OQC入库终检的原始数据分析

一　数据描述

有两个型号的产品：1） 92s7 2）2A ，如图是数据示例。

1	2	3	4	5	6	7
1	58.19	13.71	3.91	30.11	34.85	42.07
1	58.07	13.63	3.75	29.91	35.07	41.90
1	58.13	13.72	3.92	30.13	34.91	42.11
1	58.11	13.67	3.78	29.87	35.07	41.98
1	58.21	13.70	3.80	30.13	34.87	42.07
1	58.13	13.65	3.81	29.78	35.01	41.99
1	58.24	13.76	3.87	30.15	34.92	42.11
1	58.19	13.67	3.76	29.91	35.10	41.93
1	58.27	13.71	3.85	30.17	34.95	41.80
1	58.17	13.69	3.74	29.94	35.03	41.95
1	58.14	13.71	3.81	30.07	34.85	42.07
1	58.11	13.61	3.78	29.95	35.01	41.91
1	58.11	13.74	3.87	30.11	34.91	42.13
1	58.13	13.67	3.76	29.94	35.07	41.93
1	58.17	13.75	3.89	30.07	34.95	42.10
1	58.10	13.69	3.75	29.80	35.11	41.98
1	58.16	13.78	3.93	30.13	34.95	42.13
1	58.09	13.65	3.79	29.94	35.07	41.87
1	58.21	13.79	3.91	30.07	34.93	42.17
1	58.03	13.67	3.81	29.87	35.12	41.80
1	58.31	13.70	3.92	30.11	34.95	42.07
1	58.11	13.63	3.79	29.91	35.07	41.92
1	58.27	13.72	3.84	30.12	34.91	42.11

以型号为92s7为例

1） 数据时间跨度较大，有17年8月4日（A）, 17年8月23日（A），2017年9月18日（B）, 17年10月31日（B）, 18年3月16日（A）

2） 数据有12个维度，11个未命名的尺寸变量和一个热阻变量。我们将热阻视为。在不同时期热阻的统计方法可能存在差异，如标记为（A）的数据热阻基本在一个数量级上为0.2左右，（B）为10+，或者40-50可能存在记录问题。

发现：由于这是同一种型号的产品，这说明该企业在质量控制方面存在没有统一标准情况，进一步推理可能没有质量管理SOP或者工作流程等。但是

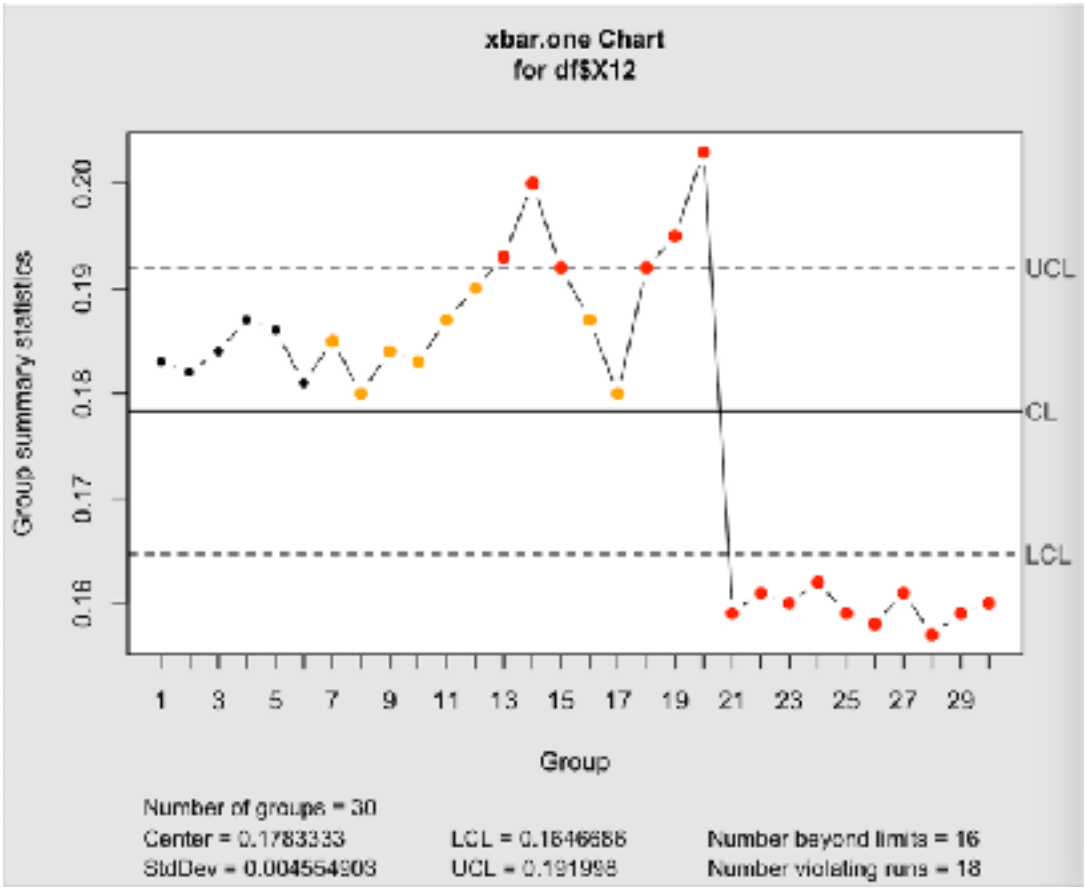
二 对A组数据进行分析

由于三组数据来源于同一型号的产品，于是将三组数据合并来分析。得到30组数据，每组12个变量。

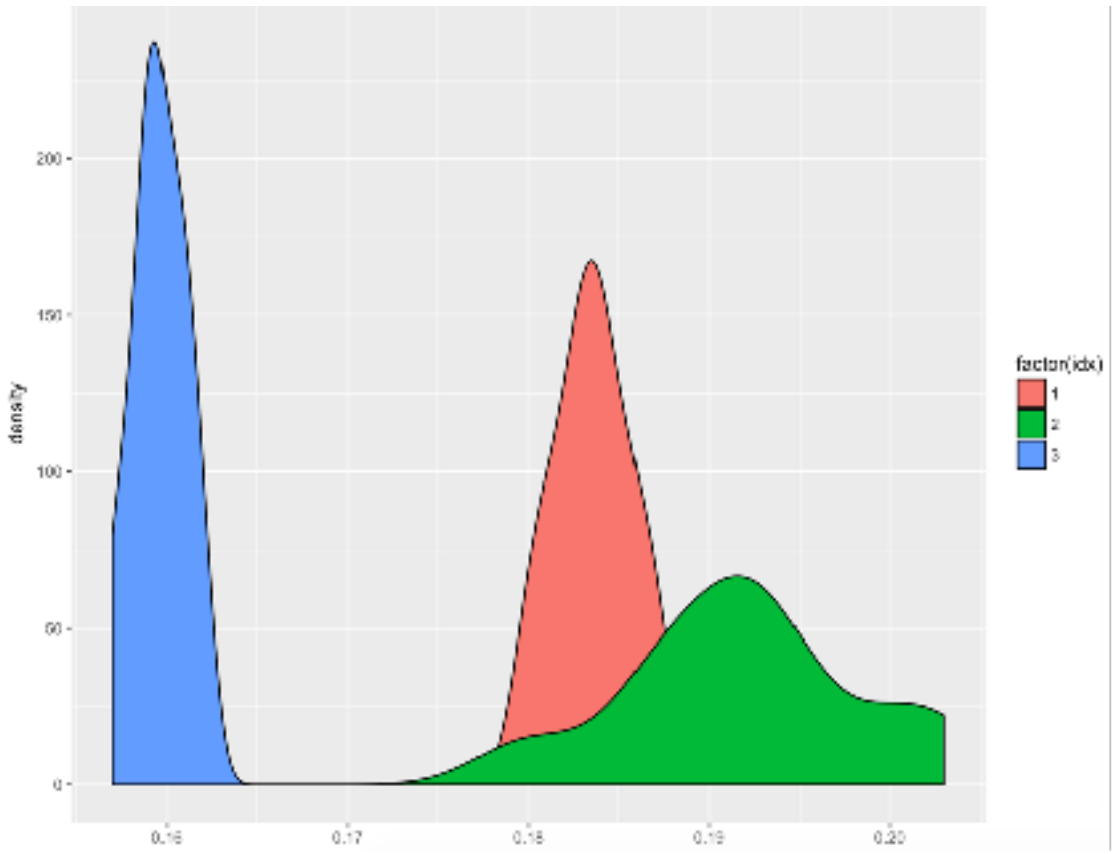
1）虽然是同一个型号的数据，但是不同时间的热阻值存在显著差异

数据日期	热阻均值	热阻方差
17年8月4日	0.184	0.00000472
17年8月23日	0.192	0.0000437
18年3月16日	0.160	0.00000227

通过控制图来可视化其模式变化

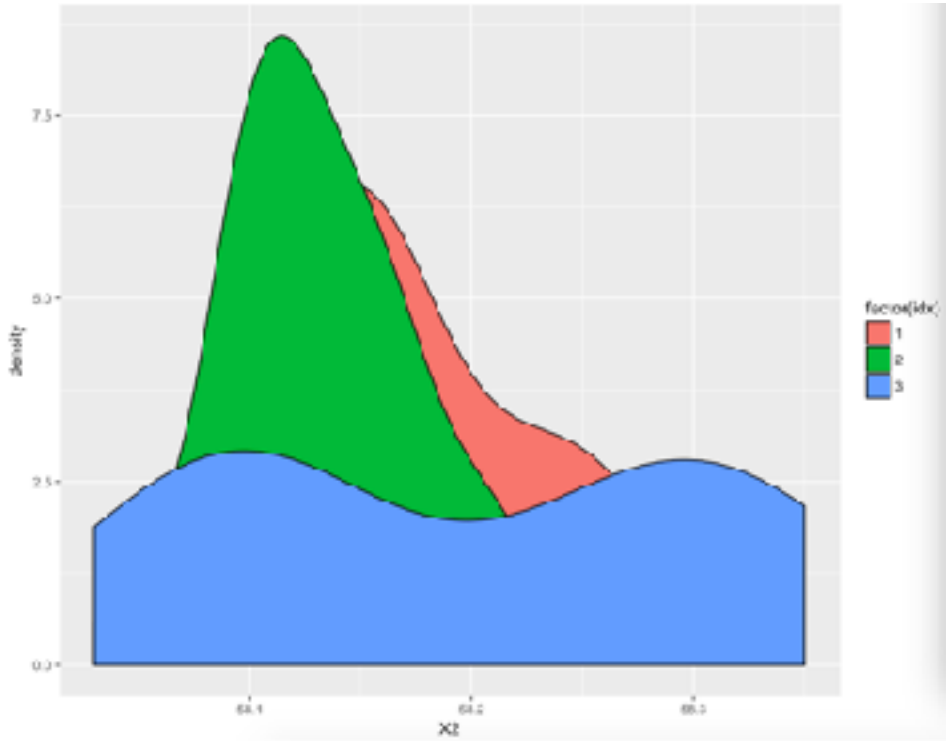


这些热阻值可能来自不同的分布。

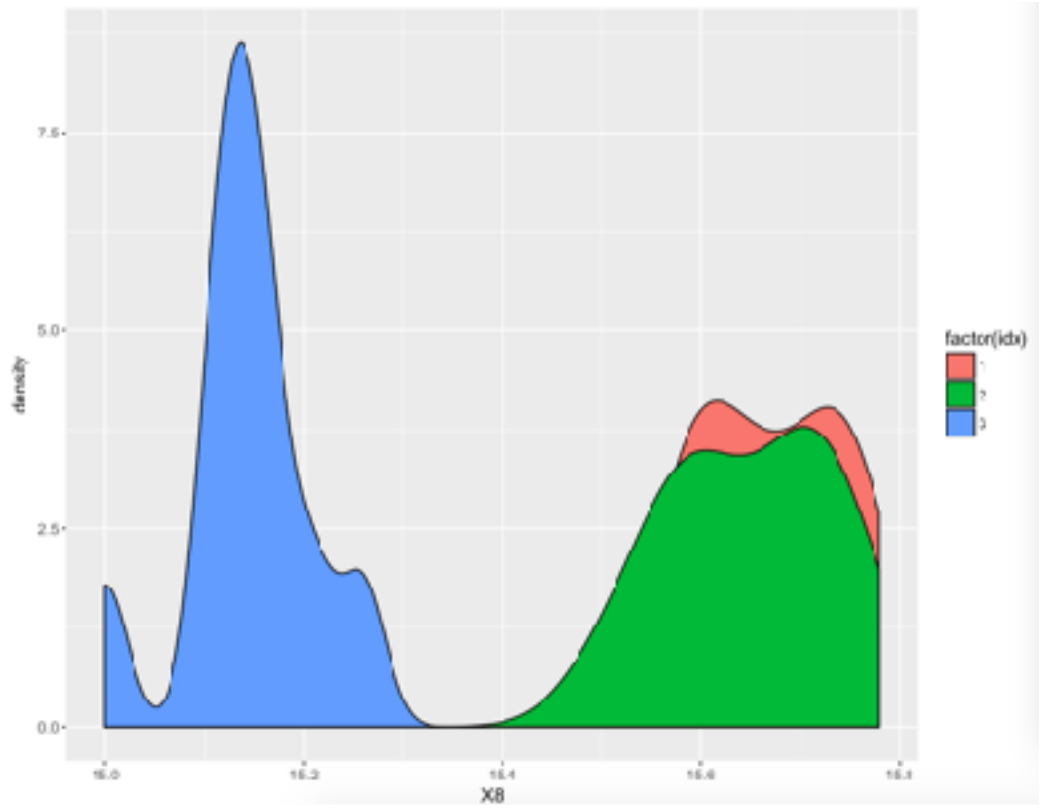


2） 其他11个自变量也存在热阻变量类似的问题

如X2变量



X8变量



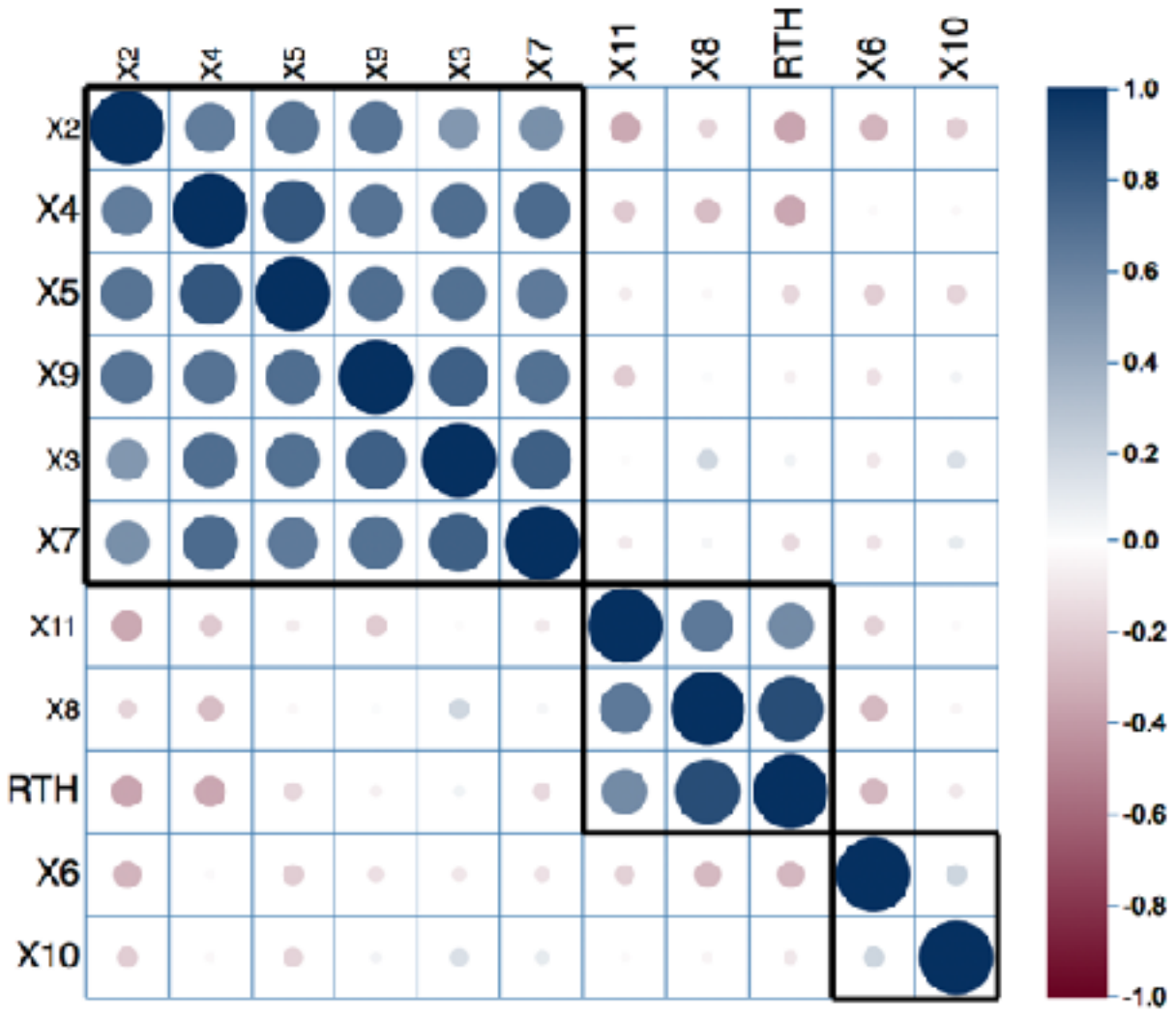
发现：经过可视化，我们有理由相信这些数据来源与不同分布，这和质量控制，控制图的的假设相左，说明该企业存在较明显的质量问题，各个工艺在不同时间波动很大。

三 挖掘显著影响热阻变化的因素

1） 相关分析

X8，X11和RTH具有较强的相关性

分别为0.857和0.559



2）线性回归

RTH作为因变量，其他11个变量作为自变量

```
Call:
lm(formula = X12 ~ ., data = df[, -13])

Residuals:
       Min         1Q     Median         3Q        Max
-0.0100340 -0.0041117 -0.0007711  0.0032292  0.0153916

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.186777   2.487850   0.477   0.6388
X1                NA         NA      NA       NA
X2          -0.072312   0.027239  -2.655   0.0156 *
X3           0.022264   0.066296   0.336   0.7407
X4           0.022742   0.048842   0.466   0.6468
X5          -0.010217   0.024013  -0.425   0.6753
X6          -0.002797   0.001700  -1.646   0.1163
X7          -0.030042   0.021452  -1.400   0.1775
X8           0.044869   0.008428   5.324 3.88e-05 ***
X9           0.047135   0.030987   1.521   0.1447
X10         -0.021329   0.018612  -1.146   0.2660
X11         -0.016876   0.027621  -0.611   0.5485
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007067 on 19 degrees of freedom
Multiple R-squared:  0.8438,    Adjusted R-squared:  0.7615
F-statistic: 10.26 on 10 and 19 DF,  p-value: 1.014e-05
```

发现： X8和X2显著影响RTH

这里可以利用R方差来筛选对方差解释更多的变量。该变量就为显著影响RTH的变量。

3）主成分分析

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
Standard deviation     2.1120118 1.3660294 1.1453107 0.8803766 0.70262639 0.5900652
Proportion of Variance 0.4460620 0.1866036 0.1311737 0.0775063 0.04936838 0.0348177
Cumulative Proportion  0.4460620 0.6326656 0.7638393 0.8413456 0.89071394 0.9255316
                           Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.51137375 0.45927597 0.39584701 0.33992830
Proportion of Variance 0.02615031 0.02109344 0.01566949 0.01155512
Cumulative Proportion  0.95168195 0.97277539 0.98844488 1.00000000
```

四个主成分可以解释84%的方差。

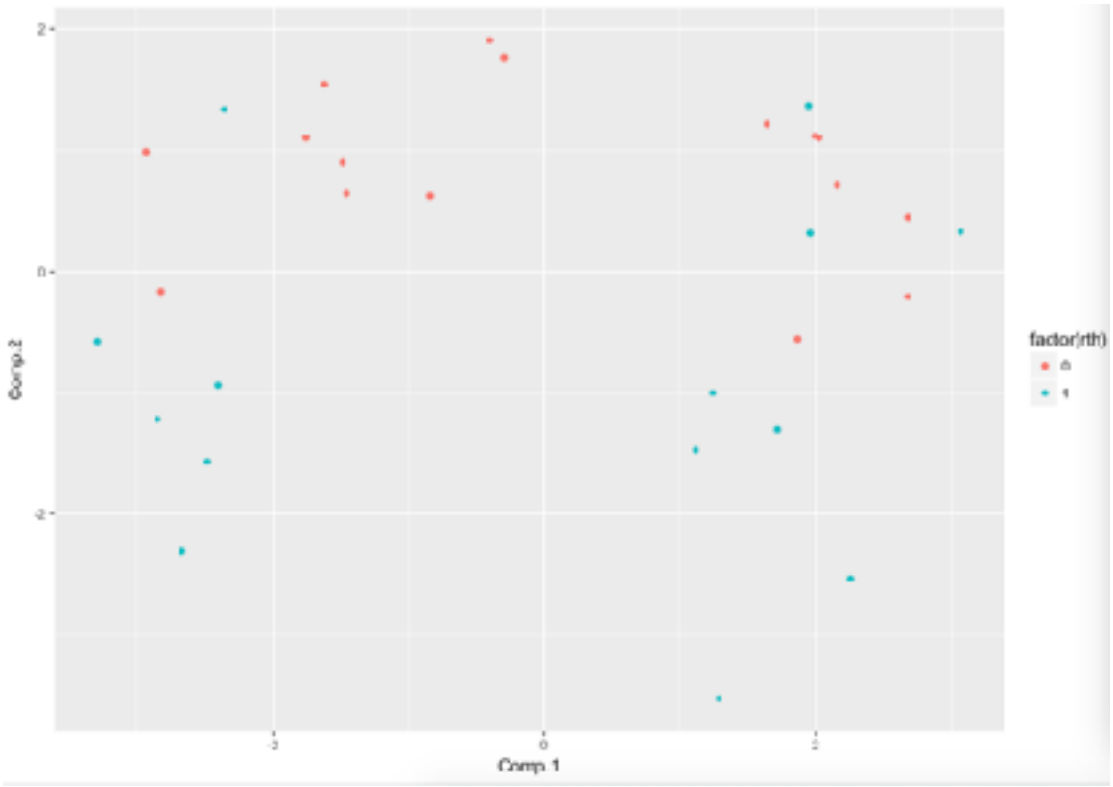
这是RTH的四分位，

```
   0%   25%   50%   75%  100%
0.157 0.161 0.183 0.187 0.203
```

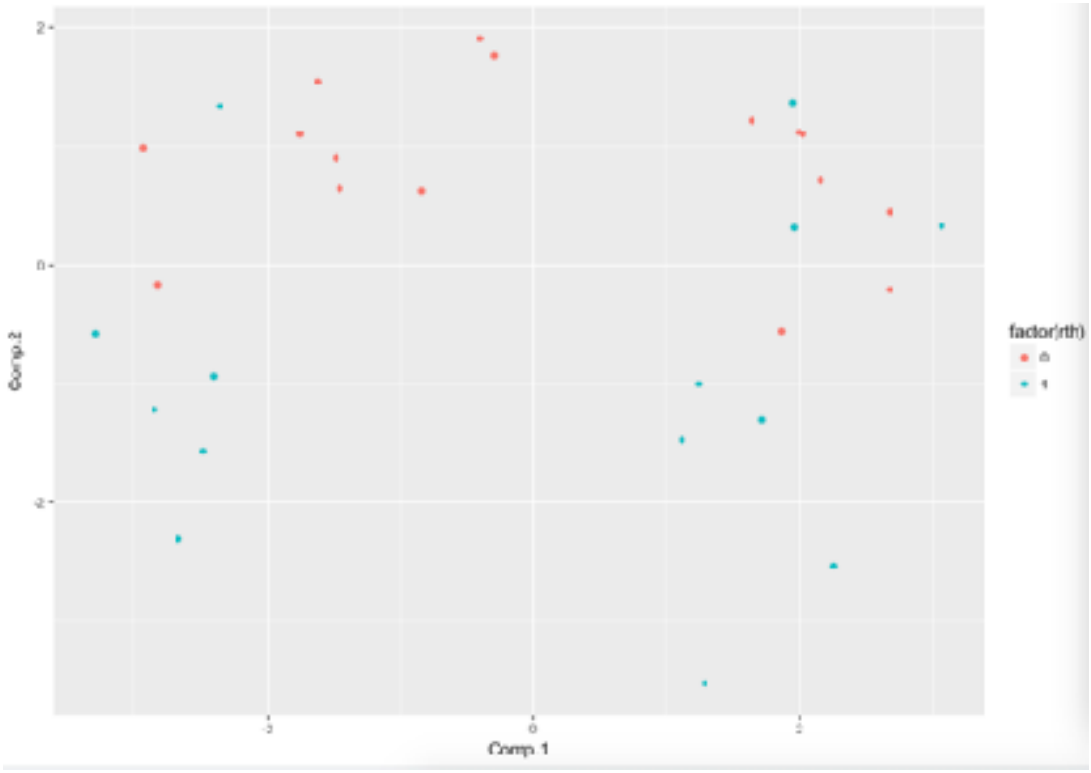
令RTH与分别与0.187，0.183盒0.161进行比较，大于等于赋值为0，视为不合格，反之为1.这样可以模拟质量标准提高时，查看哪些变量会影响RTH的质量。

将第一二个主成分作为X和Y轴进行可视化，在新的标准下是否合格作为颜色的标签。

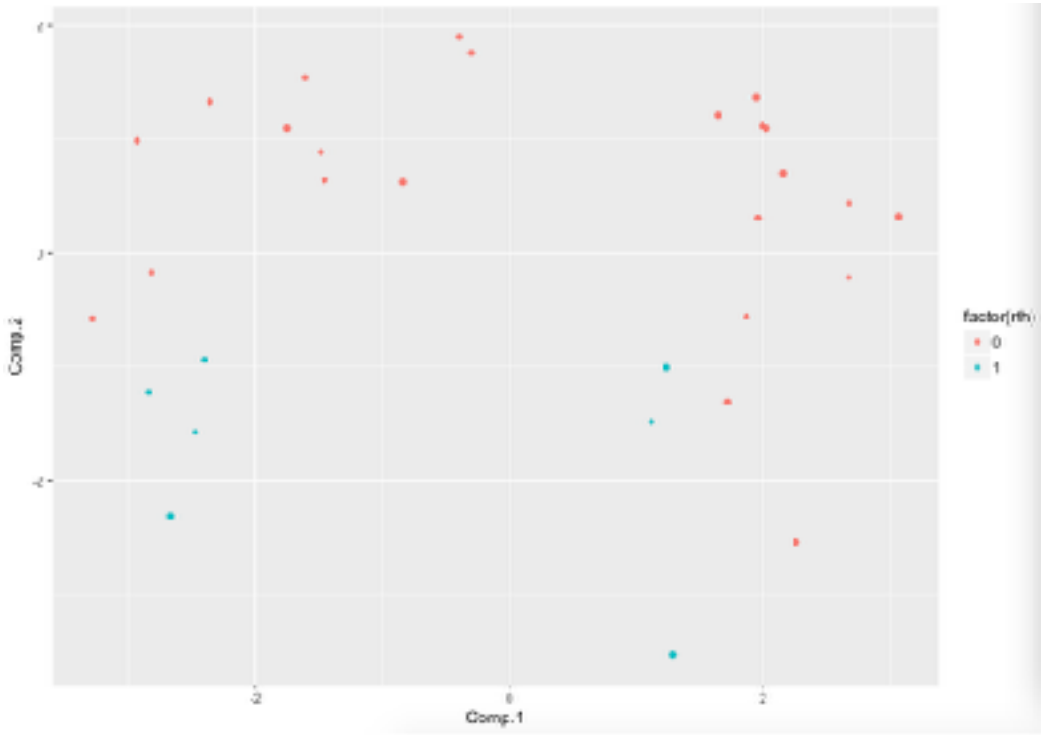
RTH小于0.187视为合格时



RTH小于0.183视为合格



RTH小于0.161视为合格



发现当RTH的合格标准提升时，第二个主成分需要显著变小。

4） 利用随即森林计算特征重要程度

利用随即森林训练一个模型，RTH所谓因变量，其他11个数据做为自变量。按照其特征重要程度排序。

X8	X11	X7	X6	X4	X10
0.0020730335	0.0010916732	0.0005859681	0.0004025036	0.0003295870	0.0003103235

X5	X7	X9	X3	X1
0.0002793270	0.0002088145	0.0001981868	0.0000991339	0.0000000000

总结:
经过数据分析有三个发现 1） 企业的质量管理缺乏统一的标准。2）生产存在不合理的波动，不同时期的生产标准发生明显的偏移。 3）影响RTH的变量