Clustering and t-test

T-test for change in Clinical Outcome with p values

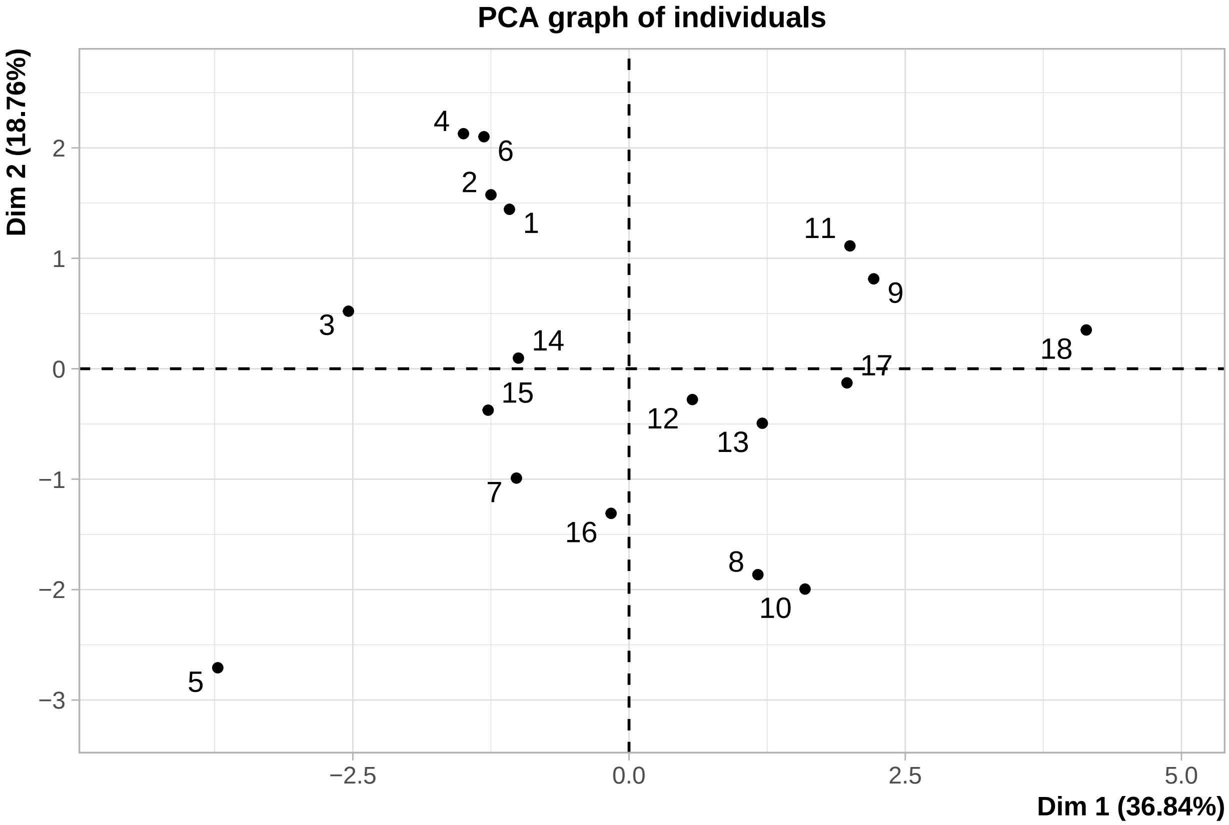
$$\% \ change \ in \ outcome = \frac{first \ read - second \ read}{first \ read}$$

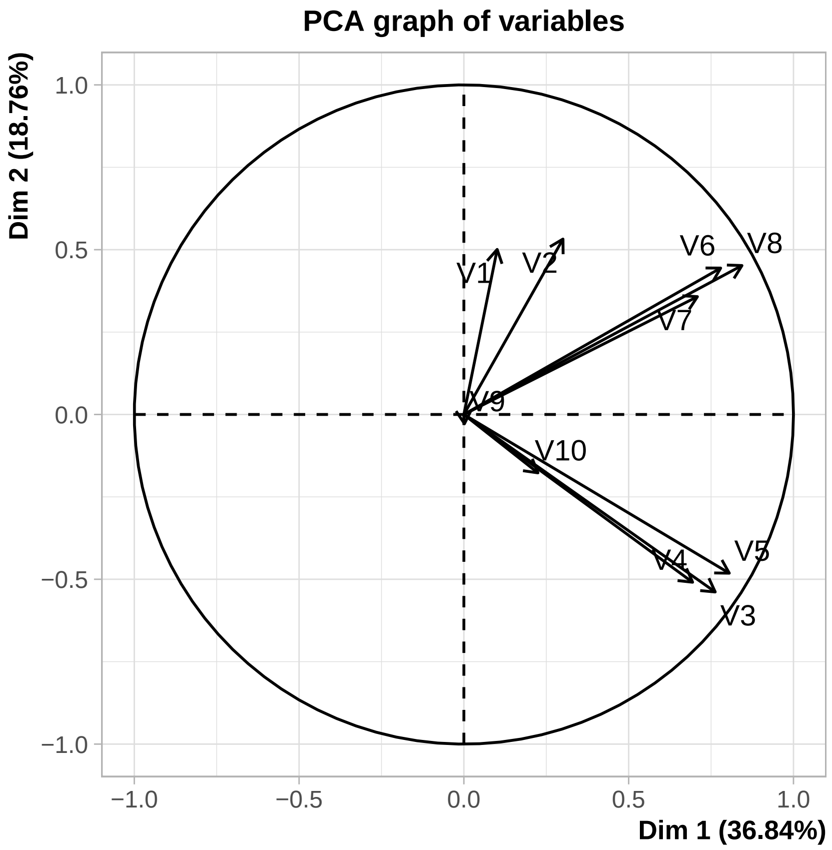
```
data <- read_excel("finaldata.xlsx", sheet = NULL, col_types = c(rep("guess", 6),
                                                                 "date", rep("guess", 2),
                                                                 rep("numeric", 10)))

status <- factor(c("low", rep("intermediate", 4),
                   "low", "intermediate", "intermediate",
                   rep("low", 2), rep("intermediate", 3),
                   "low", rep("intermediate", 2), rep("low", 2)))

status <- forcats::fct_relevel(status, c("low"))

#smoking <- as.factor(c("low", "high", "low", "high", "high", "low",
#                       rep("high", 3), rep("low", 3), "high", "NA", "high", "low"))
#smoking <- smoking[-c(8, 13, 18:20)]

smoking <-
patient <- 1:18

first_read <- data[seq(1, 54, 3), ]
second_read <- data[seq(2, 54, 3), ]

pval <- NULL
for(i in 10:19){
  change <- 1 - as.numeric(unlist(second_read[, i] / first_read[, i]))
  cat(names(second_read[, i]), "with p-value of", t.test(change ~ status)$p.value, "\n")
  pval <- c(pval, t.test(change ~ status)$p.value)
}
```

```
## Tumor Volume (mL) with p-value of 0.651912
## Largest Node Volume (mL) with p-value of 0.06978384
## SULmax Tumor with p-value of 0.2943616
## SULmedian with p-value of 0.3630501
## SULpeak with p-value of 0.3854198
## SULmax Largest Node with p-value of 0.0148899
## SULmedian node with p-value of 0.6197499
## SULpeak node with p-value of 0.03951802
## Diffusion Mean Tumor with p-value of 0.1825276
## Diffusion mean ADC Largest Node with p-value of 0.7309697
```

```
change_matrix <- array(NA, c(18, 10))

for(i in 1:10){
  change_matrix[, i] <- 1 - as.numeric(unlist(second_read[1:18, i + 9] / first_read[1:18, i + 9]))
}

change_matrix[is.na(change_matrix)] <- 0
```

```
pca <- PCA(change_matrix)
```



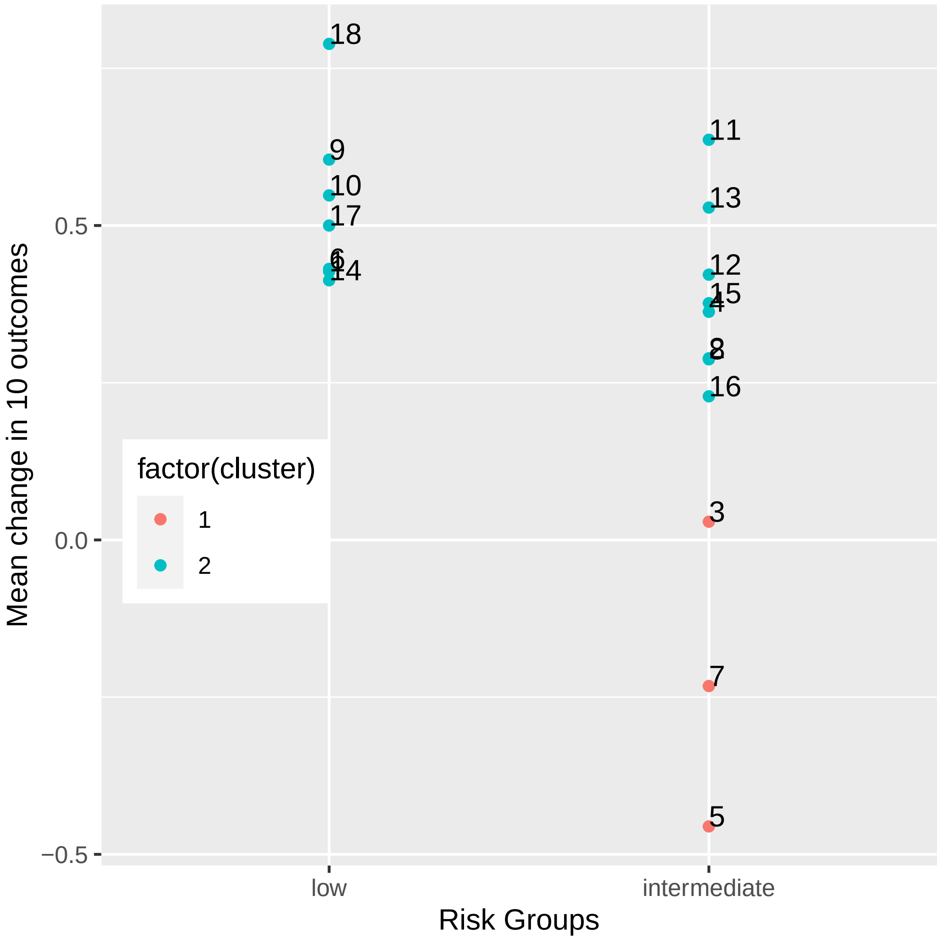


```
knitr::kable(cbind(names(data)[10:19], as.numeric(round(pca$var$coord[, 1], 3))))
```

Tumor Volume (mL)	0.101
Largest Node Volume (mL)	0.301
SULmax Tumor	0.763
SULmedian	0.694
SULpeak	0.805
SULmax Largest Node	0.779
SULmedian node	0.709
SULpeak node	0.844
Diffusion Mean Tumor	0.002
Diffusion mean ADC Largest Node	0.225

```
clust <- kmeans(change_matrix, 2, iter.max = 1000, nstart = 1000)
clust <- kmeans(change_matrix[, 6], 2, iter.max = 1000, nstart = 1000)

dat <- data.frame(risk = status, cluster = factor(clust$cluster), change = change_matrix[, 6],
                  patient = patient)
```

```
ggplot(dat, aes(risk, change, label = patient)) +
  geom_point(aes(colour = factor(cluster))) +
  geom_text(aes(label = patient), hjust = 0, vjust = 0) +
  labs(x = "Risk Groups", y = "Mean change in 10 outcomes") +
  theme(legend.position=c(0.15, 0.4),
        strip.background = element_blank())
```



```
# Patient with in cluster 1
dat$patient[dat$cluster == 1]
```

```
## [1] 3 5 7
```

```
# Patient in cluster 2
dat$patient[dat$cluster == 2]
```

```
##  [1]  1  2  4  6  8  9 10 11 12 13 14 15 16 17 18
```

```
pdf("clusterplot.pdf")
ggplot(dat, aes(risk, change, label = patient)) +
  geom_point(aes(colour = cluster)) +
  geom_text(aes(label = patient),hjust=0, vjust=0) +
  labs(x = "Risk Groups", y = "Mean change in 10 outcomes") +
  theme(legend.position=c(0.15, 0.4),
        strip.background = element_blank())
dev.off()
```

```
## pdf
##   2
```

```
knitr::kable(data.frame(Patient = dat$patient, Cluster = dat$cluster, value = round(change_matrix[, 6],
```

Patient	Cluster	value
1	2	0.427
2	2	0.289
3	1	0.029
4	2	0.363
5	1	-0.456
6	2	0.431
7	1	-0.232
8	2	0.287
9	2	0.605
10	2	0.548
11	2	0.636
12	2	0.422
13	2	0.529
14	2	0.413
15	2	0.376
16	2	0.228
17	2	0.500
18	2	0.789

Patient Characteristics

```r
patient_data <- read_excel("patient.xlsx")
patient_data <- cbind(patient_data, change_matrix)

names(patient_data)[10:19] <- names(data)[9:18]
names(patient_data)[5] <- "Smoking"
patient_data <- data.frame(cbind(patient_data, cluster = dat$cluster))

## Tumor volume and SULpeak change and smoking are related
summary(lm(Tumor.Volume..mL. ~ Smoking, data = patient_data))
```

```
##
## Call:
## lm(formula = Tumor.Volume..mL. ~ Smoking, data = patient_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73028 -0.04801  0.07703  0.16304  0.35462
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.312043   0.080201   3.891   0.0013 **
## Smoking     -0.007487   0.003088  -2.425   0.0275 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2775 on 16 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.223
## F-statistic:  5.88 on 1 and 16 DF,  p-value: 0.02752
```

```r
summary(lm(SULpeak ~ Smoking, data = patient_data))
```

```
##
## Call:
## lm(formula = SULpeak ~ Smoking, data = patient_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79289 -0.06923  0.02890  0.14945  0.34876
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.440061   0.080297   5.480 5.03e-05 ***
## Smoking     -0.006420   0.003091  -2.077   0.0543 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2779 on 16 degrees of freedom
## Multiple R-squared:  0.2123, Adjusted R-squared:  0.1631
## F-statistic: 4.313 on 1 and 16 DF,  p-value: 0.05429
```

```
## Largest node volume change, SULmax tumor, and SULmedian and gender are related
summary(lm(Largest.Node.Volume..mL. ~ Sex, data = patient_data))
```

```
##
## Call:
## lm(formula = Largest.Node.Volume..mL. ~ Sex, data = patient_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35757 -0.12107  0.01071  0.10741  0.36614
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5673     0.1403   4.043 0.000942 ***
## SexM         -0.3255     0.1488  -2.187 0.043919 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1984 on 16 degrees of freedom
## Multiple R-squared:  0.2302, Adjusted R-squared:  0.1821
## F-statistic: 4.784 on 1 and 16 DF,  p-value: 0.04392
```

```
summary(lm(SULmax.Tumor ~ Sex, data = patient_data))
```

```
##
## Call:
## lm(formula = SULmax.Tumor ~ Sex, data = patient_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52226 -0.17244 -0.01515  0.18506  0.37369
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6055     0.1713   3.535  0.00275 **
## SexM         -0.3742     0.1817  -2.060  0.05609 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2422 on 16 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1602
## F-statistic: 4.242 on 1 and 16 DF,  p-value: 0.05609
```

```
summary(lm(SULmedian ~ Sex, data = patient_data))
```

```
##
## Call:
## lm(formula = SULmedian ~ Sex, data = patient_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34530 -0.06732  0.01716  0.10974  0.37577
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5767     0.1395   4.134  0.00078 ***
## SexM         -0.2932     0.1480  -1.981  0.06505 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1973 on 16 degrees of freedom
## Multiple R-squared:  0.197,  Adjusted R-squared:  0.1468
## F-statistic: 3.924 on 1 and 16 DF,  p-value: 0.06505
```

```
tb1 <- patient_data %>%
  dplyr::select(Age, cluster, Smoking, Sex, Tumor.site, Risk.grouping) %>%
  tbl_summary(statistic = list(all_continuous() ~ "{mean} ({sd})",
                   all_categorical() ~ "{n} ({p}%)"))%>%
  bold_labels()
tb1
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 18**
Age	61 (7)
cluster	
1	3 (17%)
2	15 (83%)
Smoking	15 (22)
Sex	
F	2 (11%)
M	16 (89%)
Tumor.site	
BOT	9 (50%)
Tonsil	9 (50%)
Risk.grouping	
Intermediate	11 (61%)
Low	7 (39%)

```
tb2 <- patient_data %>%
    dplyr::select(Age, cluster, Smoking, Sex, Tumor.site, Risk.grouping) %>%
  tbl_summary(by = cluster,
    statistic = list(all_continuous() ~ "{mean} ({sd})",
                     all_categorical() ~ "{n} ({p}%)")) %>%
    add_p(test = list(all_categorical() ~ "fisher.test",
                    all_continuous() ~ "aov")) %>%
  bold_p() %>%
  bold_labels()
tb2
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**1**, N = 3	**2**, N = 15	**p-value**
Age	68 (5)	59 (7)	0.057
Smoking	30 (40)	12 (17)	0.2
Sex			>0.9
F	0 (0%)	2 (13%)	
M	3 (100%)	13 (87%)	
Tumor.site			>0.9
BOT	2 (67%)	7 (47%)	
Tonsil	1 (33%)	8 (53%)	
Risk.grouping			0.2
Intermediate	3 (100%)	8 (53%)	
Low	0 (0%)	7 (47%)	

```
tbl_merge(list(tb1, tb2), tab_spanner = c(NA_character_, "**Cluster**"))
```

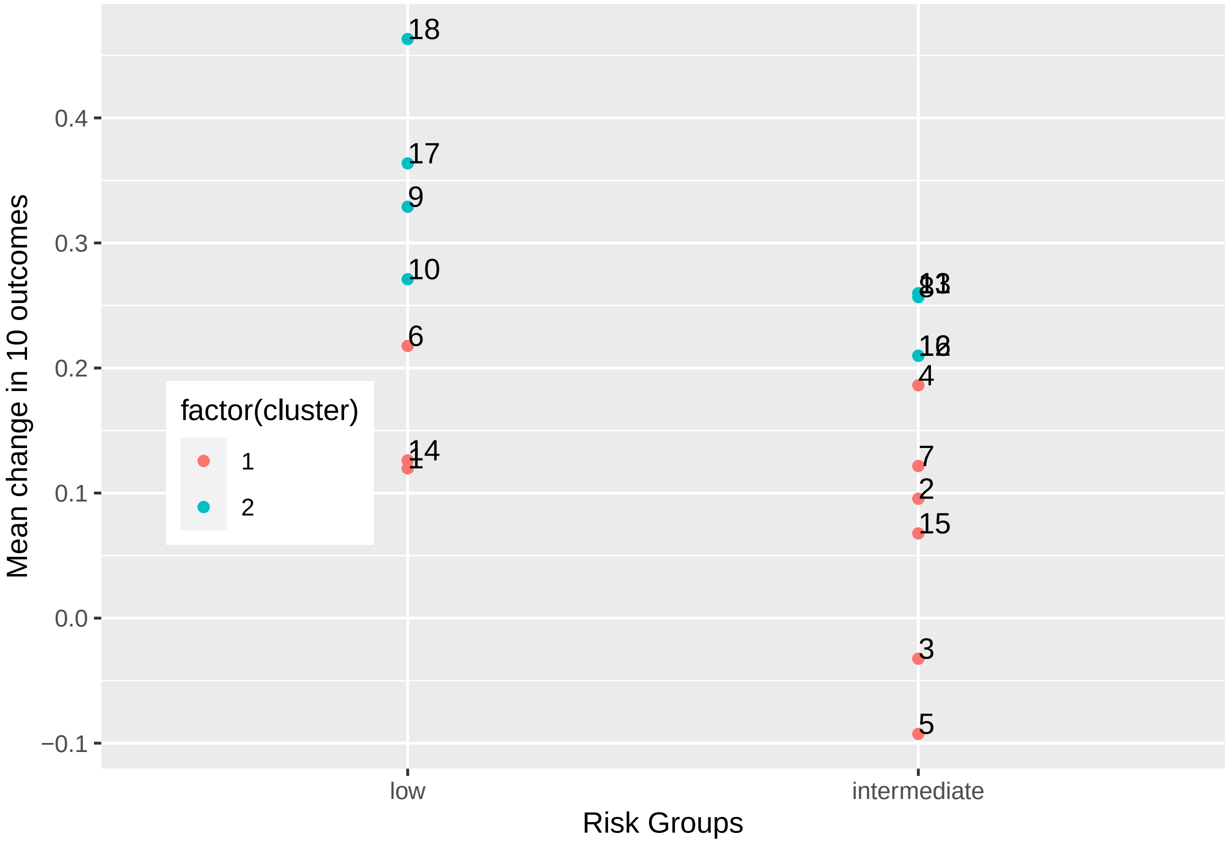
```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 18**	**1**, N = 3	**2**, N = 15	**p-value**
Age	61 (7)	68 (5)	59 (7)	0.057
cluster				
1	3 (17%)			
2	15 (83%)			
Smoking	15 (22)	30 (40)	12 (17)	0.2
Sex				>0.9
F	2 (11%)	0 (0%)	2 (13%)	
M	16 (89%)	3 (100%)	13 (87%)	
Tumor.site				>0.9
BOT	9 (50%)	2 (67%)	7 (47%)	
Tonsil	9 (50%)	1 (33%)	8 (53%)	
Risk.grouping				0.2
Intermediate	11 (61%)	3 (100%)	8 (53%)	
Low	7 (39%)	0 (0%)	7 (47%)	

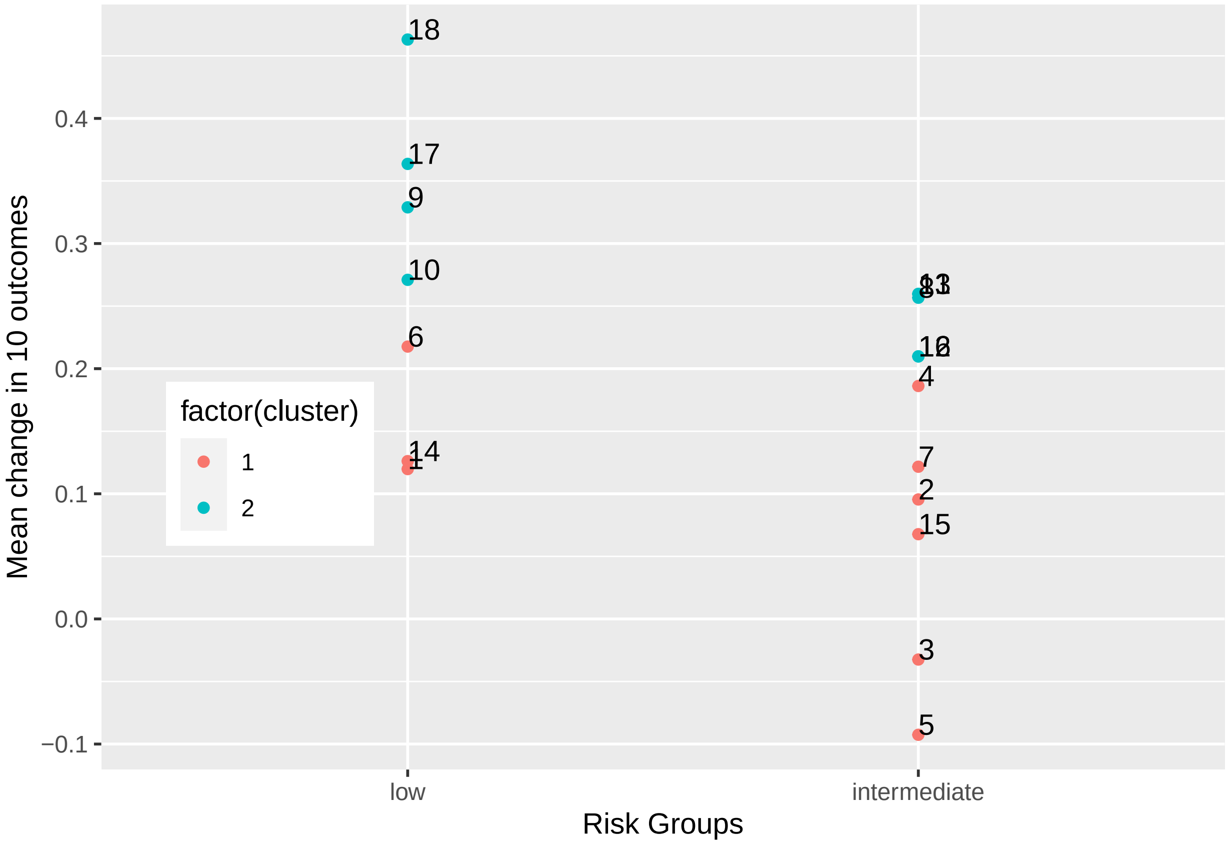
Cross validation

```
grouping  <- array(NA, c(18, 10))

meanchange <- apply(change_matrix, 1, function(x){mean(x, na.rm = T)})

for(i in 1:10){
  clust <- kmeans(change_matrix[, -i], 2, iter.max = 1000, nstart = 1000)
  dat <- data.frame(risk = status, cluster = factor(clust$cluster),
                    change = meanchange, patient = patient)
  cat("Removing ", names(second_read[i + 9]), "\n")
  print(ggplot(dat, aes(risk, change, label = patient)) +
  geom_point(aes(colour = factor(cluster))) +
  geom_text(aes(label = patient), hjust = 0, vjust = 0) +
  labs(x = "Risk Groups", y = "Mean change in 10 outcomes") +
  theme(legend.position=c(0.15, 0.4),
        strip.background = element_blank()))

  grouping[, i] <- clust$cluster
}
```

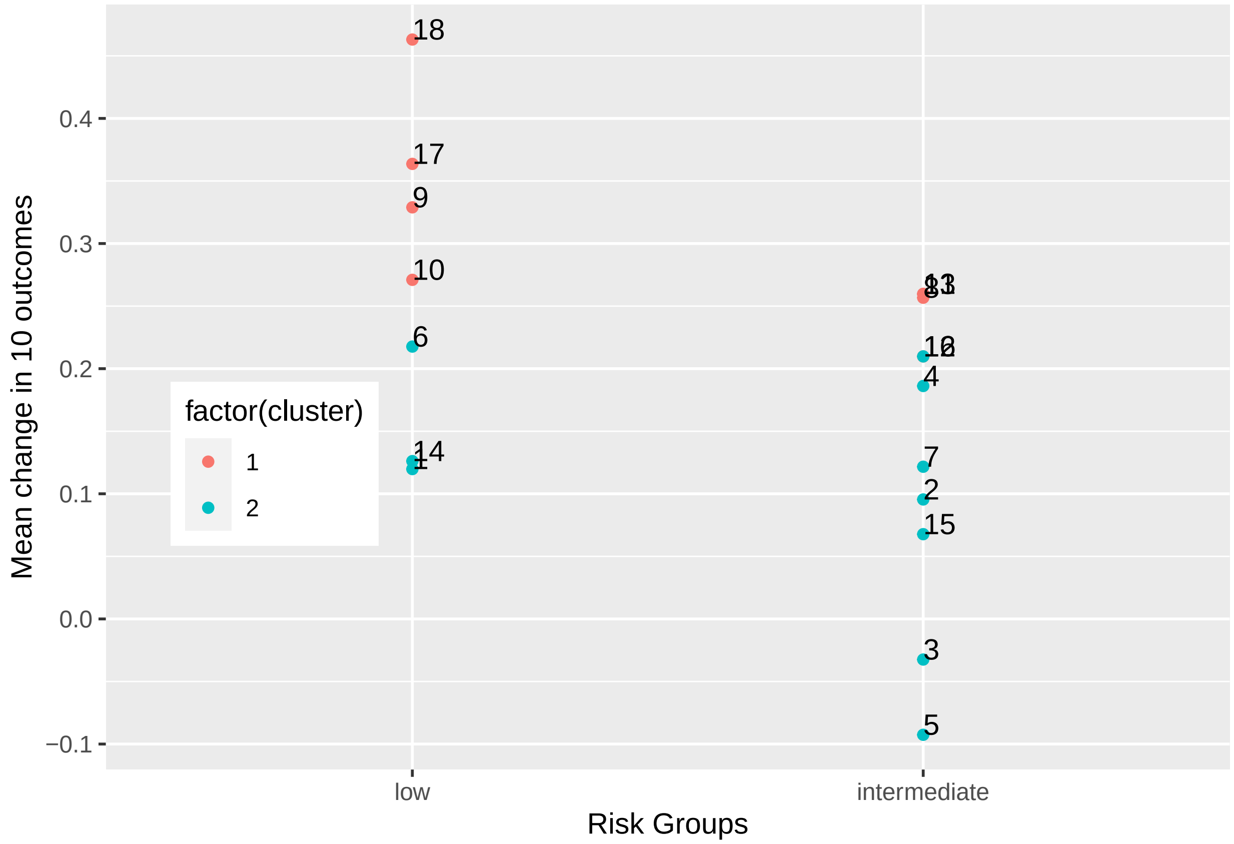
```
## Removing  Tumor Volume (mL)
```



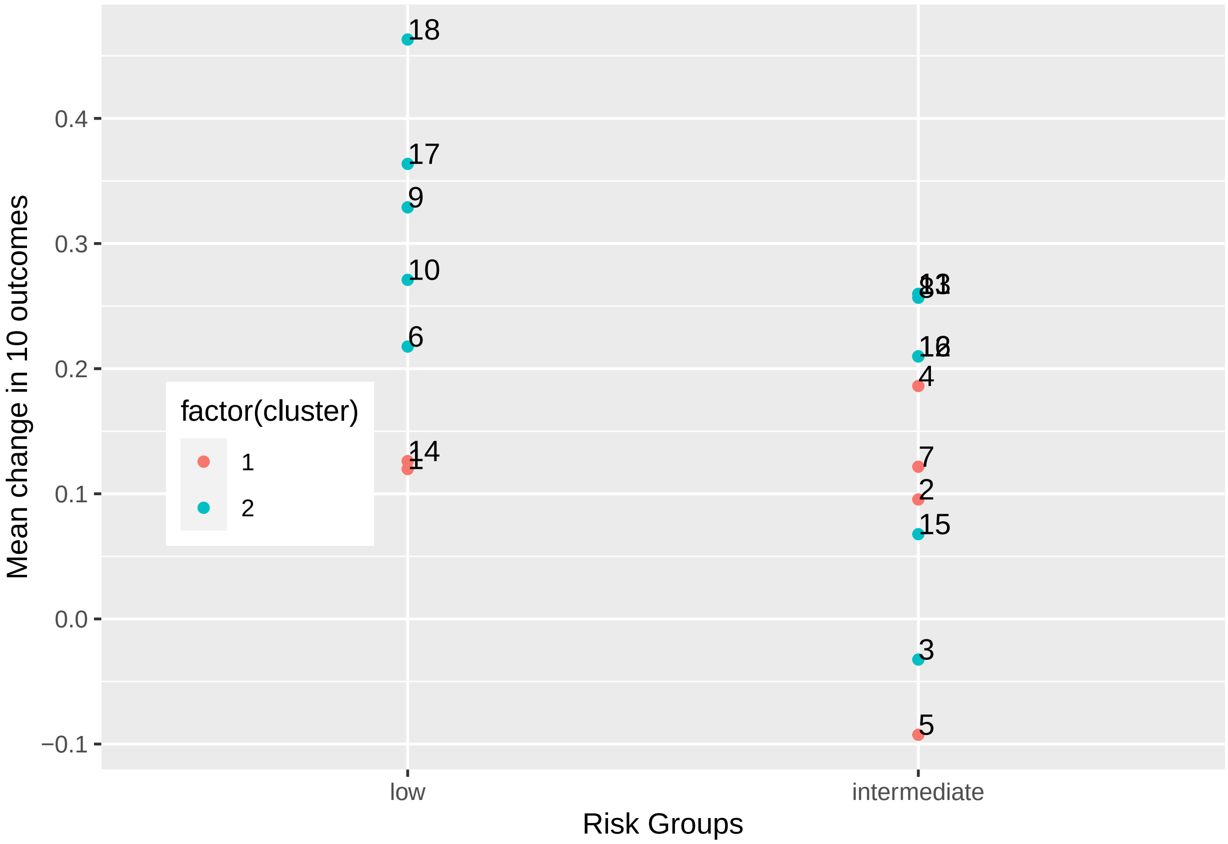
```
## Removing  Largest Node Volume (mL)
```



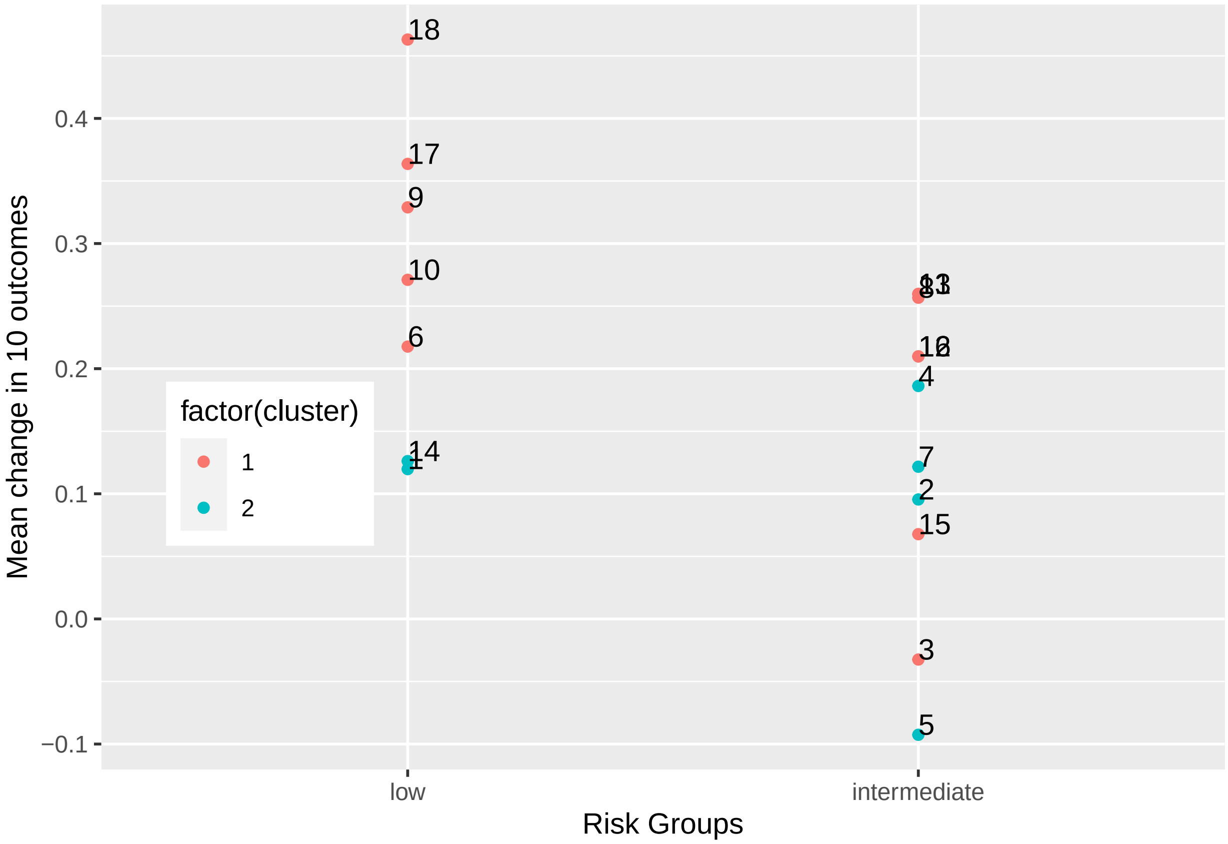
```
## Removing  SULmax Tumor
```



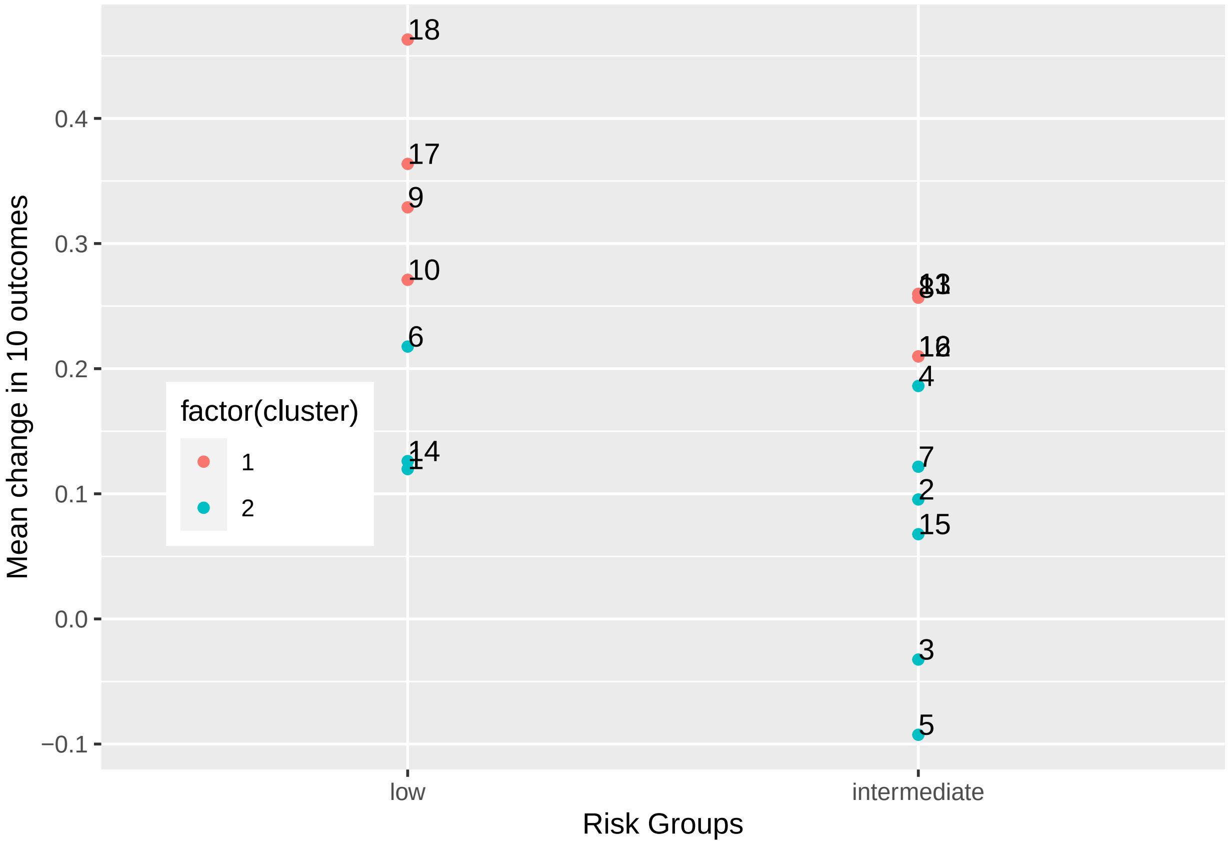
```
## Removing  SULmedian
```



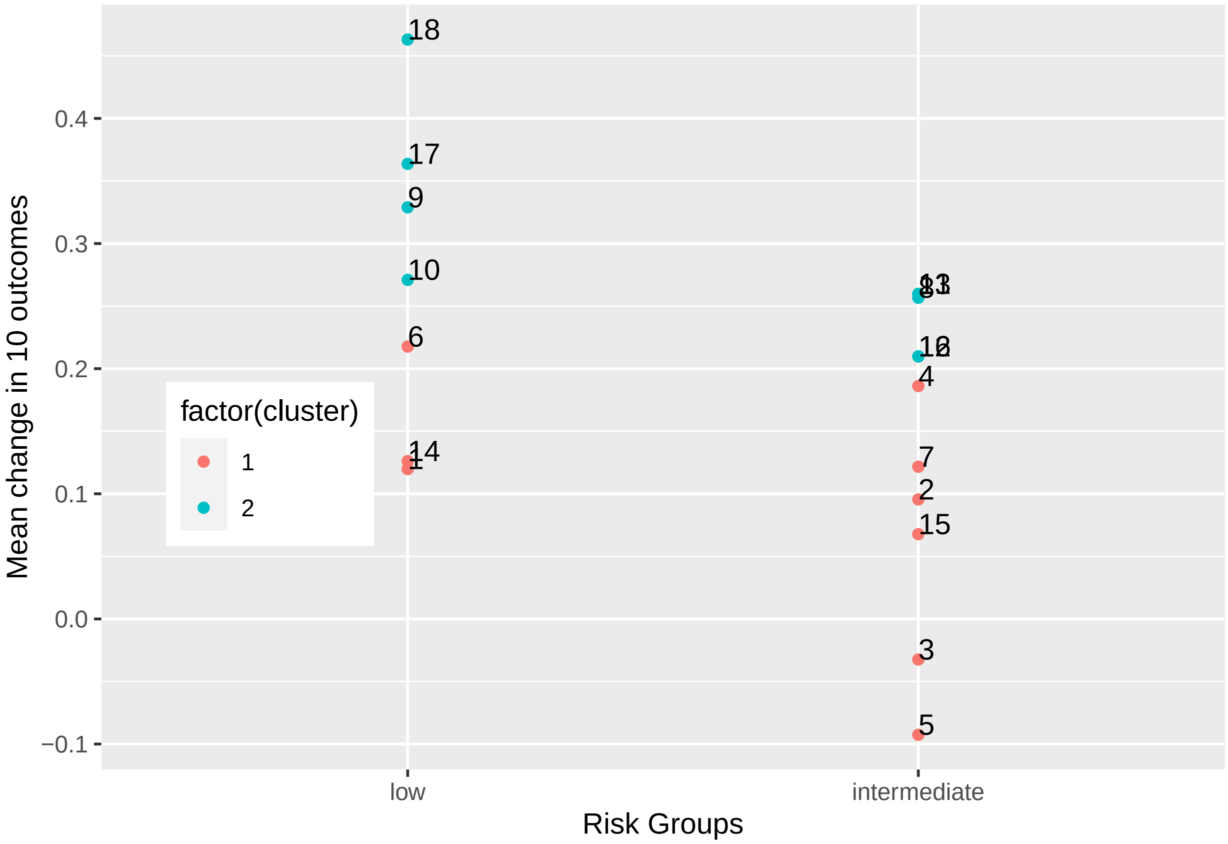
```
## Removing  SULpeak
```



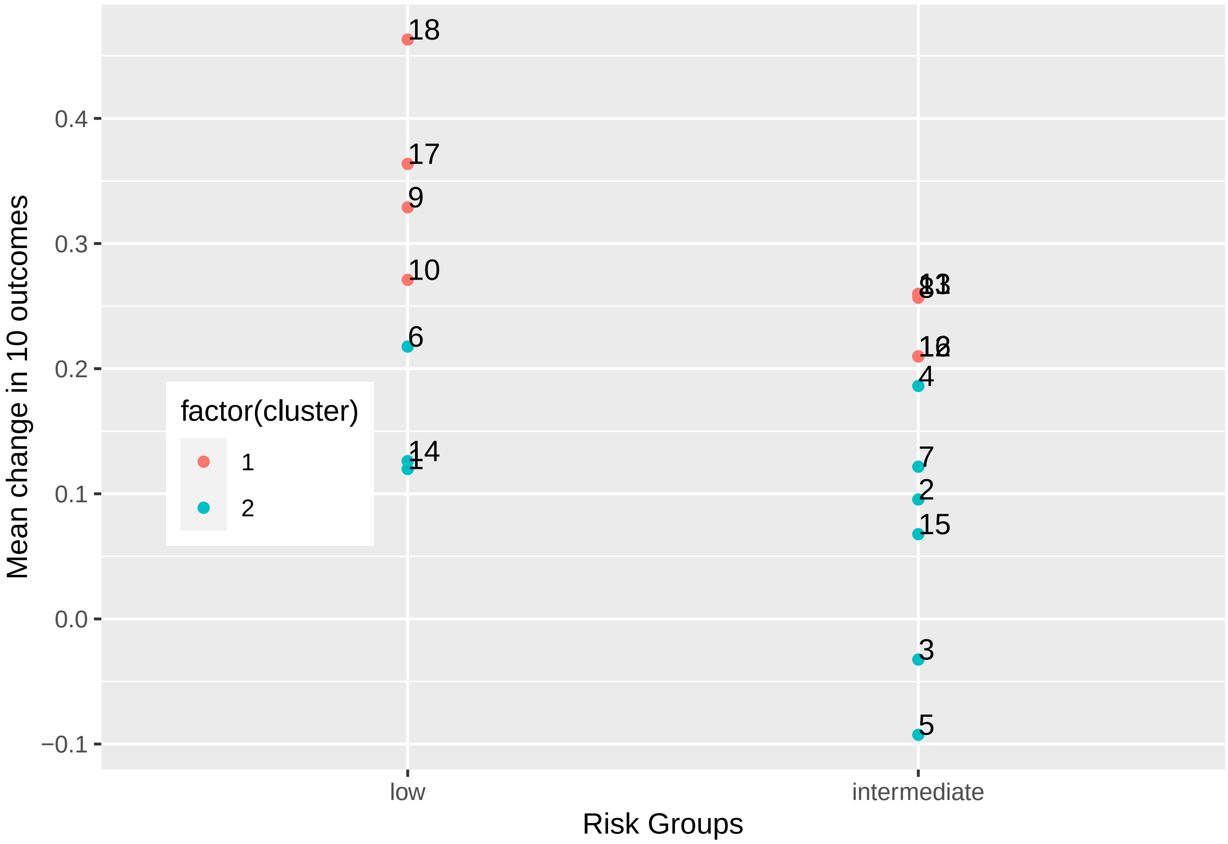
```
## Removing  SULmax Largest Node
```



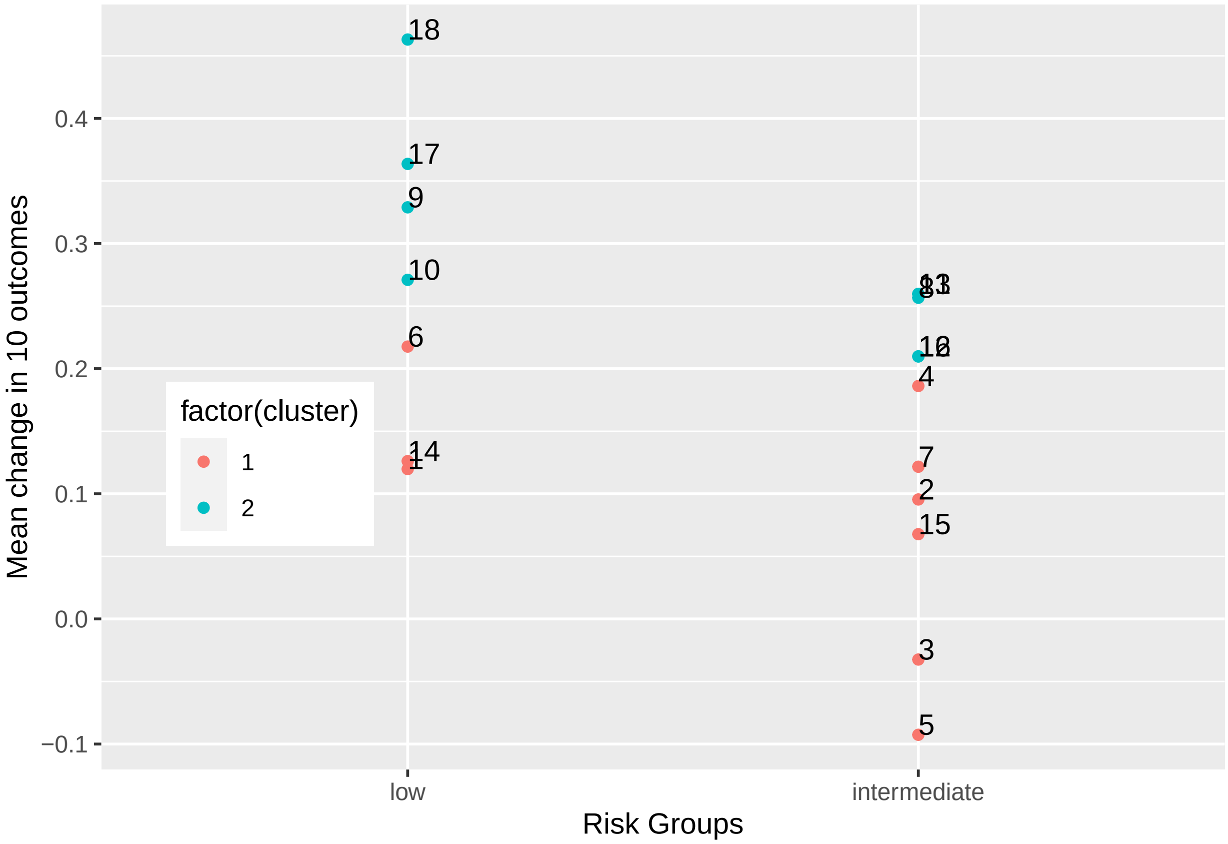
```
## Removing  SULmedian node
```



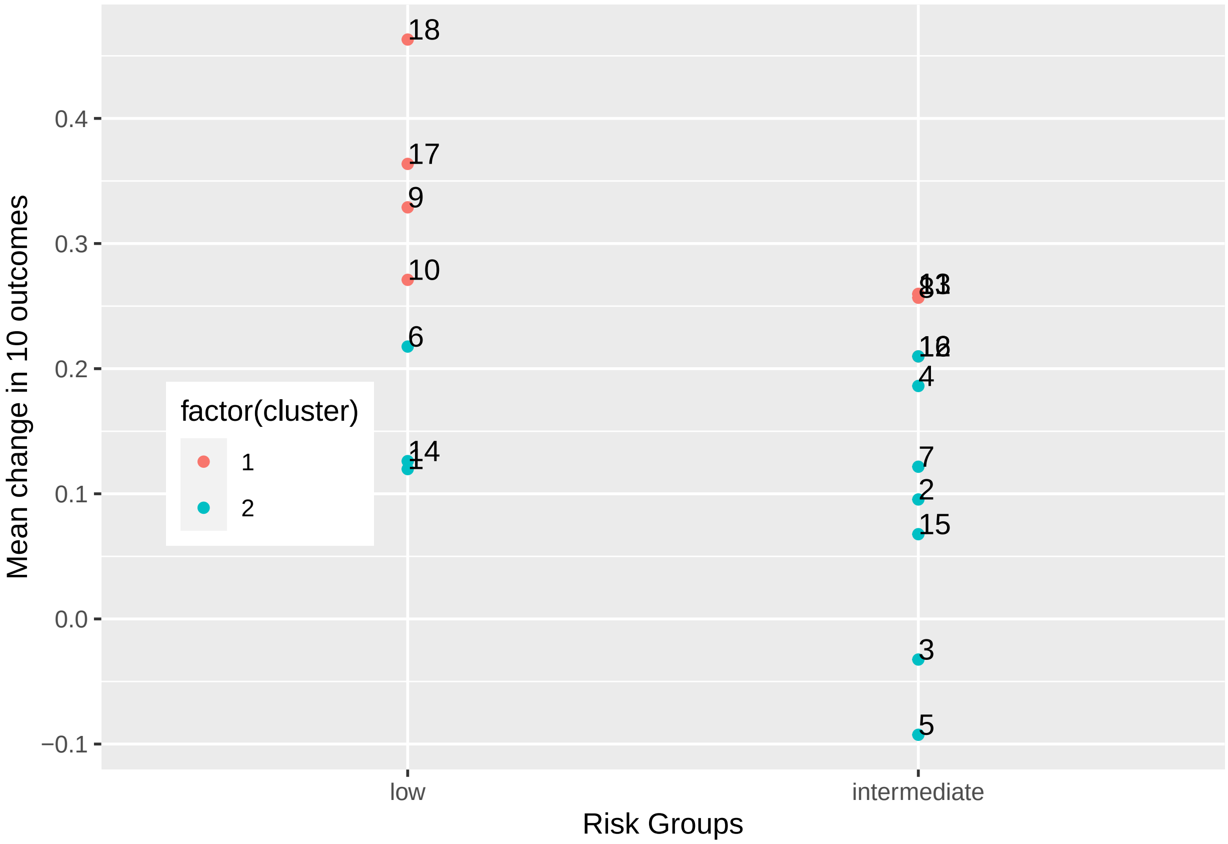
```
## Removing  SULpeak node
```



```
## Removing  Diffusion Mean Tumor
```



```
## Removing  Diffusion mean ADC Largest Node
```



Clustering table after removing each column

```r
colnames(grouping) <- names(second_read[10:19])

knitr::kable(grouping)
```

Tumor Volume (mL)	Largest Node Volume (mL)	SULmax Tumor	SULmedian	SULpeak	SULmax Largest Node	SULmedian node	SULpeak node	Diffusion Mean Tumor	Diffusion mean ADC Largest Node
1	1	2	1	2	2	1	2	1	2
1	1	2	1	2	2	1	2	1	2
1	1	2	2	1	2	1	2	1	2
1	1	2	1	2	2	1	2	1	2
1	1	2	1	2	2	1	2	1	2
1	1	2	2	1	2	1	2	1	2
1	1	2	1	2	2	1	2	1	2
2	2	1	2	1	1	2	1	2	1
2	2	1	2	1	1	2	1	2	1
2	2	1	2	1	1	2	1	2	1
2	2	1	2	1	1	2	1	2	1
2	2	1	2	1	1	2	1	2	1
2	2	1	2	1	1	2	1	2	1
1	1	2	1	2	2	1	2	1	2
1	1	2	2	1	2	1	2	1	2
2	2	2	2	1	1	2	1	2	2
2	2	1	2	1	1	2	1	2	1

Tumor Volume (mL)	Largest Node Volume (mL)	SULmax Tumor	SULmedian	SULpeak	SULmax Largest Node	SULmedian node	SULpeak node	Diffusion Mean Tumor	Diffusion mean ADC Largest Node
2	2	1	2	1	1	2	1	2	1