Exploratory Multivariate Data Analysis – Hw 2

Yu-Xiang Lin(ID: B06305030)

10/6/2020

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0
```

```
## Warning: package 'tibble' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Problem 1

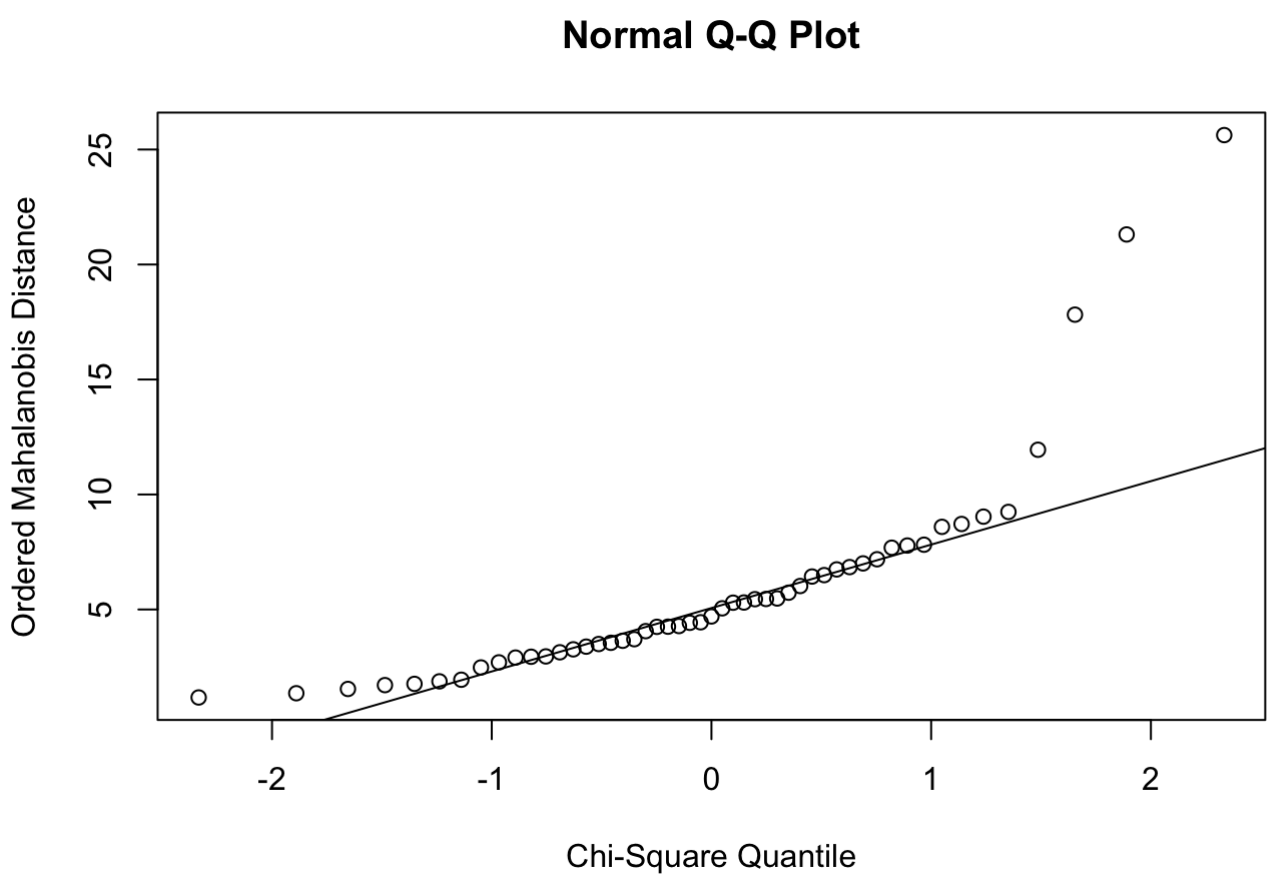
Six hematology variables were measured on 51 workers. (hematology.dat) y1=hemoglobin concentration, y2=packed cell volume, y3= white blood cell count, y4=lymphocyte count, y5=neutrophil count, y6=serum lead concentration. (a) Calculate D_i^2 for each observation. Draw the Q-Q plots and qqline for D_i^2 to identify outliers. (b) Draw qq-plots and qqline for y1 to y5 and comment on their normality. (c) Calculate Wilks' static and perform F test to identify outliers.

(a)

```
# import data
df1 <- read.table("/Users/linyuxiang/Desktop/多變量/Data/hematology.dat")

p <- ncol(df1)
Sx <- cov(df1)
D2 <- mahalanobis(df1,colMeans(df1),Sx)
  p <- ncol(df1)
  n <- nrow(df1)
  Sx <- cov(df1)
  D2 <- mahalanobis(df1,colMeans(df1),Sx)
  index <-  ((1:n)-0.5)/n
  quant <- qchisq(index,p)
  df1 <- bind_cols(df1,data_frame(D2 = D2,quant = quant))
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
  qqnorm(df1$D2,ylab = "Ordered Mahalanobis Distance",xlab = "Chi-Square Quantile");q
qline(df1$D2)
```



- 從Q-Q plot中可以看出觀察樣本存在outlier

```
which((D2 / p) >2.5)
```

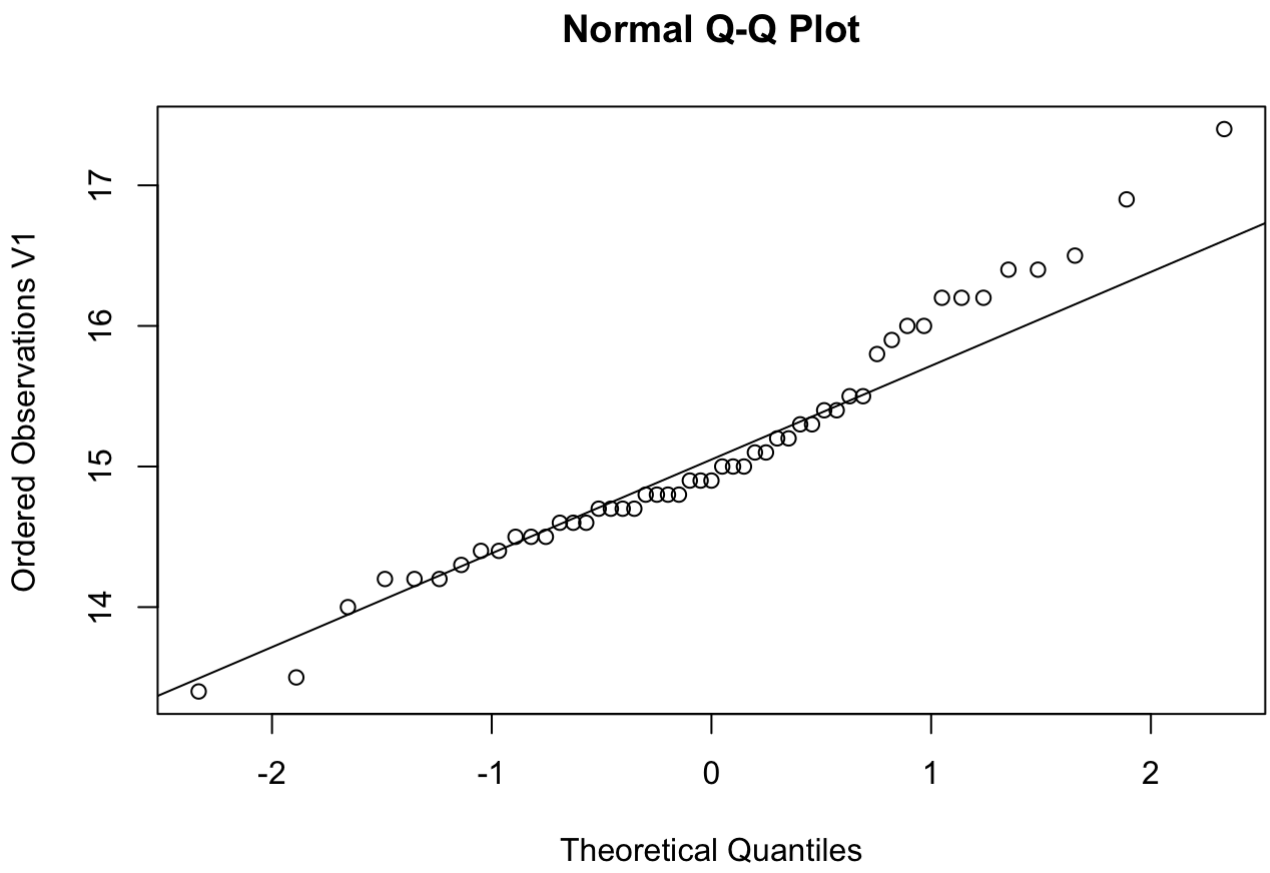
```
## [1] 10 47 50
```

- Ans: 使用(D2 / p) > 2.5為標準判斷outlier得到outlier為第10, 47, 50個樣本

(b)

V1

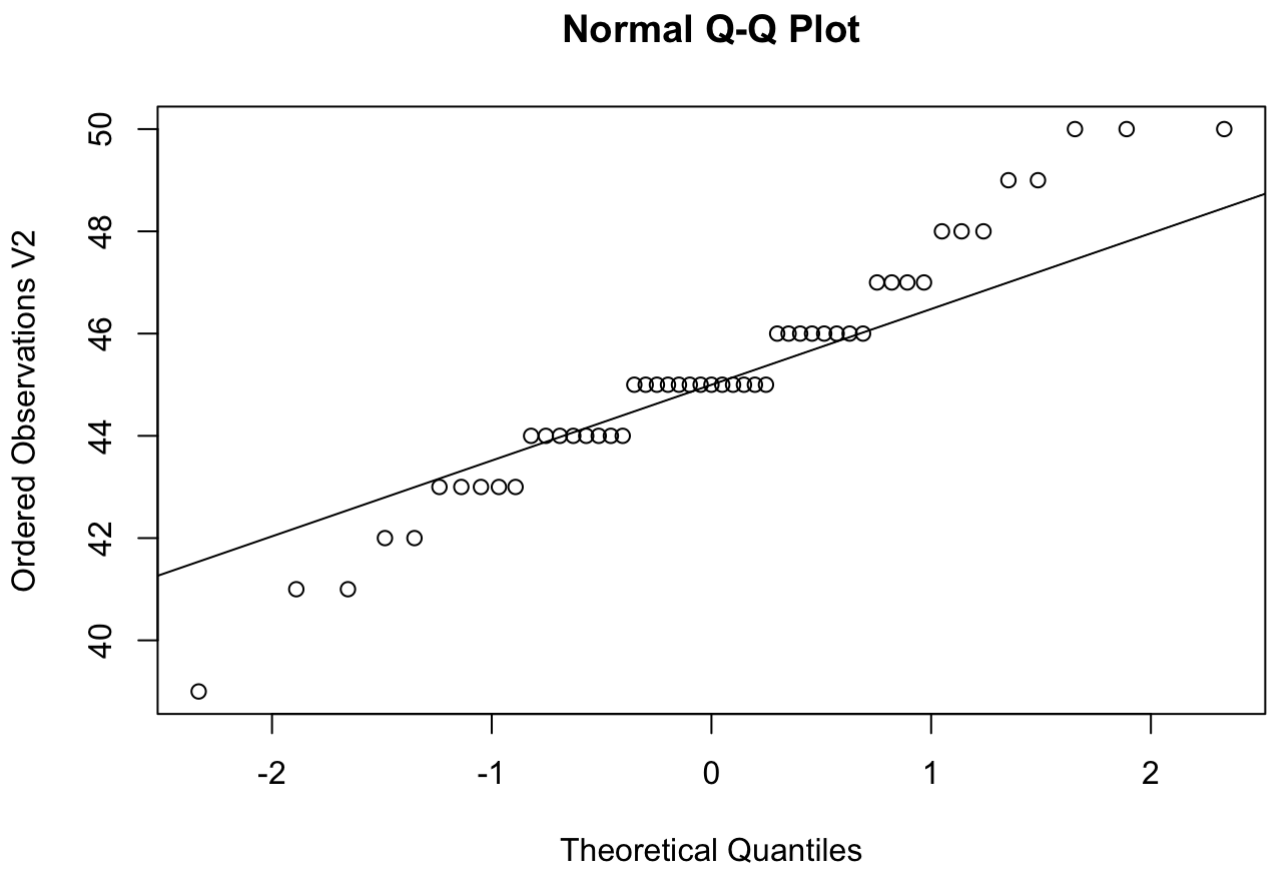
```
qqnorm(df1$V1, ylab="Ordered Observations V1");qqline(df1$V1)
```



- Ans: ## 從Q-Q plot中V1看起來並非常態分佈，有部分觀察高於Q-Q line

V2

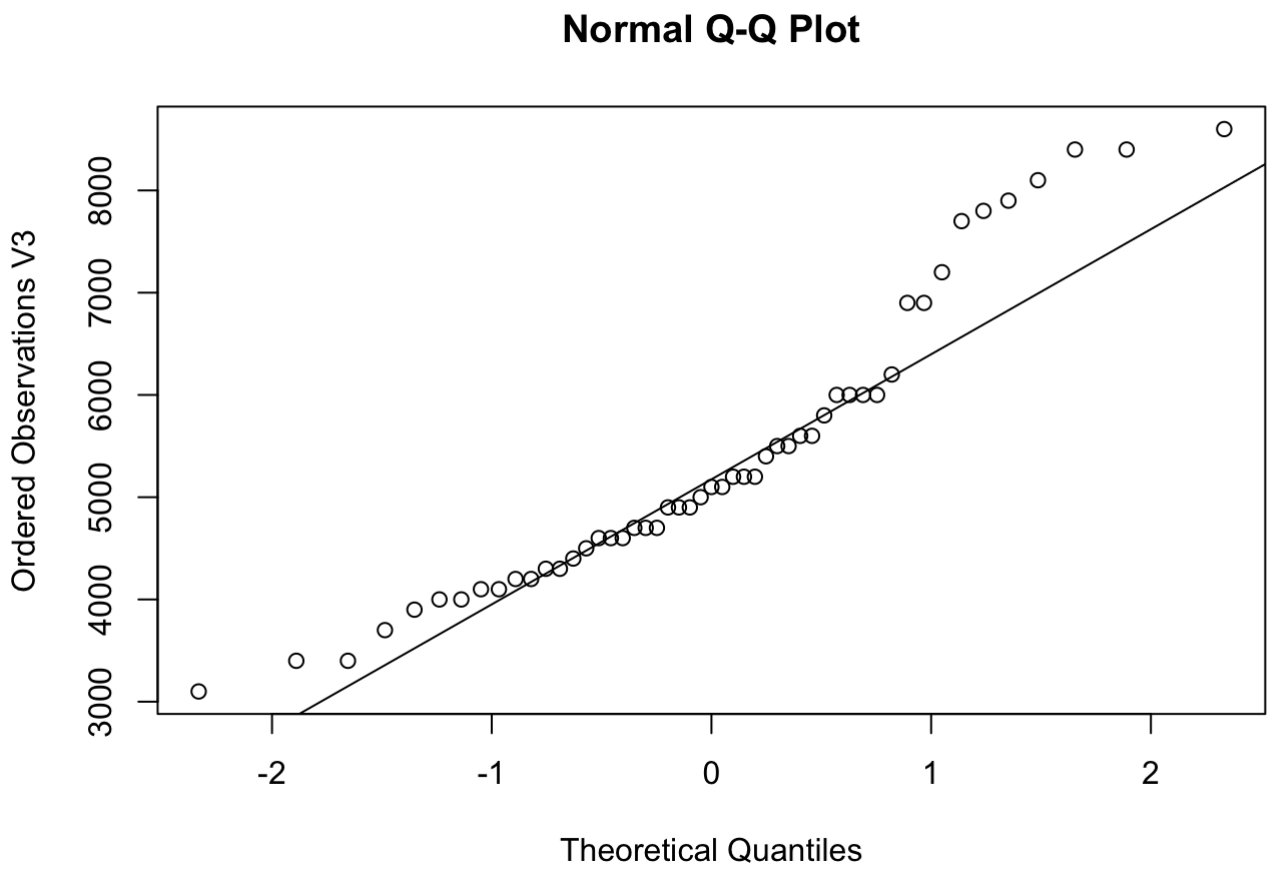
```
qqnorm(df1$V2, ylab="Ordered Observations V2");qqline(df1$V2)
```



- Ans: 從Q-Q plot中V2看起來並非常態分佈，只有少數觀察值座落於Q-Q line

V3

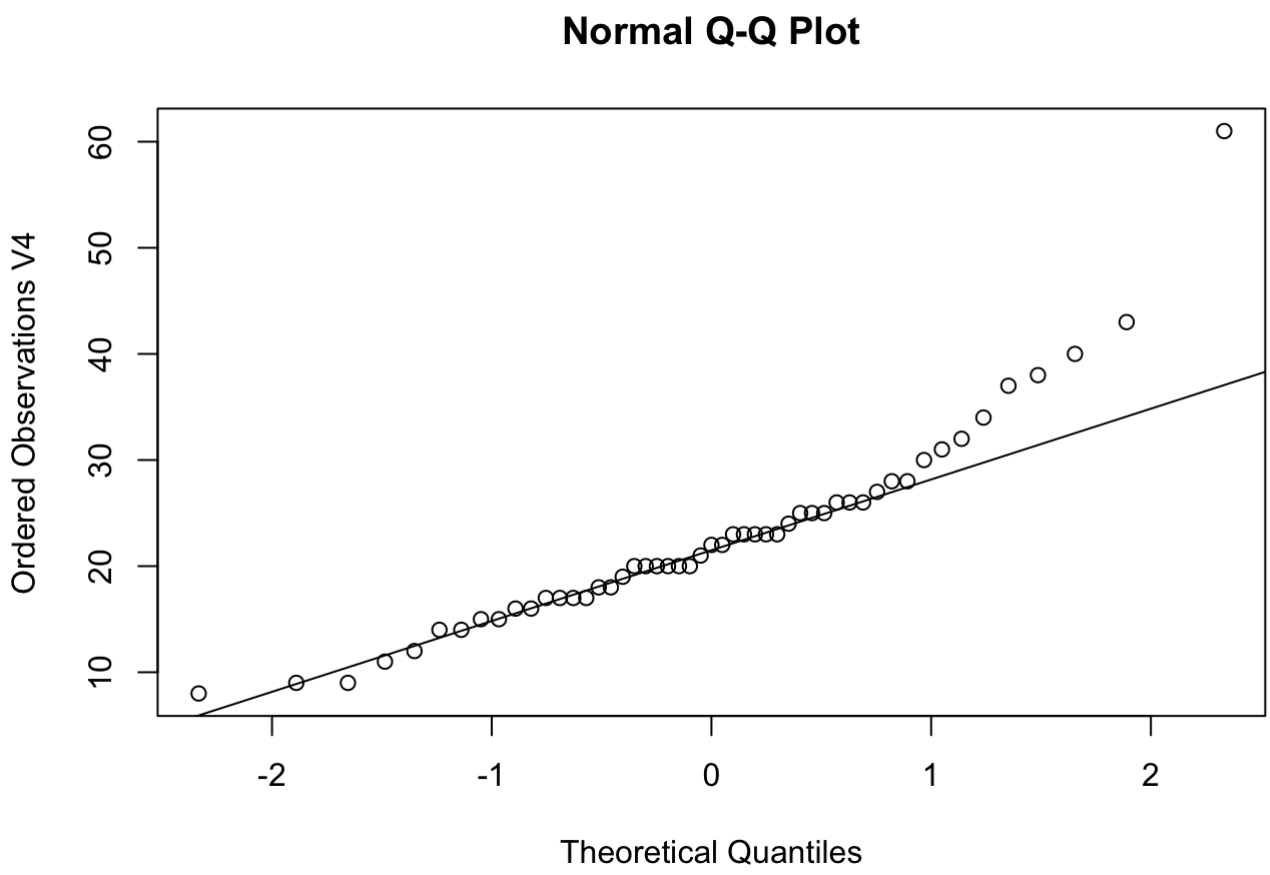
```
qqnorm(df1$V3, ylab="Ordered Observations V3");qqline(df1$V3)
```



- Ans: 從Q-Q plot中V3看起來並非常態分佈，在樣本V3數值小及大時觀察值並沒有座落於Q-Q line

V4

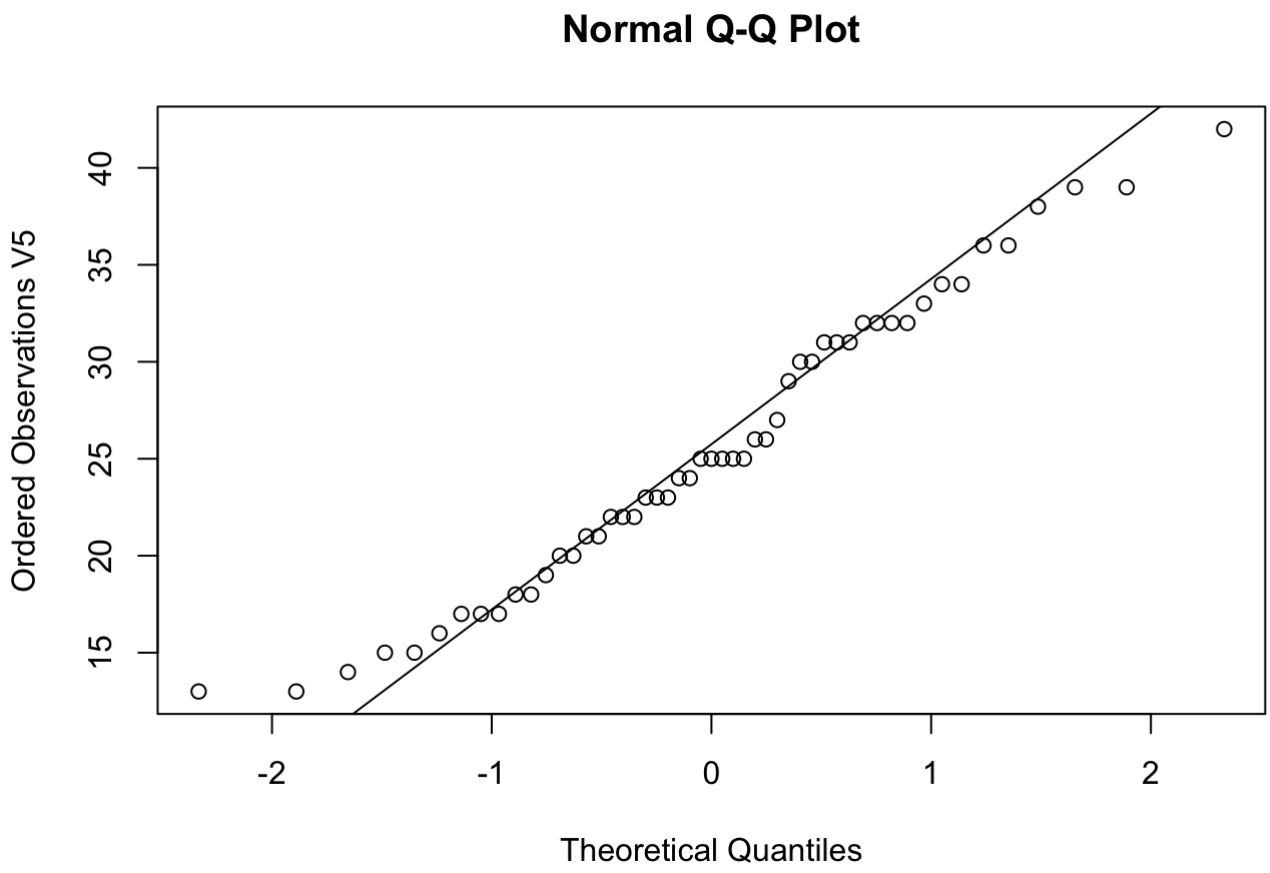
```
qqnorm(df1$V4, ylab="Ordered Observations V4");qqline(df1$V4)
```



- Ans: 從Q-Q plot中V4看起來並非常態分佈，在樣本V4數值大時觀察值並沒有座落於Q-Q line

V5

```
qqnorm(df1$V5, ylab="Ordered Observations V5");qqline(df1$V5)
```



- Ans: 從Q-Q plot中V5看起來可能是常態分佈，只有在樣本V5數值大時及V5約等於25時偏離Q-Q line

(c)

```
w <- 1-((n*(D2)) / (n-1)^2)
#Fi <-  ((n-p-1)/p) * (1/(1-((n*D2/(n-1)^2)))-1)
Fi <- ((n-p-1)/p) * ((1/w)-1)
which(Fi > qf(0.995,p,n-p-1))
```

```
## [1] 10 47 50
```

- Ans: 使用Wilks' statisc進行outlier檢定可以得到以上樣本為outlier, 和起初使用D2 / p > 2.5的標準相同，outlier為第10, 47, 50個觀察值

Problem 2

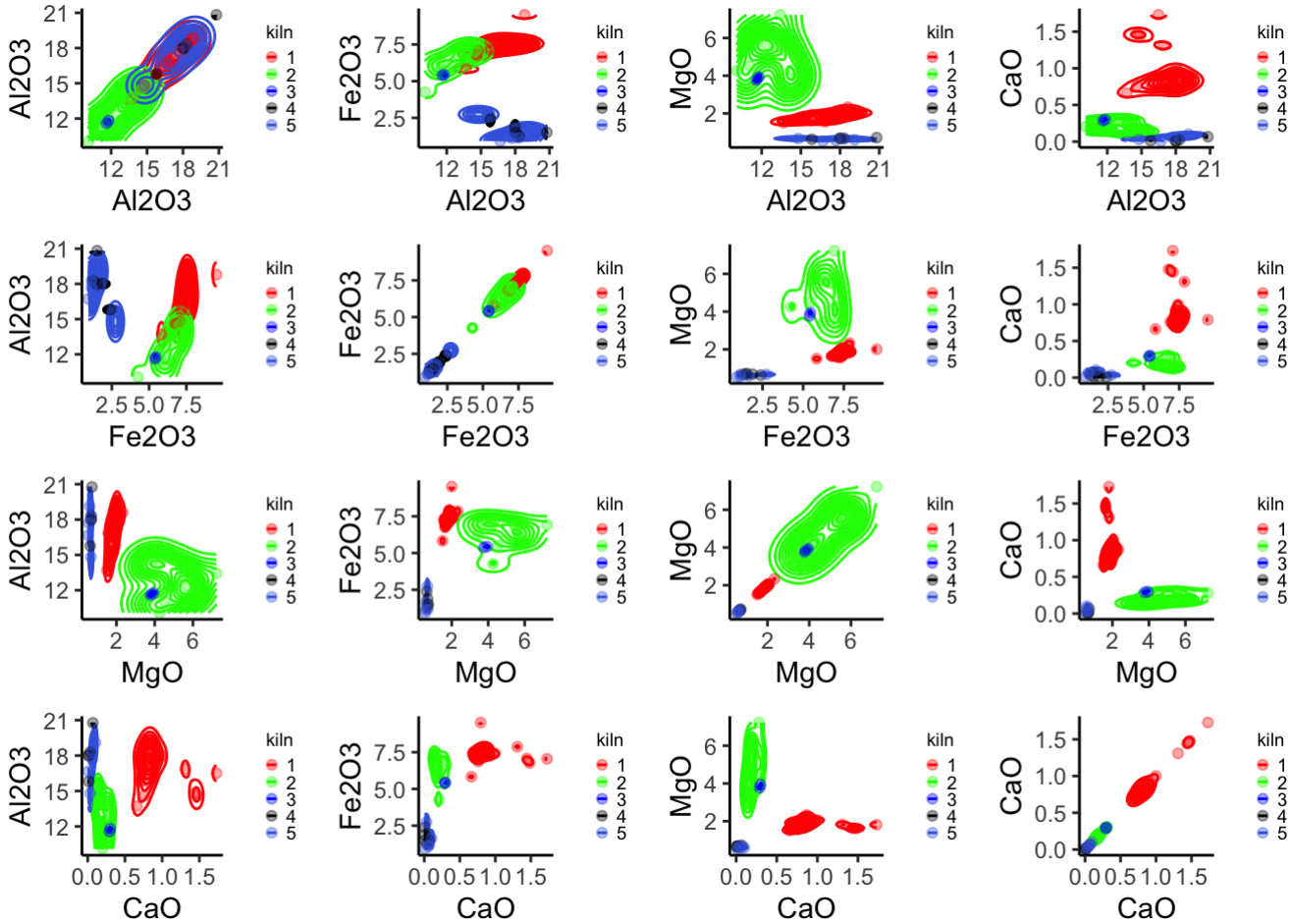
The data (pottery.csv) consists of the results of chemical analysis on on Romano-British pottery made in three different regions (region 1 contains kiln 1, region 2 contains kilns 2 and 3, and region 3 contains kilns 4 and 5). Construct a scatterplot matrix of the chemical composition of Romano-British pottery and identify each unit by its kiln number and showing the estimated bivariate density on each panel. What does the resulting diagram tell you?

```
# import data
df2 <- read.csv("/Users/linyuxiang/Desktop/多變量/Data/pottery.csv")
# plot 2d density and scatter function
plot <- function(i,ii){
p <- ggplot(data = df2, aes(x = df2[,i], y = df2[,ii],col = as.character(kiln)))+
  geom_density_2d()+
  geom_jitter(alpha=0.35)+
  labs(color = "kiln")+
  scale_color_manual(values = c("red","green","blue","black","royalblue"))+
  theme_classic()+
  xlab(paste0(names(df2)[i]))+ylab(paste0(names(df2)[ii]))+
  theme(legend.title = element_text(size = 7),
              legend.text  = element_text(size = 7),
              legend.key.size = unit(0.3, "lines"))
return(p)
}
p_11 <- plot(1,1);p_12 <- plot(1,2);p_13 <- plot(1,3);p_14 <- plot(1,4)
p_21 <- plot(2,1);p_22 <- plot(2,2);p_23 <- plot(2,3);p_24 <- plot(2,4)
p_31 <- plot(3,1);p_32 <- plot(3,2);p_33 <- plot(3,3);p_34 <- plot(3,4)
p_41 <- plot(4,1);p_42 <- plot(4,2);p_43 <- plot(4,3);p_44 <- plot(4,4)

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(grid)

grid.arrange(p_11, p_12, p_13, p_14,
             p_21, p_22, p_23, p_24,
             p_31, p_32, p_33, p_34,
             p_41, p_42, p_43, p_44,nrow=4)
```



- Ans: 從上圖中可以看出不論是哪個地區生產的陶器，主要成分都是Al2O3，且比例最少的是CaO。另外從第二行及第四行圖表可以看出kiln2 使用Fe2O3較多，CaO的使用比例較少；從第二行及第三行圖表則可以以看出kiln1 是Fe2O3使用較多，MgO使用較少。從第二行圖表野可以看出kiln3 Fe2O3使用較少；Al2O3在各地區的使用比例則是看不大出差別。

no run

```
#
# pairs(df2[,1:2],pch = 15,col = 1:5, cex = 1, oma = c(5,5,5,15),panel = panel.contou
r)
# pairs(df2[,1:2],pch = 15,col = 1:5, cex = 1, oma = c(5,5,5,15))
# par(xpd = TRUE)
# legend("topright",col = 1:5, pch = 15, legend = paste0("klin",1:5))
#
# pairs(df2[,1:4],panel = panel.contour,diag.panel = panel.histogram)


# draw <- function(i,ii){
# ggplot(data = df2, aes(x = df2[,i],y =df2[,ii],color = as.character(kiln)))+
#   geom_point()+
#   scale_colour_manual(name = "klin",values = c("red","blue","green","black","royalb
lue"))+
#     xlab(names(df2)[i])+ylab(names(df2)[ii])
# }
# par(mfrow = (c(1,2)))
# draw(1,2);draw(1,3)
```