```r
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(reshape2)
library(knitr)
library(psych)
library(broom)
library(ggfortify)
```

\section*{\centering Executive summary}

This report summarizes the statistical analysis results associated with the study of element composition level in cannabis leaves sampled from different types of soil. The purpose of this report is to provide insights for future study of determining the soil type in which cannabis grown, based only on elemental differences. Rstudio(Version 1.1.456) and SAS(9.4)\footnote{Reproducible codes are provided in Appendixes.} are used to perform statistical analysis and produce tables, plots, and this report.

Section 1 presents basic descriptive statistics and plots. This report select two elements **K** and **Ti** to analyze differences among groups in section 2. To investigate the elemental differences, this report provides results of ANOVA analysis as well as tests for assumptions. In subsection post-hoc test, this report presents the results of pair-wise comparison of sample means using Tukey's Honest Significant Difference method. Moreover, this report conducts correlation analysis for three random chosen pairs of elements **Ca-Mg, Ti-Ce, Ga-Rb**. Combine with previous results, the correlation analysis results in section 3 can also aid in determining in which soil cannabis were originally grown.

In conclusion, this report proves that there are significant differences in the elemental composition of cannabis leaves grown in different types of soil. Note that, this report is only to provide some preliminary results to shed lights on future studies on classification of cannabis grown in different sites. The readers should bear in mind that the sample size used is relatively small and the sampling methods are not provided(second-hand source), and hence the robustness of results and conclusion is not guaranteed.

Introduction

Researchers have been trying to discover the relationships between elemental composition of the leaves of cannabis and in which type of soil cannabis were grown. This report intends to investigate whether there are any significant differences in the elemental composition of Cannabis leaves grown in different types of soil, based on the given data set with 56 observations from 4 different types of soil.

The variables are:

- **Sample Name**: Reference number of the samples, which will be ignore in this report.
- **Group The four soil types**:

\begin{enumerate} \item Potting mix (pm) \item Blockhouse Bay (bhb) \item Mission Bay (mb) \item Northland (nth) \end{enumerate}

- **Mg-Th**: The elements measured in the leaves. Note the units of measurement are not given(assumed to be same).In this report, symbioses are used for simplicity.

```r
#reading dataset into R############################################

potplants <- read_csv("~/R/assignments/potplants_MT5762.csv",
col_types = cols(Group = col_factor(levels = c("mb",
"pm", "bhb", "nth")), `Sample Name` = col_skip()))
```

This report focuses on answering three primary questions:

1. Do the data indicate differences in the elemental composition of Cannabis leaves grown in different soil types?
2. Are some of the elements related to one another in terms of their levels in the sampled leaves?
3. Can the results ultimately allow determination of what soil the plants were grown in ?

To answer the first question, this report focuses on 2 elements found in the leaves of cannabis. First, one-way ANOVA is used to investigate whether elements in cannabis' leaves differ among groups and a post-hoc test is used to find out how the levels of each element influenced by different types of soil. Moreover , to answer the second question, this report analyzes the correlations of three pairs of elements and further analyze how their correlations affected by different types of soil.

Elemental composition of Cannabis leaves in different soil

As shown in Table 1, there are total 56 observations from 4 groups. Each observation gives value of the levels of 38 elements. In Table 1, elements, whose levels with the highest standard deviation and widest range, are shown on the top. **Ca** is the most disperse element compared to the others. However, this table is not an accurate way to show how elements differ among groups because outliers within each group may cause some distraction.

```r
#gennerating summary statistics############################################

dtable<-describe(potplants[,2:39])
dtable$vars<-rownames(dtable)
options(scipen = 200)
dtable<-dtable%>%
  select(vars,n,min,max,mean,sd,range)%>%
  arrange(desc(sd,range))

kable(dtable,caption = "Summary statistics for all elements",digits = 3)
```

Figure 1 and figure 2 are box-plot for all elements in the data set, different soil types are filled with different colors. **K, Ti** are chosen for further investigation.

```r
#generating box-plots for all elements############################################

p1<-melt(potplants,id.vars ="Group",variable.name = "elements")
plo <- ggplot(data = p1[1:840,], aes(x=elements, y=value)) +
  geom_boxplot(aes(fill=Group),outlier.shape = NA)+
  facet_wrap( ~ elements, scales="free")+
  theme(legend.position="bottom")+
  ylab(NULL)
plo
```

```r
plo2 <- ggplot(data = p1[841:2128,], aes(x=elements, y=value)) +
  geom_boxplot(aes(fill=Group),outlier.shape = NA)+
  facet_wrap( ~ elements, scales="free")+
  theme(legend.position="bottom")+
  ylab(NULL)
plo2
```

ANOVA and diagnostic test

In this section and all following statistical tests, this report uses type-1 error of 5% to determine statistical significance.

P-values shown in the following table are much smaller than 0.05, indicating strong evidences against the null hypothesis of equal means in each group. In other words, element Ti's level or K's level are not the same in all groups. Though can not determine to what extent each group differs from another, such results prove that, in terms of elemental composition, cannabis grown in different types of soil are significantly different.

```r
#apply ANOVA############################################

test_pot<-potplants%>%
dplyr::select(Group,K,Ti)

alist<-lapply(test_pot[2:3],function(x){aov(x~Group,data = test_pot)})

aov.ti<- lapply(alist, tidy)
aovdf<-do.call("rbind",aov.ti)
aovdf<-na.omit(aovdf)
aovdf$p.value<-"<<0.05"
aovdf$elements<-c("K","Ti")
kable(aovdf,caption = "ANOVA for element K and Ti",digits = 2)
```

An ANOVA model has three underlying assumption:

1. Independence
2. Normality
3. Constant spread

Because no detailed sampling method is provided, this report assume independence of each samples.

Table 3 shows the results of normality test. The Shapiro test fails to reject the hypothesis of normal distribution for the residuals because p-value is bigger than the chosen significant level, and the qq-plots presented in Figure 3 and Figure 4 also prove the normality of residual distribution.

```r
#normality test for residuals############################################

resi.list<-lapply(alist, function(x){x$resi})
nor.list<-lapply(resi.list, shapiro.test)
nor.list<-lapply(nor.list,function(x){x$p.value})
nordf<-do.call("rbind",nor.list)
nordf<-as.data.frame(nordf)
colnames(nordf)<-"pvalue"

kable(nordf,caption = "Normality test for selected elements")

#examine constant spread############################################
resi.tbl<-do.call("cbind",resi.list)
resi.tbl<-as.data.frame(resi.tbl)
resi.tbl<-data.frame(Group=test_pot$Group,K=resi.tbl$K,Ti=resi.tbl$Ti)
resisd<-resi.tbl%>%group_by(Group)%>%summarise_all(sd)
kable(resisd,caption = "Constant spread test for selected elements",digits = 2)
```

```r
#Diagnostic plots for K############################################

autoplot(alist$K, label.size = 3)
```

```r
#Diagnostic plots for Ti############################################

autoplot(alist$Ti, label.size = 3)
```

The third assumption is constant variance of residuals. As shown in Table 4, the residuals from the model using levels of Ti as dependent variable have constant spread, but residuals from model using levels of K show non-constant variance. Further, Figure 4 shows the plotting of residuals against fitted value. Residuals show heteroscedasticity, a violation of the assumption.

A common remedial measure is log transformation. After the log transformation, the testing subject is no longer the levels of elements,but the log of the original level. Since the main objective is to analyze soil differences, such transformation does not affect the test’s validity. Table 5 presents the ANOVA results using the transformed data. P-value is much lower than 0.05. In other words, ANOVA reject the null hypothesis that the mean of the log of K's levels are the same among every groups. Therefore, it is fair to conclude that cannabis from different types of soil are different in terms of K's level. Moreover the diagnostic plots for the residuals of model shows that the residuals are normally distributed and have constant variance.

```r
#log transformation for K############################################

new=test_pot
new$K=log(new$K)
ak=aov(K~Group,data=new)
akdf<-tidy(ak)
akdf$p.value[1]<-"<<0.05"
kable(akdf,caption = "ANOVA for log(K)",digits = 2)

autoplot(lm(K~Group,data=new), label.size = 3)
```

Post-hoc test

To confirm by how much each group differ from each other, this report uses a post-hoc test(Tukey 1991). Table 6 below is the results of post-hoc test(Tukey adjusted). The test for nth-pm fails to reject the null hypothesis of equivalent means, while the other pairs all have strong evidence against the null hypothesis. That is, in terms of the levels of Ti in cannabis' leaves, cannabis grown in northland are equivalent to those grown in potting mix.

```r
#POST-HOC test############################################

kdf<-tidy(TukeyHSD(ak))
tdf<-tidy(TukeyHSD(alist$Ti))
tdf[1:4,6]<-"<<0.05"
tdf[5,6]<-"0.1084"
tdf[6,6]<-"0.0009"
options(scipen=200)
kdf$pvalue<-round(kdf$adj.p.value,4)
kdf$adj.p.value[1:3]<-"<<0.05"
kdf$adj.p.value[4:6]<-as.character(kdf$pvalue[4:6])
kdf<-kdf[,-7]
kable(tdf,caption = "Post-hoc test for Ti",digits = 2)
kable(kdf,caption = "Post-hoc test for log(K)",digits = 2)
```

In table 7, the test for bhb-pm fails to reject the null hypothesis of equivalent means. In terms of the means of the log of K's levels, cannabis from blackhouse bay are equivalent to cannabis from potting mix. These results suggest that determination of originality of cannabis can be achieved through examnition of both of these elements' levels.

Correlations among elements

In this section, this report randomly choose three pairs of elements, **Ca-Mg, Ti-Ce, Ga-Rb**, using the correlation coefficient and p-value calculated to examine their correlations in cannabis grown in different types of soil.

This section first examines the correlation of Ti and Ce. Based on a significant level of 0.05, the correlation between Ti and Ce is not statistically significant in the last two group. This result can be useful when using previous results can not distinguish northland or blackhouse bay's cannabis from potting mix's. Cannabis leaves grown in potting mix area may have a more significant correlation between Ti and Ce .

```r
#Correlation of Ti-Ce############################################

ti.ce<-potplants%>%
  select(Group,Ti,Ce)%>%
  group_by(Group)%>%
  summarise(correlations=cor(Ti,Ce),p.value=cor.test(Ti,Ce)$p.value)
kable(ti.ce,caption = "Correlation of Ti-Ce",digits = 2)
```

Calcium and Magnesium are two of the most important secondary macro nutrients for plants. Therefor, it is not surprise to see that statistically significant positive correlations exist in observations in almost all groups, except for cannabis grown in northland(though insignificant).

```r
#Correlation of Ca-Mg############################################

ca.mg<-potplants%>%
  select(Group,Ca,Mg)%>%
  group_by(Group)%>%
  summarise(correlations=cor(Ca,Mg),p.value=cor.test(Ca,Mg)$p.value)
kable(ca.mg,caption = "Correlation of Ca-Mg",digits = 2)
```

Table 10 also provide some extra insights on determining in which type of soil cannabis were grown. The correlation between Ga and Rb vary among groups. Significant positive correlation is found in the first group but the correlation is insignificant in the second group. Significant negative correlation of Ga and Rb is found in the sample of northland-grown cannabis, but correlation is insignificant in the sample of Blockhouse Bay-grown cannabis. Apart from examining levels of element Ti or K in leaves, testing the correlation between Ga and Rb may aid in classifying cannabis from different site.

```r
#Correlation of Ga-Rb############################################

Ga.Rb<-potplants%>%
  select(Group,Ga,Rb)%>%
  group_by(Group)%>%
  summarise(correlations=cor(Ga,Rb),p.value=cor.test(Ga,Rb)$p.value)
kable(Ga.Rb,caption = "Correlation of Ga-Rb",digits = 2)
```

Conclusion

Overall, this report shows that cannabis grown in different types of soil have statistical significant differences, in terms of elemental composition. Testing the levels of both element Ti and K is quite enough to distinguish cannabis grown in potting mix from those grown in other sites. Moreover, follow-up analysis of correlations show that there are significant correlations in each of the three pairs of elements, Ca and Mg, Ti and Ce, and Ga and Rb. Furthermore, the correlations between elements in cannabis are affected by where cannabis were grown. **In conclusion, examing only elemental composition of cannabis leaves is sufficient for the determination of where the plants were grown.** However, with a small sample size, caution must be applied, as the findings might not be robust. If circumstances permit, using a larger sample or performing some further robustness tests may yield more reliable results.

Reference

\begin{enumerate} \item David Robinson and Alex Hayes (2018). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.0. https://CRAN.R-project.org/package=broom

\item Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

\item Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

\item Masaaki Horikoshi and Yuan Tang (2016). ggfortify: Data Visualization Tools for Statistical Analysis Results. https://CRAN.R-project.org/package=ggfortify \item Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.4.

\item R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

\item SAS Institute Inc. (2011). Base SAS® 9.3 Procedures Guide. Cary, NY: SAS Institute Inc.

\item Tukey, J. (1991). The Philosophy of Multiple Comparisons. Statistical Science, 6(1), 100-116. Retrieved from http://www.jstor.org/stable/2245714

\item Yihui Xie (2018). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.20.

\item Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

\end{enumerate}

Appendix

R code

SAS code

```sas
*import data;
PROC IMPORT OUT= WORK.pot1
            DATAFILE= "C:\Users\jiong\Documents\R\assignments\potplants_
MT5762.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

************************************************************************;
*summary statistics;

proc means mean std min max range data=work.pot1;
output out=stats1(drop=_type_ _freq_);
run;

**************************************************
*data manipulation before box-ploting;

proc transpose data=stats1 out=st2
name=column_that_was_transposed;
ID _STAT_;
run;

PROC sort data=st2 out=st1;
by descending std;
run;

proc print data=st1;
run;
PROC sort data=pot1 out=pot2;
by Sample_Name;
run;

proc transpose data=pot2
   out=long(rename=(Col1=Value))
   name=elements;
   by Sample_Name;
run;


data pot22;
    set pot2(keep=Sample_Name Group);
run;
*changing names of variables;

data my;
    set long;
    LENGTH elements2 $30.;
    elements2=SCAN (elements,-1,'_');
run;

data melt;
   merge my pot22;
   by Sample_Name;
run;

************************************************************************;
*box-plotting ;

proc sgpanel data=melt;
panelby elements2 / rows=4 columns=2 ;
vbox value / category= Group;
run;
proc transpose data=my out=wide ;
    by Sample_Name;
    id elements2;
    var Value;
run;
data new;
merge pot22 wide;
by Sample_Name;
run;

************************************************************************;
*get a new set of data;

data subset;
set new(keep=Group Sample_Name K  Ti);
run;
proc transpose data=subset
   out=long2(rename=(Col1=Value))
   name=elements;
   by Sample_Name;
run;
data new2;
merge long2 pot22;
by Sample_Name;
run;

************************************************************************;
*examine constant spread;

proc summary data=new2 nway;
     class Group elements;
     var value;
     output out=WANT(drop=_:) mean=mean std=std;
run;

proc sort data=want out=spread;
by elements;
run;

*conduct ANOVA and post-hoc analysis;

proc glm data=subset;
   class Group;
   model Ti K= Group;
   means Group/tukey;
   output out=diag r=resid;
run;

************************************************************************;
symbol1 v=dot c=blue;
*diagnostic plots
proc gplot data=diag;
plot resid*GROUP;
run;
proc univariate noprint;
var resid;
qqplot resid/normal (L=1 mu=est sigma=est);
run;
   ************************************************************************;
*test for normality;

proc univariate data=diag normaltest;
   var resid;
   probplot resid / normal(mu=est sigma=est);
run;

proc sort data=new out=sortednew;
by group;
run;

**log transformation****************************************************;

data subset2;
set subset;
lk=log(K);
run;

 **ANOVA for log(k)****************************************************;

proc glm data=subset2;
   class Group;
   model lk= Group;
   means Group/tukey;
   output out=diagk r=residk;
run;

symbol1 v=dot c=blue;
*diagnostic plots check for normality;
proc gplot data=diagk;
plot residk*GROUP;
run;

proc univariate data=diagk normaltest;
   var residk;
   probplot residk / normal(mu=est sigma=est);
run;


************************************************************************;
*correlation analysis;

ods graphics on;
title 'correlation analysis for Ti-Ce';
proc corr data=sortednew plots=scatter(alpha=.20 .30);
   var Ti Ce;
   by group;
 run;
ods graphics off;

ods graphics on;
title 'correlation analysis for Ca-Mg';
proc corr data=sortednew plots=scatter(alpha=.20 .30);
   var Ca Mg;
   by group;
 run;
ods graphics off;

ods graphics on;
title 'correlation analysis for Ga-Rb';
proc corr data=sortednew plots=scatter(alpha=.20 .30);
   var Ga Rb;
   by group;
 run;
ods graphics off;
```

alexis2113/R-Stuff- documentation built on May 10, 2019, 8:24 a.m.