Homework 2

Problem 1 Use the bivariate boxplot on the scatterplot of pairs of variables ((temp, wind), (temp, precip)) in the air pollution data to identify any outliers. Calculate the correlation between each pair of variables using all the data and the data with any identified outliers removed. Comment on the results.

library(HSAUR2)

## Warning: package 'HSAUR2' was built under R version 4.0.3

## Loading required package: tools

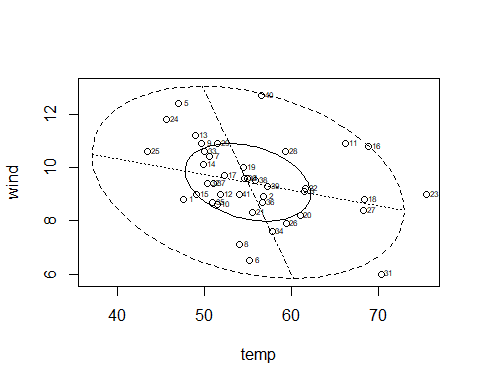
library(MVA)

## Warning: package 'MVA' was built under R version 4.0.3

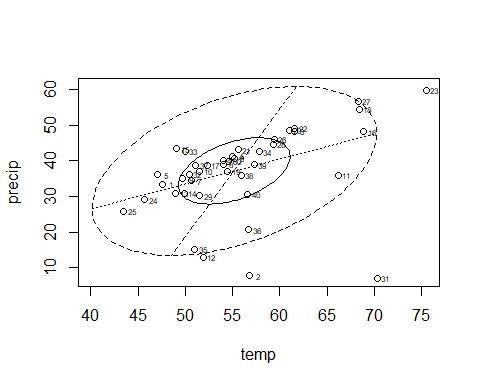
data("USairpollution", package = "HSAUR2")  
head(USairpollution)

## SO2 temp manu popul wind precip predays  
## Albany 46 47.6 44 116 8.8 33.36 135  
## Albuquerque 11 56.8 46 244 8.9 7.77 58  
## Atlanta 24 61.5 368 497 9.1 48.34 115  
## Baltimore 47 55.0 625 905 9.6 41.31 111  
## Buffalo 11 47.1 391 463 12.4 36.11 166  
## Charleston 31 55.2 35 71 6.5 40.75 148

# Bivariate Boxplot  
bvbox(cbind(USairpollution$temp, USairpollution$wind), xlab="temp", ylab = "wind")  
  
# Labeling each point according to its row number  
text(x=USairpollution$temp+0.9, y=USairpollution$wind+0.06, labels=seq(nrow(USairpollution)), cex=0.5)



# From this plot 31st row and 23rd row are outliers as they lie outside the 75th %ile circle   
  
bvbox(cbind(USairpollution$temp, USairpollution$precip), xlab = "temp", ylab = "precip")  
  
text(x=USairpollution$temp+0.9, y=USairpollution$precip+0.06, labels=seq(nrow(USairpollution)), cex=0.5)



# From this plot 2nd, 31st, and 23rd row are outliers   
  
  
cor(USairpollution$temp, USairpollution$wind)

## [1] -0.3497396

# Correlation of all temperature and wind data = -0.34   
cor(USairpollution$temp[c(-31,-23)], USairpollution$wind[c(-31,-23)])

## [1] -0.2587808

# Correlation of all temperature and wind data except outliers = -0.25  
# When we removed the outliers, the correlation decreased. Therefore the temp and wind are not highly correlated  
  
  
cor(USairpollution$temp, USairpollution$precip)

## [1] 0.3862534

# Correlation of all temp and precip data is 0.38  
cor(USairpollution$temp[c(-2,-31,-23)], USairpollution$precip[c(-2,-31,-23)])

## [1] 0.6227856

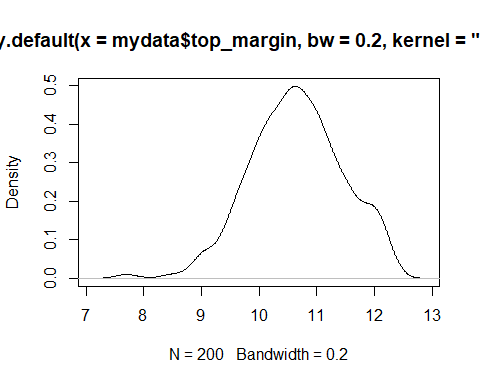
# Correlation of all temperature and precipitation data except outliers is 0.62  
  
# When we removed the outliers the correlation increased, therefore the temperature and precipitation are highly correlated

Problem 2 The banknote dataset contains measurements on 200 Swiss banknotes: 100 genuine and 100 counterfeits. The variables are the status of the “note,” length of the bill, width of the left edge, width of the right edge, bottom margin width, and top margin width. All measurements are in millimeters. Read the data and pick the variables: “note,” “top\_margin,” and “diag\_length.” banknote <- read.csv(“<http://westfall.ba.ttu.edu/isqs6348/Rdata/swiss.csv>”) mydata <- banknote[,c(1,6,7)]

# Reading data  
banknote <- read.csv("http://westfall.ba.ttu.edu/isqs6348/Rdata/swiss.csv")  
  
  
mydata <- banknote[,c(1,6,7)]  
head(mydata)

## note top\_margin diag\_length  
## 1 real 9.7 141.0  
## 2 real 9.5 141.7  
## 3 real 9.6 142.2  
## 4 real 10.4 142.0  
## 5 real 7.7 141.8  
## 6 real 10.1 141.4

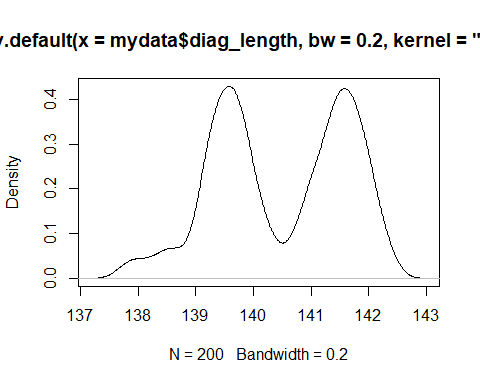
# a   
# Calculating Densities   
density\_top\_margin <- density(mydata$top\_margin, bw = .20, kernel = "gaussian")  
plot(density\_top\_margin)



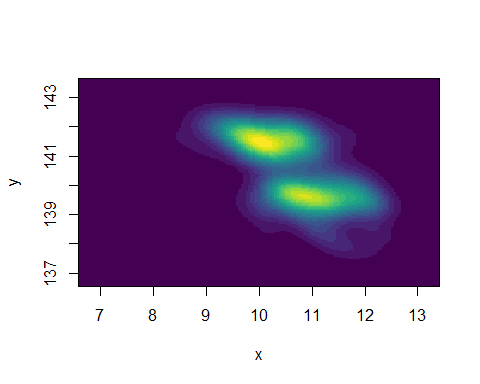
density\_diag\_length <- density(mydata$diag\_length, bw = .20, kernel = "gaussian")  
plot(density\_diag\_length)  
  
  
# b

library(ks)

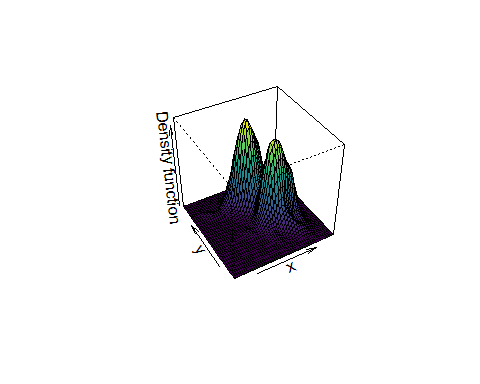
## Warning: package 'ks' was built under R version 4.0.3



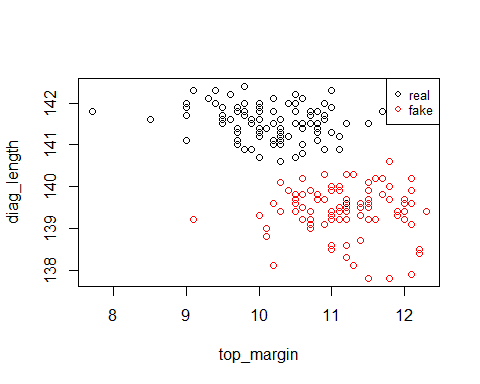
kde <- kde(mydata[,c(2,3)])  
  
plot(kde, display = "image", xlab = "x", ylab = "y", col = viridisLite::viridis(20))



plot(kde, display = "persp", col.fun = viridisLite::viridis, xlab = "x", ylab = "y")



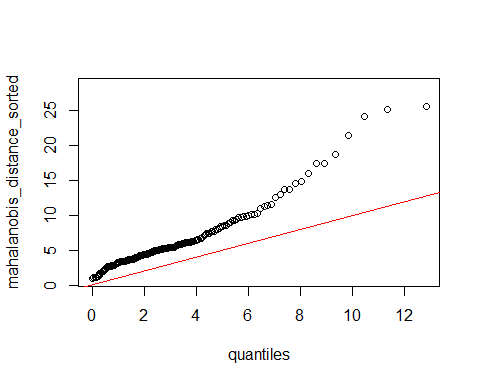
# c   
plot(mydata[,2:3], col = ifelse(mydata[,1] == "real", "black", "red"))  
legend("topright",legend = c("real", "fake"),col = c("black", "red"), pch = 1, cex = .8)



# Based on all the plots we can see that there is a clear distinction in the values for fake notes  
# and original notes. We can easily say with confidence if a note is fake or note based on its  
# top margin length and diagonal length

Problem 3 Examine the multivariate normality (MVN) of the banknote data (excluding the “note” variable) by creating the chi-square plot of the data. Load the data as follow. Follow the listed steps to examine the multivariate normality.

banknote <- read.csv("http://westfall.ba.ttu.edu/isqs6348/Rdata/swiss.csv")   
mydata2 <- banknote[,-1]  
  
# a  
# Calculating the column means  
colmeans\_vector <- colMeans(mydata2)   
  
# b  
# Calculating the covariance  
cov\_mydata<- cov(mydata2) # calculating the covariance   
  
# c  
# calculating the mahalanobis distance   
mahalanobis\_distance <- mahalanobis(mydata2, center = colmeans\_vector, cov = cov\_mydata)  
  
# d  
# sorting the distance   
mahalanobis\_distance\_sorted <- sort(mahalanobis\_distance)  
  
# e  
# finding the quantiles   
quantiles <- qchisq(seq(0,1,by=1/(nrow(mydata2)-1)), df=ncol(mydata))  
  
# plotting them  
plot(quantiles, mahalanobis\_distance\_sorted)  
abline(a = 0, b = 1, col="red")



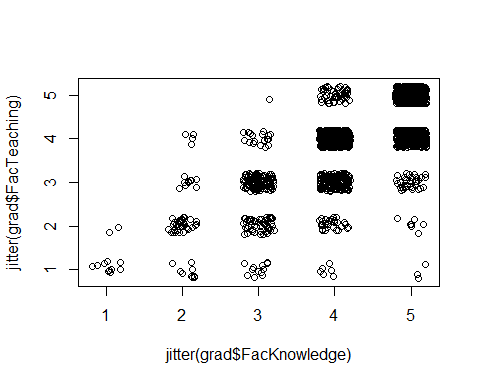
# Most the of the data is aligned closely with the red-line, hence we can say that for the most   
# part data shows strong MVN form , so yes data is MVN

Problem 4 Use the TTU graduate student exit survey data

grad <- read.csv("http://westfall.ba.ttu.edu/isqs6348/Rdata/pgs.csv")  
# a   
sum(!is.na(grad$GenRating)) # all the rows where rating is valid

## [1] 1976

# There are 1976 students with a valid rating   
  
# b   
# Using Jitter because the plot looks odd and is missing data.   
plot(jitter(grad$FacKnowledge),jitter(grad$FacTeaching))



# c   
mydata3 <- subset(grad, select = c("FacTeaching", "FacKnowledge", "Housing"))  
head(mydata3)

## FacTeaching FacKnowledge Housing  
## 1 3 3 4  
## 2 3 4 3  
## 3 4 4 4  
## 4 3 3 2  
## 5 4 4 NA  
## 6 4 5 4

# d   
#d.i   
cor(mydata3[complete.cases(mydata),])

## FacTeaching FacKnowledge Housing  
## FacTeaching 1 NA NA  
## FacKnowledge NA 1 NA  
## Housing NA NA 1

#d.ii  
pair1<-cor(mydata3[complete.cases(mydata[,c(1,2)]), c(1,2)])  
pair2<-cor(mydata3[complete.cases(mydata[,c(1,3)]), c(1,3)])  
pair3<-cor(mydata3[complete.cases(mydata[,c(2,3)]), c(2,3)])  
  
pair1

## FacTeaching FacKnowledge  
## FacTeaching 1 NA  
## FacKnowledge NA 1

pair2

## FacTeaching Housing  
## FacTeaching 1 NA  
## Housing NA 1

pair3

## FacKnowledge Housing  
## FacKnowledge 1 NA  
## Housing NA 1

#d.iii

library(norm)

## Warning: package 'norm' was built under R version 4.0.3

# using the norm package get the correlation   
pre <- prelim.norm(as.matrix(mydata3))   
em <- em.norm(pre)

## Iterations of EM:   
## 1...2...3...4...5...6...

getparam.norm(pre,em,corr=TRUE)$r

## [,1] [,2] [,3]  
## [1,] 1.0000000 0.7120454 0.1541005  
## [2,] 0.7120454 1.0000000 0.2103328  
## [3,] 0.1541005 0.2103328 1.0000000

# There is no significant difference between the methods. Based on the results we can choose any  
# for the example data. In real cases the choice will depend on the data availability  
# in cases where we have less NA values then complete.cases will be best,   
# where we have less NA values per column but overall they become more then available-cases becomes # more suitable  
# mle is suitable when we want to input the data so that any value does not get discarded.