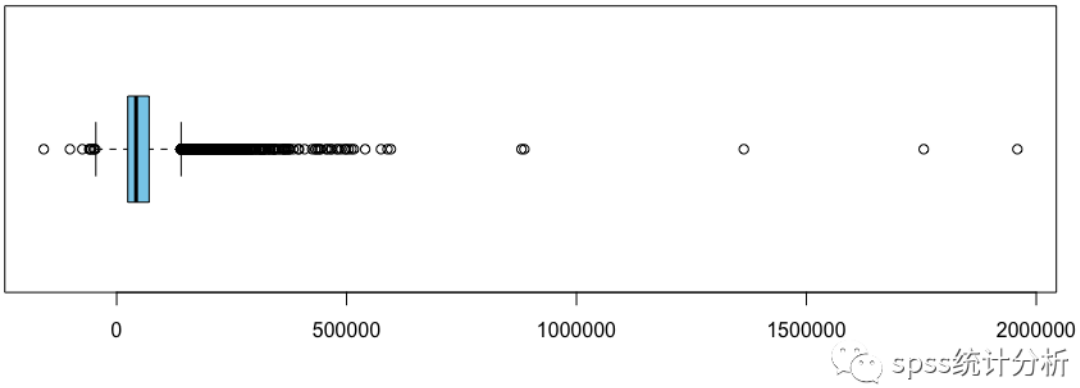
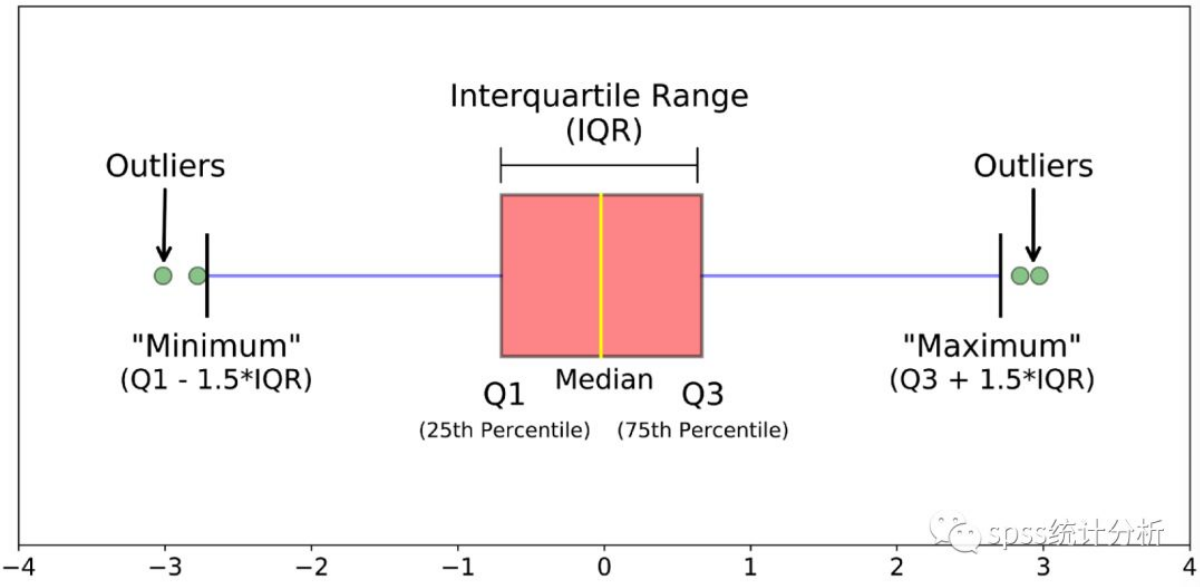
# 82.08 剔除箱线图boxplot中的异常值

如果绘制包含异常值的箱线图boxplot，可能画出如下形态。可见，下图包含大量异常数据点，偏离中间的数据主体。




看到此图，你可能会很自然的问出一个问题：箱线图显示异常数据的依据是什么？我们来看下面这个简单的箱线图：箱线图中间是一个箱体，也就是粉红色部分，箱体左边，中间，右边分别有一条线，左边是下四分位数（Q1），右边是上四分位数（Q3），中间是中位数（Median）,上下四分位数之差是四分位距（IQR），用 Q1–1.5IQR 得到下边缘（最小值），Q3+1.5IQR 得到上边缘（最大值）。在上边缘之外的数据就是极大异常值，在下边缘之外的数据极小异常值，总之在上下边缘之外的数据就是异常值。




搞清楚异常数据的产生原理之后，想要剔除它们就十分简单了，这里给出一个简单的步骤：

求出这组数据的四分位数，其中上下四分位数分别为 Q3,Q1；

求出这组数据的四分位距 IQR；

Q1– 1.5IQR 得到这组数据的下边缘；

Q3+1.5IQR 得到这组数据的上边缘；

过滤掉上下边缘之外的异常数据；

用R剔除箱线图中的异常值

下面给出用R 代码实现的数据异常值处理函数，该函数实现功能为：输入一个数据集，指定该数据集中的某个数值型变量，返回根据上述规则剔除异常值后的数据。需要注意的是，代码中使用了管道符号（%>%）以及 dplyr 包，因此在使用这段代码时，你需要导入 tidyverse 这个包。

```r
# 将数据中偏大，偏小的“异常数据”过滤掉
drop_outliers <- function(df_name,dep_col) {
  a <- df_name[,dep_col]
  iqr <- IQR(a)
  q1 <- as.numeric(quantile(a,0.25))
  q3 <- as.numeric(quantile(a,0.75))
  bottom <- q1-1.5*iqr
  top <- q3+1.5*iqr
  df_return <- df_name %>%
    filter(df_name[,dep_col]>=bottom&df_name[,dep_col]<=top)
  return(df_return)
}
```



我们剔除其中的异常数据后，得到箱线图如下图所示。你会发现，这个箱线图依然存在异常值，这是为什么呢?