Horse-colic-analysis

### horse\_colic

### 数据摘要

* 对标称属性，给出每个可能取值的频数

table(horse.colic$surgery)

##   
## ? 1 2   
## 1 180 119

table(horse.colic$age)

##   
## 1 9   
## 276 24

table(horse.colic$temperature\_of\_extremities)

##   
## ? 1 2 3 4   
## 56 78 30 109 27

table(horse.colic$peripheral\_pulse)

##   
## ? 1 2 3 4   
## 69 115 5 103 8

table(horse.colic$mucous\_membranes)

##   
## ? 1 2 3 4 5 6   
## 47 79 30 58 41 25 20

table(horse.colic$capillary\_refill\_time)

##   
## ? 1 2 3   
## 32 188 78 2

table(horse.colic$pain)

##   
## ? 1 2 3 4 5   
## 55 38 59 67 39 42

table(horse.colic$peristalsis)

##   
## ? 1 2 3 4   
## 44 39 16 128 73

table(horse.colic$abdominal\_distension)

##   
## ? 1 2 3 4   
## 56 76 65 65 38

table(horse.colic$nasogastric\_tube)

##   
## ? 1 2 3   
## 104 71 102 23

table(horse.colic$nasogastric\_reflux)

##   
## ? 1 2 3   
## 106 120 35 39

table(horse.colic$rectal\_examination)

##   
## ? 1 2 3 4   
## 102 57 13 49 79

table(horse.colic$abdomen)

##   
## ? 1 2 3 4 5   
## 118 28 19 13 43 79

table(horse.colic$abdominocentesis\_appearance)

##   
## ? 1 2 3   
## 165 41 48 46

table(horse.colic$outcome)

##   
## ? 1 2 3   
## 1 178 77 44

table(horse.colic$surgical\_lesion)

##   
## 1 2   
## 191 109

table(horse.colic$site\_of\_lesion)

## < table of extent 0 >

# table(horse.colic$type\_of\_lesion\_1)  
table(horse.colic$cp\_data)

##   
## 1 2   
## 99 201

* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数

## Warning: 强制改变过程中产生了NA  
  
## Warning: 强制改变过程中产生了NA  
  
## Warning: 强制改变过程中产生了NA  
  
## Warning: 强制改变过程中产生了NA  
  
## Warning: 强制改变过程中产生了NA  
  
## Warning: 强制改变过程中产生了NA  
  
## Warning: 强制改变过程中产生了NA

summary(horse.colic$rectal\_temperature)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 35.40 37.80 38.20 38.17 38.50 40.80 60

summary(horse.colic$pulse)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 30.00 48.00 64.00 71.91 88.00 184.00 24

summary(horse.colic$respiratory\_rate)

## Length Class Mode   
## 300 character character

summary(horse.colic$nasogastric\_lreflux\_PH)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.000 3.000 5.000 4.708 6.500 7.500 247

summary(horse.colic$packed\_cell\_volume)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 23.0 38.0 45.0 46.3 52.0 75.0 29

summary(horse.colic$total\_protein)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 3.30 6.50 7.50 24.46 57.00 89.00 33

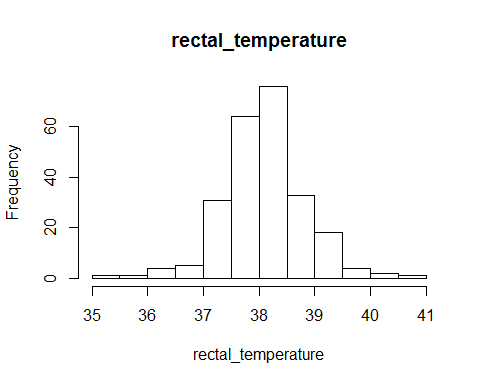
summary(horse.colic$abdomcentesis\_total\_protein)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.10 2.00 2.25 3.02 3.90 10.10 198

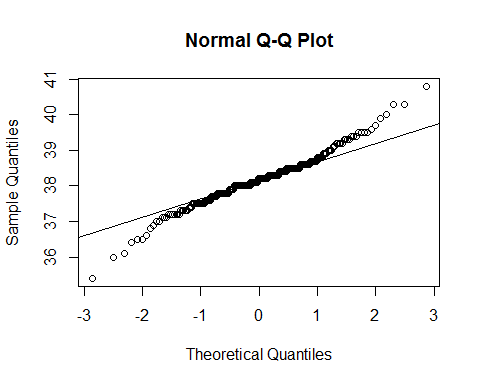
### 数据的可视化

* 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。以rectal\_temperature为例

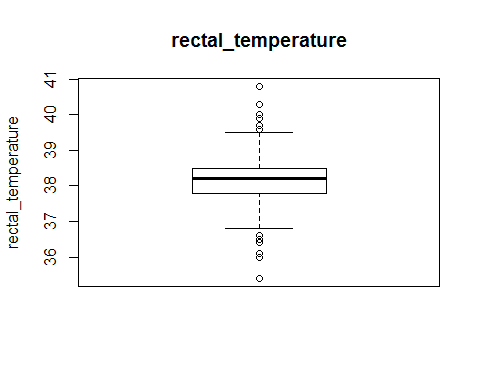
hist(horse.colic$rectal\_temperature, main = "rectal\_temperature", xlab = "rectal\_temperature")



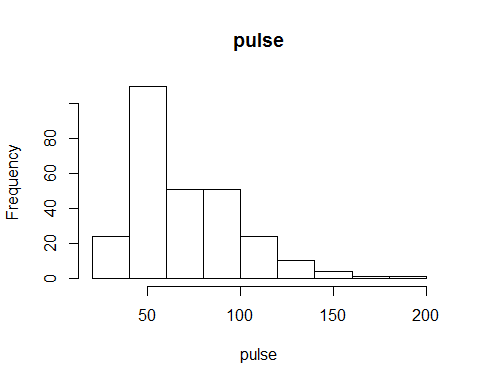
qqnorm(horse.colic$rectal\_temperature)  
qqline(horse.colic$rectal\_temperature)



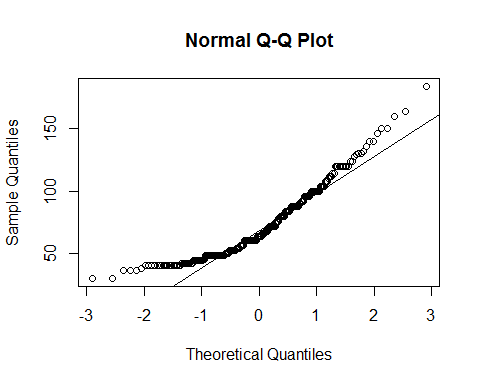
boxplot(horse.colic$rectal\_temperature, main = "rectal\_temperature", ylab = "rectal\_temperature")

 \* pluse：

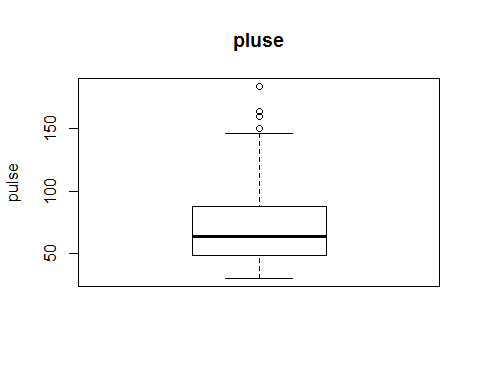
hist(horse.colic$pulse, main = "pulse", xlab = "pulse")



qqnorm(horse.colic$pulse)  
qqline(horse.colic$pulse)



boxplot(horse.colic$pulse, main = "pluse", ylab = "pulse")



### 数据缺失的处理

* 分别使用下列四种策略对缺失值进行处理: 1.将缺失部分剔除 去除缺失值超过20%的行

library(zoo)

## Warning: package 'zoo' was built under R version 3.3.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(xts)

## Warning: package 'xts' was built under R version 3.4.0

library(quantmod)

## Warning: package 'quantmod' was built under R version 3.3.3

## Loading required package: TTR

## Warning: package 'TTR' was built under R version 3.3.3

## Version 0.4-0 included new data defaults. See ?getSymbols.

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.3.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.3.3

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(DMwR)

## Warning: package 'DMwR' was built under R version 3.3.3

## Loading required package: lattice

## Loading required package: grid

newcolic <- as.matrix(horse.colic)  
newcolic <- apply(newcolic, 2, as.numeric)

## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA

## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA  
  
## Warning in apply(newcolic, 2, as.numeric): 强制改变过程中产生了NA

newcolic1 <- newcolic[-manyNAs(newcolic,0.3),]

2.对分类变量，用最高频率值来填补缺失值

newcolic2 <- newcolic1[,c(7:15,17:18,21,23:24,28)]  
newcolic2 <- centralImputation(newcolic2)

3.通过属性的相关关系来填补缺失值

newcolic3 <- newcolic1[,c(4:6,16,19:20,22)]  
c3 <- cor(newcolic1,use="complete.obs")

## Warning in cor(newcolic1, use = "complete.obs"): 标准差为零

#发现相关关系 以下为计量资料中与之相关系数最大的  
#abdomcentesis\_total\_protein 用与 respiratory\_rate 的相关关系填补   
#rectal\_temperature 用与 pulse 的相关关系来填补   
#pulse 用与 total\_protein 的相关关系填补   
#respiratory\_rate 用与 total\_protein 的相关关系来填补  
#nasogastric\_lreflux\_PH 用与 respiratory\_rate 的相关关系填补   
#packed\_cell\_volume 用与 total\_protein 的相关关系填补   
  
newcolic4 <- as.data.frame(newcolic3)  
  
lm(newcolic4$rectal\_temperature~newcolic4$pulse,data = newcolic4)

##   
## Call:  
## lm(formula = newcolic4$rectal\_temperature ~ newcolic4$pulse,   
## data = newcolic4)  
##   
## Coefficients:  
## (Intercept) newcolic4$pulse   
## 37.824413 0.004536

fillrt <- function(x){  
if(is.na(x))  
return(NA)  
else return(37.824413 + 0.004536\*x)}  
newcolic4[is.na(newcolic4$rectal\_temperature),"rectal\_temperature"] <- sapply(newcolic4[is.na(newcolic4$rectal\_temperature),"pulse"],fillrt)  
  
lm(pulse~total\_protein,data = newcolic4)

##   
## Call:  
## lm(formula = pulse ~ total\_protein, data = newcolic4)  
##   
## Coefficients:  
## (Intercept) total\_protein   
## 73.63308 -0.09722

fillrt <- function(x){  
if(is.na(x))  
return(NA)  
else return( 73.63308 - 0.09722\*x)}  
newcolic4[is.na(newcolic4$pulse),"pulse"] <- sapply(newcolic4[is.na(newcolic4$pulse),"total\_protein"],fillrt)  
  
lm(respiratory\_rate~total\_protein,data = newcolic4)

##   
## Call:  
## lm(formula = respiratory\_rate ~ total\_protein, data = newcolic4)  
##   
## Coefficients:  
## (Intercept) total\_protein   
## 31.77490 -0.08504

fillrt <- function(x){  
if(is.na(x))  
return(NA)  
else return( 31.77490 - 0.08504\*x)}  
newcolic4[is.na(newcolic4$respiratory\_rate),"respiratory\_rate"] <- sapply(newcolic4[is.na(newcolic4$respiratory\_rate),"total\_protein"],fillrt)  
  
lm(packed\_cell\_volume~total\_protein,data = newcolic4)

##   
## Call:  
## lm(formula = packed\_cell\_volume ~ total\_protein, data = newcolic4)  
##   
## Coefficients:  
## (Intercept) total\_protein   
## 46.63269 -0.01005

fillrt <- function(x){  
if(is.na(x))  
return(NA)  
else return( 46.63269 - 0.01005\*x)}  
newcolic4[is.na(newcolic4$packed\_cell\_volume),"packed\_cell\_volume"] <- sapply(newcolic4[is.na(newcolic4$packed\_cell\_volume),"total\_protein"],fillrt)  
  
lm(nasogastric\_lreflux\_PH~respiratory\_rate,data = newcolic4)

##   
## Call:  
## lm(formula = nasogastric\_lreflux\_PH ~ respiratory\_rate, data = newcolic4)  
##   
## Coefficients:  
## (Intercept) respiratory\_rate   
## 4.33626 0.01386

fillrt <- function(x){  
if(is.na(x))  
return(NA)  
else return( 4.33626 - 0.01386\*x)}  
newcolic4[is.na(newcolic4$nasogastric\_lreflux\_PH),"nasogastric\_lreflux\_PH"] <- sapply(newcolic4[is.na(newcolic4$nasogastric\_lreflux\_PH),"respiratory\_rate"],fillrt)  
  
lm(abdomcentesis\_total\_protein~respiratory\_rate,data = newcolic4)

##   
## Call:  
## lm(formula = abdomcentesis\_total\_protein ~ respiratory\_rate,   
## data = newcolic4)  
##   
## Coefficients:  
## (Intercept) respiratory\_rate   
## 2.73026 0.01137

fillrt <- function(x){  
if(is.na(x))  
return(NA)  
else return( 2.73026 + 0.01137\*x)}  
newcolic4[is.na(newcolic4$abdomcentesis\_total\_protein),"abdomcentesis\_total\_protein"] <- sapply(newcolic4[is.na(newcolic4$abdomcentesis\_total\_protein),"respiratory\_rate"],fillrt)  
  
newcolic3 <- cbind(newcolic2,newcolic4)

4.通过数据对象之间的相似性来填补缺失值

newcolic4 <- knnImputation(newcolic3,k=5,meth = "median")

5.保存

write.csv(newcolic4, file="horse\_colic\_std.csv")

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.