PROJECT 1- BASIC TESTS AND INFERENCE

MT5762

180029243

Executive Summary

This report analises 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. Some packages were used, such as dplyr, tidyverse, ggplot, corrplot and ggpubr. There were also tests carried out, such as Tukey and t-test. Among other facts, these tests showed that the p-value is close to zero and that the data is chi-squared distributed. 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 code for this report is in Appendix A and complementary images are in Appendix B.

Analysis

Data Set Overview

The data set used for this report records the elemental composition of cannabis leaves across different soil types. There are 40 variables in this data set, the first of which is Sample Name, with 56 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 (Figs 1-3).

Methodology

The software used for this report is R Studio version 3.5.1.. Additional packages will be used, namely ggplot2, tidyverse, dplyr, corrplot and ggpubr.

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 44000 to a maximum of 81000 units. Potassium has a mean of 22615 and Magnesium of 39230. These two elements also deviate considerably less in value

than Calcium. Comparatively to the other elements in this soil type, the values of Aluminium and Titanium are miniscule, with means of 51.9 and 5.1 respectively (Figs 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 114 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, 9184 units, but the opposite for Potassium and Magnesium, 1092 and 1130, respectively. The values of the mean and standard deviation for Aluminium and Titanium 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 5000 units. Titanium and Aluminium remain low in this soil type.

As described above and shown at Figure 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.

	Element	Group	mean	sd
1	Al	bhb	51.923077	14.395957
2	Al	mb	17.660000	7.327301
3	Al	nth	31.777778	5.142416
4	Al	pm	28.666667	4.887132
5	Ca	bhb	63615.384615	11586.906846
6	Ca	mb	86.700000	48.256376
7	Ca	nth	54111.111111	9184.830489
8	Ca	pm	30208.333333	9069.677942
9	K	bhb	22615.384615	5140.338211
10	K	mb	114.800000	22.518141
11	K	nth	16777.777778	1092.906421
12	K	pm	25500.000000	6467.309097
13	Mg	bhb	39230.769231	8197.091417
14	Mg	mb	27.410000	21.286326
15	Mg	nth	19444.444444	1130.388331
16	Mg	pm	26958.333333	3196.182052
17	Ti	bhb	5.100000	1.318459
18	Ti	mb	0.946000	1.201473
19	Ti	nth	7.577778	1.622327
20	Ti	pm	8.850000	1.446735

Figure 1- Table, Mean and Standard Deviation of Each Element

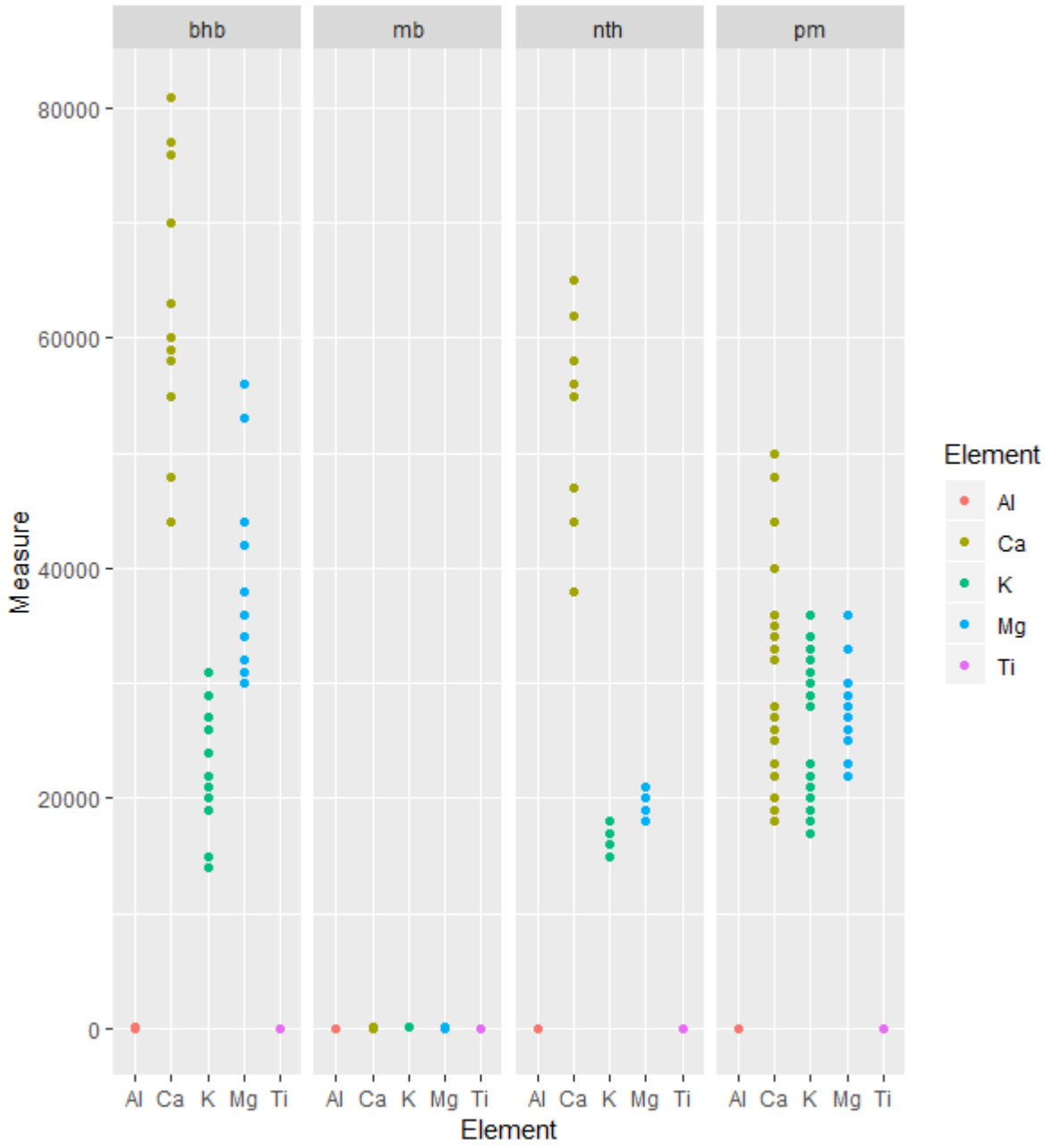

Figure 2- Scatterplot, elements in function of soil types.

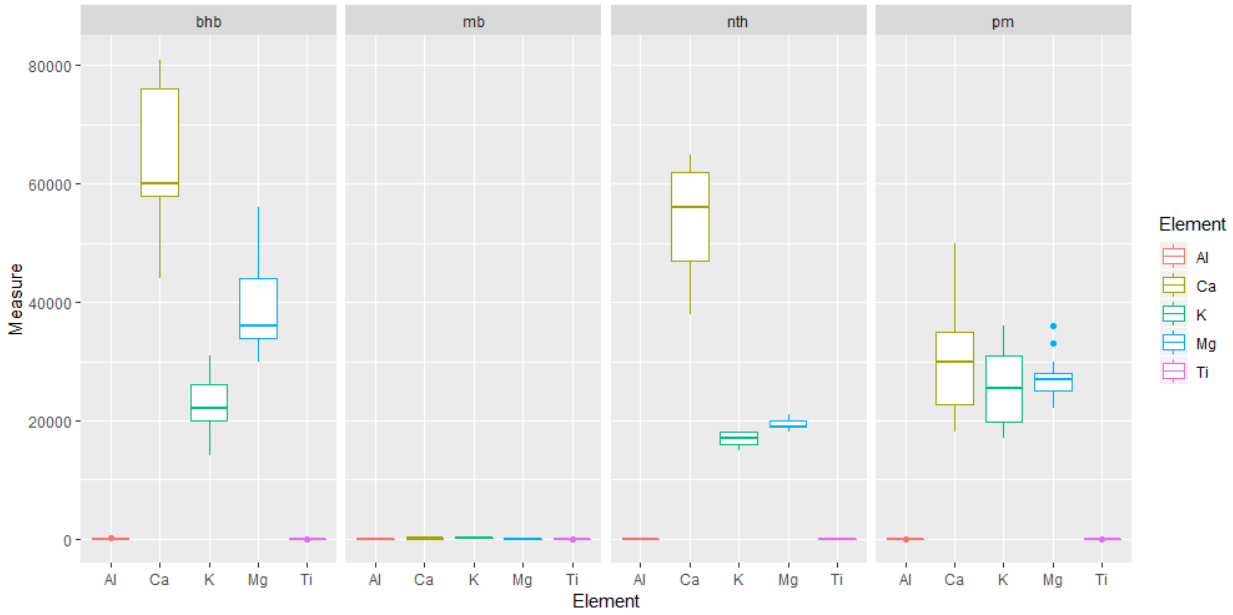

Figure 3- Boxplot, elements in function of soil types.

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 Gallium show an almost perfect correlation, 0.985. The same phenomenon is seen with the pair of elements Calcium and Strontium and their 0.982 correlation. On the other hand, Barium and Molybdenum have a negatively correlated, as well as Europium and Molybdenum. Elements such as 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 291 cases of correlation between pairs of elements, or 42% of all possible pairs.

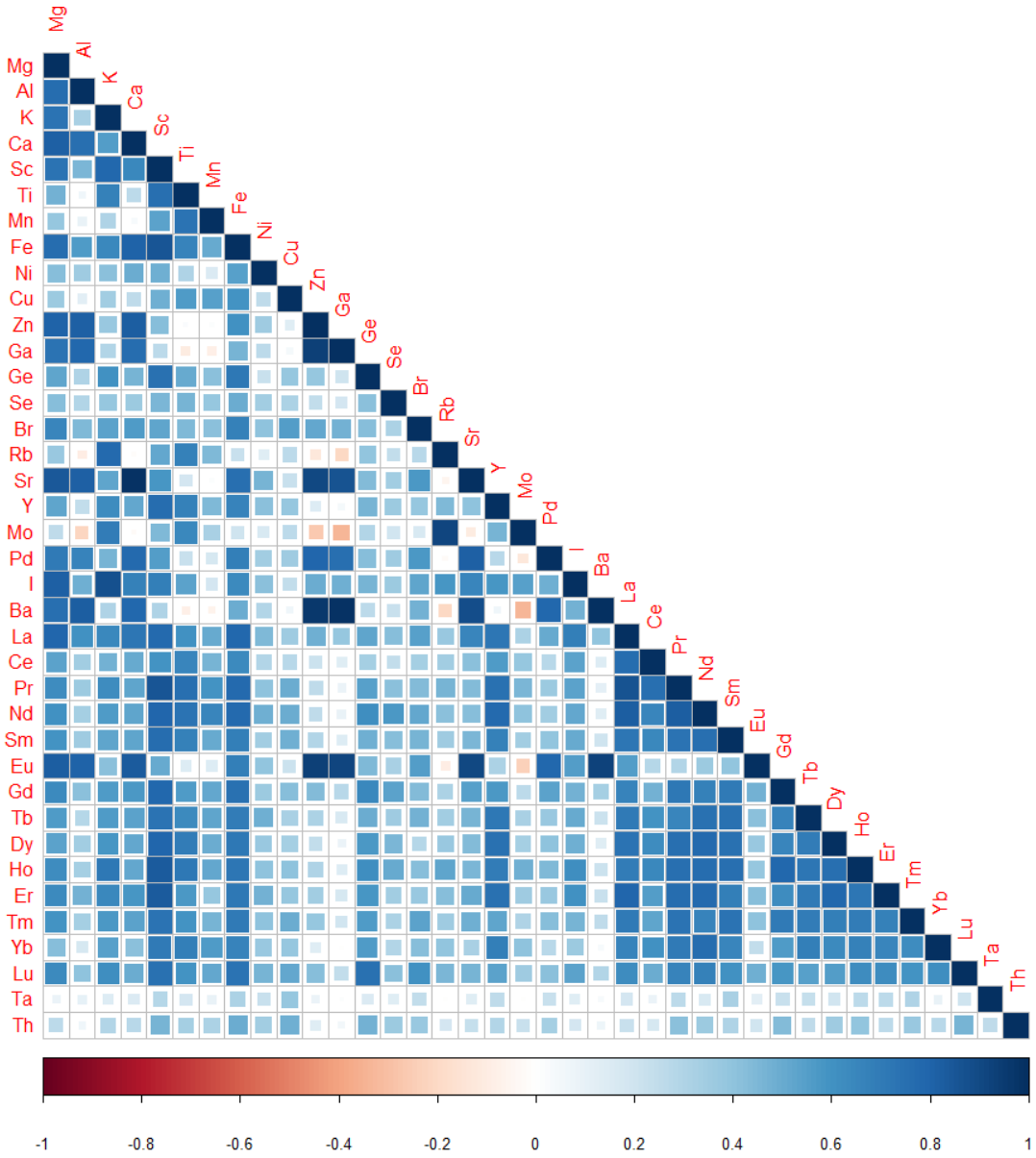

Figure 4- Correlation Plot, Correlation between the elements

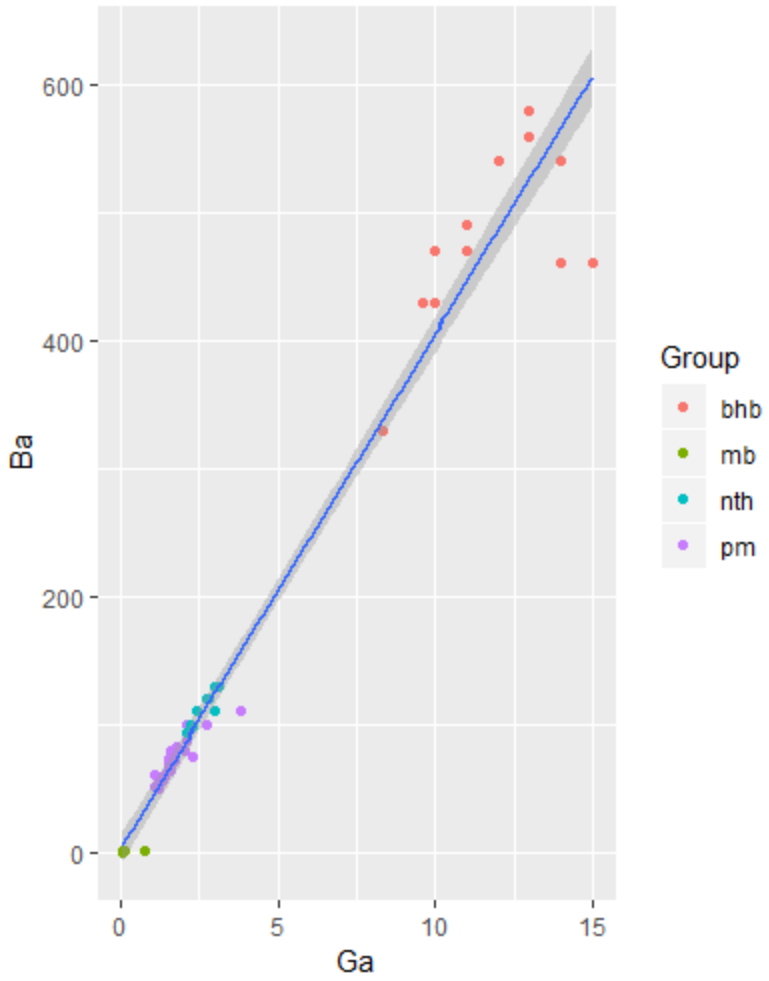

Figure 5- Scatterplot, Ba in function of Ga

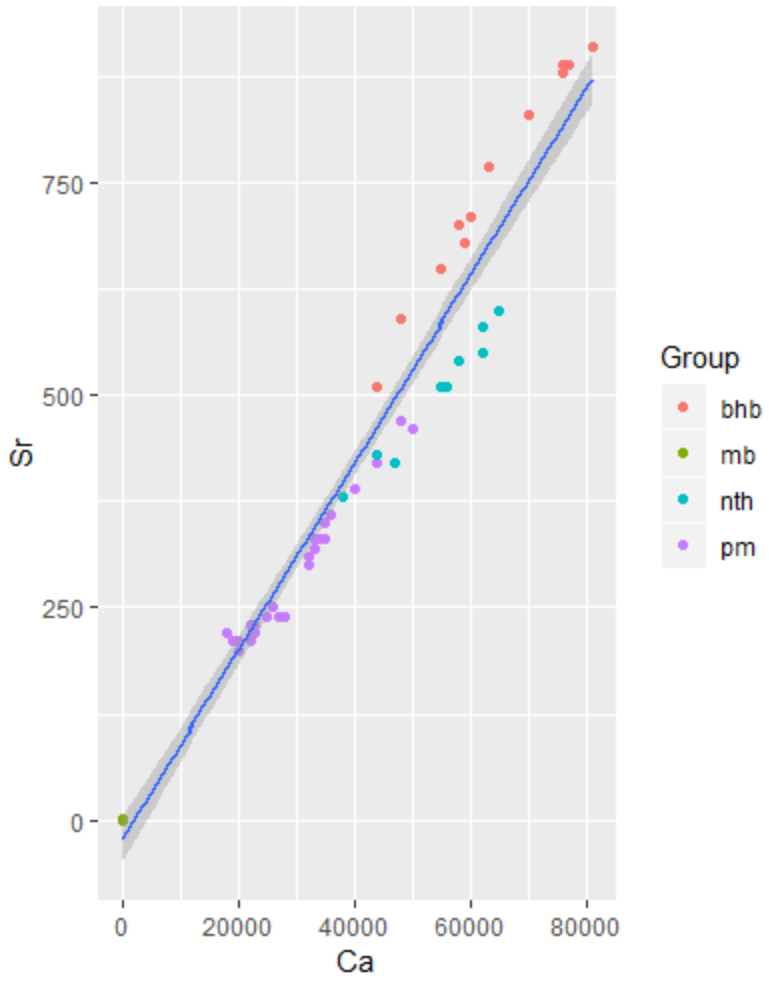

Figure 6- Scatterplot, Sr in function of Ca

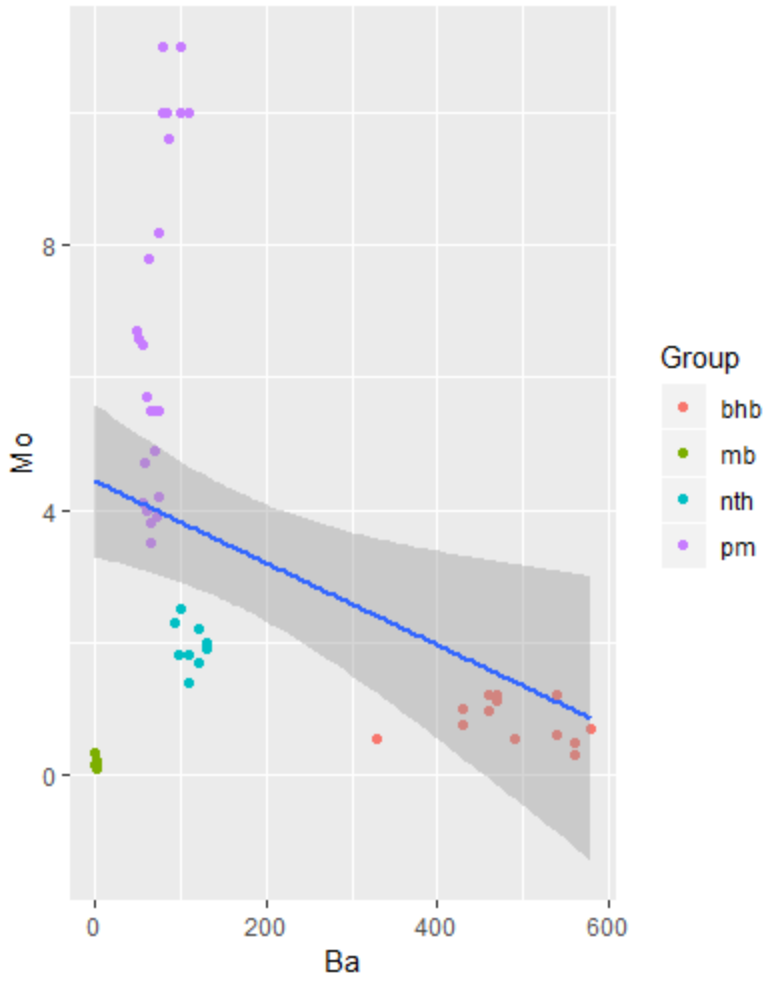

Figure 7- Scatterplot, Mo in function of Ba

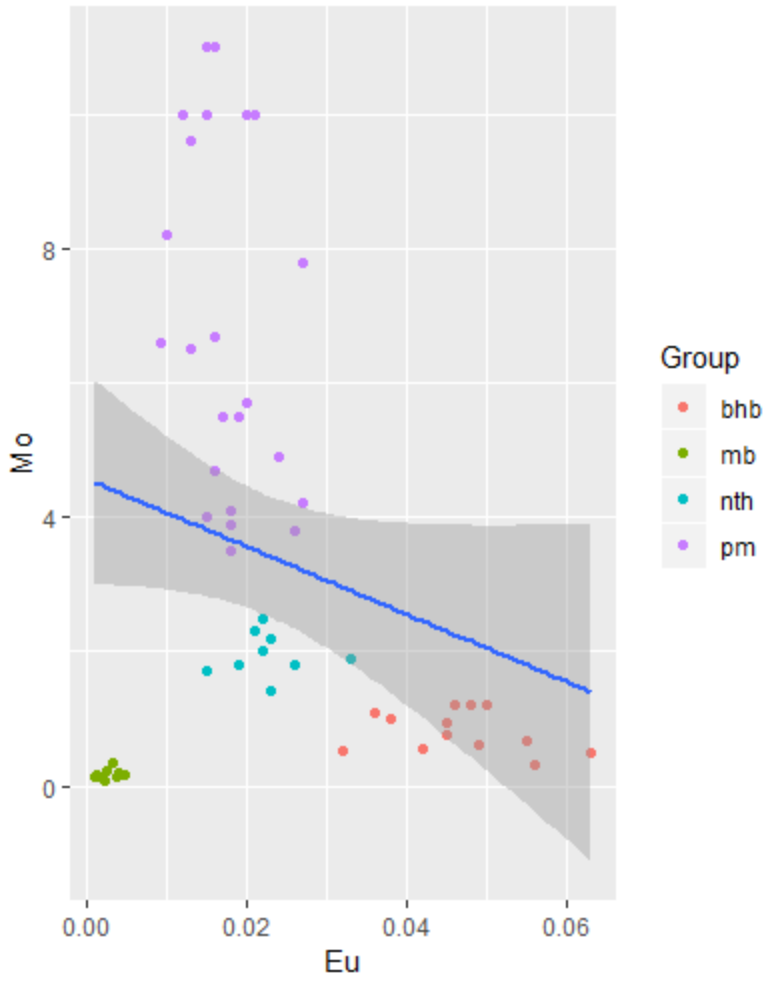

Figure 8- Scatterplot, Mo in function of Eu

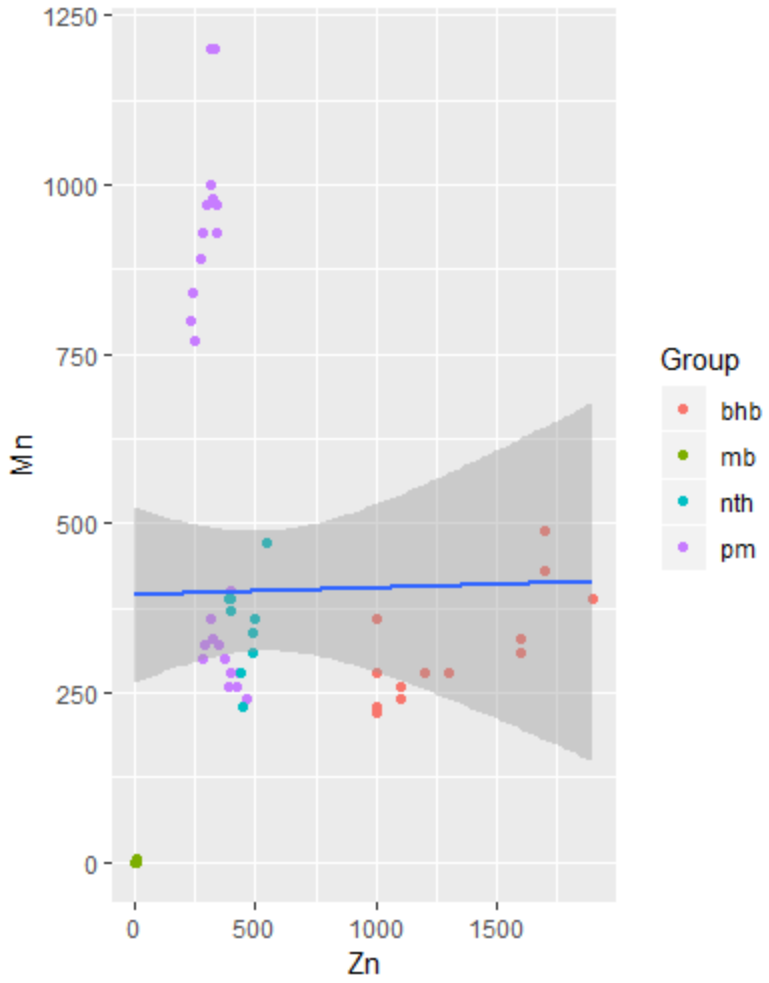

Figure 9- Scatterplot, Mn in function of Zn

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).

Tests

Tests were carried out for three elements: Magnesium, Potassium and Calcium. The tests ANOVA, Tukey, T-test, Wilcox and Normality test were used for this data set. Complementary figures for these tests are in Appendix B.

The null hypothesis, Ho for this data set is that there are no differences between soil types in their composition. Extreme changes in means for the same element in different soils seems to suggest otherwise, for example the mean of Magnesium in Mission Bay is 27.4 and in the Potting Mix is 26958.3. The t-test for Magnesium has revealed that the p-value is less than 2.2e^-16. As p-value < (0.05), the null hypothesis is rejected and it is argued with 95% confidence that there is a difference between the soil types' composition. The 95% confidence interval is 20142.14 and 27439.08 for Magnesium. Only 23 of the 56 data entries are within this 95% confidence interval for this element. The same conclusions can be drawn from other elements (Fig 18).

It is assumed that the values for the element measures are chi-squared distributed (Figure 11). This assumption can also be confirmed by the ANOVA tests. The F-value for the elements selected is high, for example Magnesium is 110,2. This high F-value strengthens the idea that the data has high variability and that the null hypothesis should be rejected.

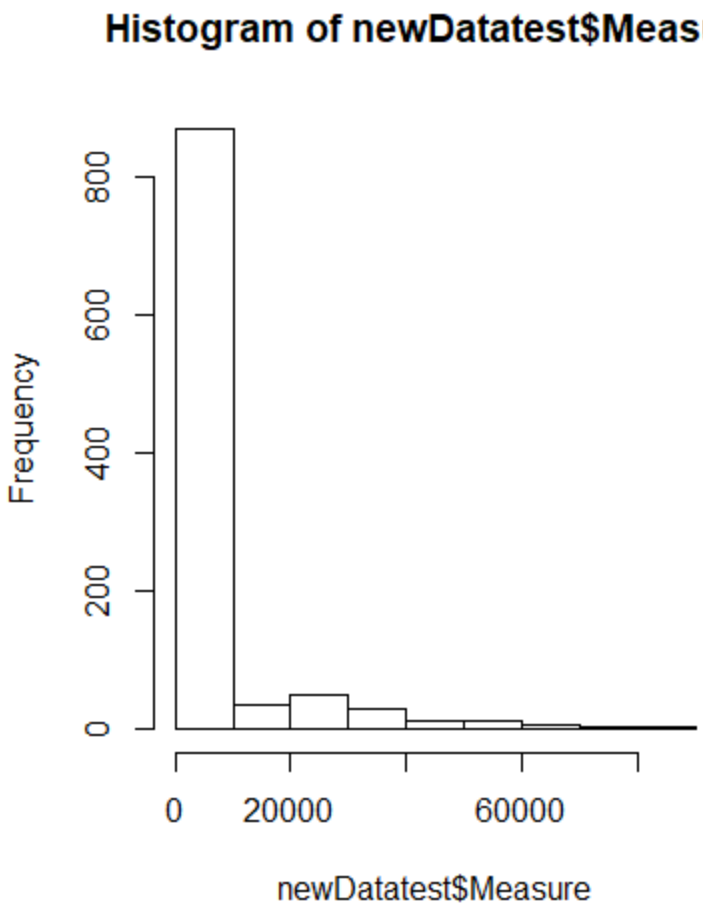

Figure 10- Histogram, Distribution of element measures

The Tukey test also confirms that there are indeed differences in the soil types (Fig 11) (Figs 15-18).

```
$`Group`
             diff         lwr         upr
mb-bhb   -39203.359 -44223.440 -34183.278
nth-bhb  -19786.325 -24961.641 -14611.009
pm-bhb   -12272.436 -16382.438  -8162.434
nth-mb    19417.034  13933.330  24900.739
pm-mb     26930.923  22438.793  31423.053
pm-nth     7513.889   2848.922  12178.856
             p adj
mb-bhb   0.0000000
nth-bhb  0.0000000
pm-bhb   0.0000000
nth-mb   0.0000000
pm-mb    0.0000000
pm-nth   0.0004635
```

Figure 11- Tukey test, results

Normality tests allow for a visual representation of the data distribution. The density of these variables varies significantly. Once again this disproves normality of the data (Appendix A) (Figs 12-14).

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.

The data tests have revealed that the p-value is extremely close to zero, smaller than α, therefore, the null hypothesis is rejected; the values of the elements have failed normality and a chi-squared distribution is apparent; and the F-value is high, therefore there is a high variability of data values (Dalgaard, 2008)(Uboe, 2017)(Tukey, 1991).

Appendices

Appendix A

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

Question 1

```
#importing the data set “potplants”

potplants <- read.csv("C:/Users/José Baltazar/Downloads/potplants_MT5762.csv", header = TRUE)

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

test <-  potplants [c("Group", "Mg", "Al", "K", "Ca", "Ti")]

#”newData” is a new function arranges the data in “test”

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

#”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(newDatatest)
```

Question 2

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

test2 <-  potplants [c("Sample.Name", "Group", "Mg", "Al", "K", "Ca", "Ti")]


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


 newData2 %>% group_by(Element, Sample.Name)

#graphs to compare the measures of a pair of elements

#Ba and Mo

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

```
#Eu and Mo

eumo <- ggplot(data = potplants, aes(x= Eu, y= Mo)) + geom_point(data = potplants, aes(x = Eu, y = Mo, colour = Group)) + geom_smooth(method = lm)

#Ga and Ba

gaba <- ggplot(data = potplants, aes(x= Ga, y= Ba)) + geom_point(data = potplants, aes(x = Ga, y = Ba, colour = Group)) + geom_smooth(method = lm))

#Fe and Mn

femn <- ggplot(data = potplants, aes(x= Fe, y= Mn)) + geom_point(data = potplants, aes(x = Fe, y =Mn, colour = Group)) + geom_smooth(method = lm)

#Ca and Sr

casr <- ggplot(data = potplants, aes(x= Ca, y = Sr)) + geom_point(data = potplants, aes(x = Ca, y = Sr, colour = Group)) + geom_smooth(method = lm)

#correlation between the elements

corr1 <- cor(potplants[3:40])

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

#How many elements have a correlation smaller than -0.5 or greater than +0.5?

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

Tests

```
#anova test

  summary(aov(Ca ~ Group, data = potplants))

  summary(aov(K ~ Group, data = potplants))

  summary(aov(Mg ~ Group, data = potplants))

 #tukey test

  fm1 <- aov(Mg ~ Group, data = potplants)

  fm2 <- aov(K ~ Group, data = potplants)

  fm3 <- aov(Ca ~ Group, data = potplants)

  TukeyHSD(fm1, "Group")

  plot(TukeyHSD(fm1, "Group"))

  TukeyHSD(fm2, "Group")

  plot(TukeyHSD(fm2, "Group"))
```

```
TukeyHSD(fm3, "Group")
plot(TukeyHSD(fm3, "Group"))
#t.test
t.test(potplants$Mg)
t.test(potplants$K)
t.test(potplants$Ca)
#wilcox.test
wilcox.test(potplants$Mg)
wilcox.test(potplants$K)
wilcox.test(potplants$Ca)
#normality test
ggqqplot(potplants$Ca, main = "qqplot Ca")
ggqqplot(potplants$K, main = "qqplot K")
ggqqplot(potplants$Mg, main = "qqplot Mg")

ggdensity(potplants$Ca, main = "Density Ca")
ggdensity(potplants$K, main = "Density K")
ggdensity(potplants$Mg, main = "Density Mg")
```

Appendix B

This appendix consists of visual representations of the testing functions used for this report.

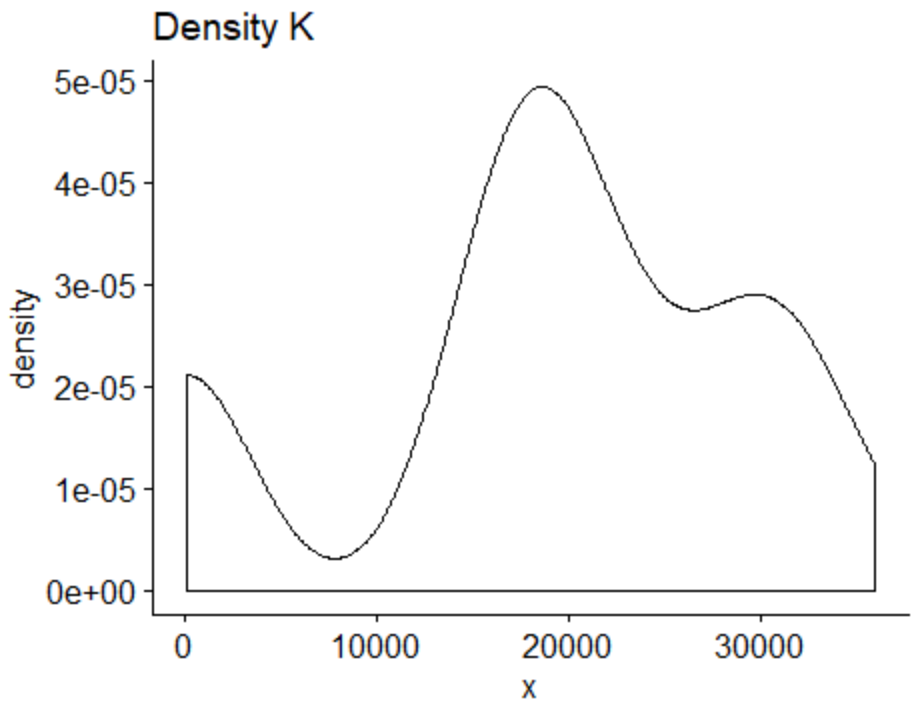

Figure 12- Density Plot, Potassium

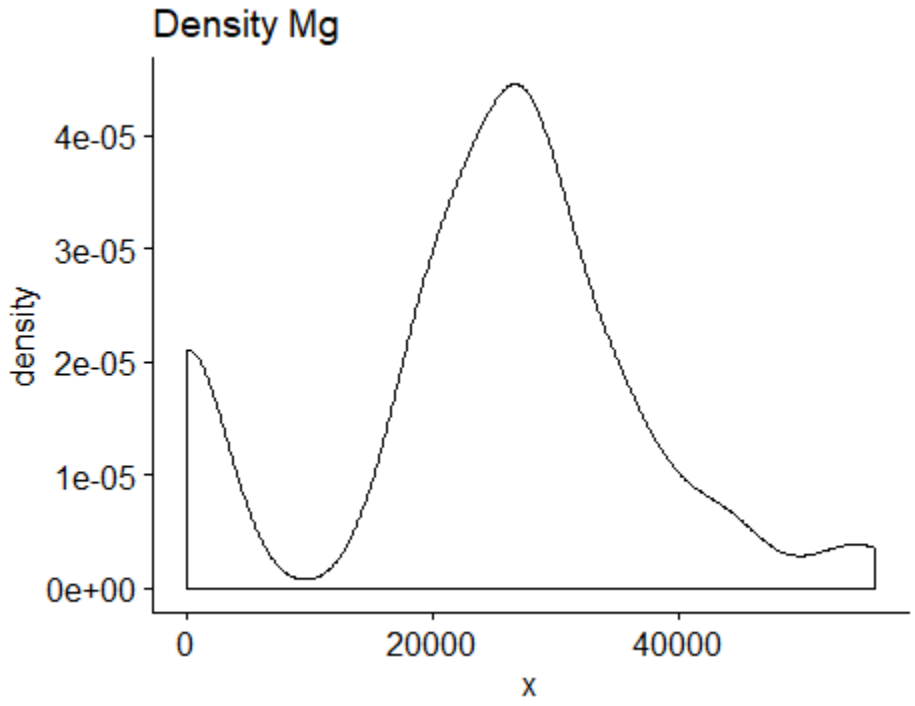

Figure 13- Density Plot, Magnesium

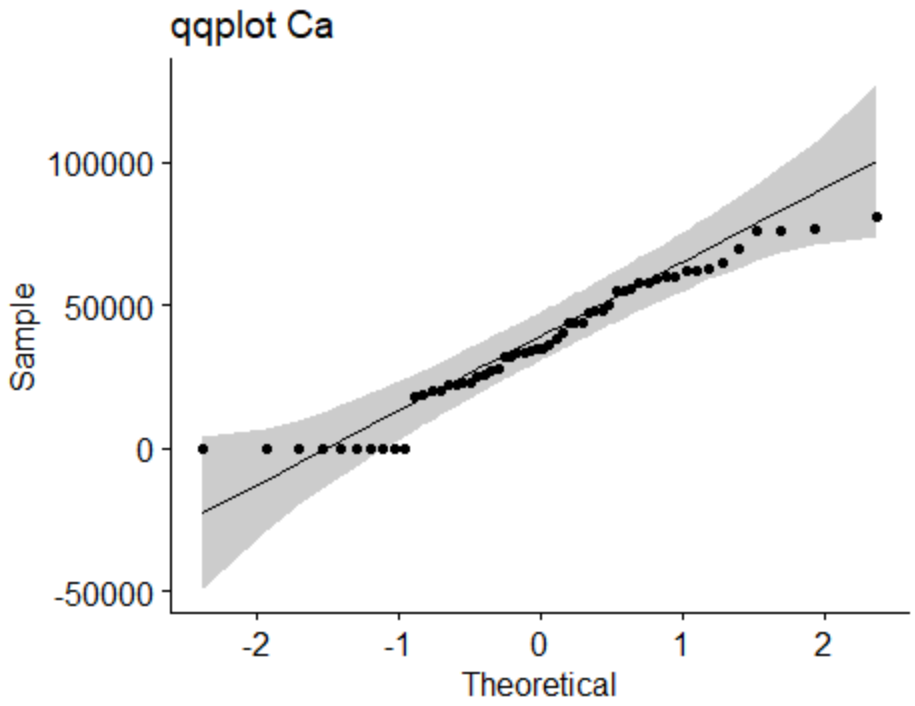

Figure 14- QQPlot, Calcium

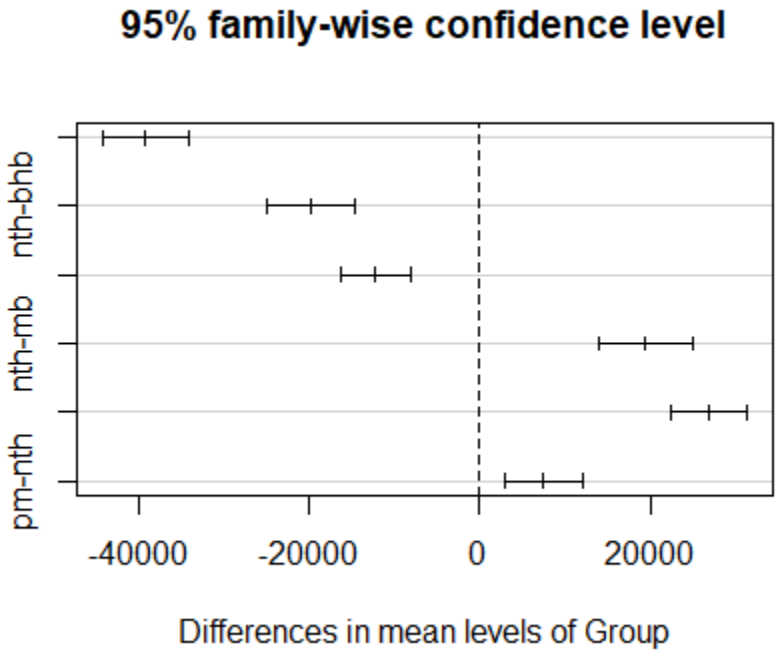

Figure 15- Tukey test, Magnesium

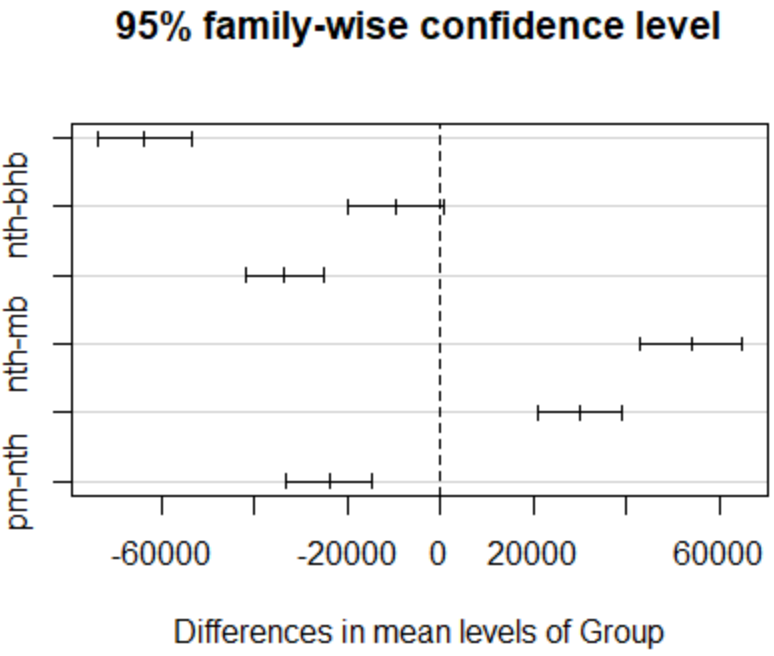

Figure 16- Tukey test, Calcium

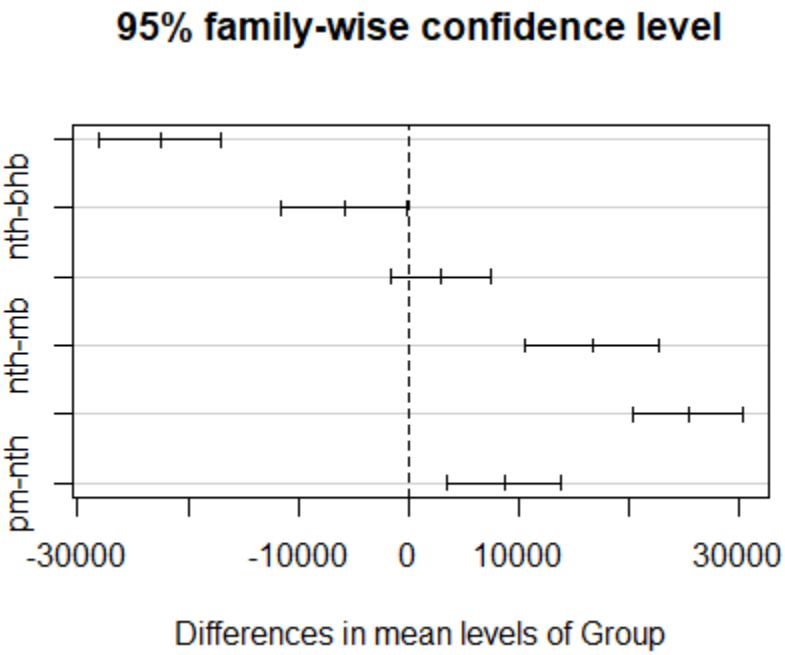

Figure 17- Tukey test, Potassium

```
        One Sample t-test

data:  potplants$Mg
t = 13.068, df = 55, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 20142.14 27439.08
sample estimates:
mean of x
 23790.61
```

Figure 18- T-test, Mg

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.

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.

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

Tukey, J. (1991). The Philosophy of Multiple Comparisons. *Statistical Science*. 6 (1), pp100-117.

Uboe, J (2017). *Introductory Statistics for Business and Economics- Theory, Exercises and Solutions*. Gewerbestrasse, Switzerland: Springer. pp70-81.

Uboe, J (2017). *Introductory Statistics for Business and Economics- Theory, Exercises and Solutions*. Gewerbestrasse, Switzerland: Springer. pp275-290.