#### 3.1 数据可视化和摘要

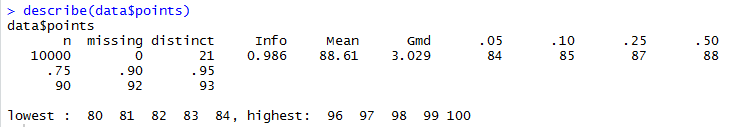
##### 数据摘要

* 对标称属性，给出每个可能取值的频数，
* 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。

1. 对属性country每个取值给出频数



2.对数值属性points给出最大、最小、均值、中位数、四分位数及缺失值的个数

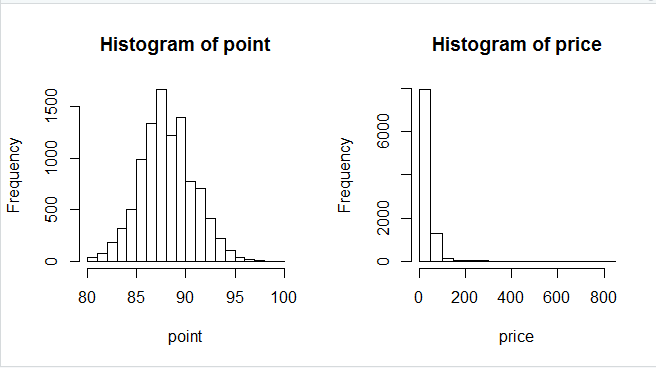


##### 数据的可视化

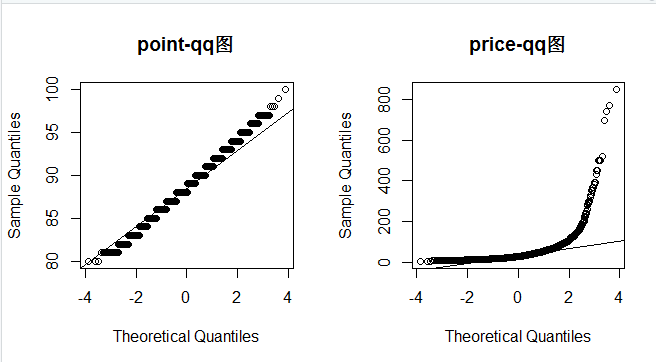
针对数值属性，

* 绘制直方图，用qq图检验其分布是否为正态分布。
* 绘制盒图，对离群值进行识别

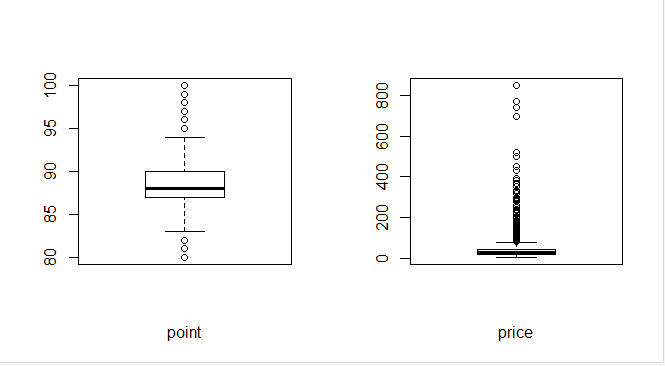
1.对points、price两个数值属性绘制直方图



1. 用qq图检验其分布是否为正态分布



1. 绘制盒图，对离群值进行识别



#### 3.2 数据缺失的处理

观察数据集中缺失数据，分析其缺失的原因。

分别使用下列四种策略对缺失值进行处理:

* 将缺失部分剔除
* 用最高频率值来填补缺失值
* 通过属性的相关关系来填补缺失值
* 通过数据对象之间的相似性来填补缺失值

处理后，可视化地对比新旧数据集。

#可视化缺失

matrixplot(data)

#1.直接剔除

newdata<-na.omit(data)

matrixplot(newdata)

#2.

par(mfrow=c(1,2))

newdata <- data

matrixplot(data)

#newdata<-newdata[-manyNAs(newdata),]

newdata <- centralImputation(newdata)

matrixplot(newdata)

#newdata$price

#3.

symnum(cor(data[,4:5],use="complete.obs"))

#4.

clean.data<-knnImputation(newdata,k=10,meth="median")

