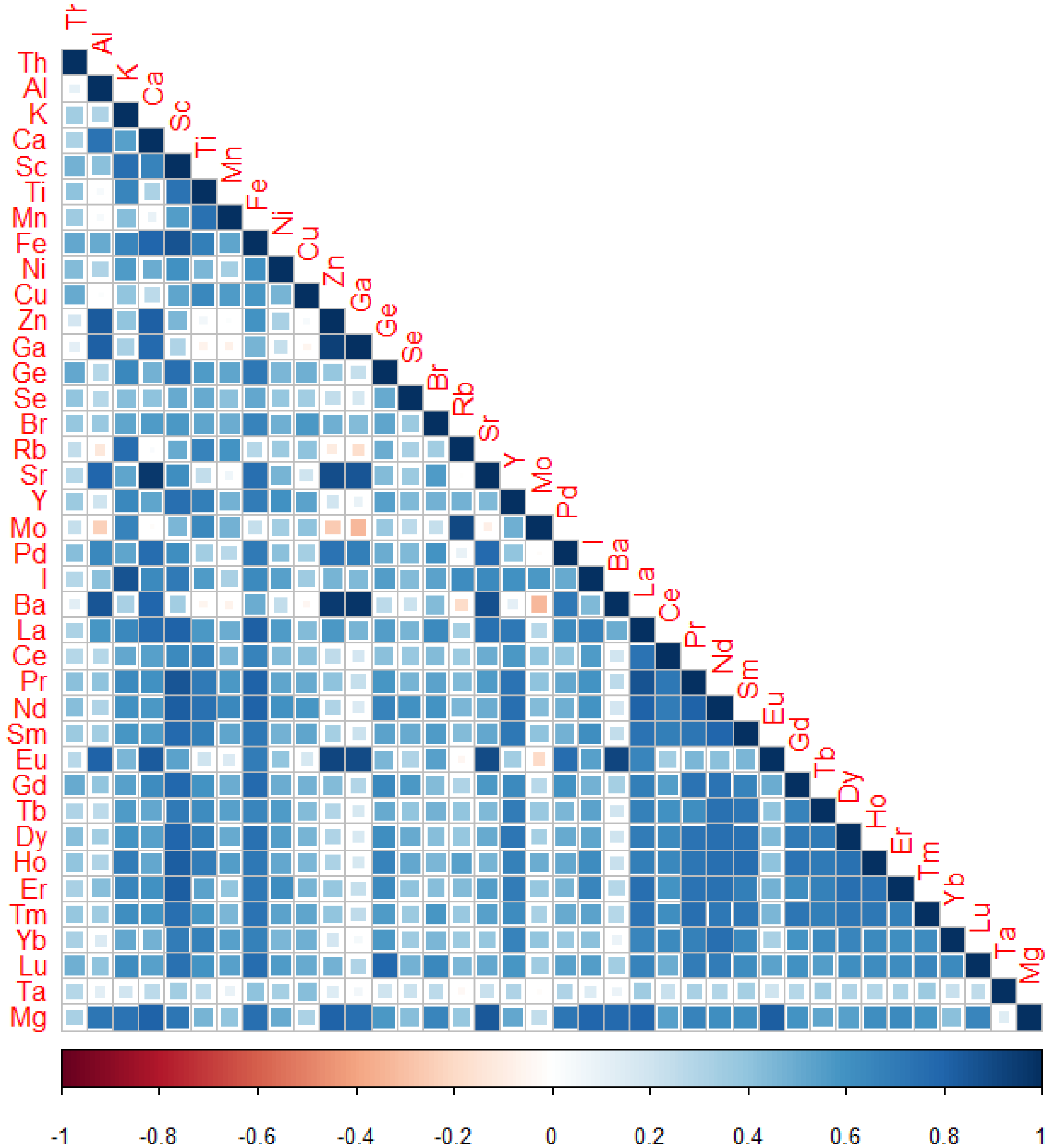

Figure 4- Correlation Plot, Correlation between the elements

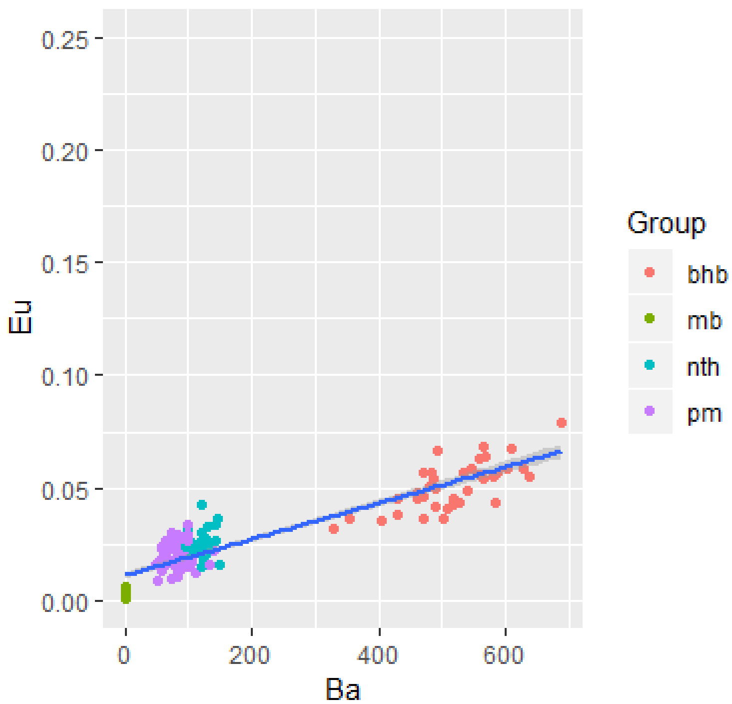

Figure 5- Scatterplot, Eu in function of Ba

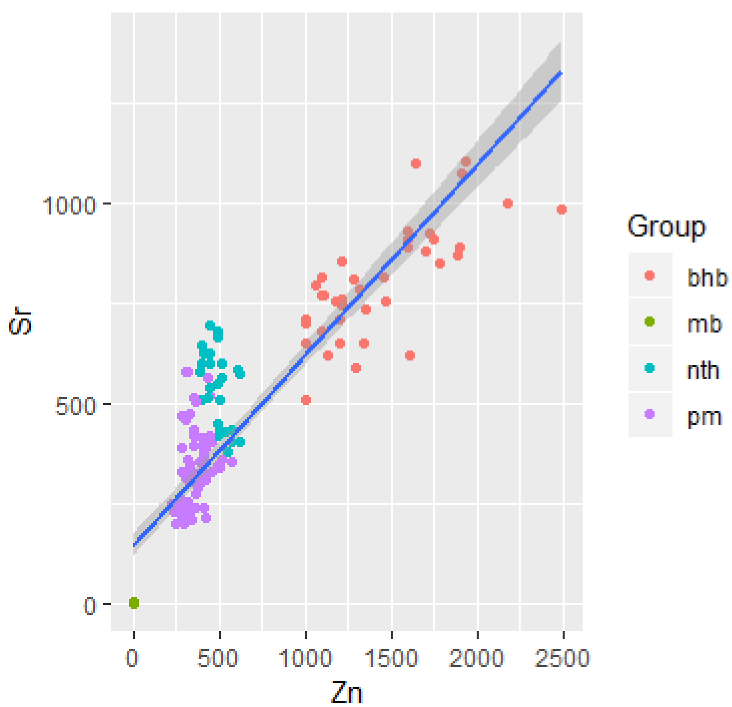

Figure 6- Scatterplot, Sr in function of Zn

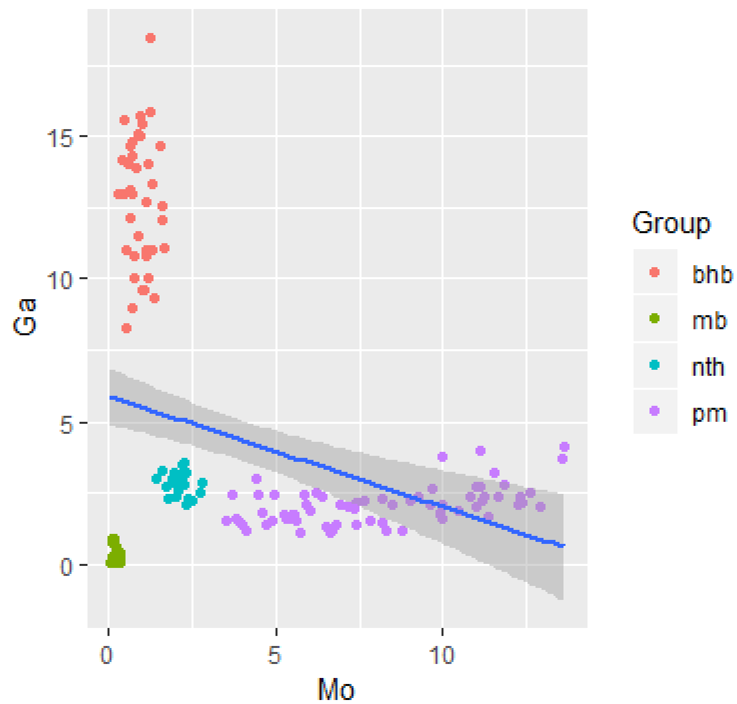

Figure 7- Scatterplot, Ga in function of Mo

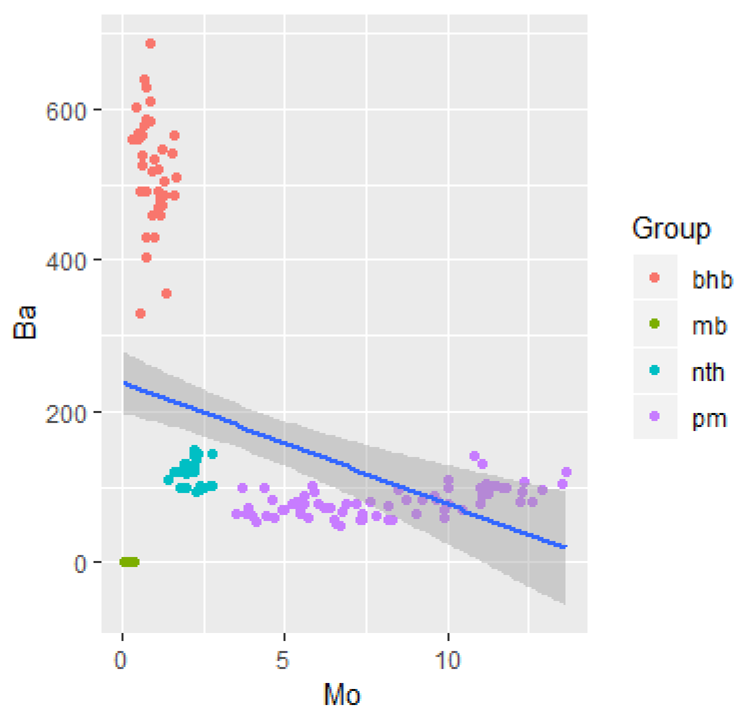

Figure 8- Scatterplot, Ba in function of Mo

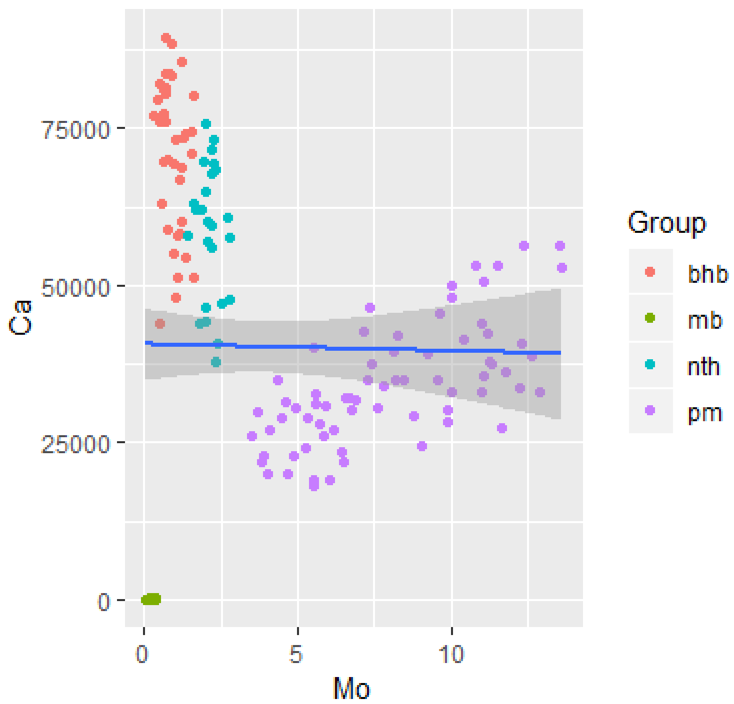

Figure 9- Scatterplot, Ca in function of Mo

Question 3

Having the ability to determine what soil the plants were grown in, just from the elemental composition of the cannabis leaves would be of great value for the aim of this research as well as its use for the police when prosecuting criminals. However, with the current variables, this seems unlikely. Other variables such as climate and soil conditions would have to be taken into account, as these variables would probably have an effect on the elemental composition of the cannabis leaves as well. Furthermore, there is not enough data available on New Zealand's soil profiles and it is not possible to control or monitor the level of anthropogenic additives in the soil. Furthermore, variations of elements observed are dependent on the genetic characteristics of the seed stock, and not so much the soil type. In addition to this, other factors such observing different parts of the same plant, the sex of the plant and its maturity all influence the results. Consequently, it is not possible to determine the soil type just from the elemental composition of the leaves (Sheppard, 2000).

Discussion/Conclusion

Regarding question 1, the differences in the mean values for different soils suggest a p-value close to zero. If this is true, then the null hypothesis would be rejected for this data set. It can be therefore concluded that there are differences elemental composition of cannabis leaves grown in different soils (Uboe, 2017).

Correlations and scatterplots have shown that most pairs of elements in this data set have a positive relationship between them, with some pairs showing an almost 1:1 correlation. Only a few have either no relationship or a negative relationship (Uboe, 2017).

It is not possible to identify the soil the plant grew in based solely on its elemental composition because there are other variables to take into account, and such elements would have an effect on the elemental composition of the cannabis leaves, such as sex or maturity (Sheppard, 2000).

Appendices

Appendix A

This appendix has all of the R code implemented for this report.

```
#importing sheet1 and cleaning data

sheet1 <- read.csv("C:/Users/José Baltazar/Desktop/potplants sheets/pp_sheet1.csv", header = TRUE)

#changing "Sample.Name" to "Sample Name"

sheet1 <- (setnames(sheet1, old = c("Sample.Name"), new =c("Sample Name")))


#importing sheet2 and cleaning data

sheet2 <- read.csv("C:/Users/José Baltazar/Desktop/potplants sheets/pp_sheet2.csv", header = TRUE)


#replacing potting mix with pm

sheet2 <- sheet2 %>% mutate(Group = recode(Group, 'potting mix' = 'pm'))


#removing NA's

sheet2<- sheet2%>%na.omit


#removing mean and variance values at the bottom

sheet2 <- (sheet2[-c(53,54),])


#changing "SampleName" to "Sample Name"

sheet2 <- (setnames(sheet2, old = c("SampleName"), new =c("Sample Name")))


#importing sheet3 and cleaning data

sheet3 <- read.csv("C:/Users/José Baltazar/Desktop/potplants sheets/pp_sheet3.csv", header = TRUE)


#remove "missing" values

sheet3 <- (sheet3[-c(49),])
```

```
#converting the elements from factor to numbers.

sheet3 <- mutate_at(sheet3, vars(-Sample.Name, -Group), function(x) as.numeric(as.character(x)))


#changing "Sample.Name" to "Sample Name"

sheet3 <- (setnames(sheet3, old = c("Sample.Name"), new =c("Sample Name")))


#combining all the sheets into one data set

sheet <- rbind(sheet1, sheet2, sheet3)


#ANALYSIS

#Question 1

#Creating a function "test" to filter the data set and show only the elements I want to analize for question 1.

test <-  sheet [c("Group", "Mg", "Al", "K", "Ca", "Sc")]


#"newData" is a new function arranges the data in "test"

newData <- test %>% gather(key = Element, value = Measure, Mg, Al, K, Ca, Sc)


#"newnewdata" give gives me the mean and standard deviations of the elements above

newnewdata <- newData %>% group_by(Element, Group) %>% summarise(mean = mean(Measure), sd = sd(Measure))


#graphs separated by soil type and with different colours for each element

ggplot(data = newData) + geom_point(aes(x = Element, y = Measure, colour = Element )) + facet_grid(~ Group)

ggplot(data = newData) + geom_boxplot(aes(x = Element, y = Measure, colour = Element )) + facet_grid(~ Group)

View(newnewdata)


#Question 2
```

```
#filtering the data to be shown only the columns I want

test2 <-  sheet [c("Sample Name", "Group", "Mg", "Al", "K", "Ca", "Sc")]


newData2 <- test2 %>% gather(key = Element, value = Measure, Mg, Al, K, Ca, Sc)


newData2 %>% group_by(Element)


#graphs to compare the measures of a pair of elements

#Ba and Zn

baeu <- ggplot(data = sheet, aes(x= Ba, y= Eu)) + geom_point(data = sheet, aes(x = Ba, y = Eu, colour =
Group)) + geom_smooth(method = lm) + scale_y_continuous(limits = c(0,0.25))


#Ba and Sr

znsr <- ggplot(data = sheet, aes(x= Zn, y= Sr)) + geom_point(data = sheet, aes(x = Zn, y = Sr, colour =
Group)) + geom_smooth(method = lm)


#Mo and Ga

moga <- ggplot(data = sheet, aes(x= Mo, y= Ga)) + geom_point(data = sheet, aes(x = Mo, y = Ga, colour =
Group)) + geom_smooth(method = lm)


#Mo and Ba

moba <- ggplot(data = sheet, aes(x= Mo, y= Ba)) + geom_point(data = sheet, aes(x = Mo, y =Ba, colour =
Group)) + geom_smooth(method = lm)


#Mo and Ca

moca <- ggplot(data = sheet, aes(x= Mo, y = Ca)) + geom_point(data = sheet, aes(x = Mo, y = Ca, colour =
Group)) + geom_smooth(method = lm)


#correlation between the elements

corr1 <- cor(sheet[3:40])
```

```
corrplot(corr1, method = "square", type = "lower")


#How many elements have a correlation smaller than -0.5 or greater than +0.5? Divided by 2 to eliminate

#removed duplicates and subtracted 38 to remove correlation between the same elements (for example Mn and Mn)

sum(corr1 > 0.5 | corr1 < -0.5)/2-38


#summary

#taking the mean and standard deviation of Th, Al, K, Ca and Sc for Sample Set One; and Mg, Al, K, Ca and Sc for Sample Sets Two and Three.

#sheet1

#th

summary(sheet1$Th)

sd(sheet1$Th)

#Al

summary(sheet1$Al)

sd(sheet1$Al)

#K

summary(sheet1$K)

sd(sheet1$K)

#Ca

summary(sheet1$Ca)

sd(sheet1$Ca)

#Sc

summary(sheet1$Sc)

sd(sheet1$Sc)

#sheet2

#Mg

summary(sheet2$Mg)
```

```
sd(sheet2$Mg)
#Al
summary(sheet2$Al)
sd(sheet2$Al)
#K
summary(sheet2$K)
sd(sheet2$K)
#Ca
summary(sheet2$Ca)
sd(sheet2$Ca)
#Sc
summary(sheet2$Sc)
sd(sheet2$Sc)
#sheet3
#Mg
summary(sheet3$Mg)
sd(sheet3$Mg)
#Al
summary(sheet3$Al)
sd(sheet3$Al)
#K
summary(sheet3$K)
sd(sheet3$K)
#Ca
summary(sheet3$Ca)
sd(sheet3$Ca)
#Sc
summary(sheet3$Sc)
sd(sheet3$Sc)
```

Appendix B

This appendix contains all the code used for SAS.

```
*question 1;
*this data set was clean in R and imported to SAS using the wizard,
Due to errors;
*Scatterplot of Al in function Group;
proc gplot  data = sheet;
title "Values of Aluminium in the different soil types";
plot Al * Group;
RUN;

*Scatterplot of Ca in function Group;
proc gplot  data = sheet;
title "Values of Calcium in the different soil types";
plot Ca * Group;
RUN;

*Scatterplot of K in function Group;
proc gplot  data = sheet;
title "Values of Potassium in the different soil types";
plot K * Group;
RUN;

*Scatterplot of Mg in function Group;
proc gplot  data = sheet;
title "Values of Magnesium in the different soil types";
plot Mg * Group;
RUN;

*Scatterplot of Sc in function Group;
proc gplot  data = sheet;
title "Values of Scamdium in the different soil types";
plot Sc * Group;
RUN;


*Boxplot of Al in function Group;
Proc SGPLOT data = sheet;
VBOX Al
/ category = Group;
title "Values of Aluminium in the different soil types";
RUN;

*Boxplot of Ca in function Group;
Proc SGPLOT data = sheet;
VBOX Ca
/ category = Group;
title "Values of Calcium in the different soil types";
RUN;

*Boxplot of K in function Group;
Proc SGPLOT data = sheet;
VBOX K
/ category = Group;
title "Values of Potassium in the different soil types";
```

```
RUN;

*Boxplot of Mg in function Group;
Proc SGPLOT data = sheet;
VBOX Mg
/ category = Group;
title "Values of Magnesium in the different soil types";
RUN;

*Boxplot of Sc in function Group;
Proc SGPLOT data = sheet;
VBOX Sc
/ category = Group;
title "Values of Scandium in the different soil types";
RUN;

*question2;
*Correlations for 5 pairs of elements: 1.Ba and Eu, 2.Zn and Sr, 3.Mo and Ga,
4.Mo and Ba,5.Mo and Ca;
Proc Corr data=sheet;
var Ba;
with Eu;
RUN;

Proc Corr data=sheet;
var Zn;
with Sr;
RUN;

Proc Corr data=sheet;
var Mo;
with Ga;
RUN;

Proc Corr data=sheet;
var Mo;
with Ba;
RUN;

Proc Corr data=sheet;
var Mo;
with Ca;
RUN;
```

Appendix C

This appendix contains all visual representations of the SAS code.

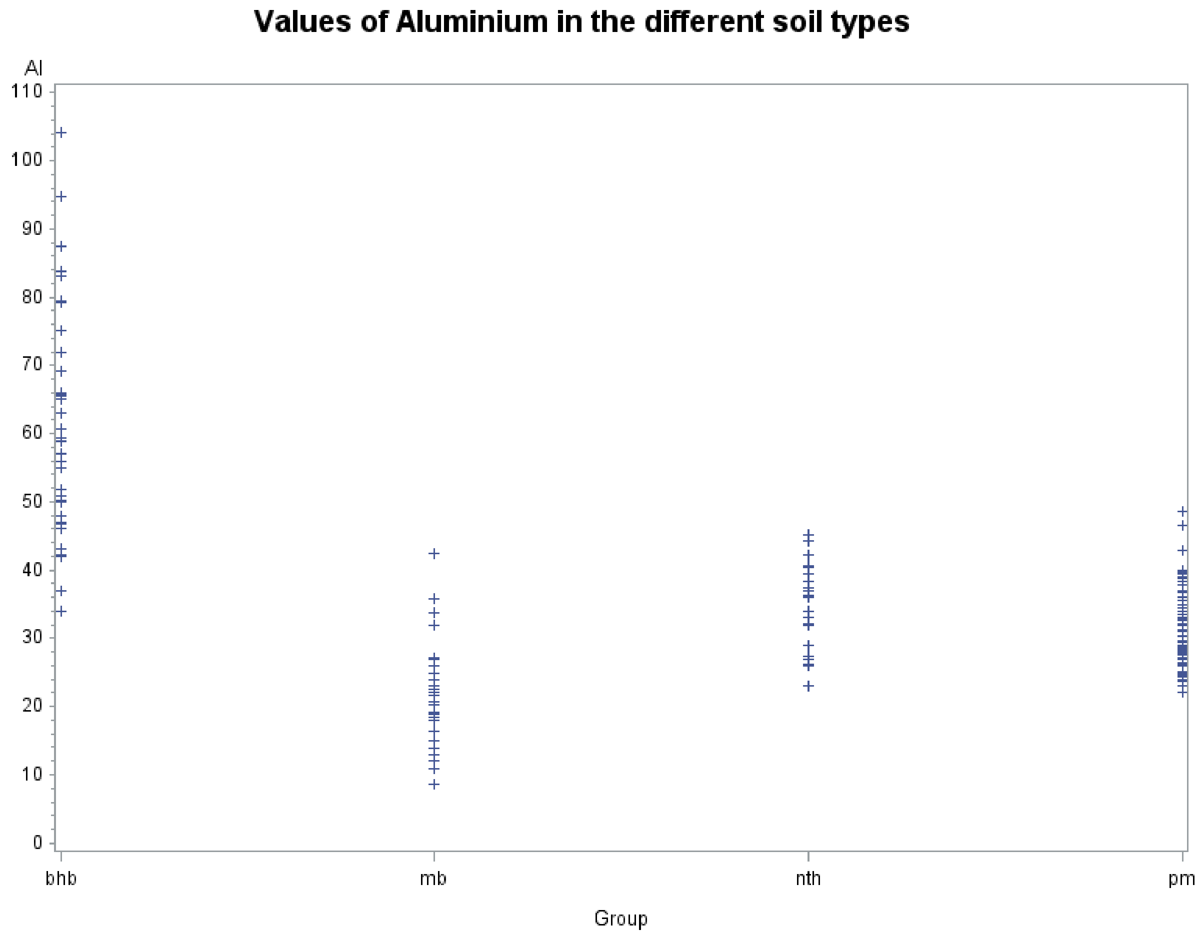

Figure 2b 1- Scatterplot, Al in function of Group

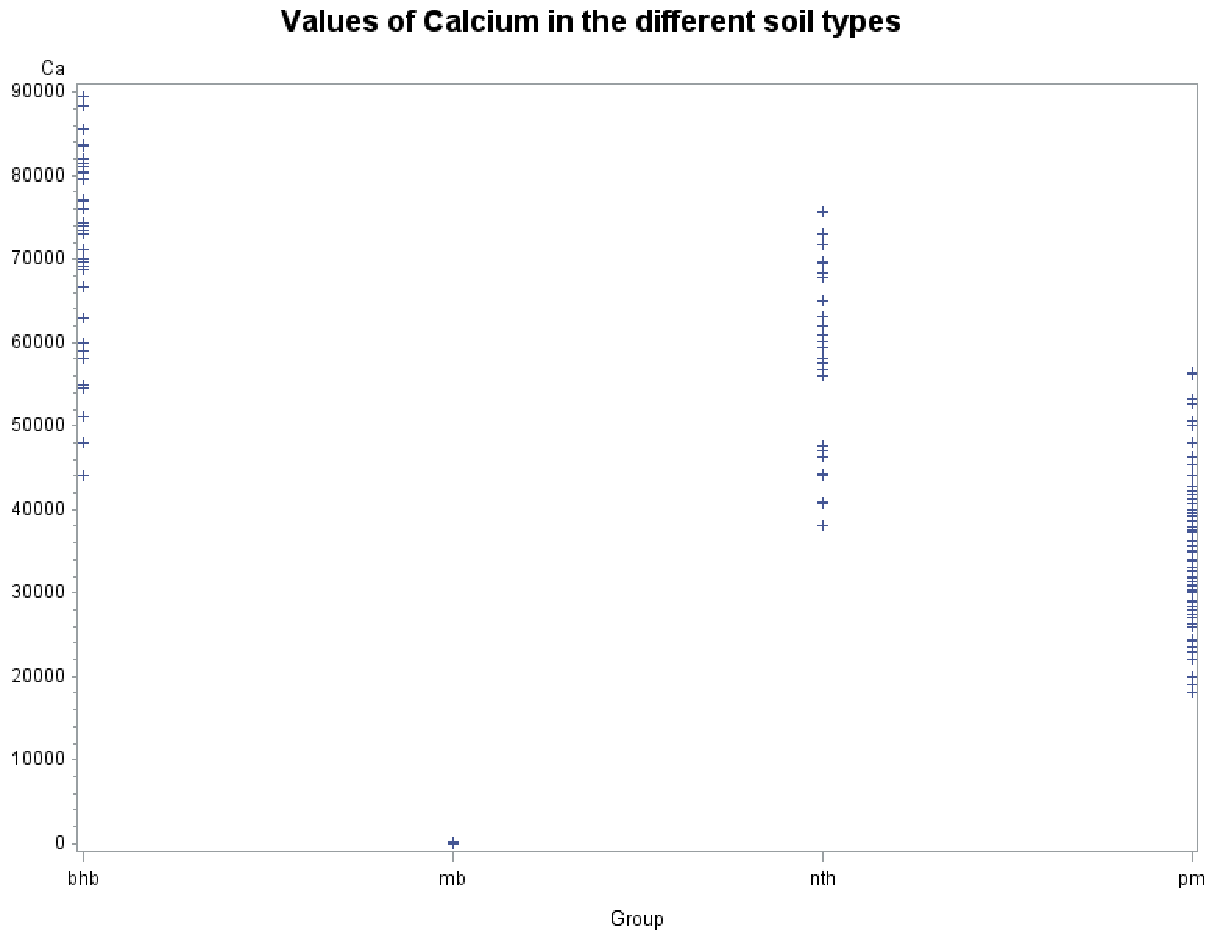

Figure 2b 2- Scatterplot, Ca in function of Group

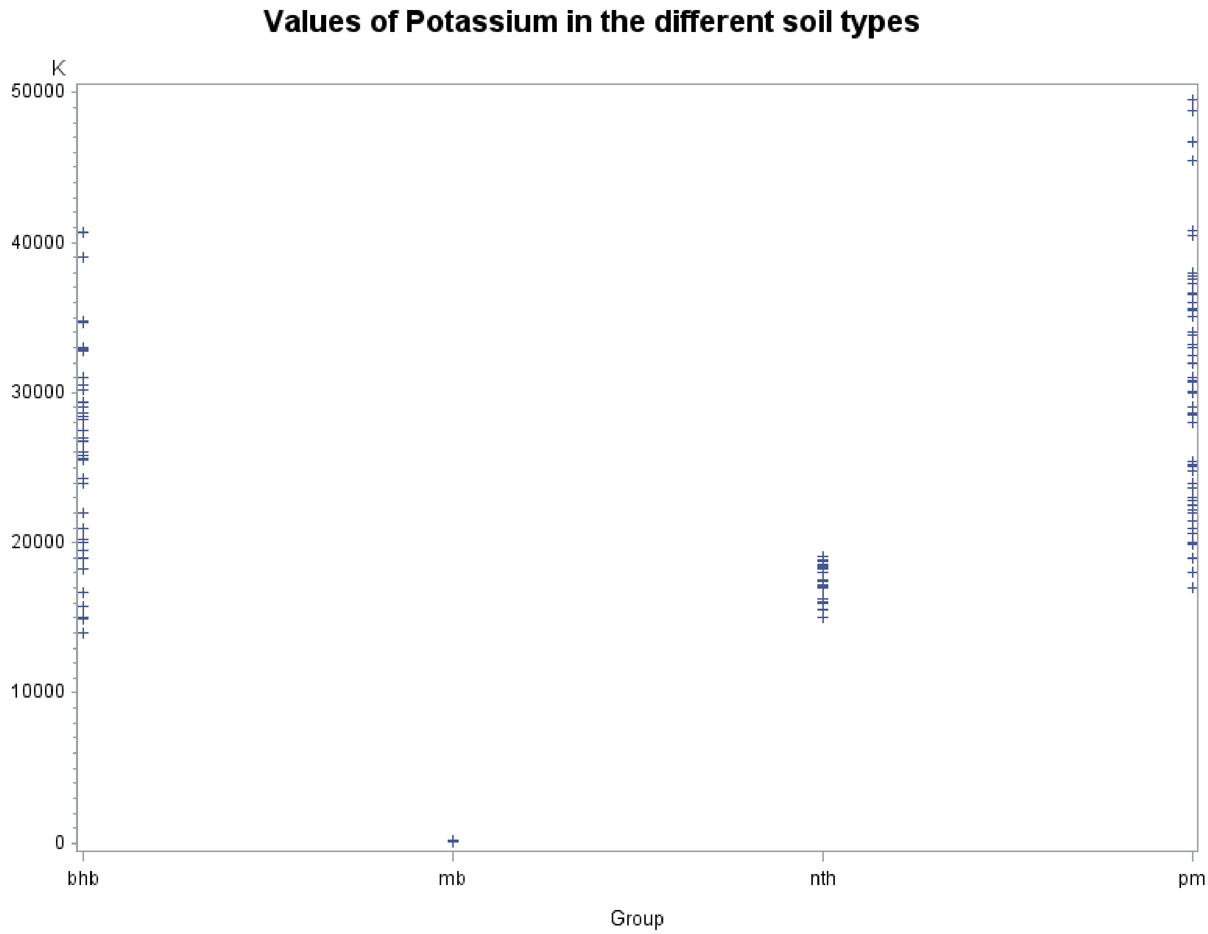

Figure 2b 3- Scatterplot, K in function of Group

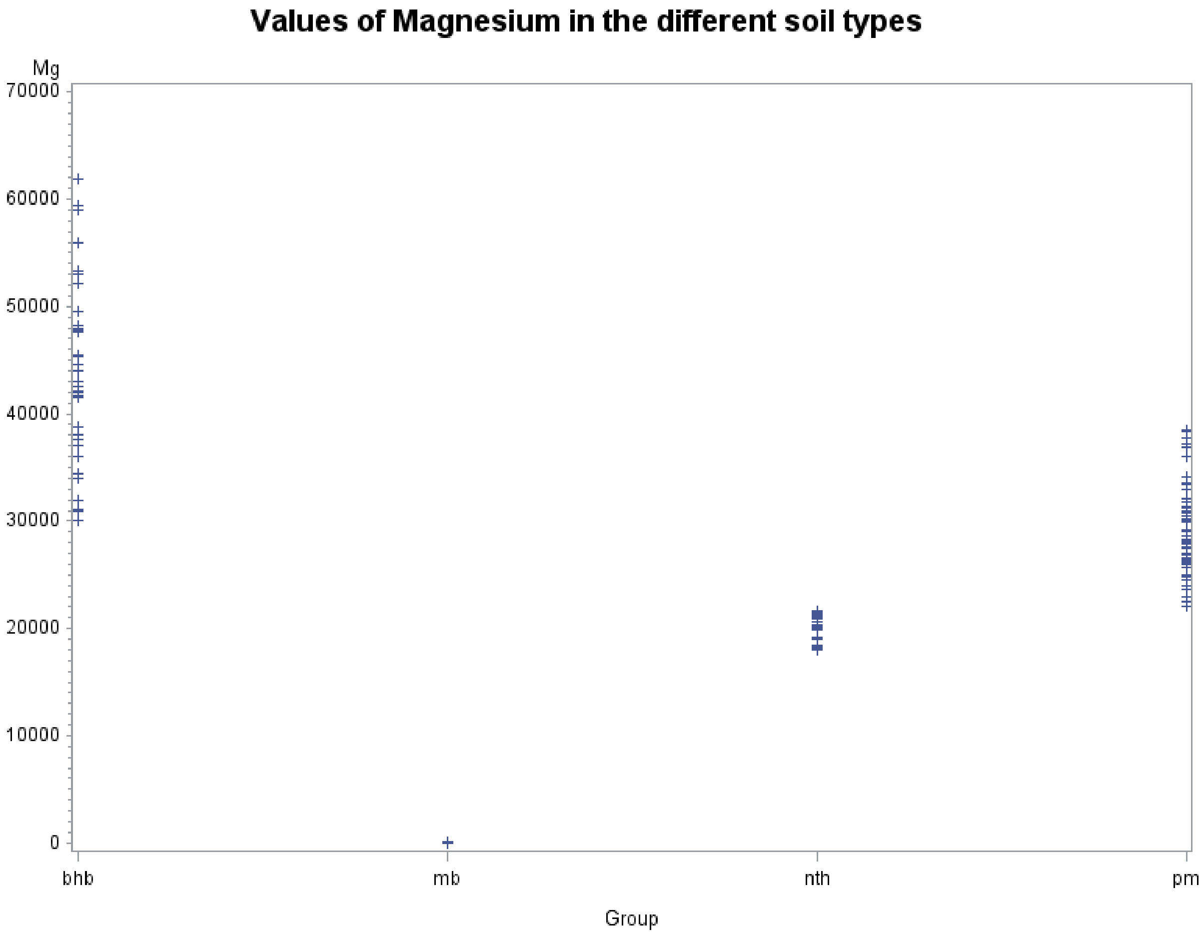

Figure 2b 4- Scatterplot, Mg in function of Group

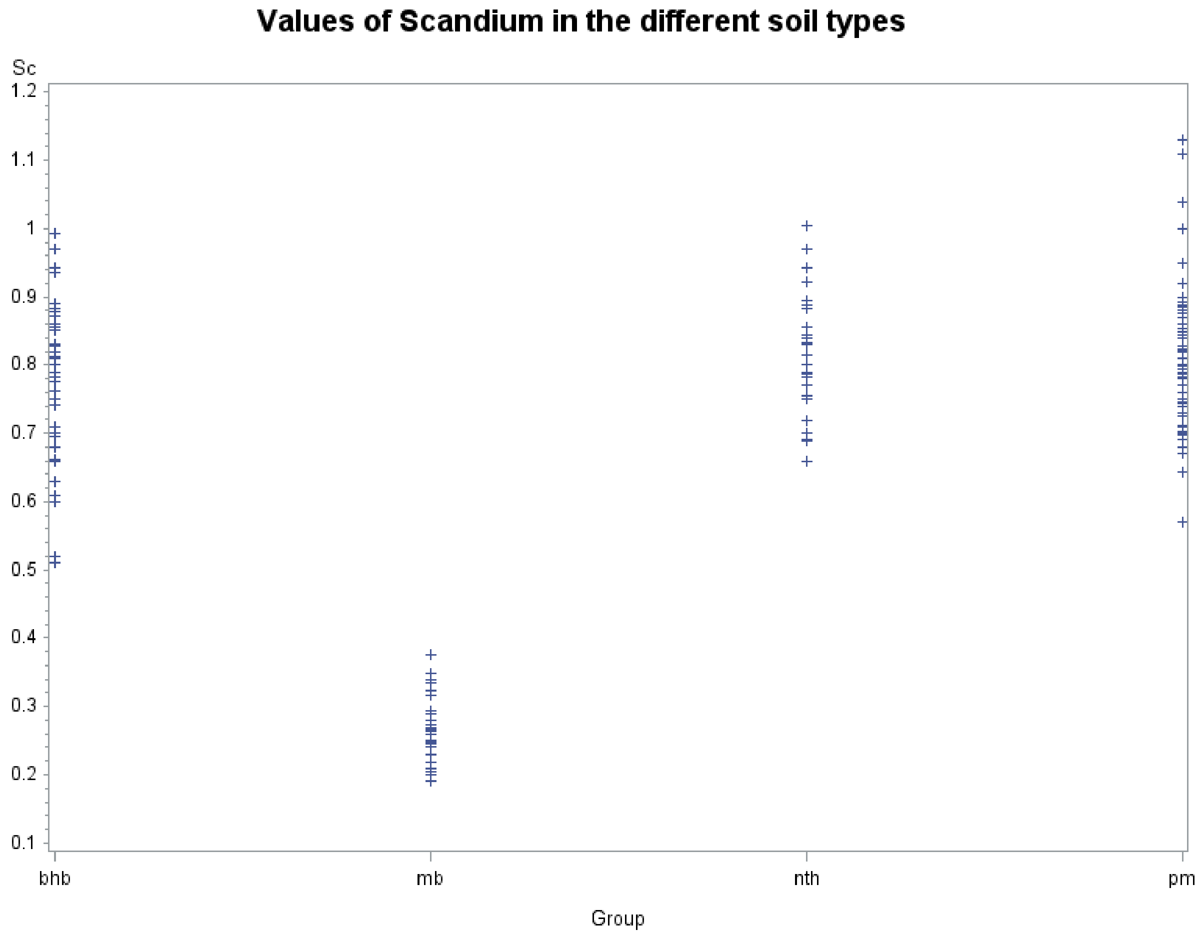

Figure 2b 5- Scatterplot, Sc in function of Group

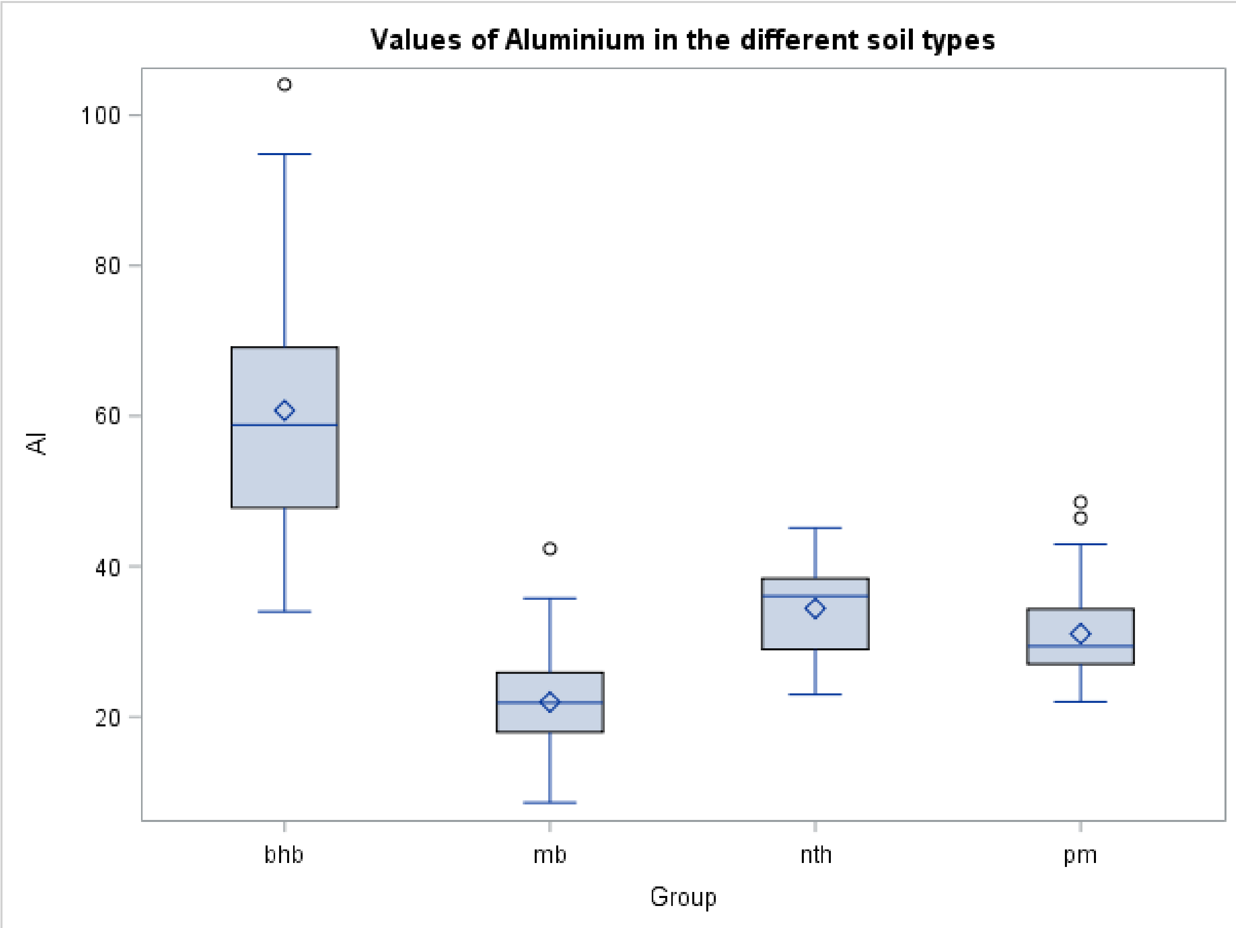

Figure 3a 1- Boxplot, Al in function of Group

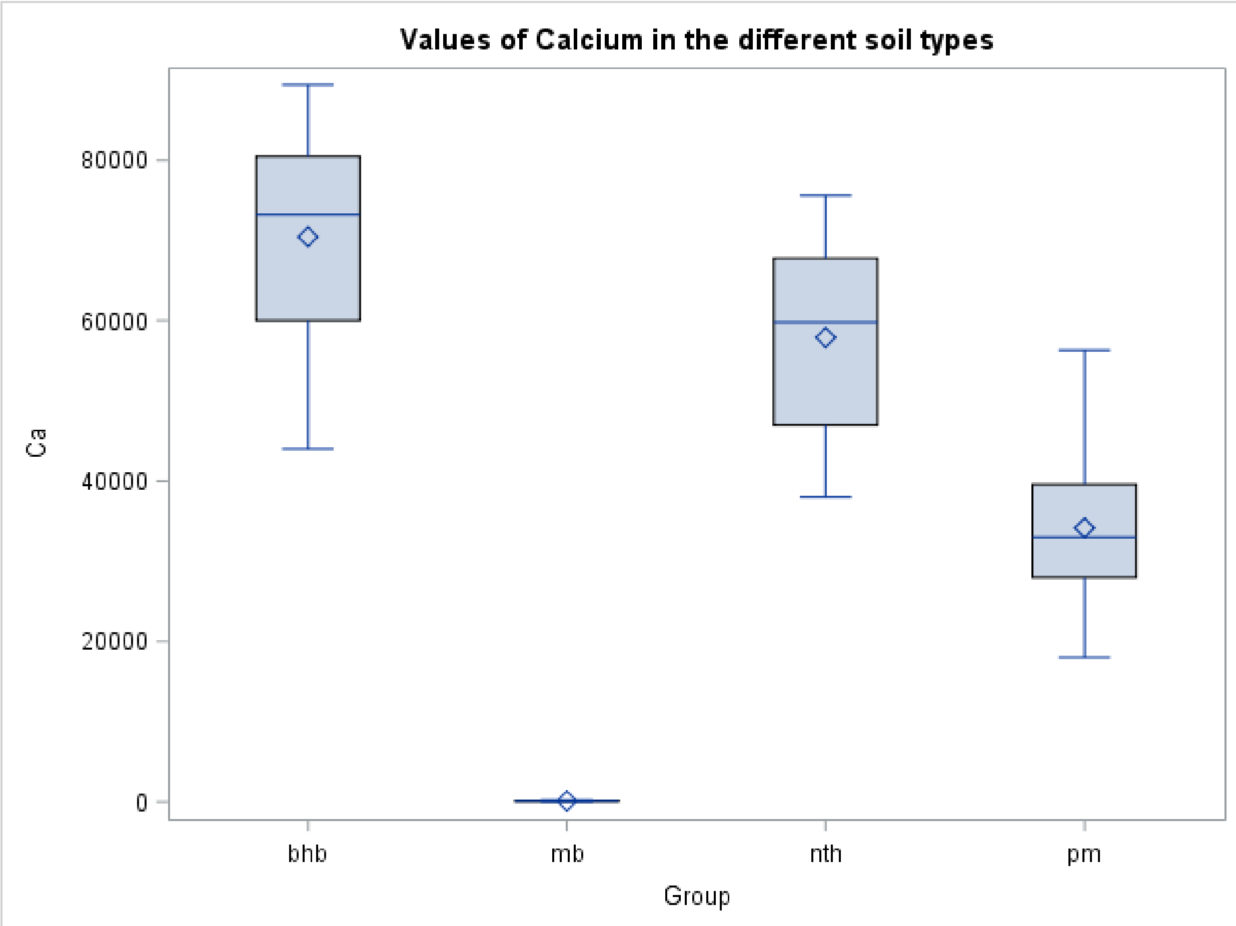

Figure 3a 2- Boxplot, Cal in function of Group

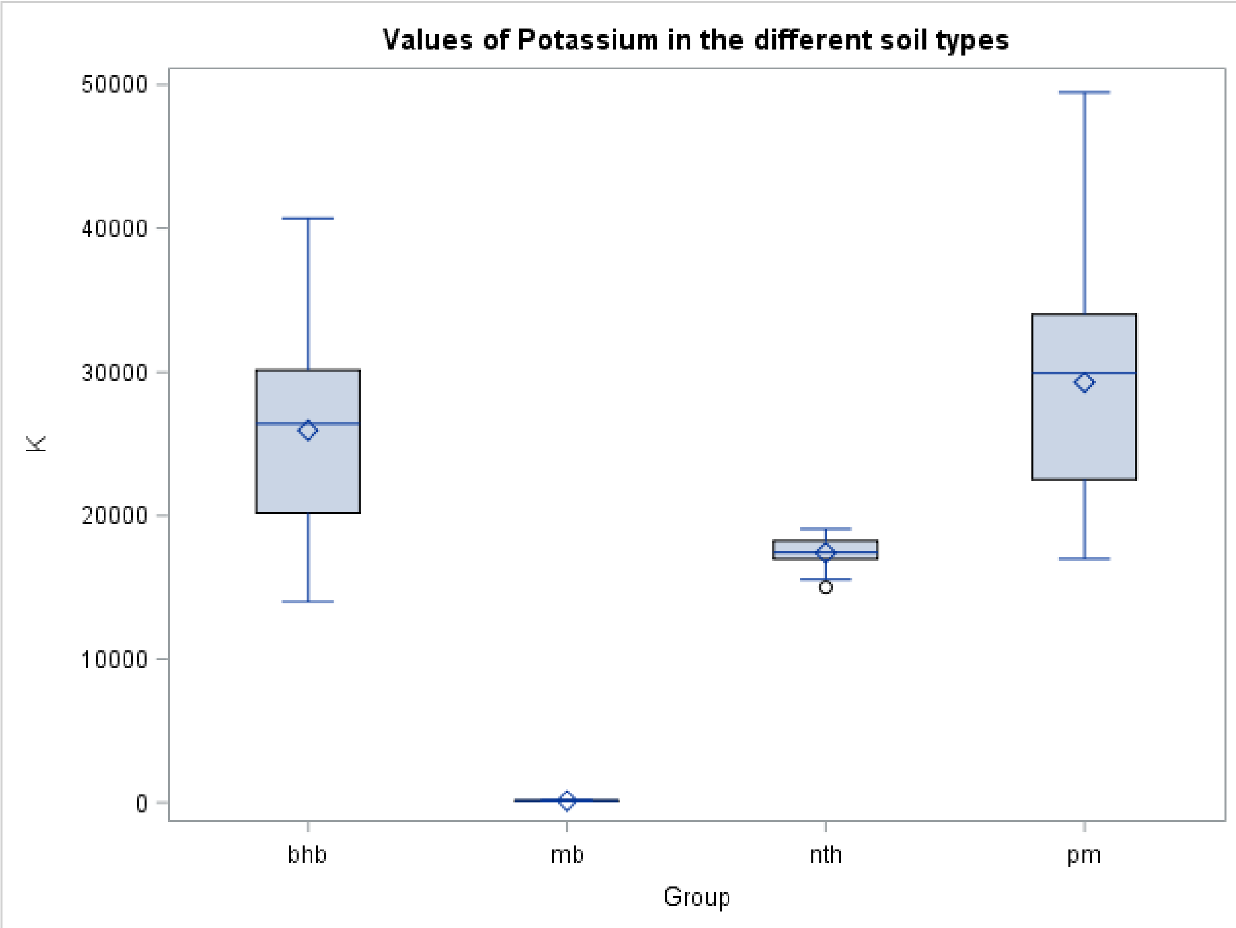

Figure 3a 3- Boxplot, K in function of Group

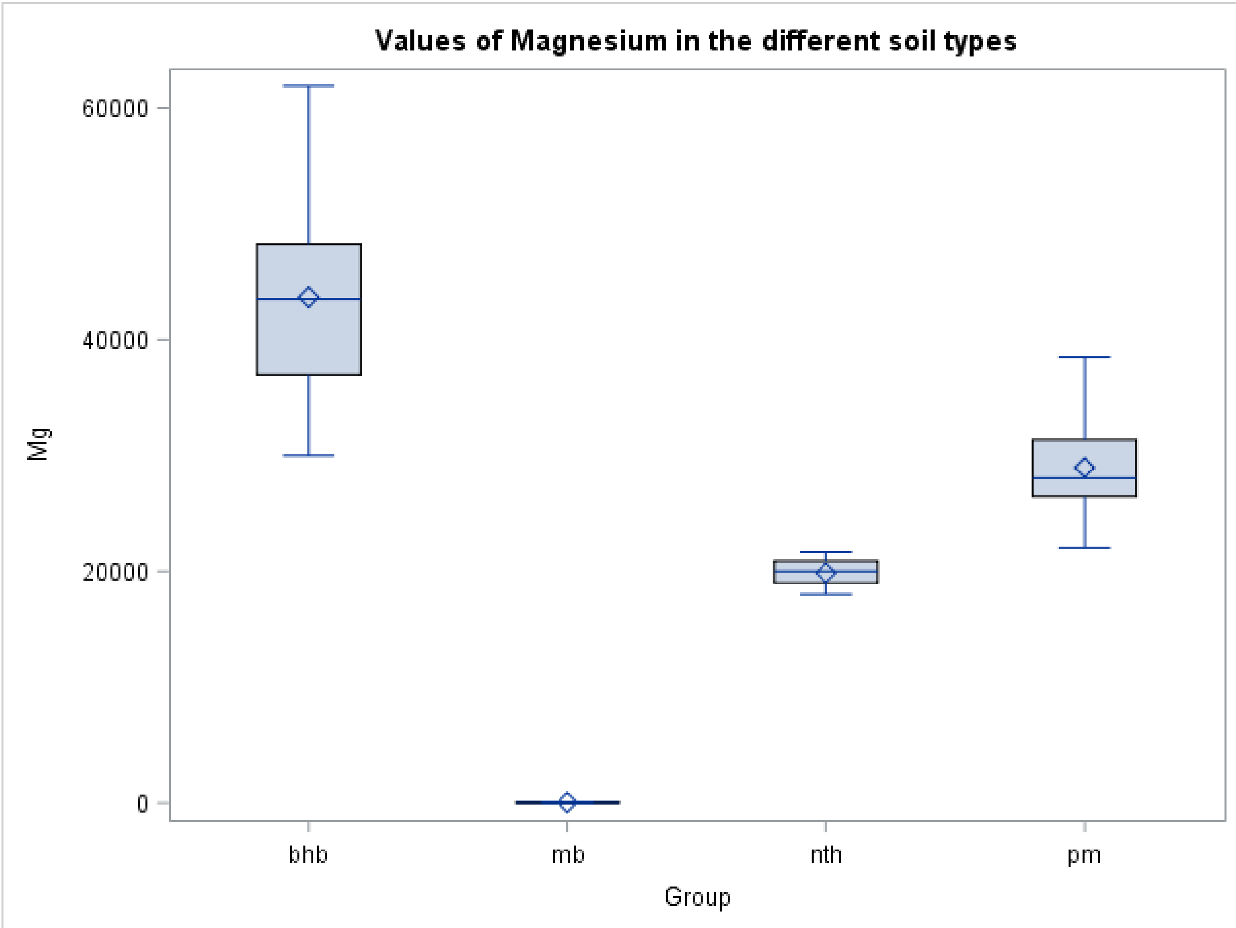

Figure 3a 4- Boxplot, Mg in function of Group

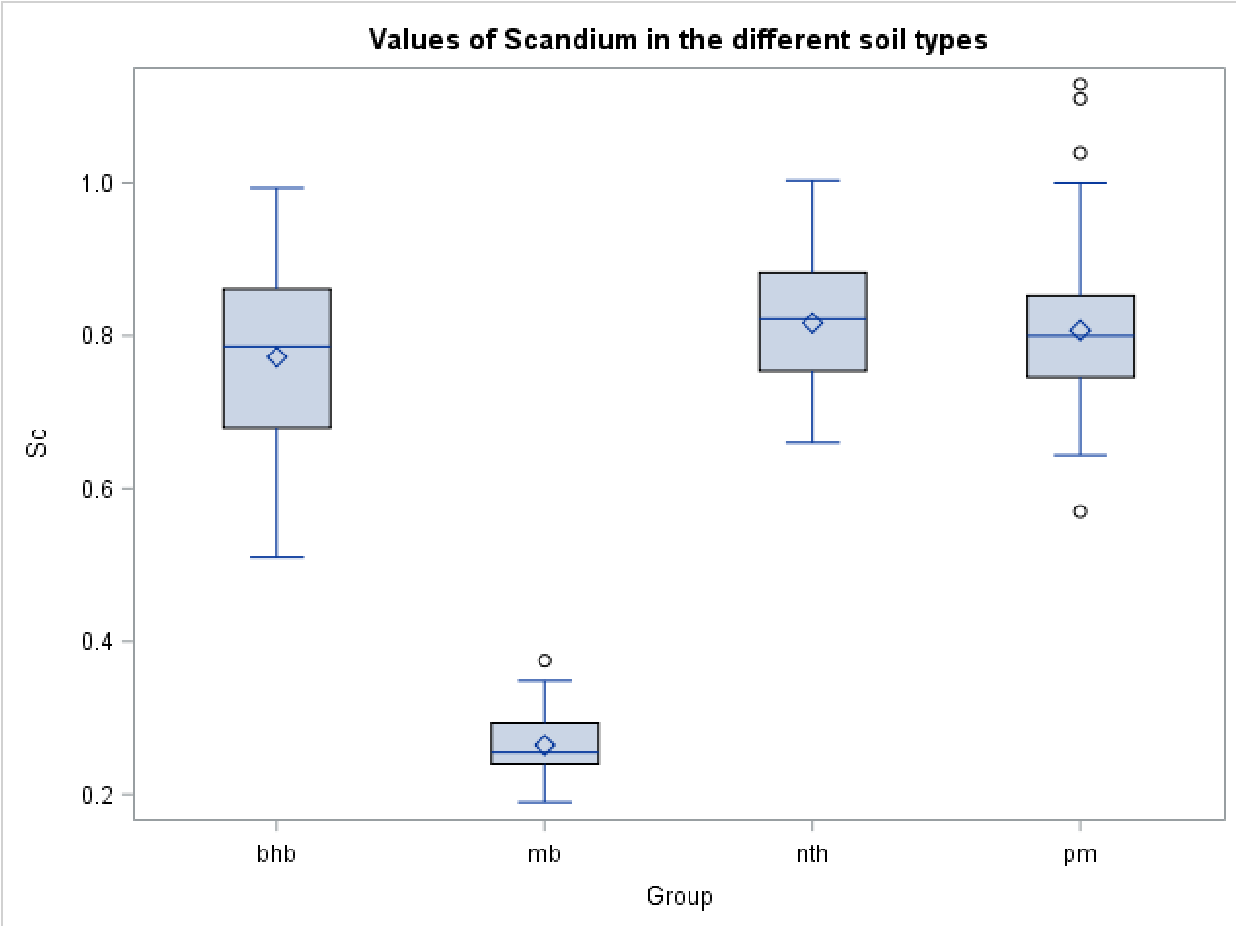

Figure 3a 5- Boxplot, Sc in function of Group

The CORR Procedure

1 With Variables:	Eu
1 Variables:	Ba

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Eu	163	0.02516	0.01735	4.10050	0.00120	0.07905
Ba	163	174.22035	197.40232	28398	0.37000	687.77468

Pearson Correlation Coefficients, N = 163 Prob > |r| under H0: Rho=0	
	Ba
Eu	0.91156 <.0001

Figure 5b- Correlation between Eu and Ba

The CORR Procedure

1 With Variables:	Sr
1 Variables:	Zn

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Sr	163	415.80540	286.94873	67776	0.10000	1105
Zn	163	568.61525	537.45531	92684	2.60000	2493

Pearson Correlation Coefficients, N = 163 Prob > |r| under H0: Rho=0	
	Zn
Sr	0.88994 <.0001

Figure 6b- Correlation between Zn and Sr

The CORR Procedure

1 With Variables:	Ga
1 Variables:	Mo

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Ga	163	4.33483	4.85924	706.57810	0.04000	18.46585
Mo	163	4.01547	3.99901	654.52121	0.07200	13.63285

Pearson Correlation Coefficients, N = 163 Prob > |r| under H0: Rho=0	
	Mo
Ga	-0.32090 <.0001

Figure 7b- Correlation between Mo and Ga

The CORR Procedure

1 With Variables:	Ba
1 Variables:	Mo

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Ba	163	174.22035	197.40232	28398	0.37000	687.77468
Mo	163	4.01547	3.99901	654.52121	0.07200	13.63285

Pearson Correlation Coefficients, N = 163 Prob > |r| under H0: Rho=0	
	Mo
Ba	-0.32843 <.0001

Figure 9b- Correlation between Mo and Ba

The CORR Procedure

1 With Variables:	Ca
1 Variables:	Mo

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Ca	163	40139	25850	6542667	28.00000	89422
Mo	163	4.01547	3.99901	654.52121	0.07200	13.63285

Pearson Correlation Coefficients, N = 163 Prob > |r| under H0: Rho=0	
	Mo
Ca	-0.01715 0.8280

Figure 10b- Correlation between Mo and Ca

Appendix D

Appendix D shows code in SAS that was attempted to be executed but was unsuccessful due to errors.

```
data sheet rename=(Sample.Name=Sample Name);
RUN;
```

References

Dalgaard, P (2008). *Introductory Statistics in R*. 2nd ed. New York, USA: Springer. pp99-108.

Ferguson, D et al. (2003). Arrests and convictions for cannabis related offences in a New Zealand birth cohort. *Drug and Alcohol Dependence*. 70 (1), pp53-63.

Randolph, K and Myers, L (2013). *Basic Statistics in Multivariate Analysis*. New York, USA: Oxford University Press. pp1-20.

Sheppard, D. (2000). *Cannabis Origin Determination Using Plant and Soil Elemental Profiles.* Available: https://researchspace.auckland.ac.nz/handle/2292/6042. Last accessed 11th Oct 2018.

R Studio. (2017). *Data Transformation with dplyr- CHEAT SHEET.*Available: https://www.rstudio.com/resources/cheatsheets/#dplyr. Last accessed 12th Oct 2018.

R Studio. (2016). *Data Visualization with ggplot2- CHEAT SHEET.*Available: https://www.rstudio.com/resources/cheatsheets/#ggplot2. Last accessed 12th Oct 2018.

SAS (2001). *Step-by-Step Programming with Base SAS® Software*. Cary, USA: SAS Publishing. pp461-470.

SAS (2001). *Step-by-Step Programming with Base SAS® Software*. Cary, USA: SAS Publishing. Pp233-240.

Smith, N. (2000). Criminals may rue pot from this plot. *The New Zealand Herald*. 1 (1), pp1.