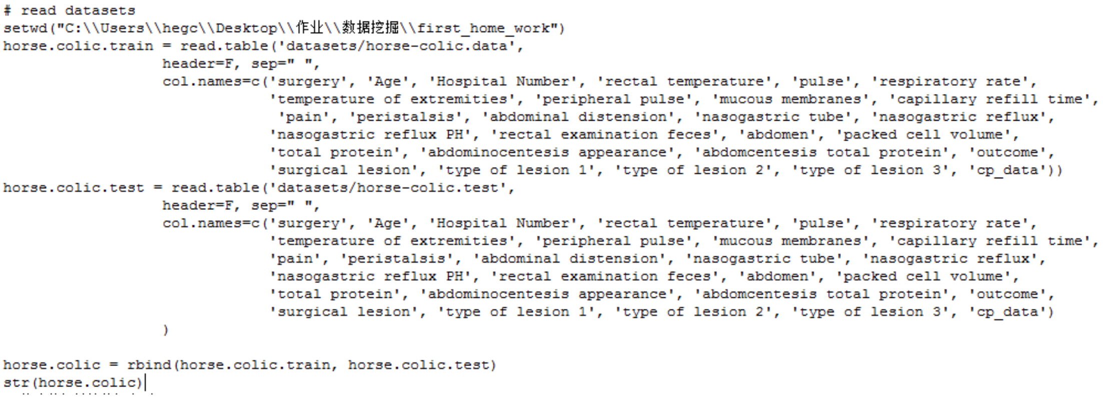
马的疝病分析

学号：2120160995 姓名：何果财
实验环境:R语言

1. 数据摘要

首先，我们需要将数据读入，然后对数据集进行预处理工作。本次作业，使用 R 语言进行数据处理，因为 R 语言有着成熟的数据分析模块，可以帮助我们快速进行处理与分析。查看数据集，发现缺失的属性值由字符 ’?’ 表示。

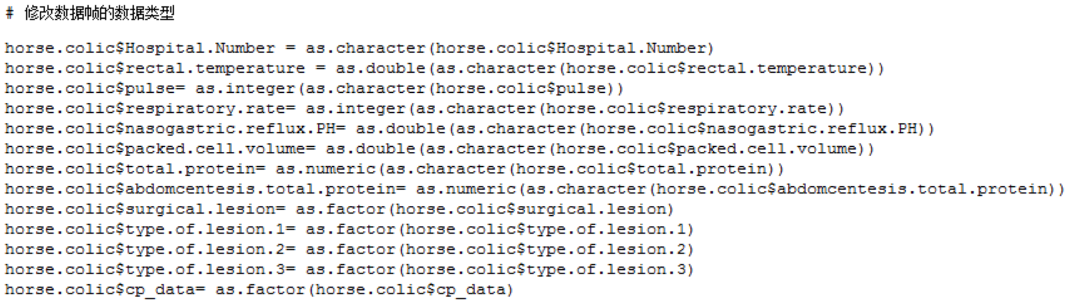
数据集共有训练数据 300 条，测试数据 68 条。每个数据点共有属性 28 个。本次作业，将把训练数据和测试数据统一处理。



```
# read datasets
setwd("C:\\Users\\hegc\\Desktop\\作业\\数据挖掘\\first_home_work")
horse.colic.train = read.table('datasets/horse-colic.data',
                  header=F, sep=" ",
                  col.names=c('surgery', 'Age', 'Hospital Number', 'rectal temperature', 'pulse', 'respiratory rate',
                              'temperature of extremities', 'peripheral pulse', 'mucous membranes', 'capillary refill time',
                               'pain', 'peristalsis', 'abdominal distension', 'nasogastric tube', 'nasogastric reflux',
                              'nasogastric reflux PH', 'rectal examination feces', 'abdomen', 'packed cell volume',
                              'total protein', 'abdominocentesis appearance', 'abdomcentesis total protein', 'outcome',
                              'surgical lesion', 'type of lesion 1', 'type of lesion 2', 'type of lesion 3', 'cp_data'))
horse.colic.test = read.table('datasets/horse-colic.test',
                  header=F, sep=" ",
                  col.names=c('surgery', 'Age', 'Hospital Number', 'rectal temperature', 'pulse', 'respiratory rate',
                              'temperature of extremities', 'peripheral pulse', 'mucous membranes', 'capillary refill time',
                              'pain', 'peristalsis', 'abdominal distension', 'nasogastric tube', 'nasogastric reflux',
                              'nasogastric reflux PH', 'rectal examination feces', 'abdomen', 'packed cell volume',
                              'total protein', 'abdominocentesis appearance', 'abdomcentesis total protein', 'outcome',
                              'surgical lesion', 'type of lesion 1', 'type of lesion 2', 'type of lesion 3', 'cp_data')
                  )

horse.colic = rbind(horse.colic.train, horse.colic.test)
str(horse.colic)
```

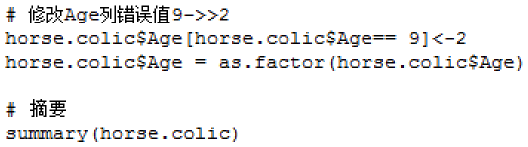
其次，我们要了解一些数据的统计特性，为后面的数据集预处理提供更多的信息。获取数据统计特性的一个方法是获取数据的描述性统计摘要。

1） 先对数据点的列属性进行修改：



```
# 修改数据帧的数据类型

horse.colic$Hospital.Number = as.character(horse.colic$Hospital.Number)
horse.colic$rectal.temperature = as.double(as.character(horse.colic$rectal.temperature))
horse.colic$pulse= as.integer(as.character(horse.colic$pulse))
horse.colic$respiratory.rate= as.integer(as.character(horse.colic$respiratory.rate))
horse.colic$nasogastric.reflux.PH= as.double(as.character(horse.colic$nasogastric.reflux.PH))
horse.colic$packed.cell.volume= as.double(as.character(horse.colic$packed.cell.volume))
horse.colic$total.protein= as.numeric(as.character(horse.colic$total.protein))
horse.colic$abdomcentesis.total.protein= as.numeric(as.character(horse.colic$abdomcentesis.total.protein))
horse.colic$surgical.lesion= as.factor(horse.colic$surgical.lesion)
horse.colic$type.of.lesion.1= as.factor(horse.colic$type.of.lesion.1)
horse.colic$type.of.lesion.2= as.factor(horse.colic$type.of.lesion.2)
horse.colic$type.of.lesion.3= as.factor(horse.colic$type.of.lesion.3)
horse.colic$cp_data= as.factor(horse.colic$cp_data)
```

2） 使用 Summary 函数分析数据摘要，注意 Age 列的数据有误，先将 Age 属性值 9 改为 2。



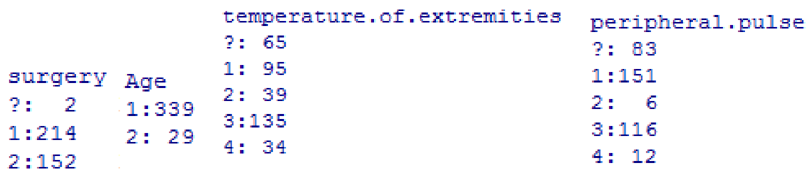
```
# 修改Age列错误值9->>2
horse.colic$Age[horse.colic$Age== 9]<-2
horse.colic$Age = as.factor(horse.colic$Age)

# 摘要
summary(horse.colic)
```

对于标称标量，可以得到每个取值的频数。例如，surgery 变量，有三个取值：？、1、2。可以看出做外科手术的比没有做过外科手术的多，值缺失的数据点有 2 个。

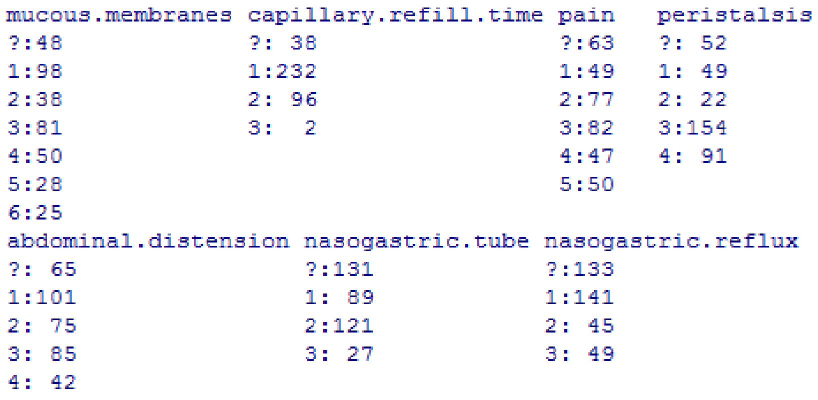
对于数值型变量，可以得到四分之一位数、中位数、均值、四分之三位数、极值等信息。

这些统计信息提供了变量值分布的初步信息，在变量有统计缺失的情况下，NA或？对应值表示缺失值的个数。通过中位数，均值，四分位数的信息，我们可以了解数据分布的偏度和分散情况，且这些信息大多数都可以通过图形来表达出来。

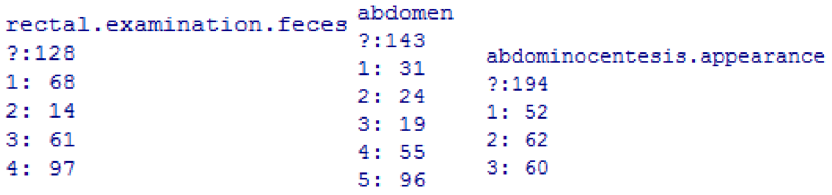
1.1 标称属性：



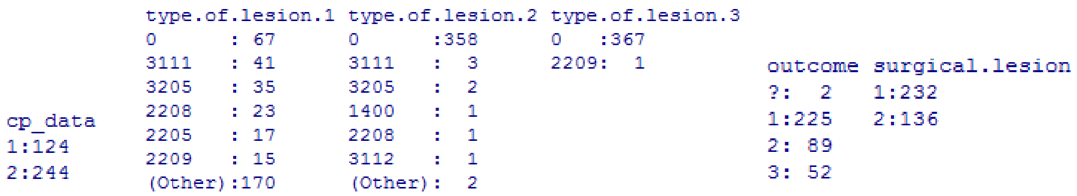
```
                      temperature.of.extremities  peripheral.pulse
                      ?: 65                       ?: 83
 surgery   Age        1: 95                       1:151
 ?:  2     1:339      2: 39                       2:  6
 1:214     2: 29      3:135                       3:116
 2:152                4: 34                       4: 12
```



```
mucous.membranes capillary.refill.time pain  peristalsis
?:48             ?: 38                 ?:63  ?: 52
1:98             1:232                 1:49  1: 49
2:38             2: 96                 2:77  2: 22
3:81             3:  2                 3:82  3:154
4:50                                   4:47  4: 91
5:28                                   5:50
6:25
abdominal.distension nasogastric.tube nasogastric.reflux
?: 65                ?:131            ?:133
1:101                1: 89            1:141
2: 75                2:121            2: 45
3: 85                3: 27            3: 49
4: 42
```

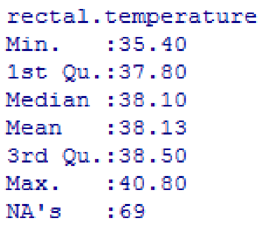


```
                          abdomen
rectal.examination.feces  ?:143
?:128                     1: 31   abdominocentesis.appearance
1: 68                     2: 24   ?:194
2: 14                     3: 19   1: 52
3: 61                     4: 55   2: 62
4: 97                     5: 96   3: 60
```

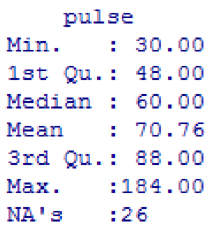


```
         type.of.lesion.1 type.of.lesion.2 type.of.lesion.3
         0       : 67     0       :358     0   :367
         3111    : 41     3111    :  3     2209:  1          outcome surgical.lesion
         3205    : 35     3205    :  2                       ?:  2   1:232
cp_data  2208    : 23     1400    :  1                       1:225   2:136
1:124    2205    : 17     2208    :  1                       2: 89
2:244    2209    : 15     3112    :  1                       3: 52
         (Other):170      (Other):  2
```

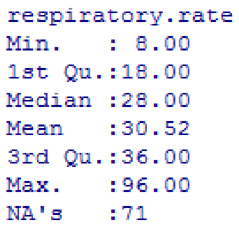
1.2 数值属性：



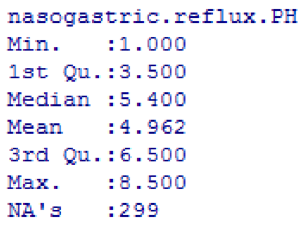
```
rectal.temperature
Min.   :35.40
1st Qu.:37.80
Median :38.10
Mean   :38.13
3rd Qu.:38.50
Max.   :40.80
NA's   :69
```



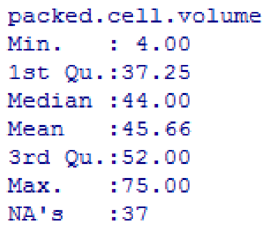
```
    pulse
Min.   : 30.00
1st Qu.: 48.00
Median : 60.00
Mean   : 70.76
3rd Qu.: 88.00
Max.   :184.00
NA's   :26
```



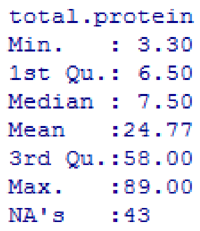
```
respiratory.rate
Min.   : 8.00
1st Qu.:18.00
Median :28.00
Mean   :30.52
3rd Qu.:36.00
Max.   :96.00
NA's   :71
```



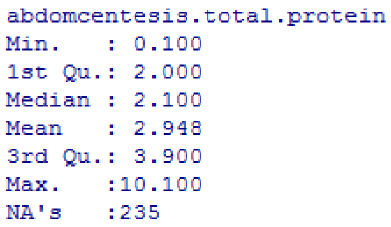
```
nasogastric.reflux.PH
Min.   :1.000
1st Qu.:3.500
Median :5.400
Mean   :4.962
3rd Qu.:6.500
Max.   :8.500
NA's   :299
```



```
packed.cell.volume
Min.   : 4.00
1st Qu.:37.25
Median :44.00
Mean   :45.66
3rd Qu.:52.00
Max.   :75.00
NA's   :37
```



```
total.protein
Min.   : 3.30
1st Qu.: 6.50
Median : 7.50
Mean   :24.77
3rd Qu.:58.00
Max.   :89.00
NA's   :43
```



```
abdomcentesis.total.protein
Min.   : 0.100
1st Qu.: 2.000
Median : 2.100
Mean   : 2.948
3rd Qu.: 3.900
Max.   :10.100
NA's   :235
```

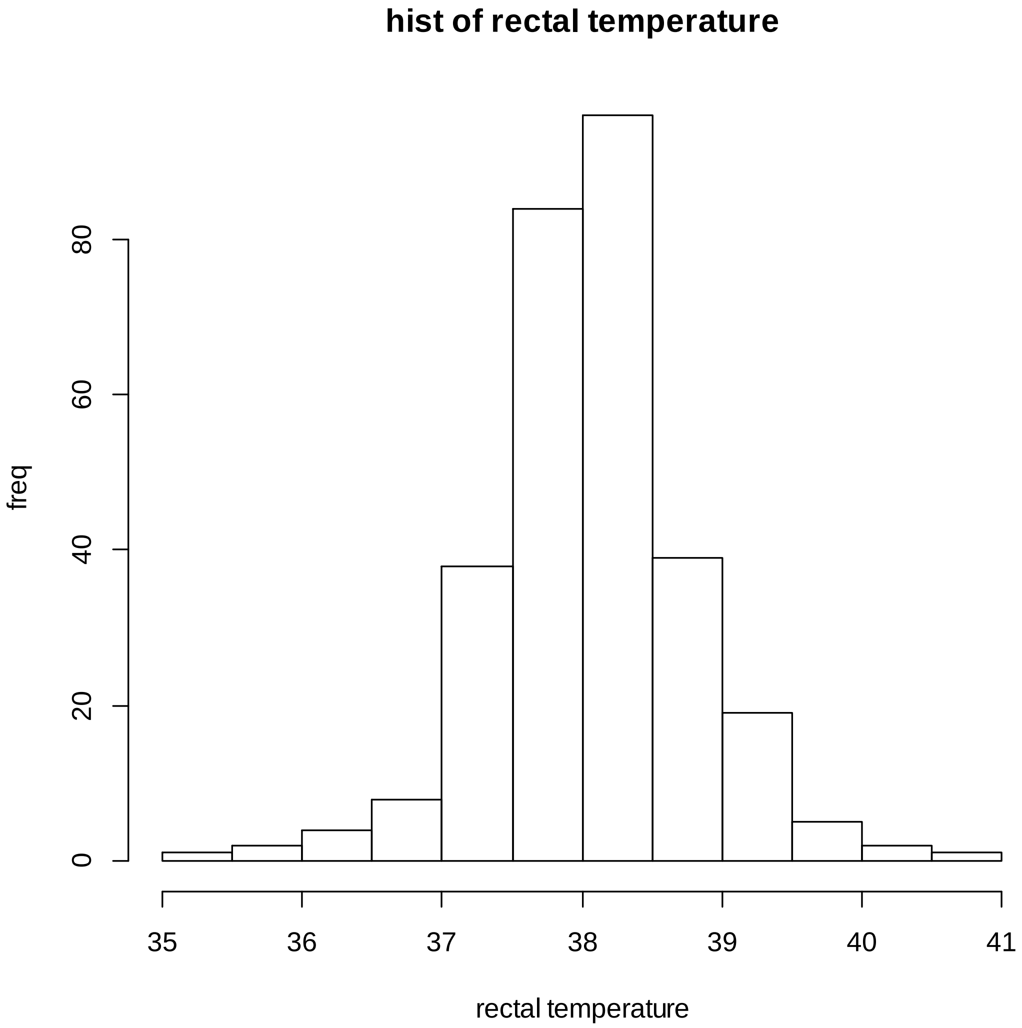
2. 数据可视化

2.1 直方图与 QQ 图

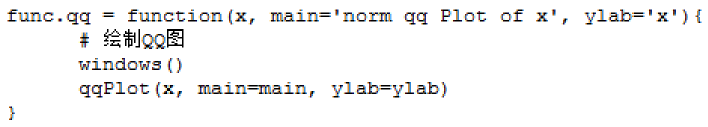
针对数值属性，绘制直方图与QQ图检验其是否符合正态分布，以Rectal temperature为例对实验结果进行分析。

绘制直方图：

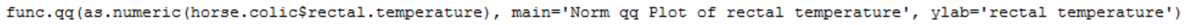
```
# 加载绘图库
library(car)

func.hist = function(x, main='hist of x', xlab='x', ylab='freq'){
       # 绘制直方图
       windows()
       hist(x, main=main, xlab=xlab, ylab=ylab)
}

func.hist(as.numeric(horse.colic$rectal.temperature), main='hist of rectal temperature', xlab='rectal temperature')
```


hist of rectal temperature
freq
0
20
40
60
80
35
36
37
38
39
40
41
rectal temperature

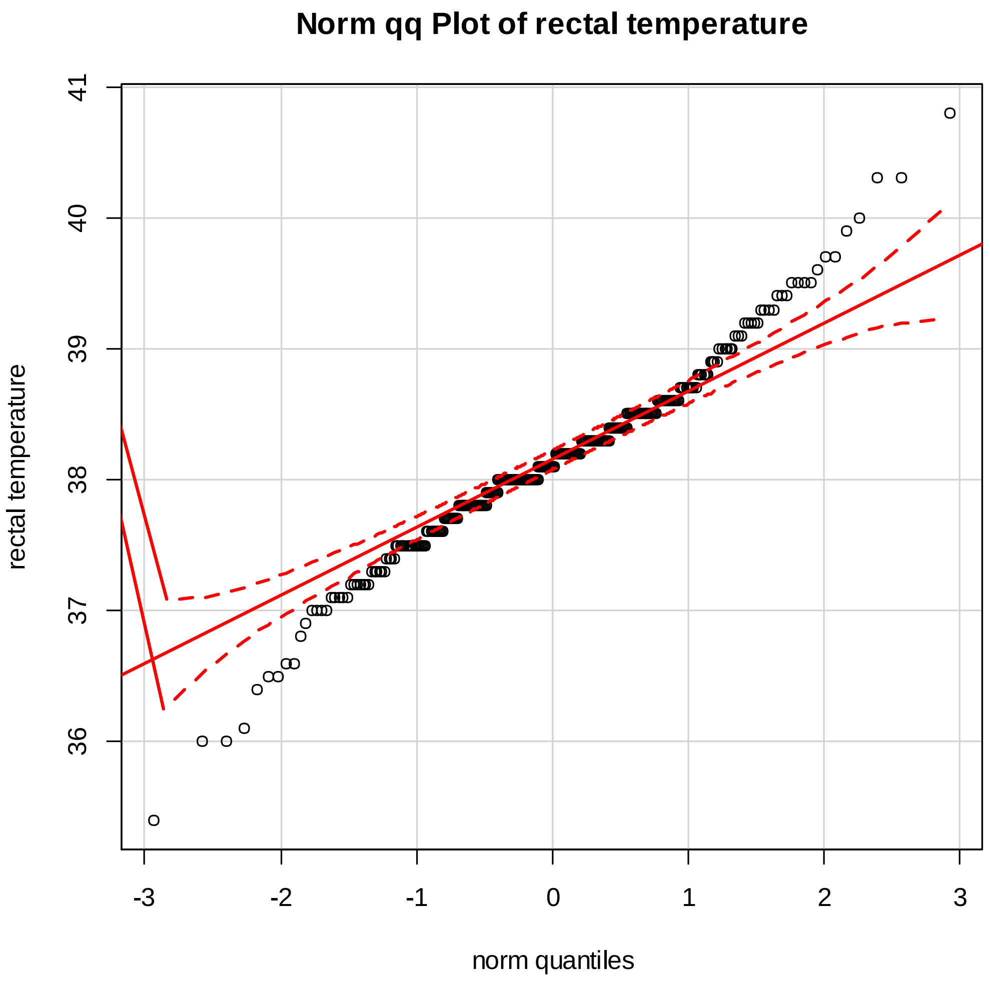
绘制 qq 图：



```
func.qq = function(x, main='norm qq Plot of x', ylab='x'){
	# 绘制QQ图
	windows()
	qqPlot(x, main=main, ylab=ylab)
}
```



```
func.qq(as.numeric(horse.colic$rectal.temperature), main='Norm qq Plot of rectal temperature', ylab='rectal temperature')
```

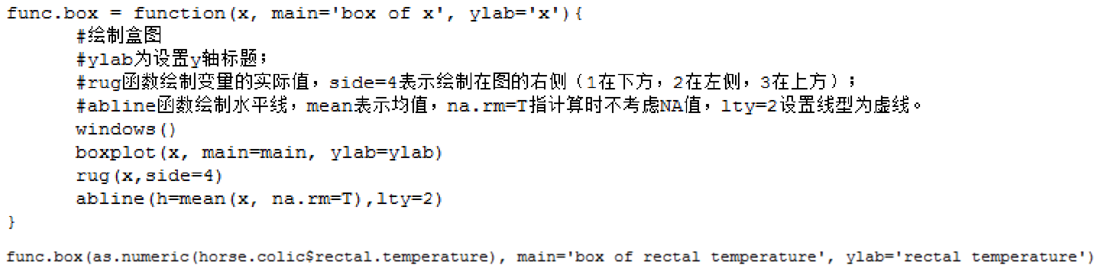


绘制出的直方图纵轴是其频数，横轴是其分布区间。QQ图中，红色实线为其QQ线，虚线为95%置信度的置信区间。

结果表明直方图显示变量Rectal temperature的分布非常接近正态分布，它的值大多数都集中在变量的均值附近；Q-Q图绘制了变量值与正态分布的理论分位数的散点图，同时他给出正态分布的95%的置信区间的带状图，除去有几个小的值明显在95%置信区间之外，基本服从正态分布。由数据图表分析可以看出，Rectal temperature基本符合正态分布。

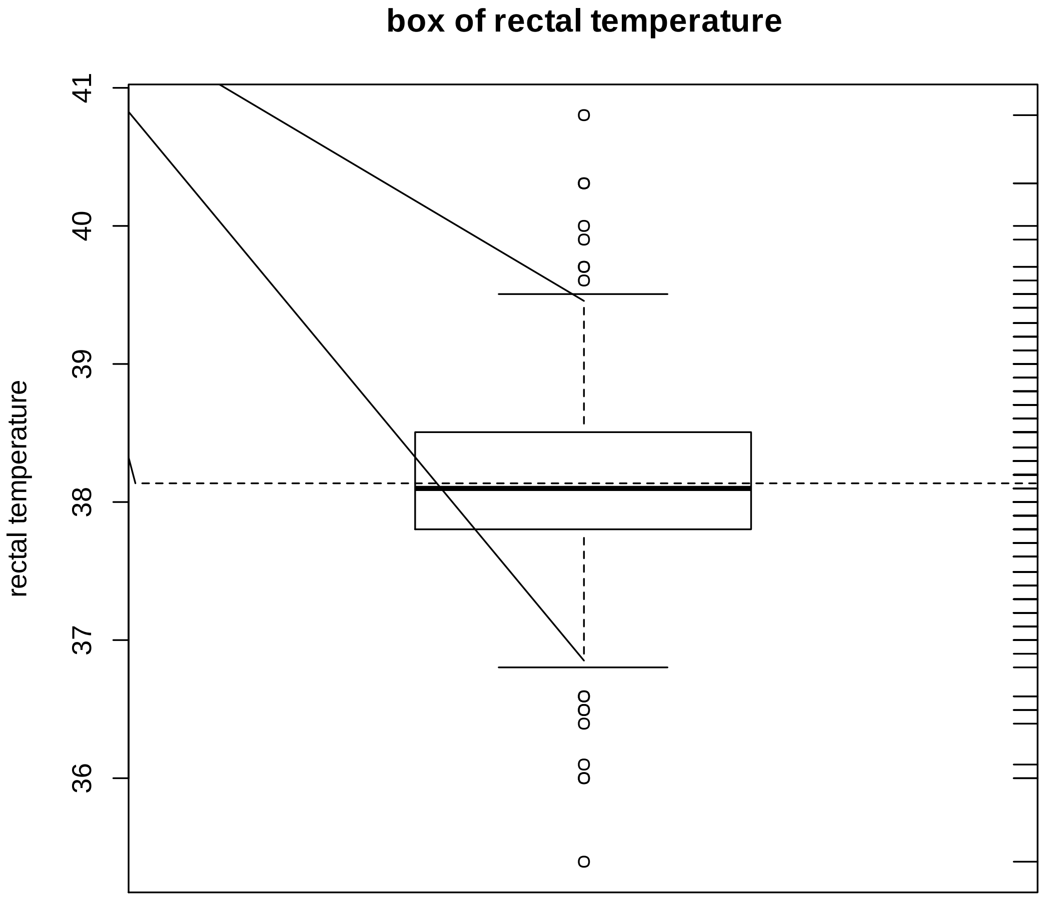
2.2 盒图:

针对数值属性，绘制盒图，识别离群点，以Rectal temperature为例对实验结果进行分析。



```
func.box = function(x, main='box of x', ylab='x'){
	#绘制盒图
	#ylab为设置y轴标题；
	#rug函数绘制变量的实际值，side=4表示绘制在图的右侧（1在下方，2在左侧，3在上方）；
	#abline函数绘制水平线，mean表示均值，na.rm=T指计算时不考虑NA值，lty=2设置线型为虚线。
	windows()
	boxplot(x, main=main, ylab=ylab)
	rug(x,side=4)
	abline(h=mean(x, na.rm=T),lty=2)
}
func.box(as.numeric(horse.colic$rectal.temperature), main='box of rectal temperature', ylab='rectal temperature')
```

结果如下：



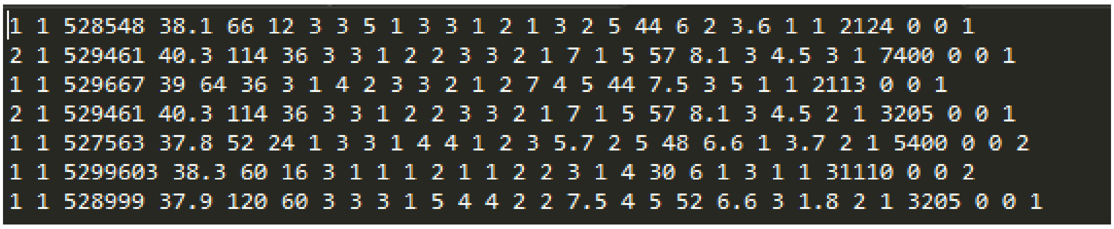
结果表明，离群点较少。小的离群点和大的离群点均为 6 个。

3. 数据缺失的处理

数据集中有很多值是缺失的，因此需要先处理数据中的缺失值。这种情形在现实问题中非常普遍，这会导致一些不能处理缺失值的分析方法无法应用。以下通过四种方式处理缺失数据。

3.1 将缺失部分剔除

1） 若将所有含缺失值得行删掉，将只剩 7 个数据点：



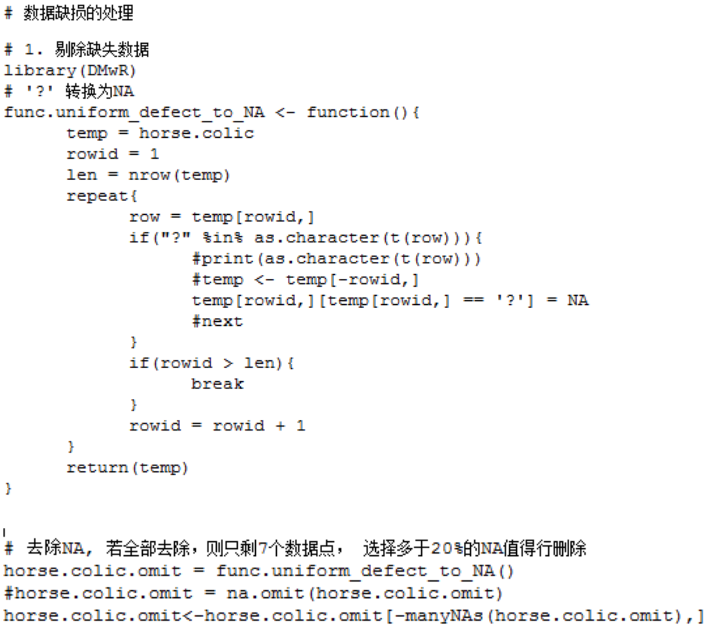
```
1 1 528548 38.1 66 12 3 3 5 1 3 3 1 2 1 3 2 5 44 6 2 3.6 1 1 2124 0 0 1
2 1 529461 40.3 114 36 3 3 1 2 2 3 3 2 1 7 1 5 57 8.1 3 4.5 3 1 7400 0 0 1
1 1 529667 39 64 36 3 1 4 2 3 3 2 1 2 7 4 5 44 7.5 3 5 1 1 2113 0 0 1
2 1 529461 40.3 114 36 3 3 1 2 2 3 3 2 1 7 1 5 57 8.1 3 4.5 2 1 3205 0 0 1
1 1 527563 37.8 52 24 1 3 3 1 4 4 1 2 3 5.7 2 5 48 6.6 1 3.7 2 1 5400 0 0 2
1 1 5299603 38.3 60 16 3 1 1 1 2 1 1 2 2 3 1 4 30 6 1 3 1 1 31110 0 0 2
1 1 528999 37.9 120 60 3 3 3 1 5 4 4 2 2 7.5 4 5 52 6.6 3 1.8 2 1 3205 0 0 1
```

2） 因此我们选择删除满足以下条件的行：

1> 包含缺失值

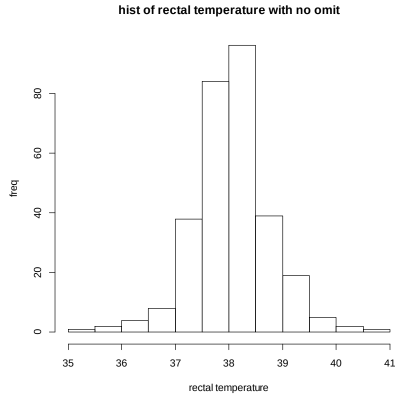
2> 缺失值个数超过 20%

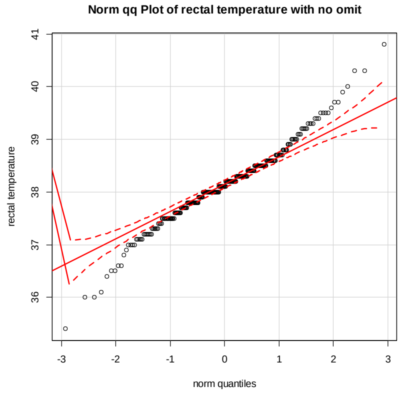
经过处理之后，得到以下结果，用原始数据与处理之后的数据以 rectal temperature 属性为例进行可视化比对。

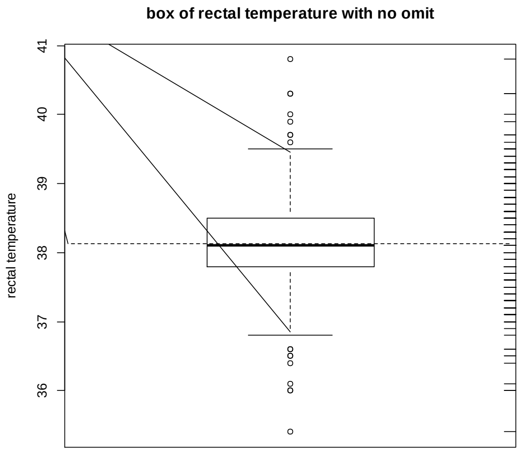


```
# 数据缺损的处理

# 1. 剔除缺失数据
library(DMwR)
# '?' 转换为NA
func.uniform_defect_to_NA <- function(){
	temp = horse.colic
	rowid = 1
	len = nrow(temp)
	repeat{
		row = temp[rowid,]
		if("?" %in% as.character(t(row))){
			#print(as.character(t(row)))
			#temp <- temp[-rowid,]
			temp[rowid,][temp[rowid,] == '?'] = NA
			#next
		}
		if(rowid > len){
			break
		}
		rowid = rowid + 1
	}
	return(temp)
}

# 去除NA, 若全部去除，则只剩7个数据点， 选择多于20%的NA值得行删除
horse.colic.omit = func.uniform_defect_to_NA()
#horse.colic.omit = na.omit(horse.colic.omit)
horse.colic.omit<-horse.colic.omit[-manyNAs(horse.colic.omit),]
```

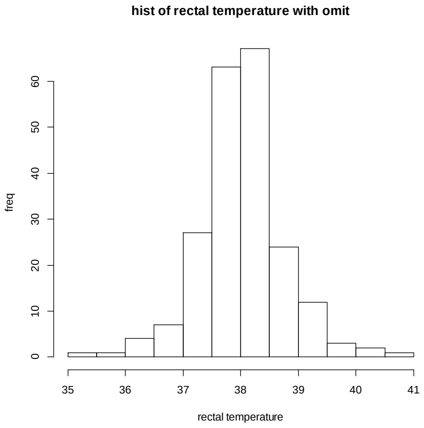
处理之前：

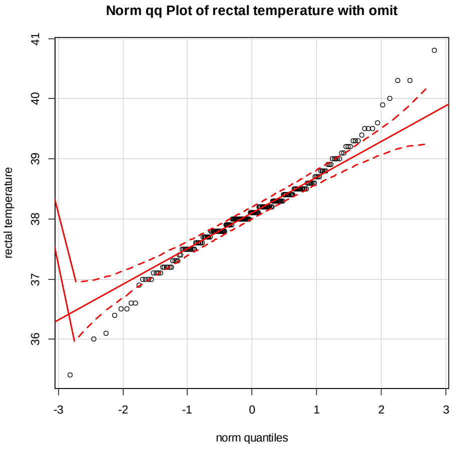


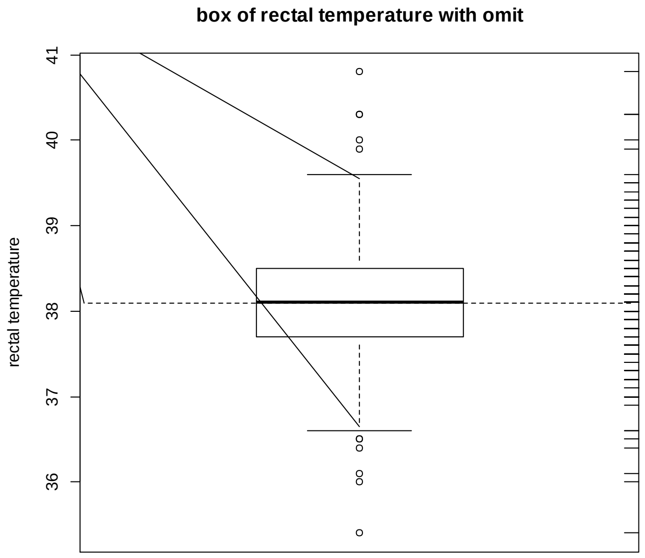



box of rectal temperature with no omit
rectal temperature
41
40
39
38
37
36

处理之后：


hist of rectal temperature with omit
freq
0
10
20
30
40
50
60
35
36
37
38
39
40
41
rectal temperature

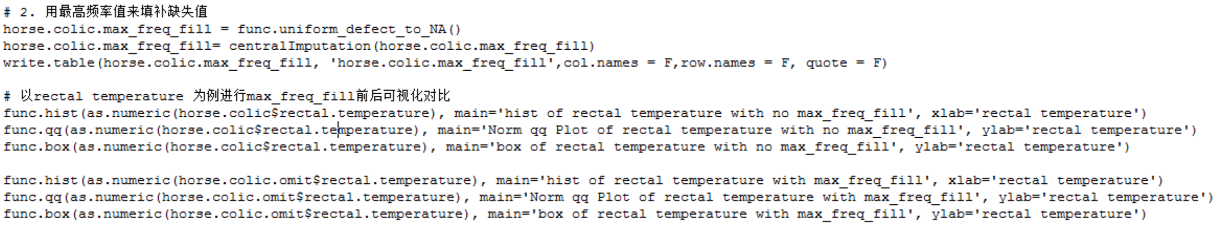

Norm qq Plot of rectal temperature with omit
rectal temperature
36
37
38
39
40
41
-3
-2
-1
0
1
2
3
norm quantiles


box of rectal temperature with omit
rectal temperature
41
40
39
38
37
36

可以看到经过处理之后，离群值变少，有利于数据的进一步分析。

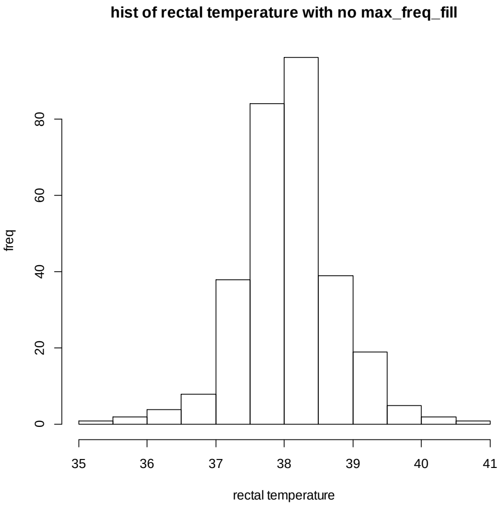
3.2 用最高频率值来填补缺失值

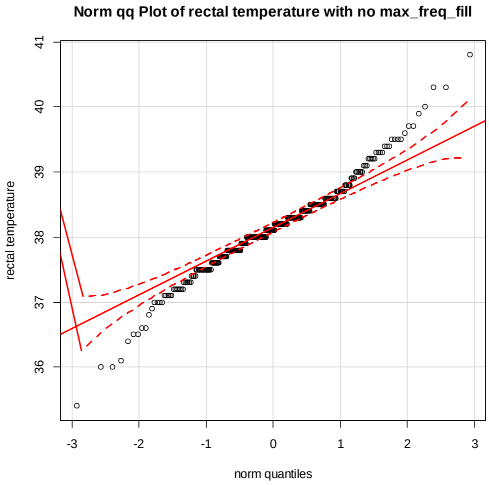
使用最高频率值来代替空缺值，使得数据分析更加的健壮。
以下是处理结果的可视化对比图，以 rectal temperature 为例。

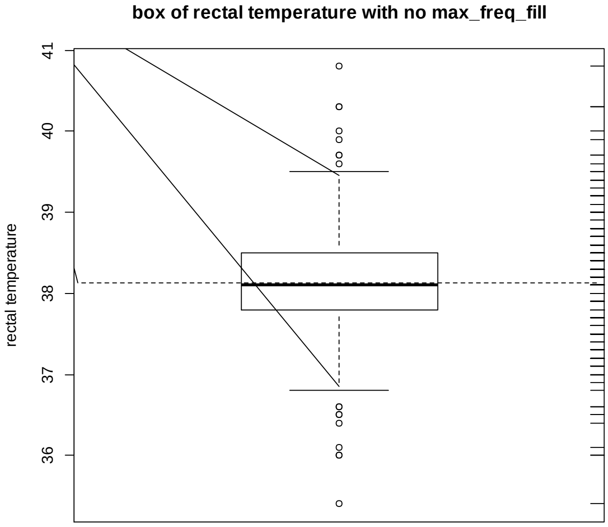


```
# 2. 用最高频率值来填补缺失值
horse.colic.max_freq_fill = func.uniform_defect_to_NA()
horse.colic.max_freq_fill= centralImputation(horse.colic.max_freq_fill)
write.table(horse.colic.max_freq_fill, 'horse.colic.max_freq_fill',col.names = F,row.names = F, quote = F)

# 以rectal temperature 为例进行max_freq_fill前后可视化对比
func.hist(as.numeric(horse.colic$rectal.temperature), main='hist of rectal temperature with no max_freq_fill', xlab='rectal temperature')
func.qq(as.numeric(horse.colic$rectal.temperature), main='Norm qq Plot of rectal temperature with no max_freq_fill', ylab='rectal temperature')
func.box(as.numeric(horse.colic$rectal.temperature), main='box of rectal temperature with no max_freq_fill', ylab='rectal temperature')

func.hist(as.numeric(horse.colic.omit$rectal.temperature), main='hist of rectal temperature with max_freq_fill', xlab='rectal temperature')
func.qq(as.numeric(horse.colic.omit$rectal.temperature), main='Norm qq Plot of rectal temperature with max_freq_fill', ylab='rectal temperature')
func.box(as.numeric(horse.colic.omit$rectal.temperature), main='box of rectal temperature with max_freq_fill', ylab='rectal temperature')
```

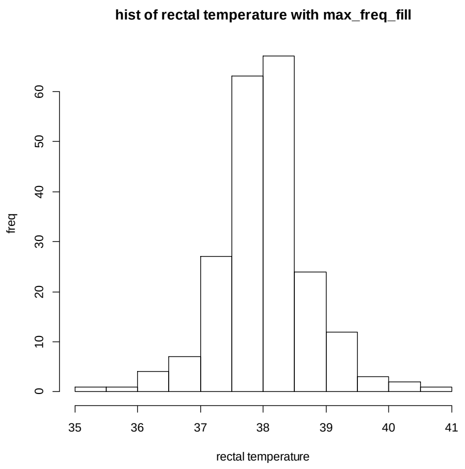
处理之前：

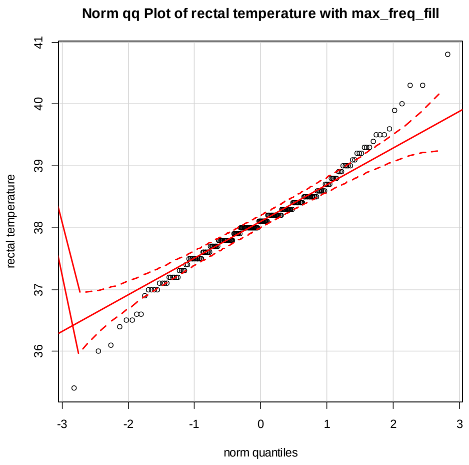


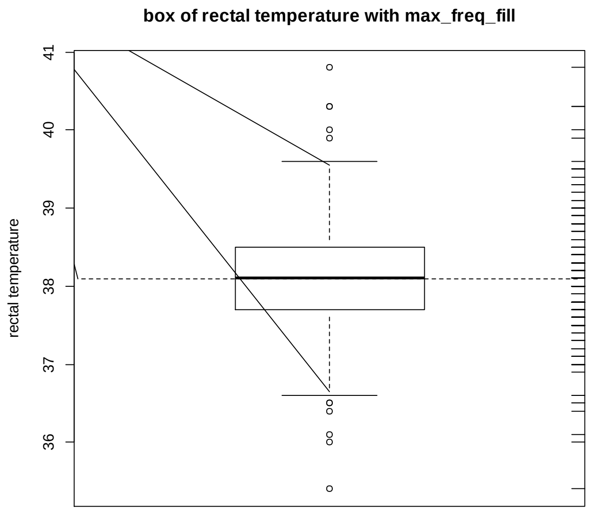




处理之后：



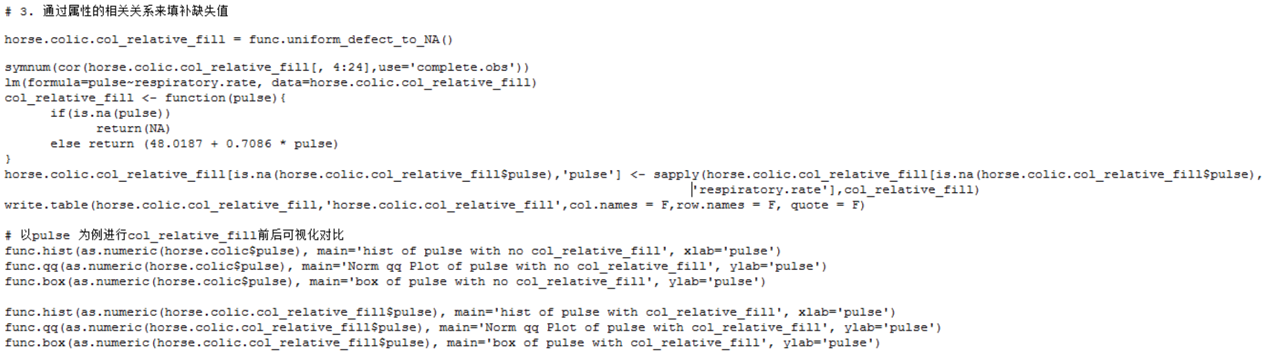




可以看到，在数据量较小的情况下，使用这种方法要优于删除缺省值的行的方法。

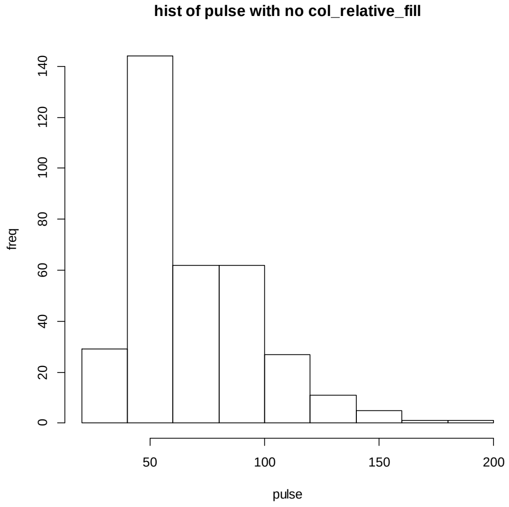
3.3 通过属性的相关关系来填补缺失值

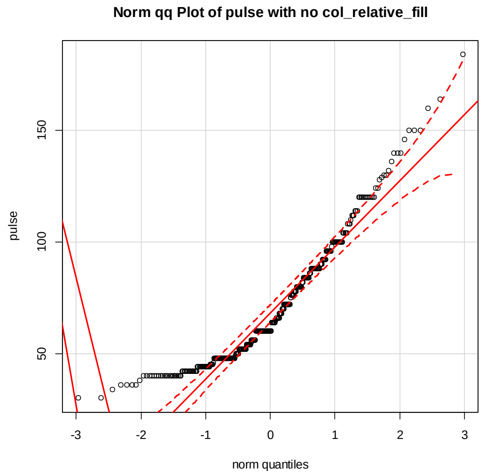
以脉搏（pulse）和呼吸频次(respiratory rate)为例：



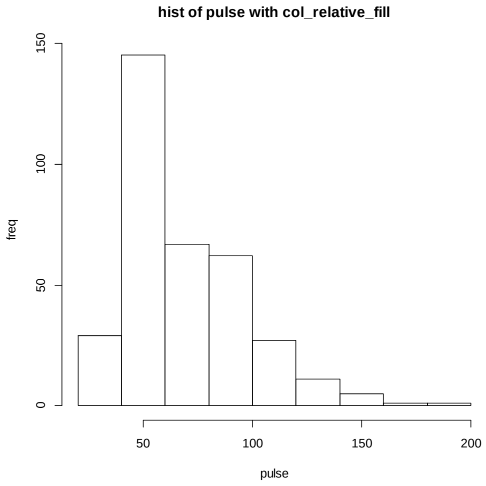
```
# 3. 通过属性的相关关系来填补缺失值

horse.colic.col_relative_fill = func.uniform_defect_to_NA()

symnum(cor(horse.colic.col_relative_fill[, 4:24],use='complete.obs'))
lm(formula=pulse~respiratory.rate, data=horse.colic.col_relative_fill)
col_relative_fill <- function(pulse){
      if(is.na(pulse))
            return(NA)
      else return (48.0187 + 0.7086 * pulse)
}
horse.colic.col_relative_fill[is.na(horse.colic.col_relative_fill$pulse),'pulse'] <- sapply(horse.colic.col_relative_fill[is.na(horse.colic.col_relative_fill$pulse),
                                                                                            'respiratory.rate'],col_relative_fill)
write.table(horse.colic.col_relative_fill,'horse.colic.col_relative_fill',col.names = F,row.names = F, quote = F)

# 以pulse 为例进行col_relative_fill前后可视化对比
func.hist(as.numeric(horse.colic$pulse), main='hist of pulse with no col_relative_fill', xlab='pulse')
func.qq(as.numeric(horse.colic$pulse), main='Norm qq Plot of pulse with no col_relative_fill', ylab='pulse')
func.box(as.numeric(horse.colic$pulse), main='box of pulse with no col_relative_fill', ylab='pulse')

func.hist(as.numeric(horse.colic.col_relative_fill$pulse), main='hist of pulse with col_relative_fill', xlab='pulse')
func.qq(as.numeric(horse.colic.col_relative_fill$pulse), main='Norm qq Plot of pulse with col_relative_fill', ylab='pulse')
func.box(as.numeric(horse.colic.col_relative_fill$pulse), main='box of pulse with col_relative_fill', ylab='pulse')
```

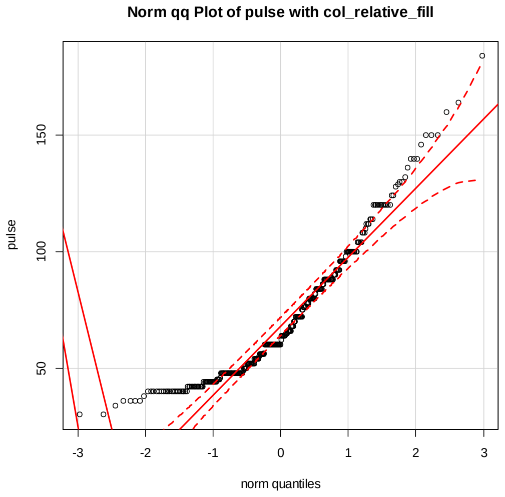
处理前：



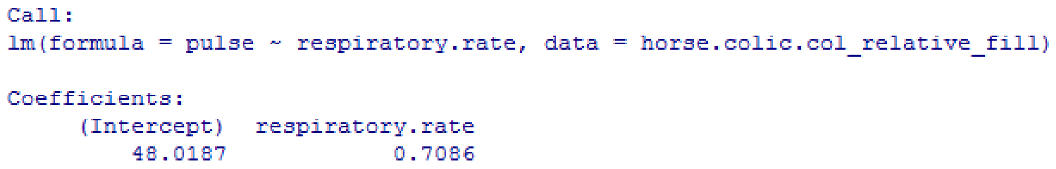


处理后：





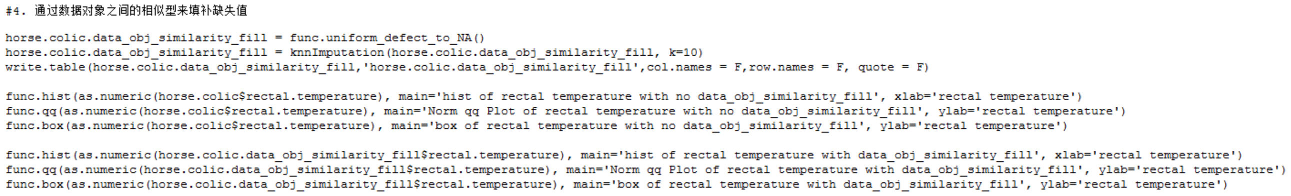
在处理的过程中，获取属性的相关性：



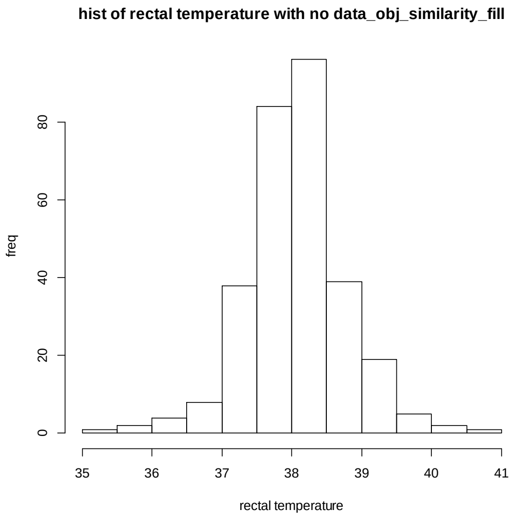
```
Call:
lm(formula = pulse ~ respiratory.rate, data = horse.colic.col_relative_fill)

Coefficients:
     (Intercept)  respiratory.rate  
         48.0187            0.7086  
```

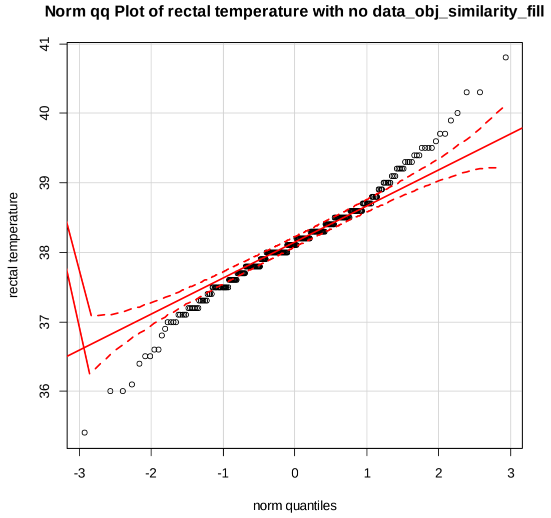
3.4 通过数据对象之间的相似性来填补缺失值

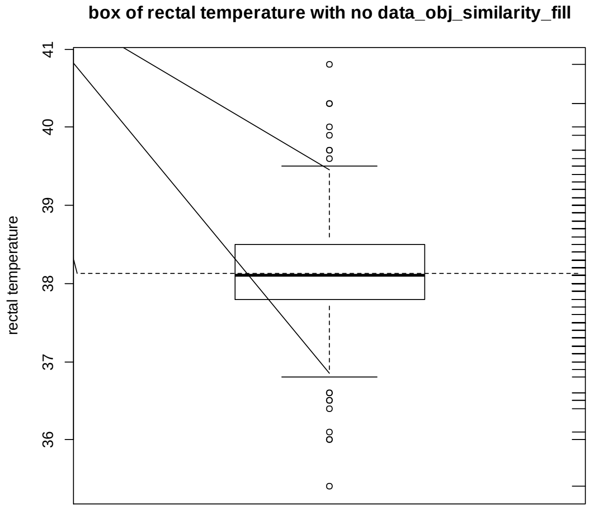
以 rectal temperature 为例，对处理前后进行可视化对比：



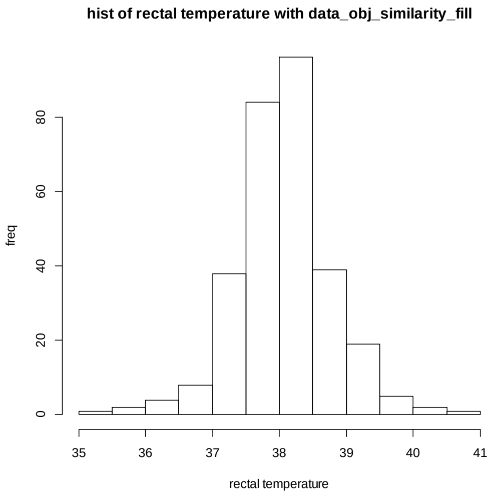
```
#4. 通过数据对象之间的相似型来填补缺失值

horse.colic.data_obj_similarity_fill = func.uniform_defect_to_NA()
horse.colic.data_obj_similarity_fill = knnImputation(horse.colic.data_obj_similarity_fill, k=10)
write.table(horse.colic.data_obj_similarity_fill,'horse.colic.data_obj_similarity_fill',col.names = F,row.names = F, quote = F)

func.hist(as.numeric(horse.colic$rectal.temperature), main='hist of rectal temperature with no data_obj_similarity_fill', xlab='rectal temperature')
func.qq(as.numeric(horse.colic$rectal.temperature), main='Norm qq Plot of rectal temperature with no data_obj_similarity_fill', ylab='rectal temperature')
func.box(as.numeric(horse.colic$rectal.temperature), main='box of rectal temperature with no data_obj_similarity_fill', ylab='rectal temperature')

func.hist(as.numeric(horse.colic.data_obj_similarity_fill$rectal.temperature), main='hist of rectal temperature with data_obj_similarity_fill', xlab='rectal temperature')
func.qq(as.numeric(horse.colic.data_obj_similarity_fill$rectal.temperature), main='Norm qq Plot of rectal temperature with data_obj_similarity_fill', ylab='rectal temperature')
func.box(as.numeric(horse.colic.data_obj_similarity_fill$rectal.temperature), main='box of rectal temperature with data_obj_similarity_fill', ylab='rectal temperature')
```

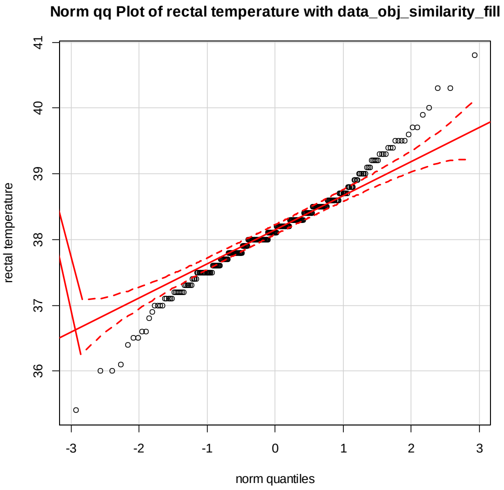

hist of rectal temperature with no data_obj_similarity_fill
freq
0
20
40
60
80
35
36
37
38
39
40
41
rectal temperature

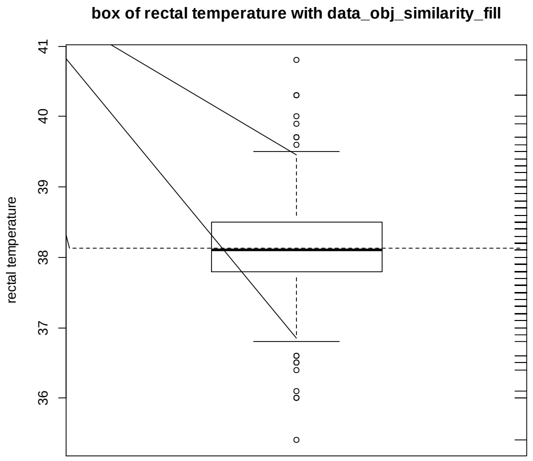

Norm qq Plot of rectal temperature with no data_obj_similarity_fill
rectal temperature
36
37
38
39
40
41
-3
-2
-1
0
1
2
3
norm quantiles


box of rectal temperature with no data_obj_similarity_fill
rectal temperature
36
37
38
39
40
41

处理后：


hist of rectal temperature with data_obj_similarity_fill
freq
0
20
40
60
80
35
36
37
38
39
40
41
rectal temperature


Norm qq Plot of rectal temperature with data_obj_similarity_fill
rectal temperature
36
37
38
39
40
41
-3
-2
-1
0
1
2
3
norm quantiles


box of rectal temperature with data_obj_similarity_fill
rectal temperature
41
40
39
38
37
36