University of
St Andrews | FOUNDED
1413

Essay cover sheet to be completed by student and attached to the front of essay.

STUDENT ID : _______180009169_____________________________

COURSE : _______MT5763 Software for Data Analysis___

ESSAY/PROJECT TITLE : _______Exploratory data analysis______________

NAME OF TUTOR(S) : _______ Carl Donovan ___________________________

SUBMISSION DEADLINE : ________2018/10/12____________________________

DATE SUBMITTED : ________2018/10/12____________________________

DECLARATION

- I confirm that I have read and understood the University's policy on plagiarism and I agree to participate in the courses provided by ELT to develop good academic practice and ways of citing and referencing the work of others.
- I confirm that this assignment is all my own work and that I have only had help from ELT staff when preparing it.
- I confirm that in preparing this piece of work I have not copied any other person's work, or any other piece of my own work.
- I confirm that this piece of work has not previously been submitted for assessment on another course.

___________________180009169__________________ (Student ID in place of signature)

___________________2018/10/12_________________ (date of submission)

1. Introduction

Statistics is not only a useful tool for scientific studies but also contribute to the process of prosecution. For instance, someone has explored the origin determination of cannabis plants with statistics, which can help police when prosecuting growers (Smith, 2000). This report is also focus on the data about cannabis plants. The data set used in this report contains information about the levels of thirty-eight elements in cannabis leaves which are grouped by four soil types which are potting mix (pm) and three locations about New Zealand. Three locations are bhb, mb, and nth, being: Blockhouse Bay (Auckland suburb), Mission Bay (Auckland suburb) and Northland (a northern region). It will first show there are differences in the elemental composition between Cannabis leaves in four groups, then prove that some elements are related in terms of their content levels in sampled leaves. Finally, this report will discuss whether the origin determination of cannabis can be made just from the elemental composition data. This investigation will use RStudio and SAS as the analysis packages. The outputs of RStudio will be show in the body of the report and the outputs will be mirrored in the appendix.

Before doing the exploratory analysis, cleaning data is conducted first to ensure the data are appropriate for analysis. The analysis process begins with summarizing the means and standard deviations, then using Analysis of Variance (ANOVA) and Tukey’s Honest Significant Difference (TukeyHSD) to achieve more details about the elemental composition differences. As for the relationships between elements, the Pearson correlation coefficient and scatter plots are used to analyze. Finally, the origin determination question will be discussed with some boxplots.

2. Clean the data

Four main things are done in this data cleaning stage, which are combing data, replacing repetitive data, converting data types, and dealing with missing values and obvious errors. Because the relationship between different samples is not important in this investigation, the first step after the data are read to RStudio is to combine all the data from three samples which are contained in different spreadsheets. Then all elements are gathered in one column and all values in another column, which could make the cleaning process more convenient because it can be easier to manipulate columns than rows. Noticing there are "potting mix" and "pm" in the Group column, which both refer to the same type of soil, so the "potting mix" is changed to "pm". In order to make the data appropriate for analysis, the types of data are also changed. For instance, the classes of "Group" is converted to factor. The factor refers to "a data structure used for fields that takes only predefined, finite number of values (categorical data)" (datamentor, n.d.). This change means RStudio will divide data into groups according to the value of Group column, which is important for the later analysis. Subsequently, there are 190 missing values are found and omitted. After these steps, the elements are spread to the first row, which revert the format like the beginning. The final step is to deal with the obvious errors. Boxplots has been drawn for each element in each group to check outliers. Boxplots have lines called whiskers, which indicate a range that contain most values. The values which are not in this range are outliers and will be plotted as individual points. Although there are some points beyond the whiskers, they are not deviate too far. It could be considered no demonstrably wrong value. Now, the data is clean and suitable for exploratory analysis.

3. Exploratory analysis of the data

3a) Different elemental composition between soil types

This investigation will not include analysis for all the elements. Instead, five elements are chosen, which are Ti, Ca, Ga, Ba and Zn. There are significant differences in the elemental composition of Cannabis leaves grown in different soil types. The table1 below shows the means and standard deviations for every element in each group. Intuitively, the elemental composition is different between groups and the difference is significant. However, this only gives people a sense but no strong evidence. Further on, the difference can be verified by using Analysis of Variance (ANOVA). ANOVA is a statistical test which can be used to test whether the means of more than two samples are equal. In this report, the ANOVA are applied to Ti and Ca elements. The most important information of an ANOVA is p-value. The elemental composition between soil types is more likely to be different if the p-value is smaller. In this case, the p-values for Ti and Ca both are less than 0.01, which is a strong evidence to prove the differences.

Table1. Summary of means and standard deviations

Elements Group	Ti (mean/sd)	Ca (mean/sd)	Ga (mean/sd)	Ba (mean/sd)	Zn (mean/sd)
bnb	5.86/1.48	70449.74/12127.61	12.72/2.33	518.04/74.90	1439.60/364.98
mb	1.50/1.32	110.12/55.72	0.27/0.25	0.98/0.39	5.37/1.63
nth	8.45/1.96	57888.48/11081.78	2.79/0.40	119.81/16.71	493.36/69.62
pm	9.61/1.71	34161.92/9470.33	2.06/0.67	80.69/19.23	362.19/67.80

3b) Related content levels between elements

This section will give evidences that there are some pairs of elements are related in terms of their levels in sampled leaves. Preliminarily, Pearson correlation coefficients are calculated to estimate which pairs are likely to be related (Table2). Pearson correlation coefficients is a common tool to measure the linear relationship between two variables. The value range of it is between -1 and 1. The signs show the directions of relationships, and the absolute values illustrate how strong the relationship are. According to the results, there are three pairs of elements (Ga ~ Ba, Ga ~ Zn, Ba ~ Zn) have correlation coefficients over 0.9. Therefore, they seem to be related in terms of their levels in the sampled leaves. This phase then followed by significance tests that all pairs have passed. Furthermore, figure1 contains some plots to visualize the correlations between elements.

Table2. Pearson Correlation coefficients between elements

	Ti	Ca	Ga	Ba	Zn
Ti	1.00	0.31	-0.06	-0.05	0.05
Ca	0.31	1.00	0.77	0.79	0.82
Ga	-0.06	0.77	1.00	0.98	0.94
Ba	-0.05	0.79	0.98	1.00	0.96
Zn	0.05	0.82	0.94	0.96	1.00

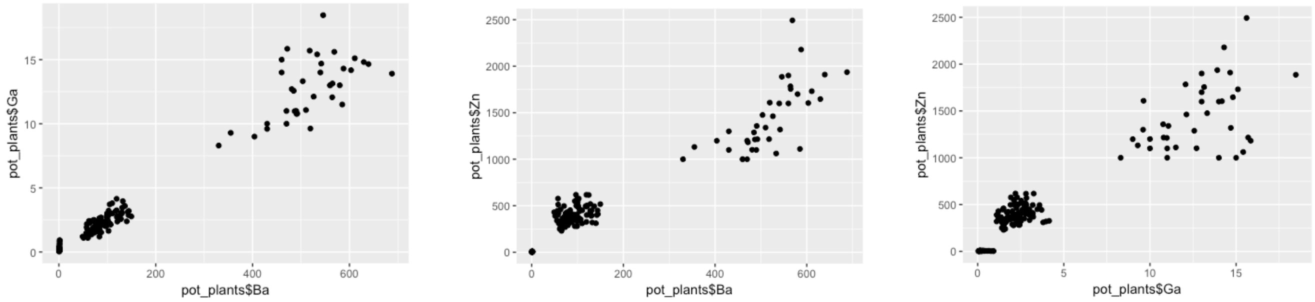


Figure1. Related elements

3c) Determination of the type of soil from the elemental composition

It might not very cautious to determinate the soil types that the plants were grown in just from the elemental composition of the leaves. Although it has been proved that the average levels of Ti and Ca are significantly different between Cannabis leaves grown in four soil types, there could still be variability in different units. In other words, the ranges of elements content for different groups can overlap. By making boxplots of Ti and Ca, it is clear that only the level of Ca in mb group is not overlapped with other groups (Figure2). Finding some unique elements in plants which were grown in particular soil types can be a strong evidence to make determination. Moreover, finding elements which have significant different ranges might also help to achieve the goal.

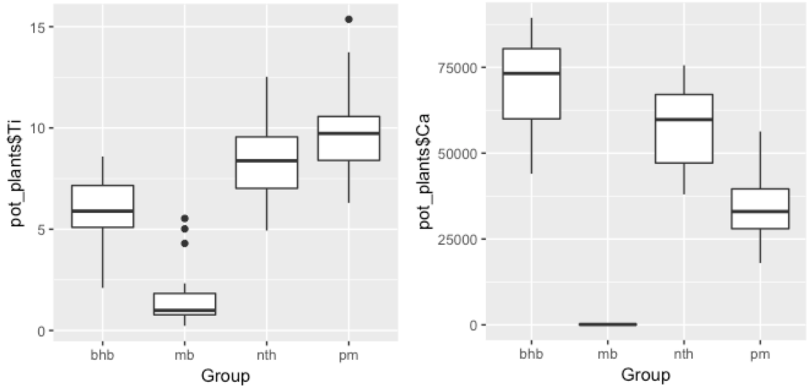


Figure2. Boxplots of Ti and Ca

4. Conclusion

To sum up, after cleaning and doing exploratory analysis of the data, three conclusions can be made. First, there are strong evidences indicate that the levels of Ti and Ca are significantly different between Cannabis leaves grown in four soil types. Second, the content levels of some pairs of elements, such as Ga and Ba, are closely related. Finally, it is not very cautious to make determinations about which places the plants were grown in just from the elemental composition of the leaves mainly because the ranges of elements content for different groups are overlapped. Finding some elements which are unique or have significant different ranges between four groups might be helpful to solve the origin determination problem.

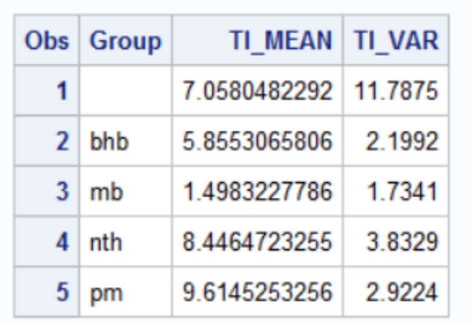
Reference:

Simth, N. (2000) *Criminals may rue pot from this plot.* Available at: https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=142910 (Accessed: 10th Oct. 2018)

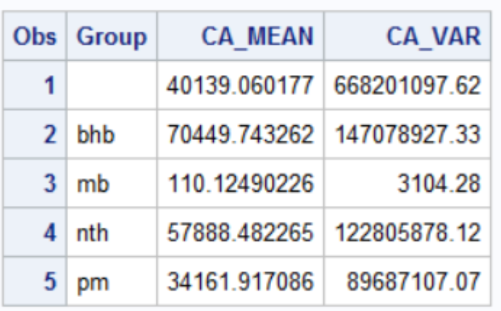
Datamentor (n.d.) *R factors.* Available at: https://www.datamentor.io/r-programming/factor/ (Accessed: 10th Oct. 2018)

Appendix A

Table 1b Summary of means and standard deviations



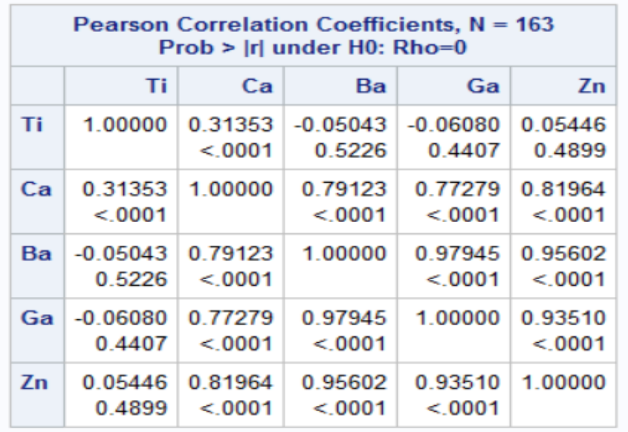
Obs	Group	TI_MEAN	TI_VAR
1		7.0580482292	11.7875
2	bhb	5.8553065806	2.1992
3	mb	1.4983227786	1.7341
4	nth	8.4464723255	3.8329
5	pm	9.6145253256	2.9224



Obs	Group	CA_MEAN	CA_VAR
1		40139.060177	668201097.62
2	bhb	70449.743262	147078927.33
3	mb	110.12490226	3104.28
4	nth	57888.482265	122805878.12
5	pm	34161.917086	89687107.07

Appendix B

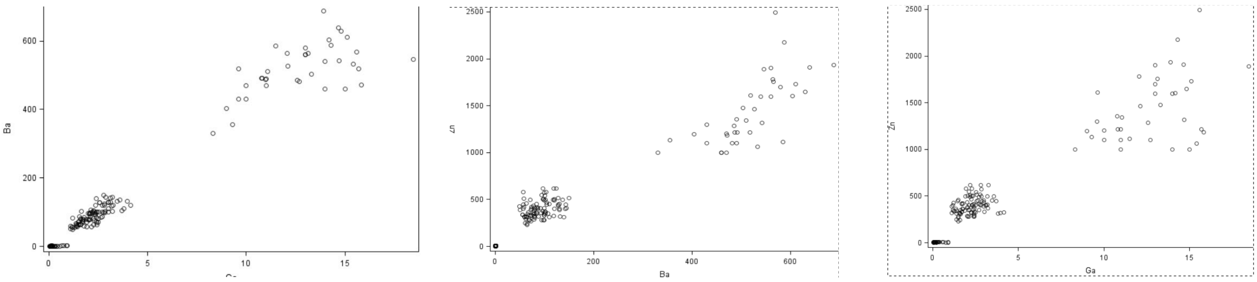
Table 2b. Pearson Correlation coefficients between elements



Pearson Correlation Coefficients, N = 163
Prob > |r| under H0: Rho=0

	Ti	Ca	Ba	Ga	Zn
Ti	1.00000	0.31353 <.0001	-0.05043 0.5226	-0.06080 0.4407	0.05446 0.4899
Ca	0.31353 <.0001	1.00000	0.79123 <.0001	0.77279 <.0001	0.81964 <.0001
Ba	-0.05043 0.5226	0.79123 <.0001	1.00000	0.97945 <.0001	0.95602 <.0001
Ga	-0.06080 0.4407	0.77279 <.0001	0.97945 <.0001	1.00000	0.93510 <.0001
Zn	0.05446 0.4899	0.81964 <.0001	0.95602 <.0001	0.93510 <.0001	1.00000

Appendix C



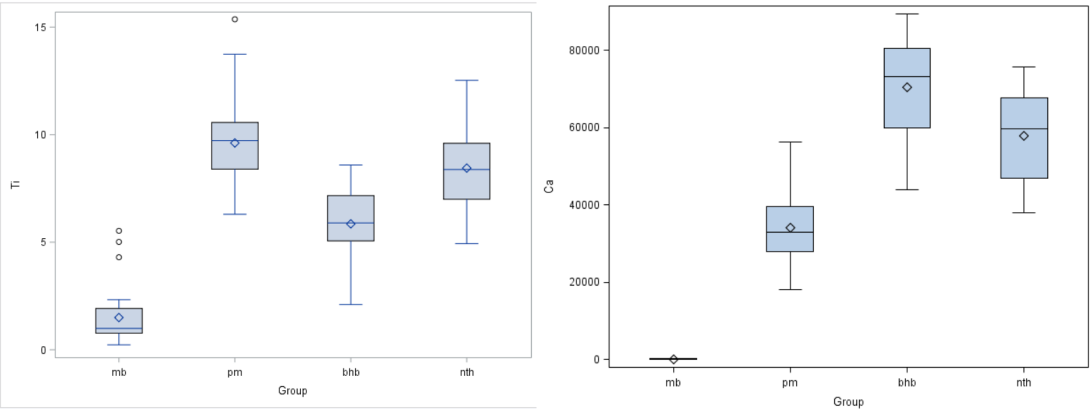
Ga~Ba

Ba~Zn

Ga~Zn

Figure 1b. Related elements

Appendix D



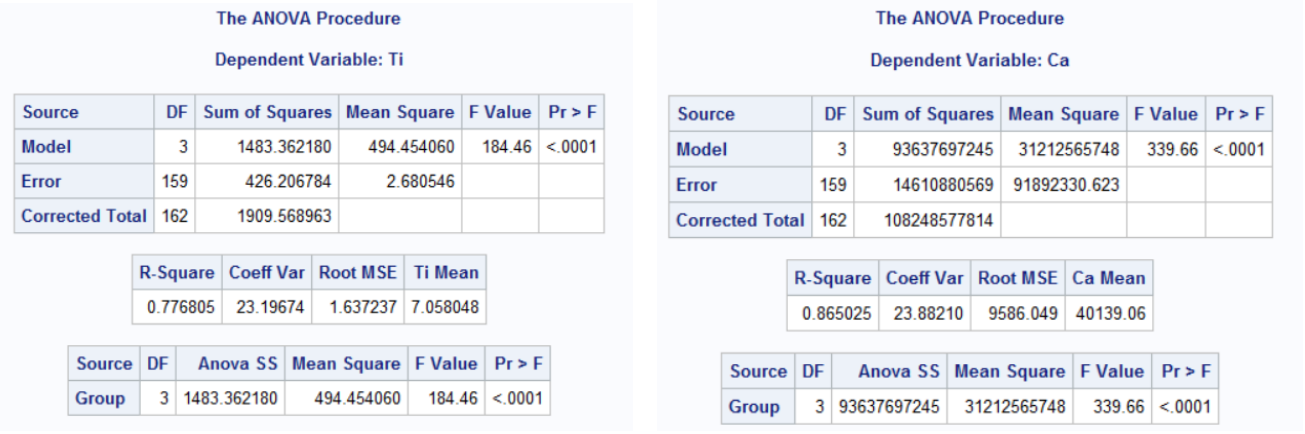
Ti

Ca

Figure 2b. Boxplots of Ti and Ca

Appendix E

Table 3 ANOVA results



The ANOVA Procedure

Dependent Variable: Ti

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1483.362180	494.454060	184.46	<.0001
Error	159	426.206784	2.680546		
Corrected Total	162	1909.568963			

R-Square	Coeff Var	Root MSE	Ti Mean
0.776805	23.19674	1.637237	7.058048

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Group	3	1483.362180	494.454060	184.46	<.0001

The ANOVA Procedure

Dependent Variable: Ca

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	93637697245	31212565748	339.66	<.0001
Error	159	14610880569	91892330.623		
Corrected Total	162	108248577814			

R-Square	Coeff Var	Root MSE	Ca Mean
0.865025	23.88210	9586.049	40139.06

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Group	3	93637697245	31212565748	339.66	<.0001