Clustering Analysis Project 2

Laura Hu

2023-04-29

Load library

```
library(ggplot2)
library(ggfortify)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(dslabs)
library(caret)
```

```
## Loading required package: lattice
```

```
library(modelr)
```

```
##
## Attaching package: 'modelr'

## The following object is masked from 'package:dslabs':
##
##     heights
```

```
library(yardstick)
```

```
## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.

##
## Attaching package: 'yardstick'

## The following objects are masked from 'package:modelr':
##
##     mae, mape, rmse

## The following objects are masked from 'package:caret':
##
##     precision, recall, sensitivity, specificity
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

Load Data

```
Bully <- read.csv("~/Documents/School /bullying_2018_cleaned.csv")
```

Initial feature edits to clean up the data

```
#Delete the first row of the data frame

Bully <-Bully %>% select(-X) %>% select(-record)
Bully <- Bully[-1,]

#Delete 'Physical_fighting_12_or_more' and 'Physically_attacked_12_or_more' because all variables are f

Bully <-Bully %>% select(-Physical_fighting_12_or_more) %>% select(-Physically_attacked_12_or_more)

#Bully %>% group_by(Miss_school_no_permission)%>%tally()
```

Convert Categorical to Numeric

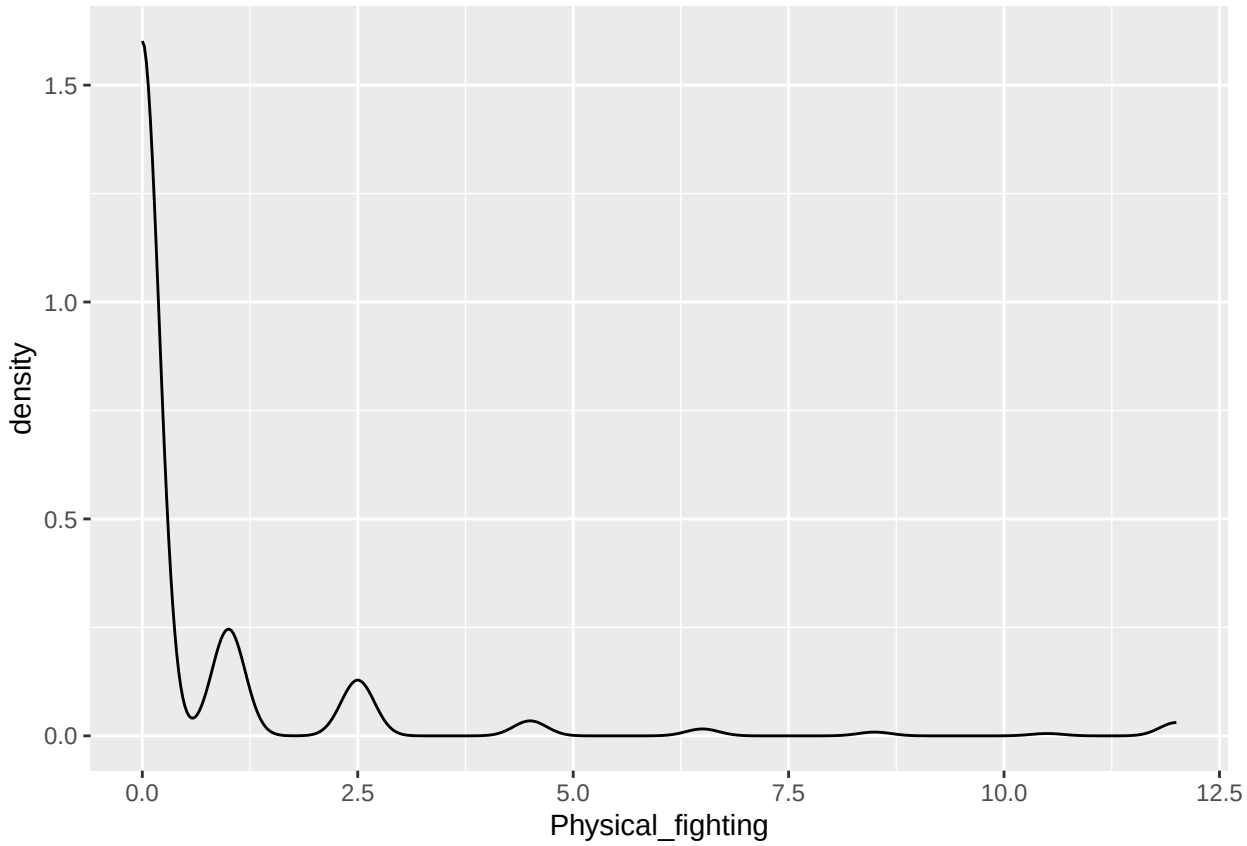
```
# Change categorical data to numeric


#Bullying
```

```
Bully$Bullied.in.School <- c(Yes = 1, No = 0)[Bully$Bullied.in.School]
Bully$Bullied.outside.school <- c(Yes = 1, No = 0)[Bully$Bullied.outside.school]
Bully$Cyber.Bullied <- c(Yes = 1, No = 0)[Bully$Cyber.Bullied]

#Sex
Bully$Sex <- c(Male = 1, Female = 0)[Bully$Sex]

#Felt lonely
Bully$Most_of_the_time_or_always_felt_lonely <- c(Yes = 1, No = 0)[Bully$Most_of_the_time_or_always_fel

#Missed school or class without permission
Bully$Missed_classes_or_school_without_permission <- c(Yes = 1, No = 0)[Bully$Missed_classes_or_school_

#Close friends
Bully$Close_friends_3_or_more <- c(True = 1, False = 0)[Bully$Close_friends_3_or_more]

#Miss school no permission
Bully$Miss_school_no_permission <- c('0 days' = 0, '1 or 2 days' = 1,
'3 to 5 days' = 2, '6 to 9 days' = 3, '10 or more days' = 4 )[Bully$Miss_school_no_permission]

#Scale data
Bully_scaled <- Bully%>% mutate(across(where(is.numeric), scale))
```

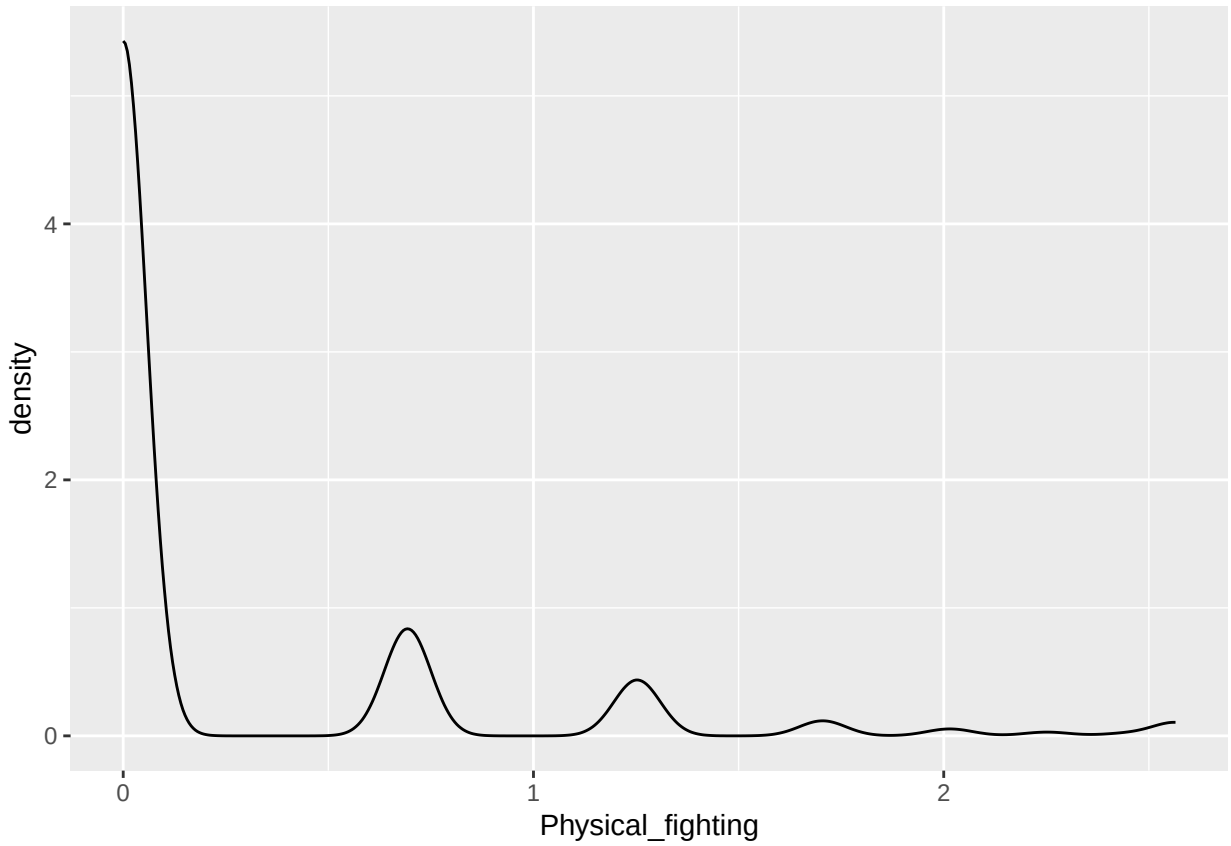
Skewness

We ran a series of plots to find out that **'Physically_attacked**', '**Physical_fighting**' are heavily skewed to the right. We need to do some transformations on them to wider the range.

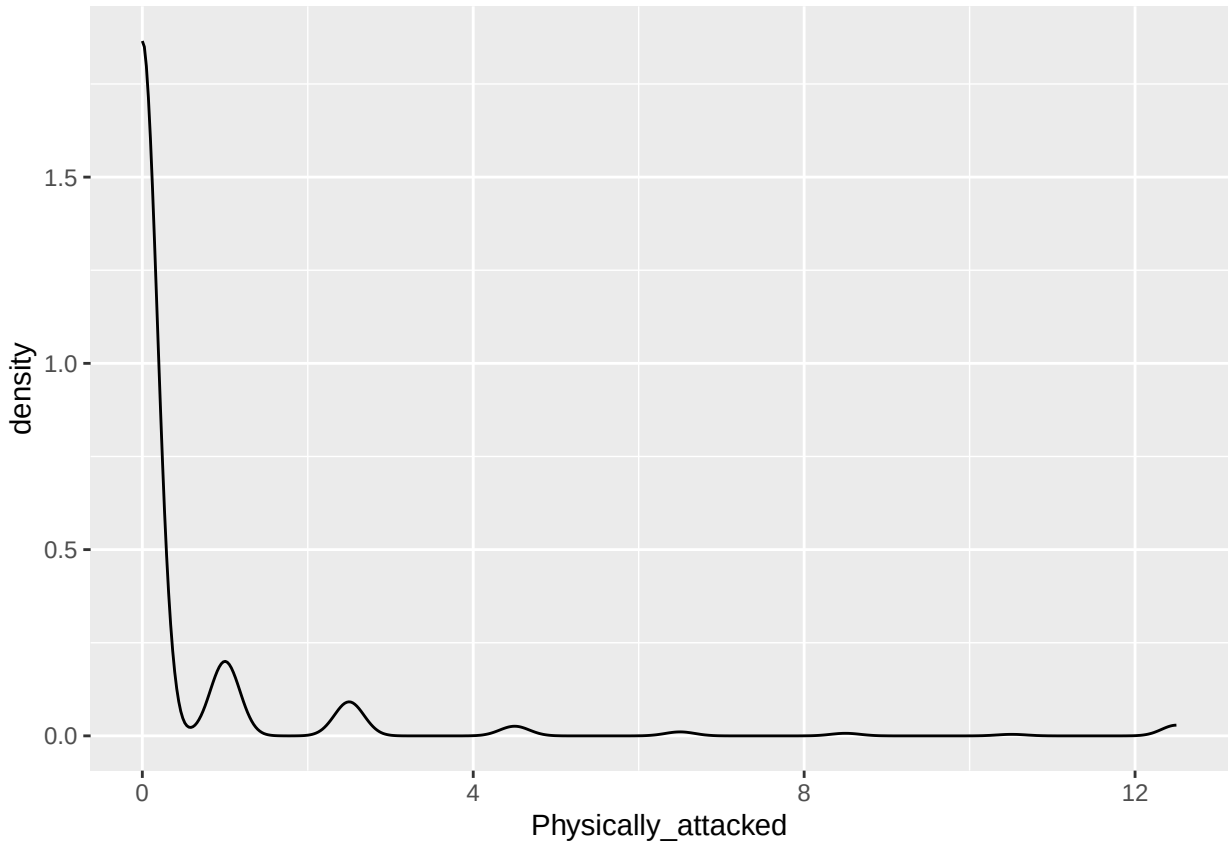
```
# Skew Physcial fighting
ggplot(Bully, aes(x = Physical_fighting)) + geom_density()
```



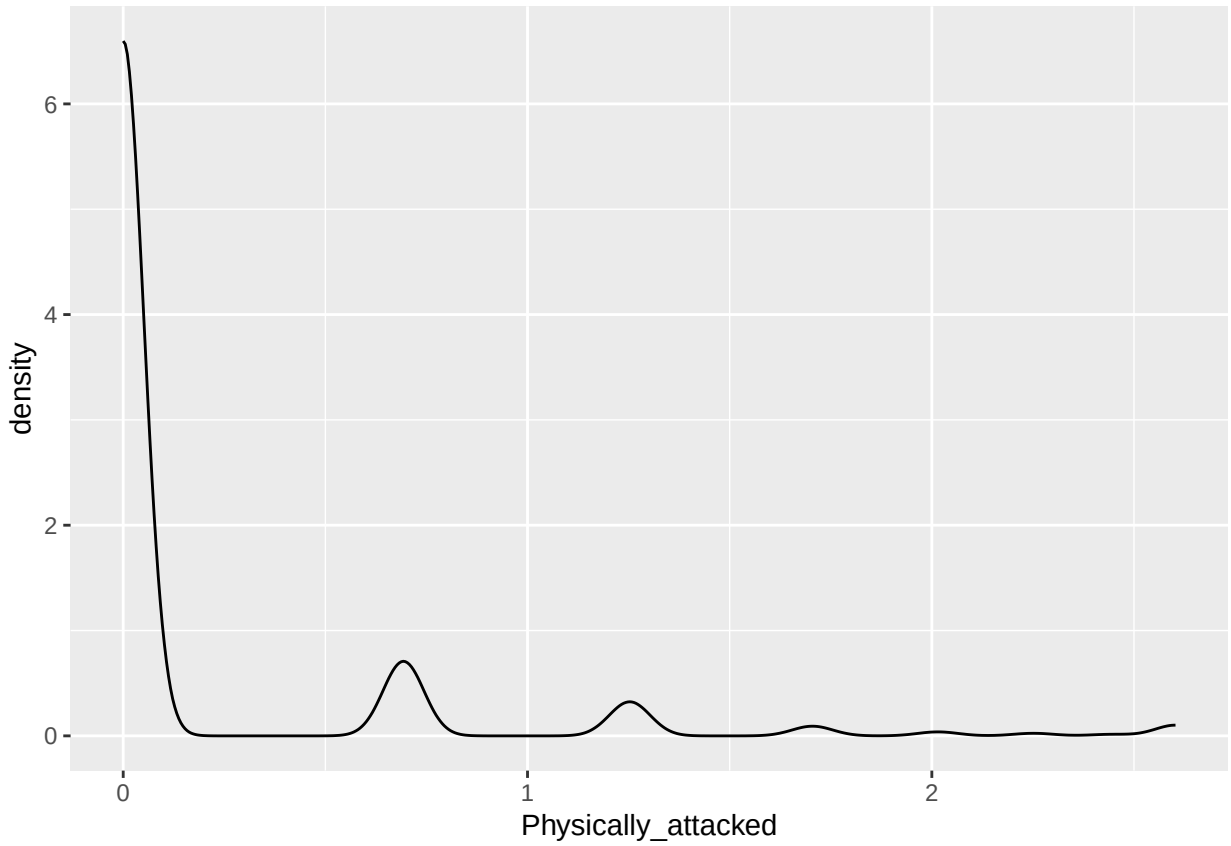

```
Bully$Physical_fighting <- log(Bully$Physical_fighting + 1)
ggplot(Bully, aes(x = Physical_fighting)) + geom_density()
```




```
#Skew Physically attacked
ggplot(Bully, aes(x = Physically_attacked)) + geom_density()
```




```
Bully$Physically_attacked <- log(Bully$Physically_attacked+1)
ggplot(Bully, aes(x = Physically_attacked )) + geom_density()
```



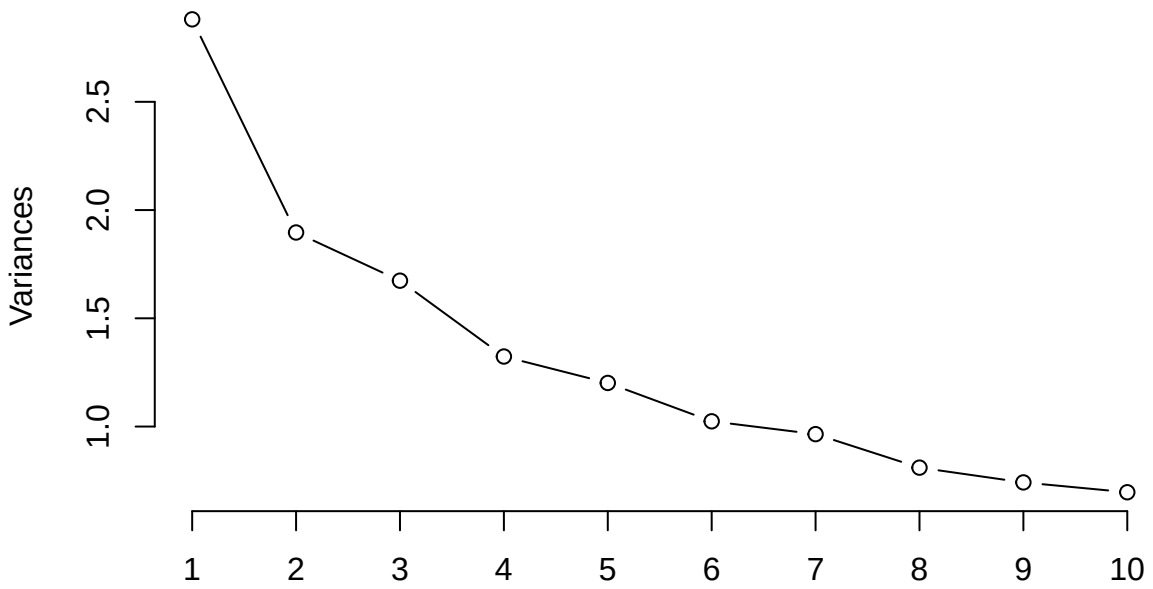
Principle Component Analysis

For the first method, we want to try the principle component analysis to see how the variables correlates.

```
pca_Bully <- Bully_scaled %>% select(where(is.numeric)) %>% prcomp(scale = TRUE)
summary(pca_Bully)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.6973 1.3771 1.2938 1.15043 1.09609 1.01201 0.98225
## Proportion of Variance 0.1921 0.1264 0.1116 0.08823 0.08009 0.06828 0.06432
## Cumulative Proportion  0.1921 0.3185 0.4301 0.51832 0.59841 0.66669 0.73101
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.90009 0.86134 0.83456 0.80607 0.77344 0.48637 0.41700
## Proportion of Variance 0.05401 0.04946 0.04643 0.04332 0.03988 0.01577 0.01159
## Cumulative Proportion  0.78502 0.83448 0.88091 0.92423 0.96411 0.97988 0.99147
##                           PC15
## Standard deviation     0.35760
## Proportion of Variance 0.00853
## Cumulative Proportion  1.00000
```

```
screeplot(pca_Bully, type = "lines")
```




```
library(broom)
```

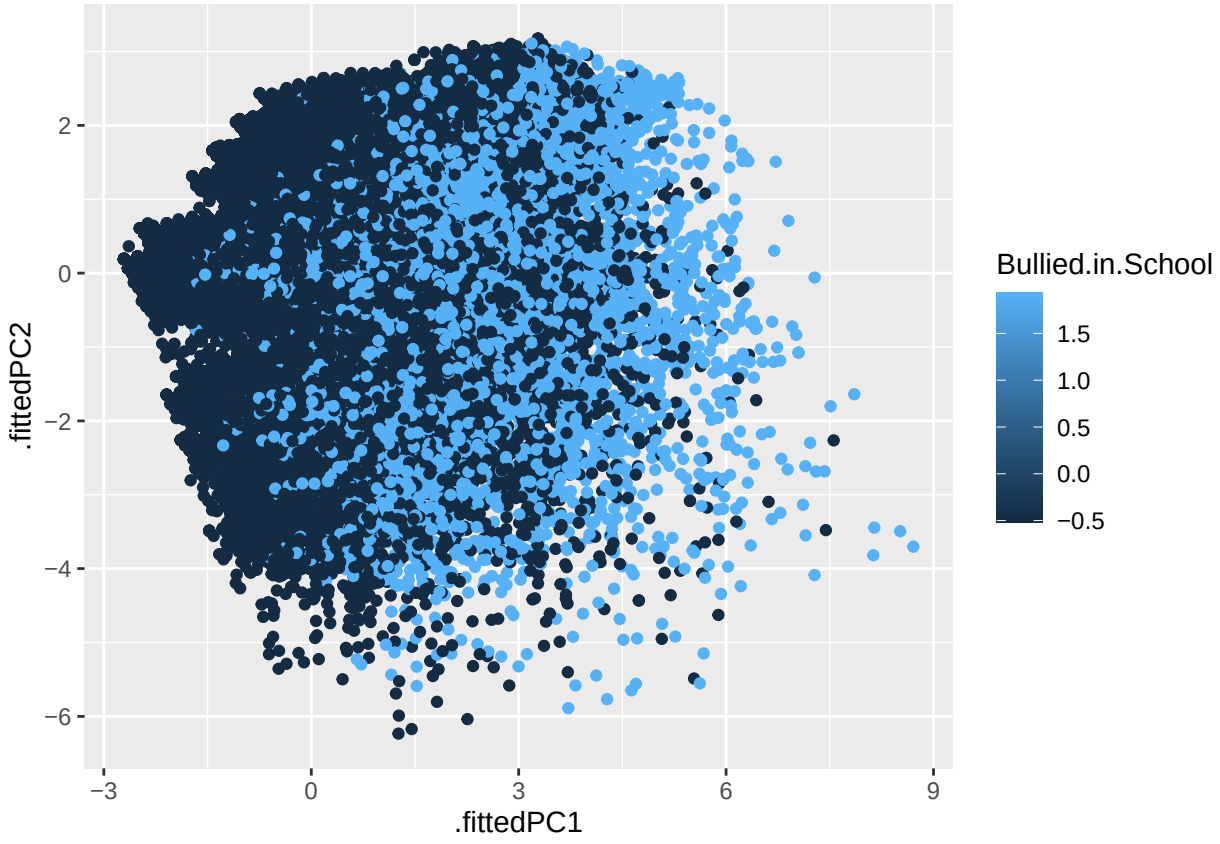
```
##
## Attaching package: 'broom'
```

```
## The following object is masked from 'package:modelr':
##
##     bootstrap
```

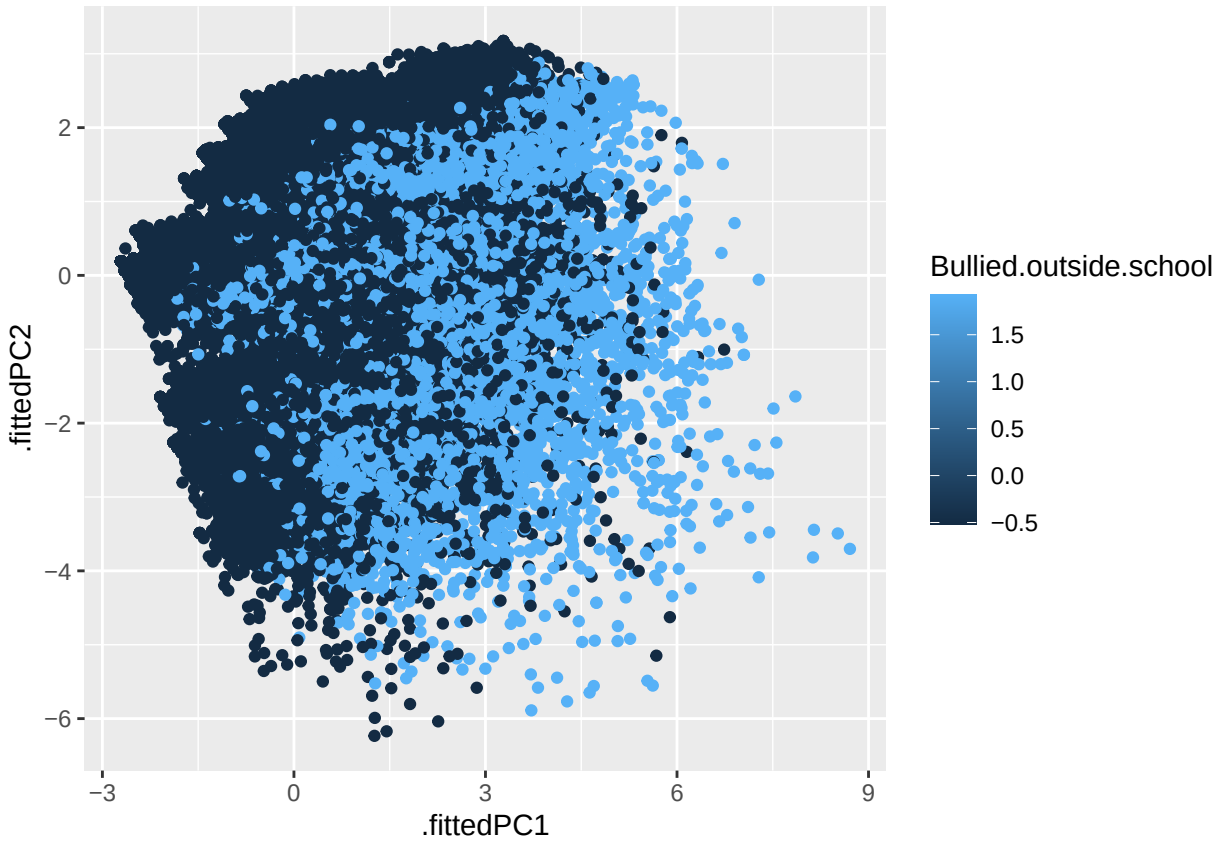
```
Bully_augmented <- pca_Bully %>% augment(Bully_scaled)
head(Bully_augmented)
```

```
## # A tibble: 6 x 31
##   .rownames Bullied.in~1 Bulli~2 Cyber~3 Custo~4 Sex[,1] Physi~5 Physi~6 Felt_~7
##   <chr>            <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 2              -0.516  -0.519  -0.521  -0.674   1.05   -0.279  -0.343  -1.14
## 2 3              -0.516  -0.519  -0.521  -1.41   -0.948  -0.279  -0.343  -0.290
## 3 4              -0.516  -0.519  -0.521  -1.41    1.05   -0.279   0.193  -1.14
## 4 5              -0.516  -0.519  -0.521  -0.674  -0.948   0.300  -0.343   0.554
## 5 6              -0.516  -0.519  -0.521  -2.15   -0.948  -0.279  -0.343  -0.290
## 6 7              -0.516  -0.519  -0.521  -1.41    1.05    0.300   0.997  -1.14
## # ... with 22 more variables: Close_friends <dbl[,1]>,
## #   Miss_school_no_permission <dbl[,1]>,
## #   Other_students_kind_and_helpful <dbl[,1]>,
## #   Parents_understand_problems <dbl[,1]>,
## #   Most_of_the_time_or_always_felt_lonely <dbl[,1]>,
## #   Missed_classes_or_school_without_permission <dbl[,1]>,
## #   Close_friends_3_or_more <dbl[,1]>, .fittedPC1 <dbl>, .fittedPC2 <dbl>, ...
```

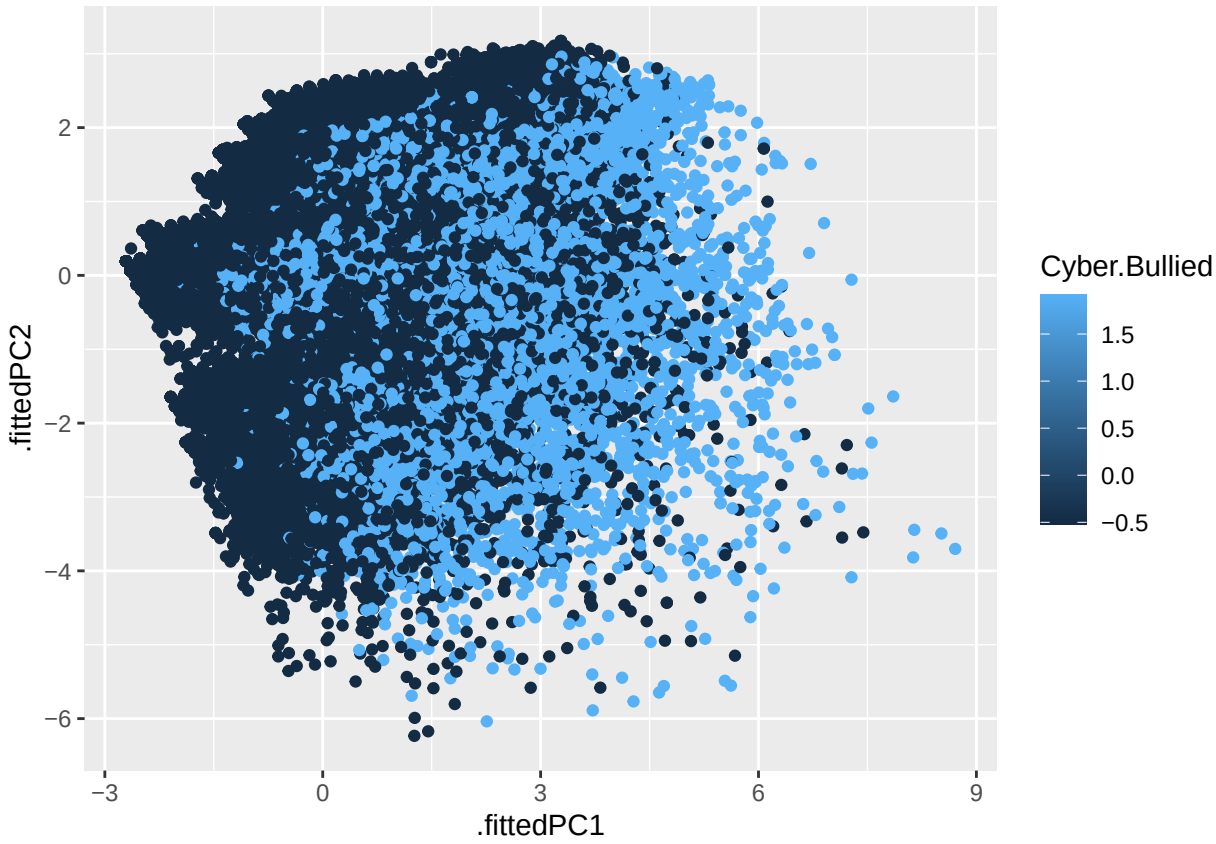
```
ggplot(Bully_augmented) + geom_point(aes(x = .fittedPC1, y = .fittedPC2 , col = Bullied.in.School))
```



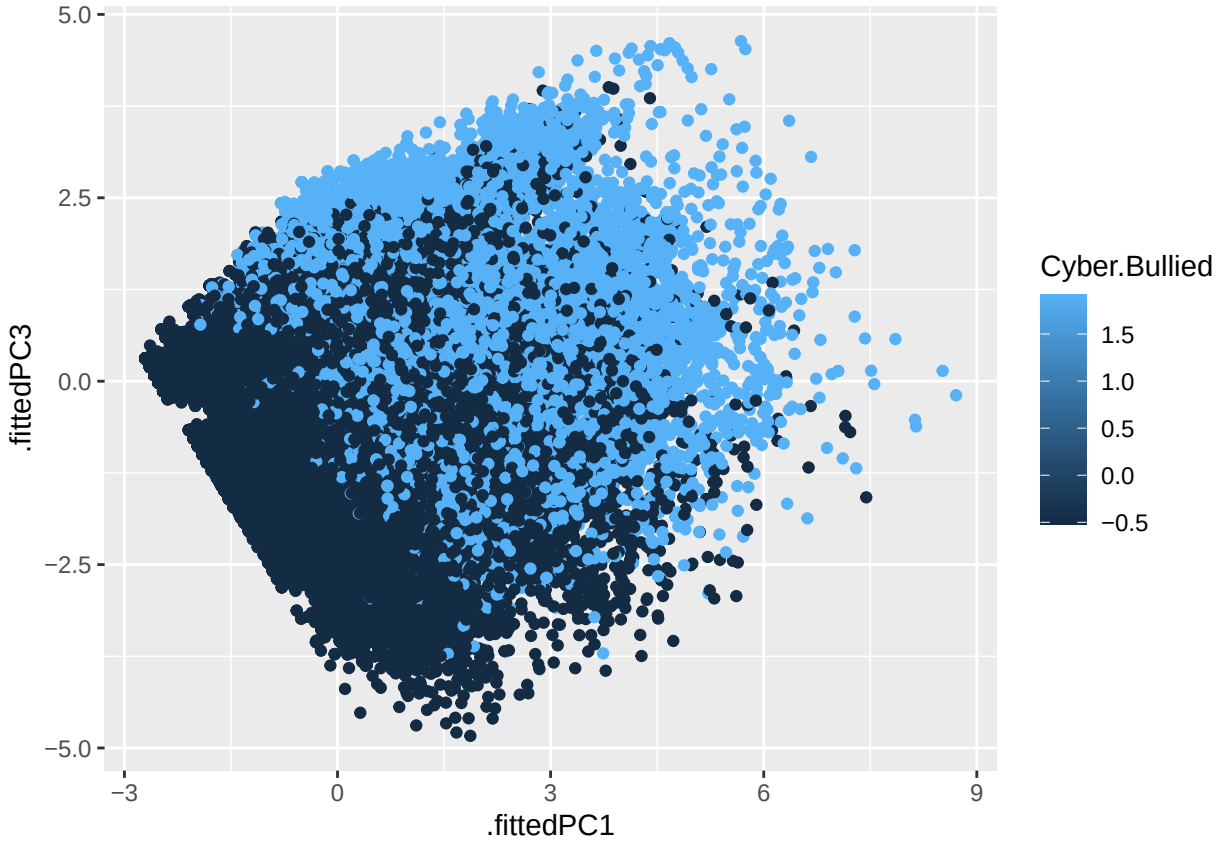

```
ggplot(Bully_augmented) + geom_point(aes(x = .fittedPC1, y = .fittedPC2 , col = Bullied.outside.school))
```




```
ggplot(Bully_augmented) + geom_point(aes(x = .fittedPC1, y = .fittedPC2 , col = Cyber.Bullied))
```



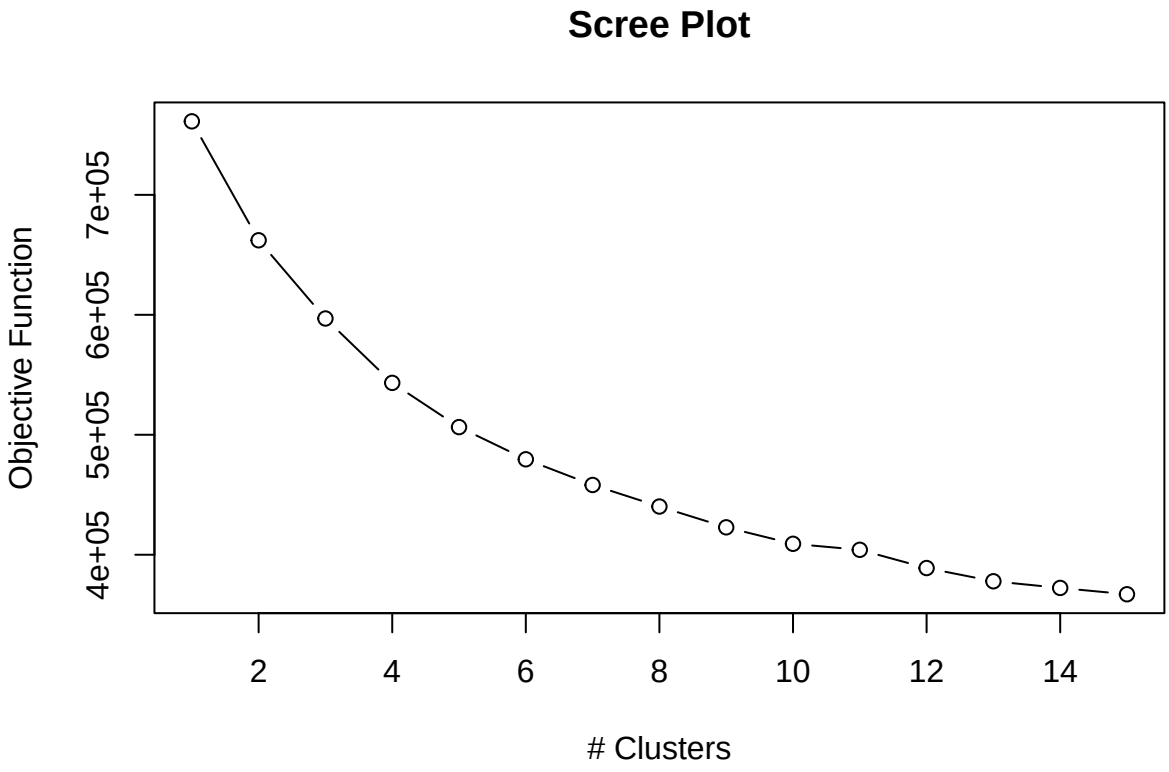

```
ggplot(Bully_augmented) + geom_point(aes(x = .fittedPC1, y = .fittedPC3 , col = Cyber.Bullied))
```



By displaying some visuals, we found that we need at least 10 PCA's to explain the variance. There is no clear division between PCA's in the graphs.

Cluster Analysis

```
set.seed(123)
within_ss <- rep(NA, 15)
for (i in 1:15){
  km_res <- kmeans(x = Bully_scaled,
                   centers = i,
                   nstart = 5)
  within_ss[i] <- km_res$tot.withinss
}

plot(1:15, within_ss, type = "b", ylab = "Objective Function", xlab = "# Clusters",
main = "Scree Plot")
```



It is hard to tell from the scree plot, so we decided that the point where it makes sense the most is k = 8

```
res_10 <- kmeans(x = Bully_scaled,
                 centers = 10,
                 nstart = 5)
```

size

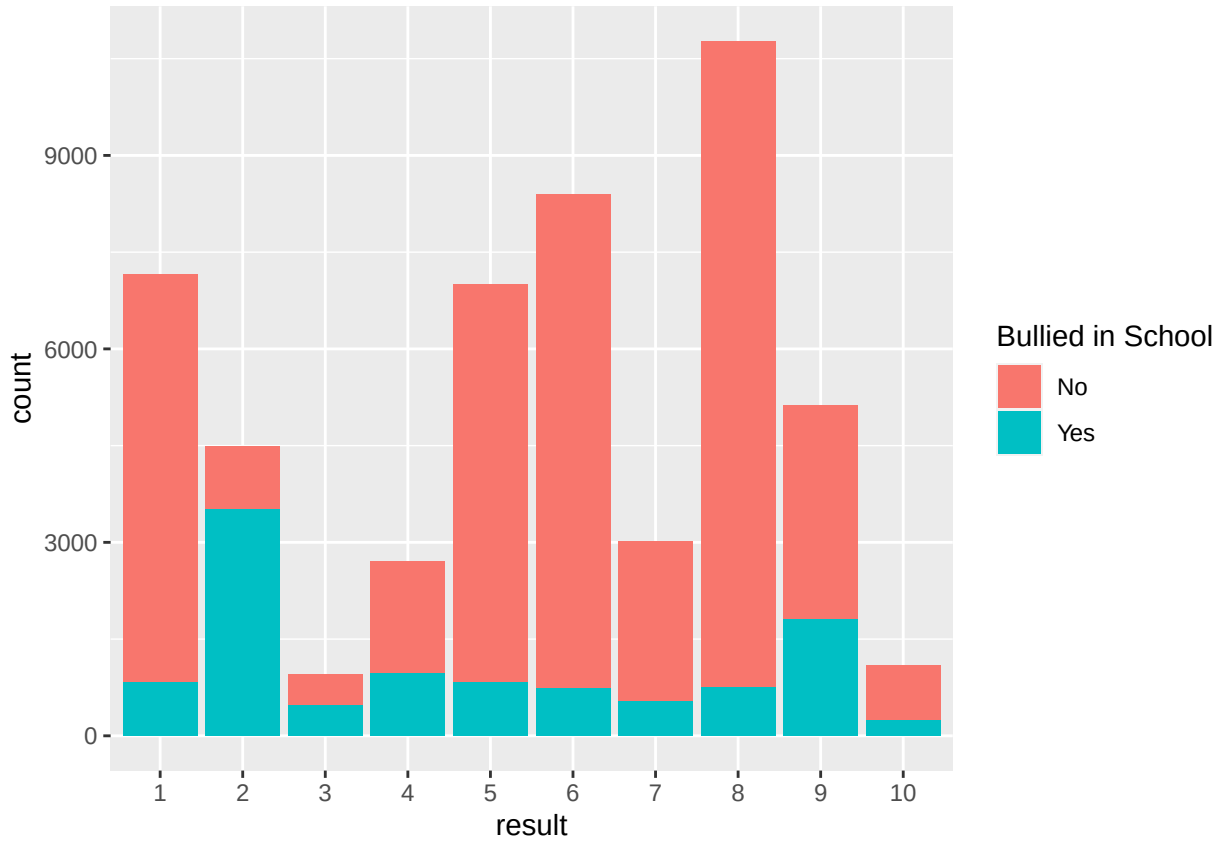
```
# Add clusters into the original data
result <- res_10$cluster
result <- as.factor(result)
size <- res_10$size
size
```

```
##  [1]  7154  4499   960  2714  7000  8397  3024 10774  5127  1104
```

```
Bully_cluster <- data.frame(Bully, result)
```

Visualization of the clusters

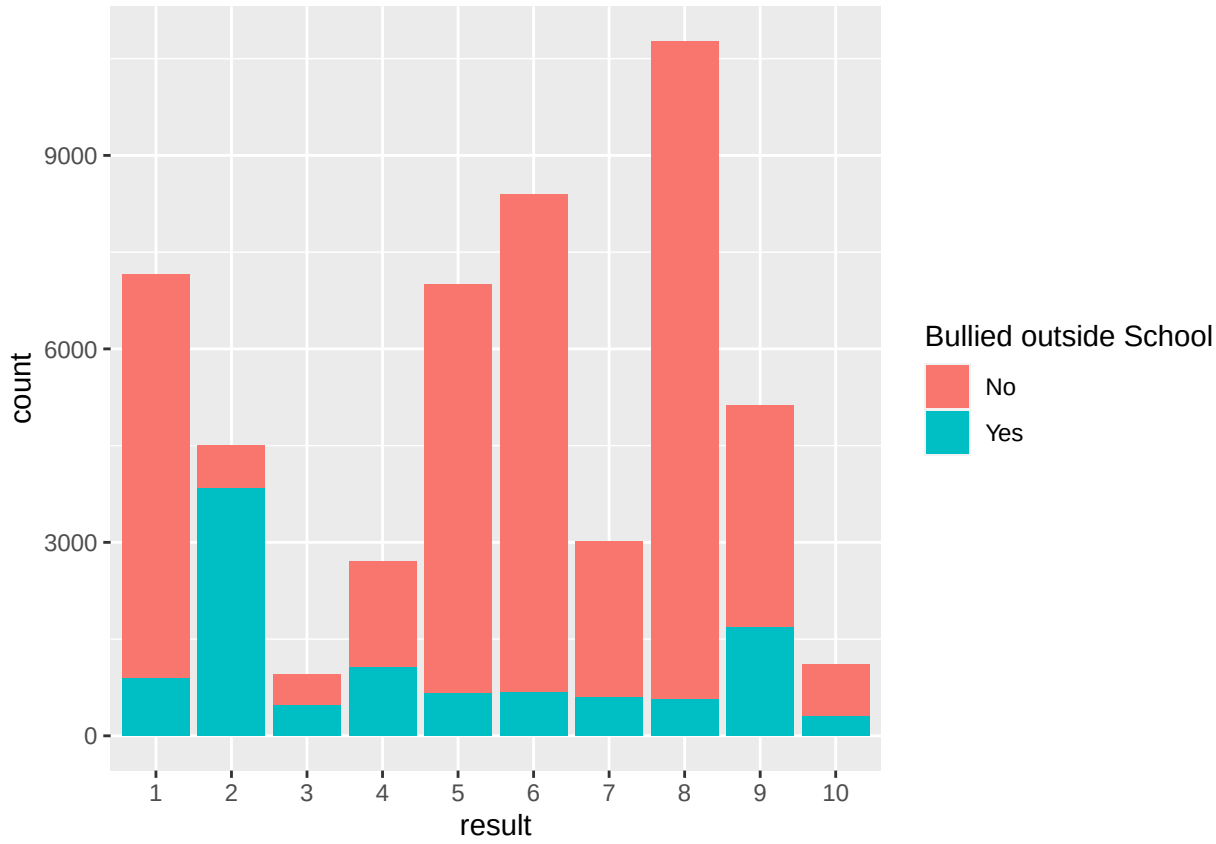
```
ggplot(Bully_cluster, aes(x = result, fill = factor(Bullied.in.School, levels = c(0,1), labels = c("No"
```




```
Bullied.in.School_table <- table(Bully_cluster$result, Bully_cluster$Bullied.in.School)
Bullied.in.School_table
```

```
##
##          0     1
##    1   6323   831
##    2    988  3511
##    3    487   473
##    4   1746   968
##    5   6163   837
##    6   7659   738
##    7   2495   529
##    8  10028   746
##    9   3325  1802
##    10   867   237
```

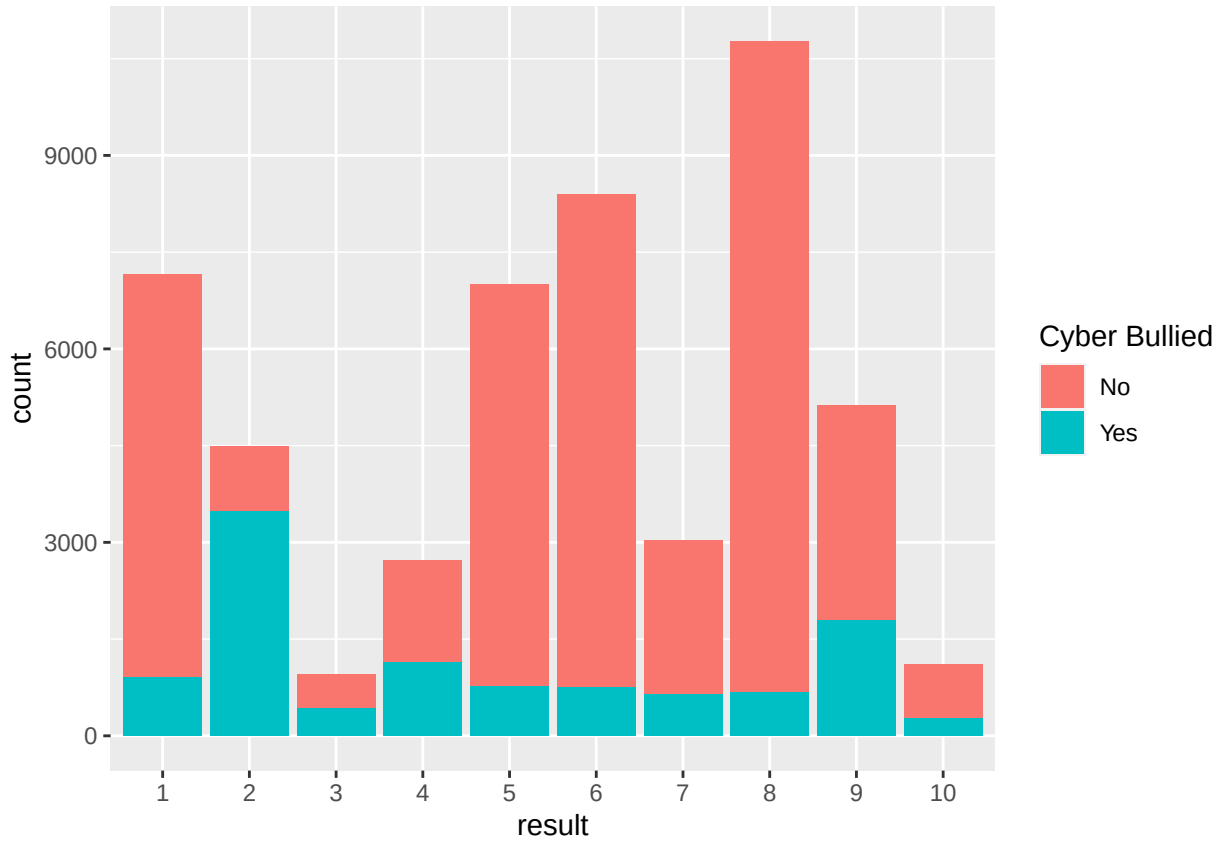
```
ggplot(Bully_cluster, aes(x = result, fill = factor(Bullied.outside.school, levels = c(0,1), labels = c
```




```
Bullied.outside.School_table <- table(Bully_cluster$result, Bully_cluster$Bullied.outside.school)
Bullied.outside.School_table
```

```
##
##          0     1
##    1   6255   899
##    2    655  3844
##    3    489   471
##    4   1650  1064
##    5   6334   666
##    6   7718   679
##    7   2430   594
##    8  10211   563
##    9   3438  1689
##    10   801   303
```

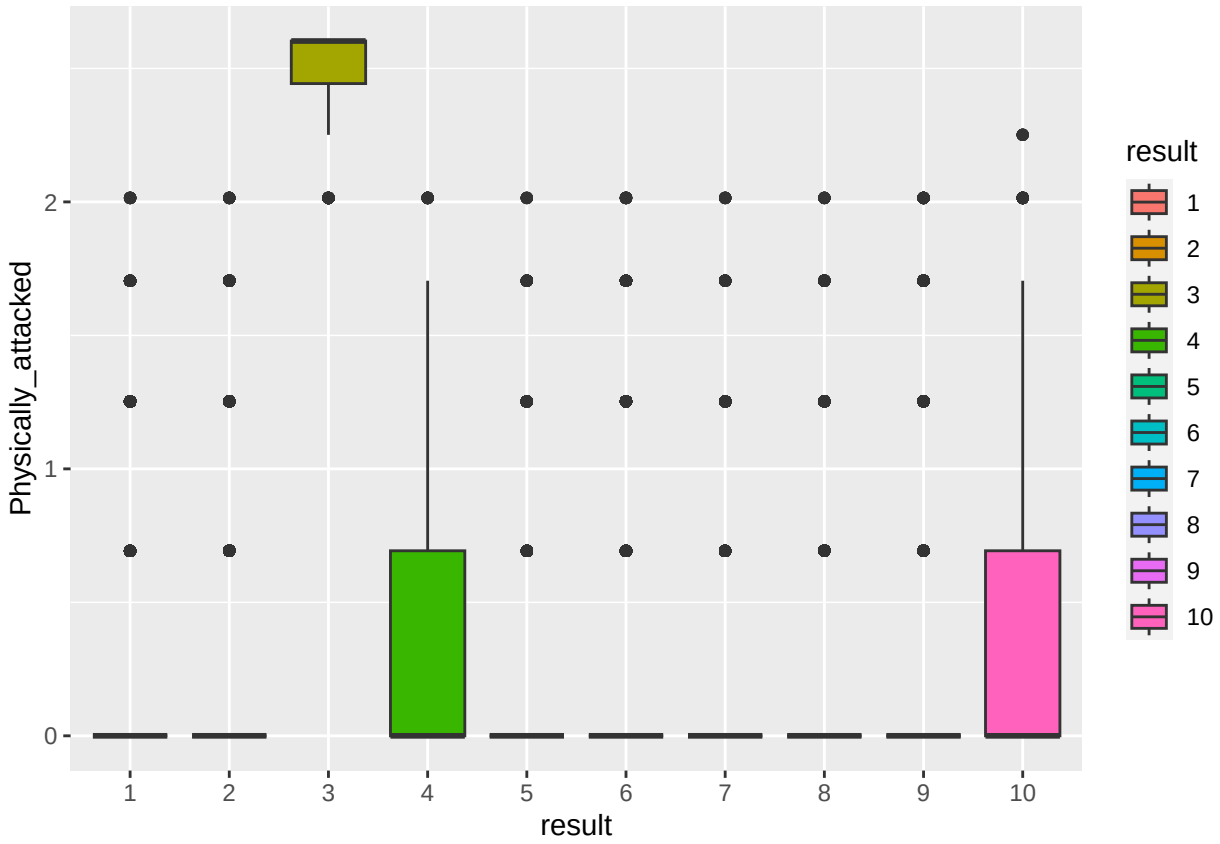
```
ggplot(Bully_cluster, aes(x = result, fill = factor(Cyber.Bullied, levels = c(0,1), labels = c("No", "Ye
```



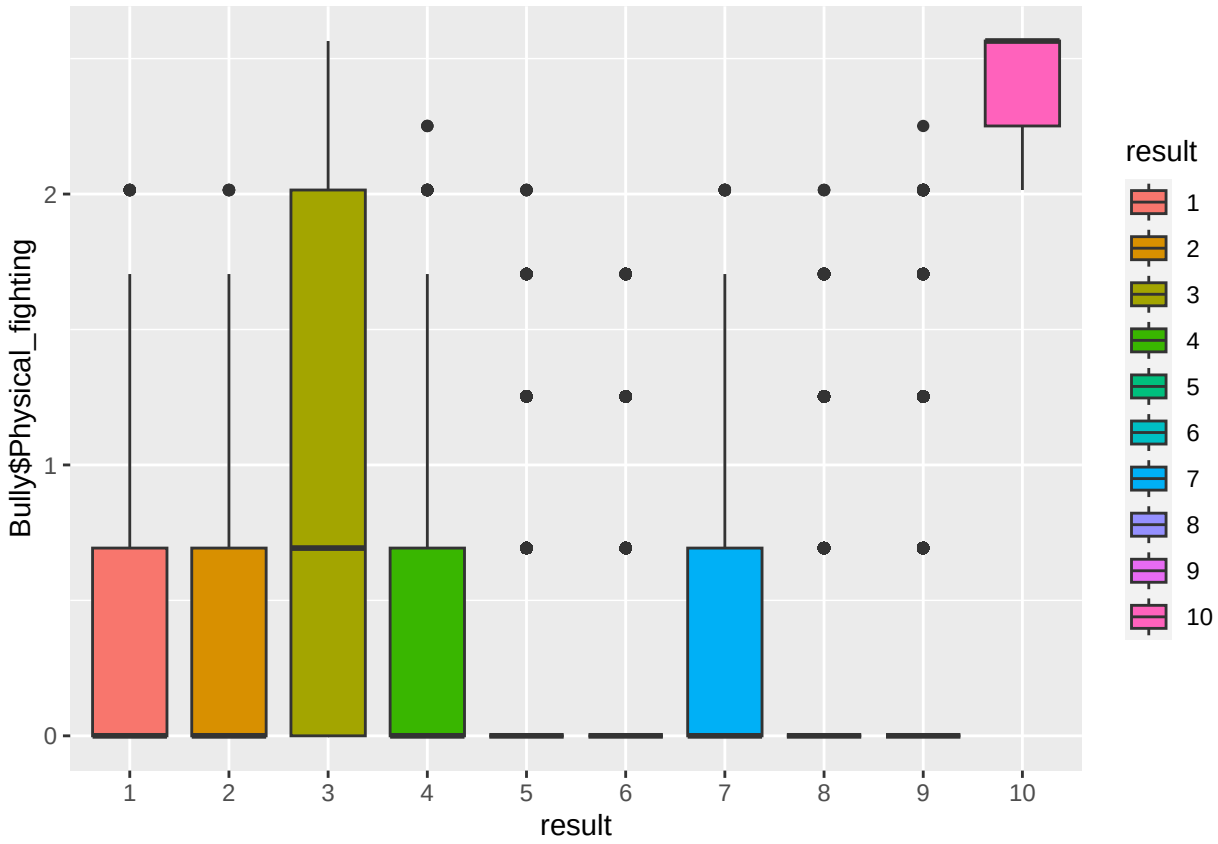

```
Cyber.Bullied_table <- table(Bully_cluster$result, Bully_cluster$Cyber.Bullied)
Cyber.Bullied_table
```

```
##
##          0     1
##    1   6252   902
##    2   1023  3476
##    3    541   419
##    4   1568  1146
##    5   6239   761
##    6   7647   750
##    7   2374   650
##    8  10096   678
##    9   3342  1785
##    10   825   279
```

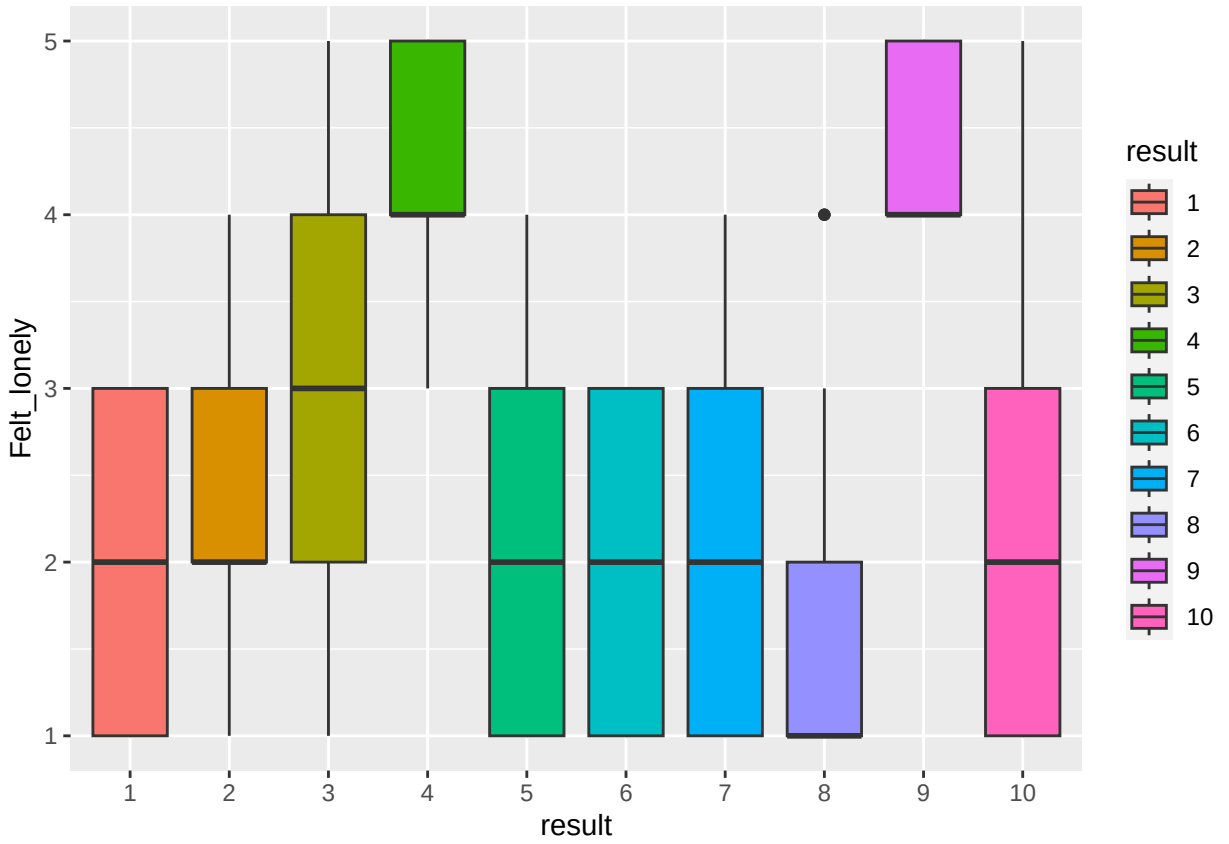
```
ggplot(Bully_cluster, aes(x = result, y = Physically_attacked, fill = result)) + geom_boxplot()
```



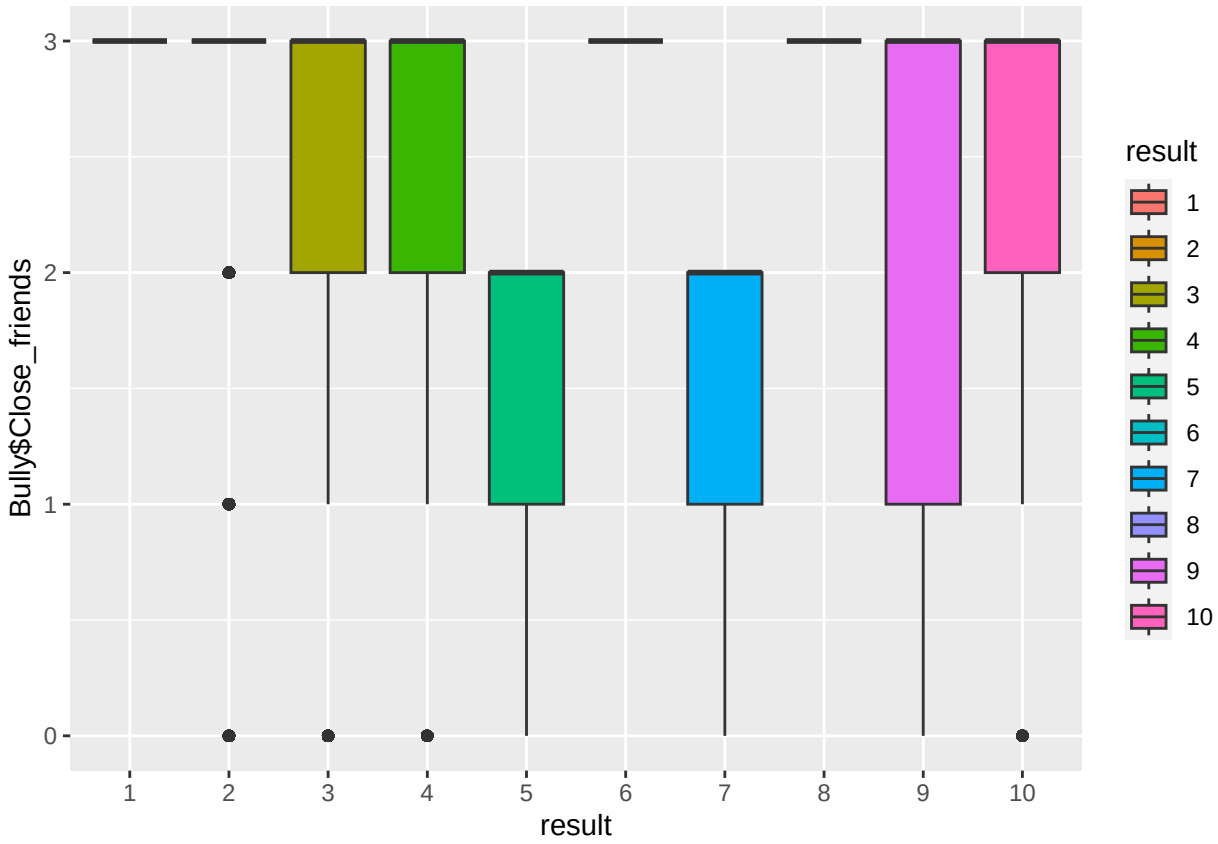

```
ggplot(Bully_cluster, aes(x = result, y = Bully$Physical_fighting, fill = result)) + geom_boxplot()
```



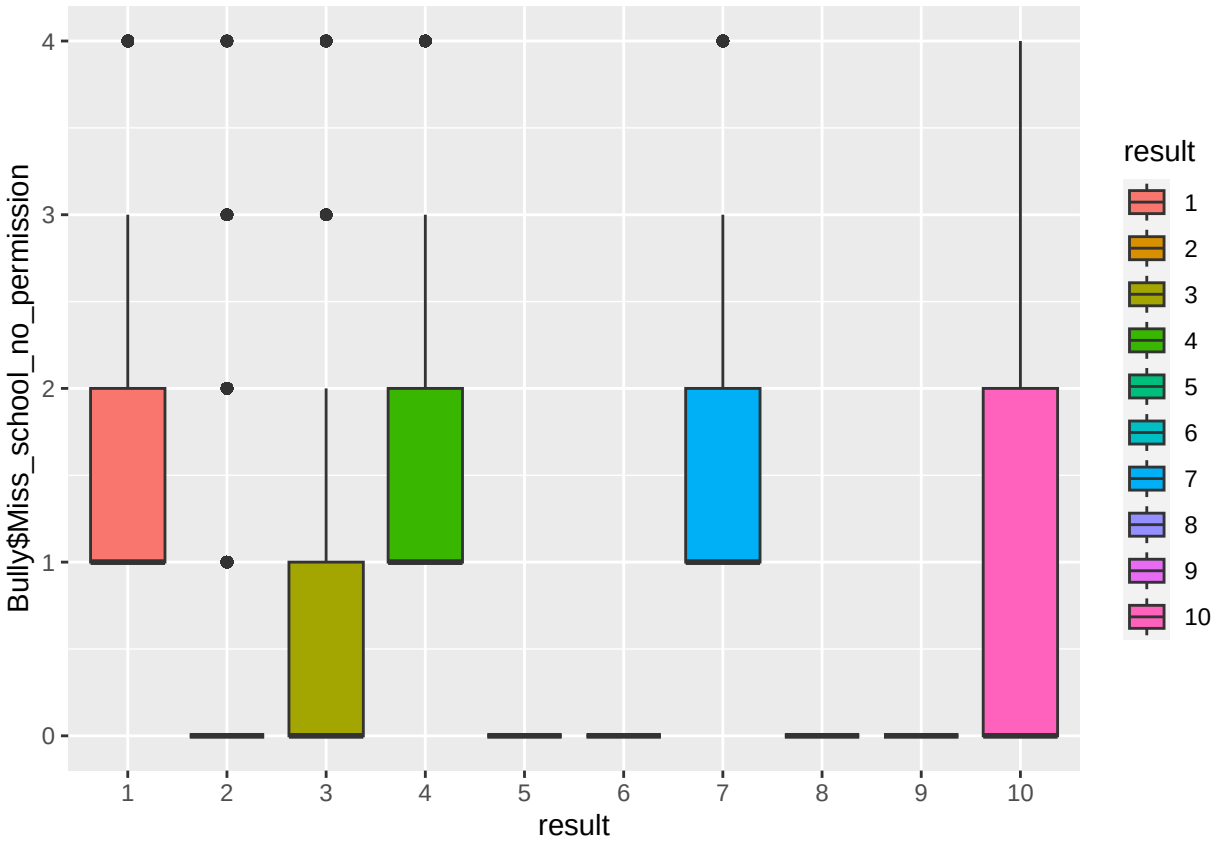

```
ggplot(Bully_cluster, aes(Bully, x = result, y =Felt_lonely, fill = result)) + geom_boxplot()
```



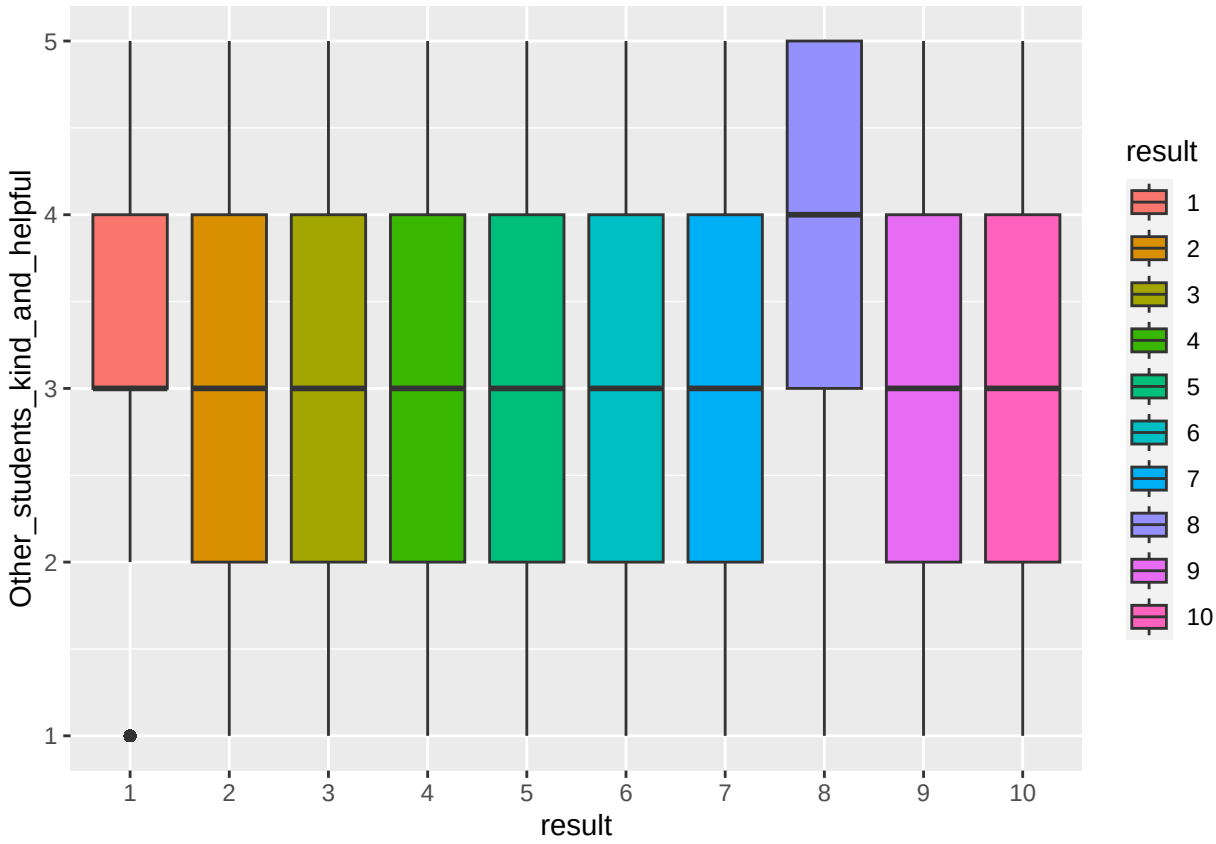

```
ggplot(Bully_cluster, aes(x = result, y = Bully$Close_friends, fill = result)) + geom_boxplot()
```



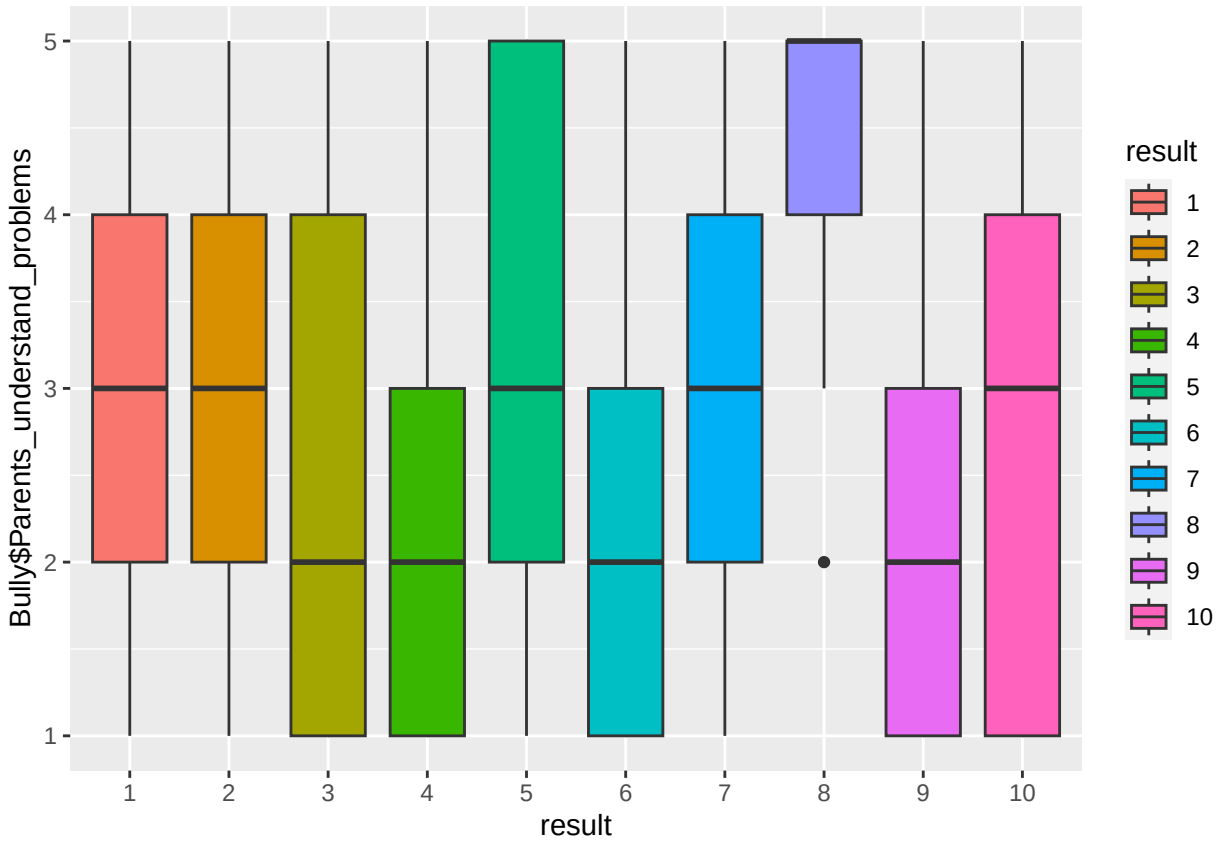

```
ggplot(Bully_cluster, aes(x = result, y = Bully$Miss_school_no_permission, fill = result)) + geom_boxpl
```



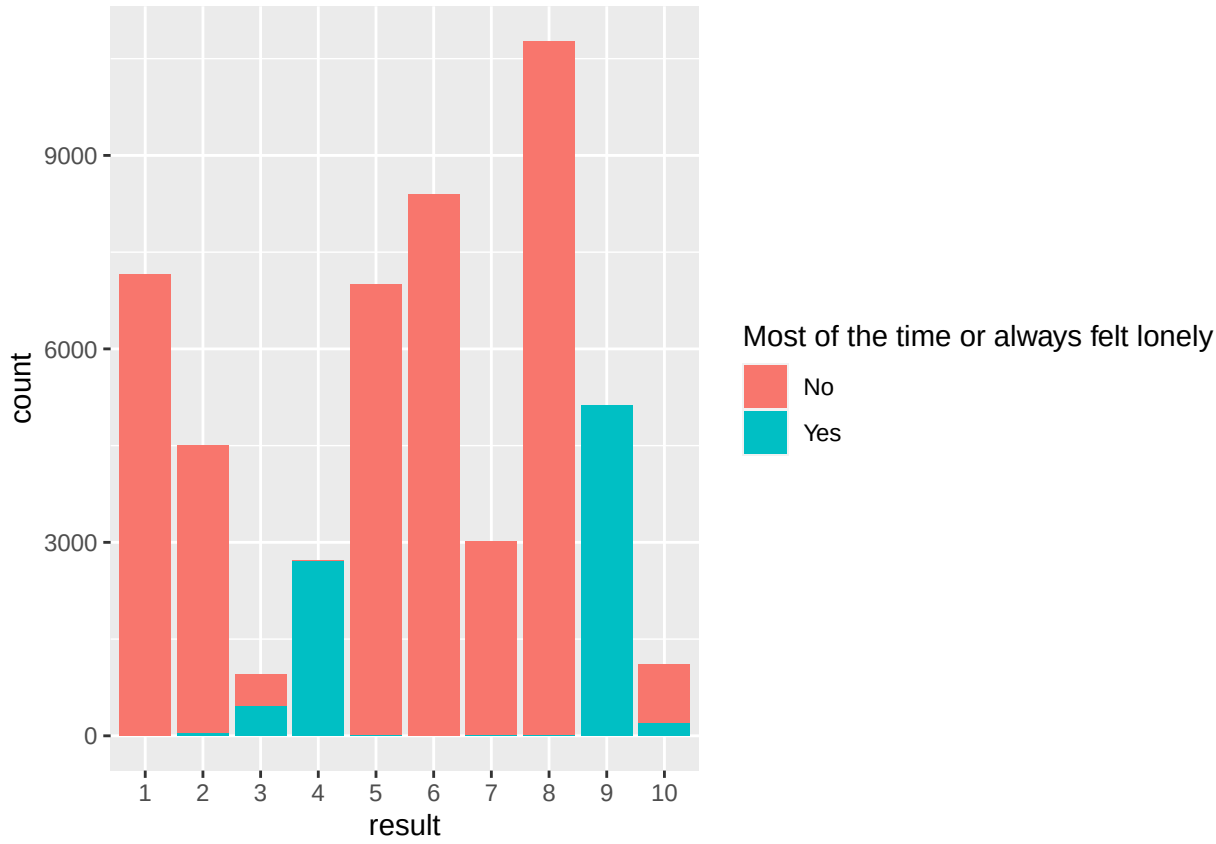

```
ggplot(Bully_cluster, aes(x = result, y = Other_students_kind_and_helpful, fill = result)) + geom_boxpl
```



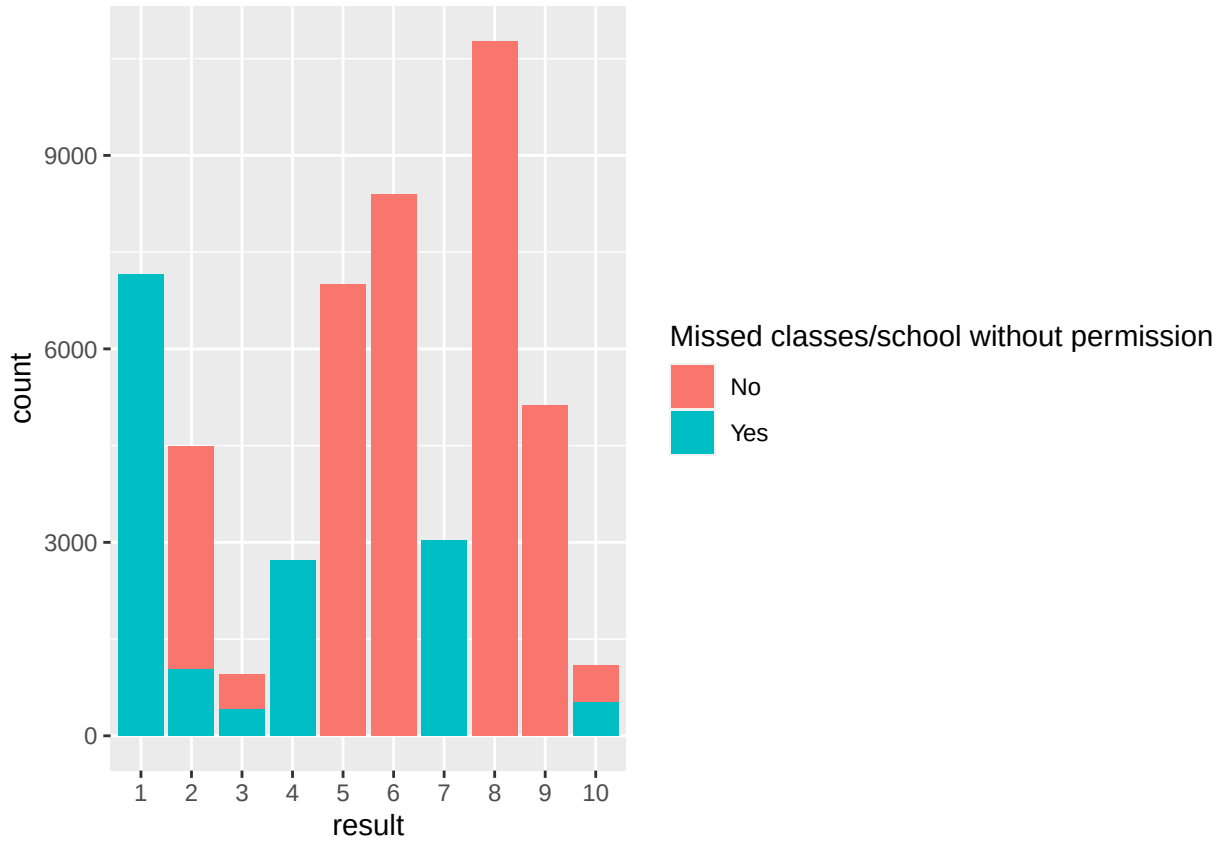

```
ggplot(Bully_cluster, aes(x = result, y = Bully$Parents_understand_problems, fill = result, )) + geom_b
```



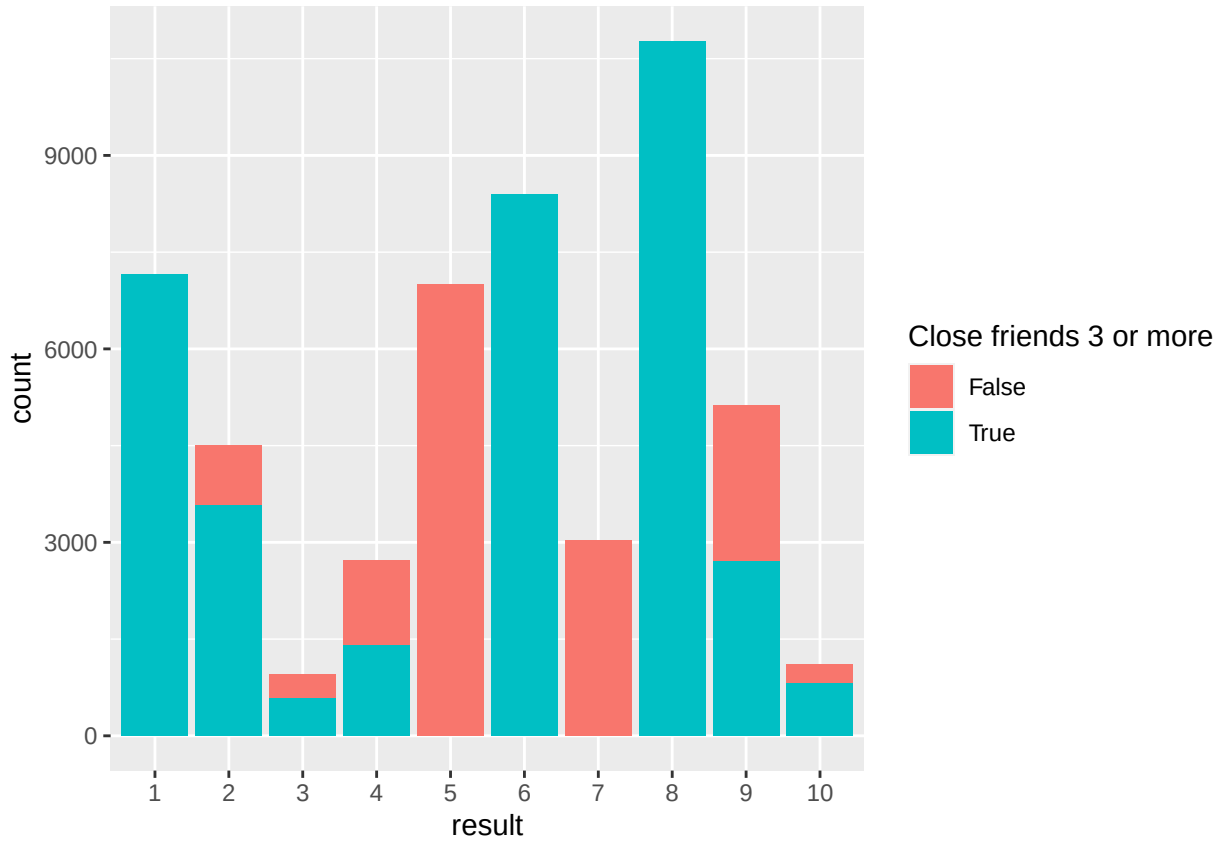

```
ggplot(Bully_cluster, aes(x = result, fill = factor(Most_of_the_time_or_always_felt_lonely, levels = c(
```



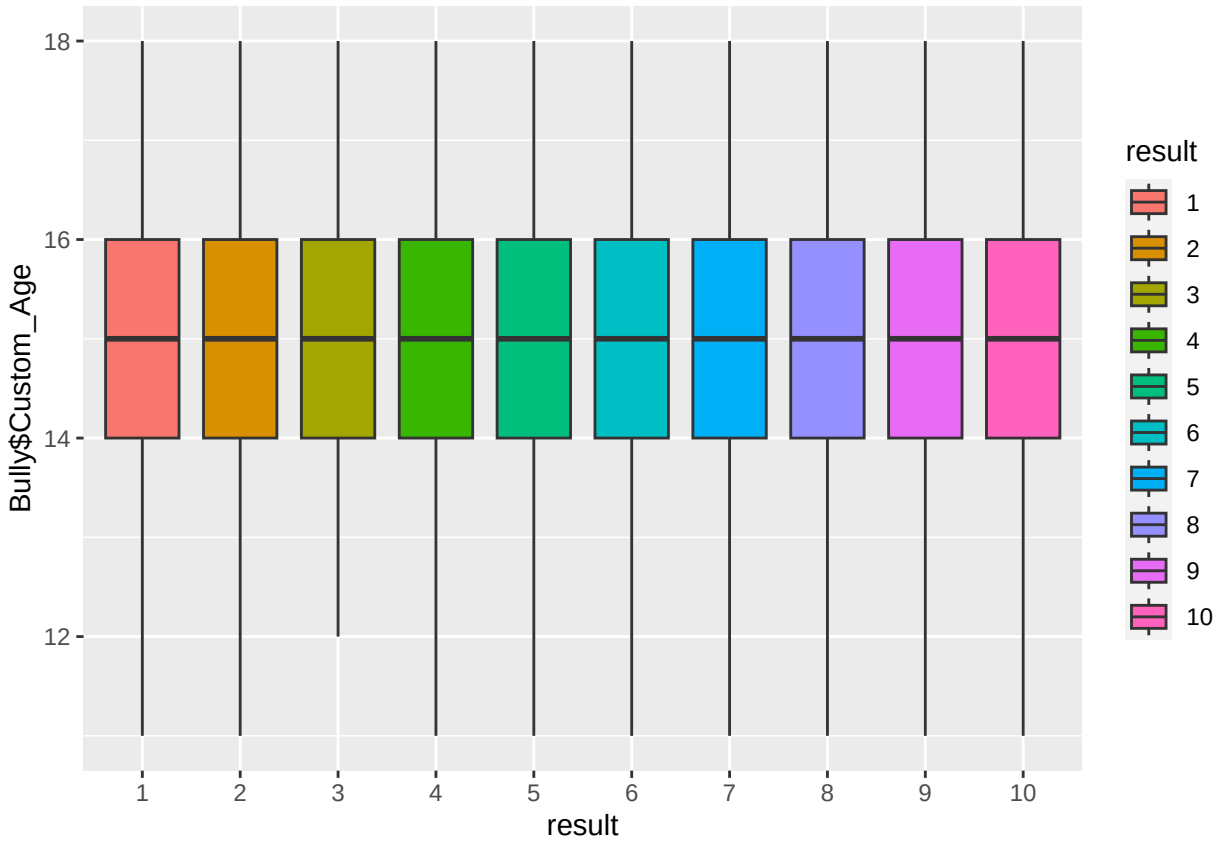

```
ggplot(Bully_cluster, aes(x = result, fill = factor(Missed_classes_or_school_without_permission, levels
```



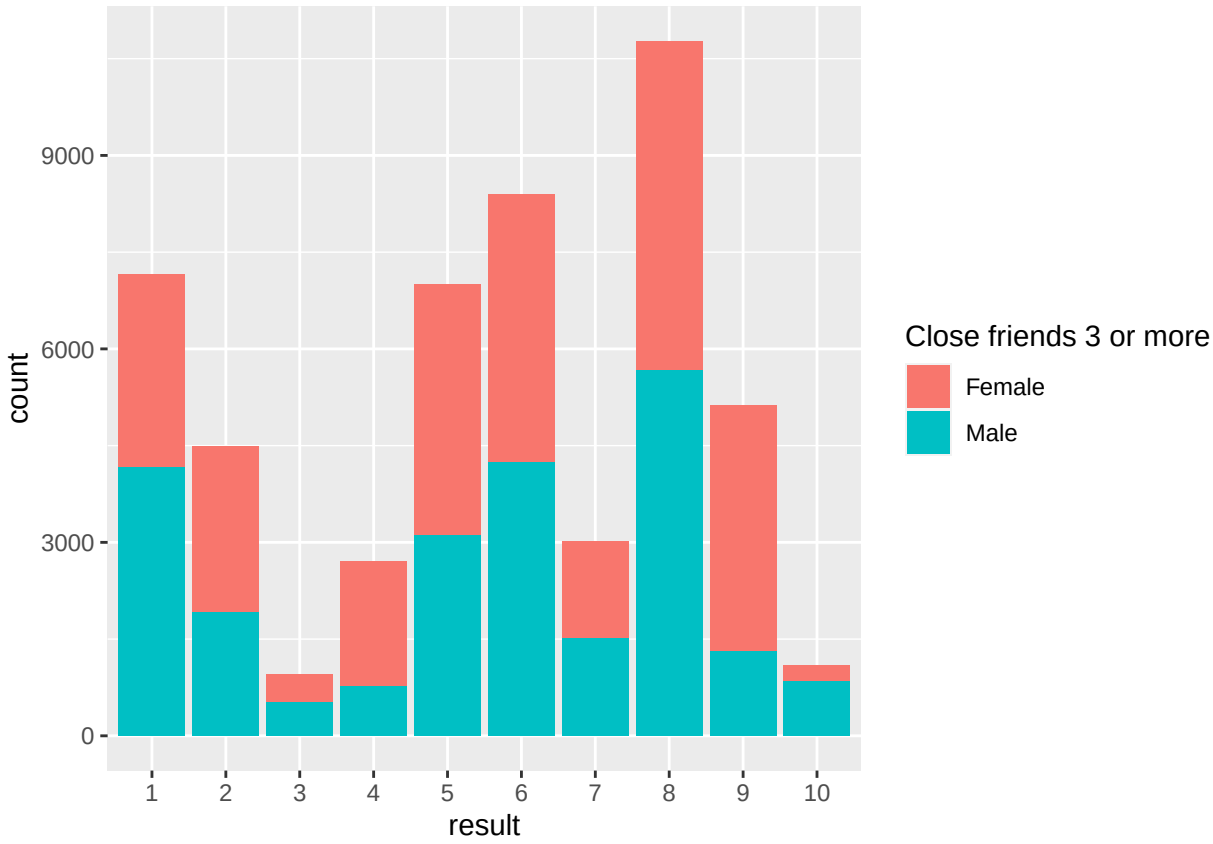

```
ggplot(Bully_cluster, aes(x = result, fill = factor(Close_friends_3_or_more, levels = c(0,1), labels = c
```




```
ggplot(Bully_cluster, aes(x = result, y = Bully$Custom_Age, fill= result)) + geom_boxplot()
```

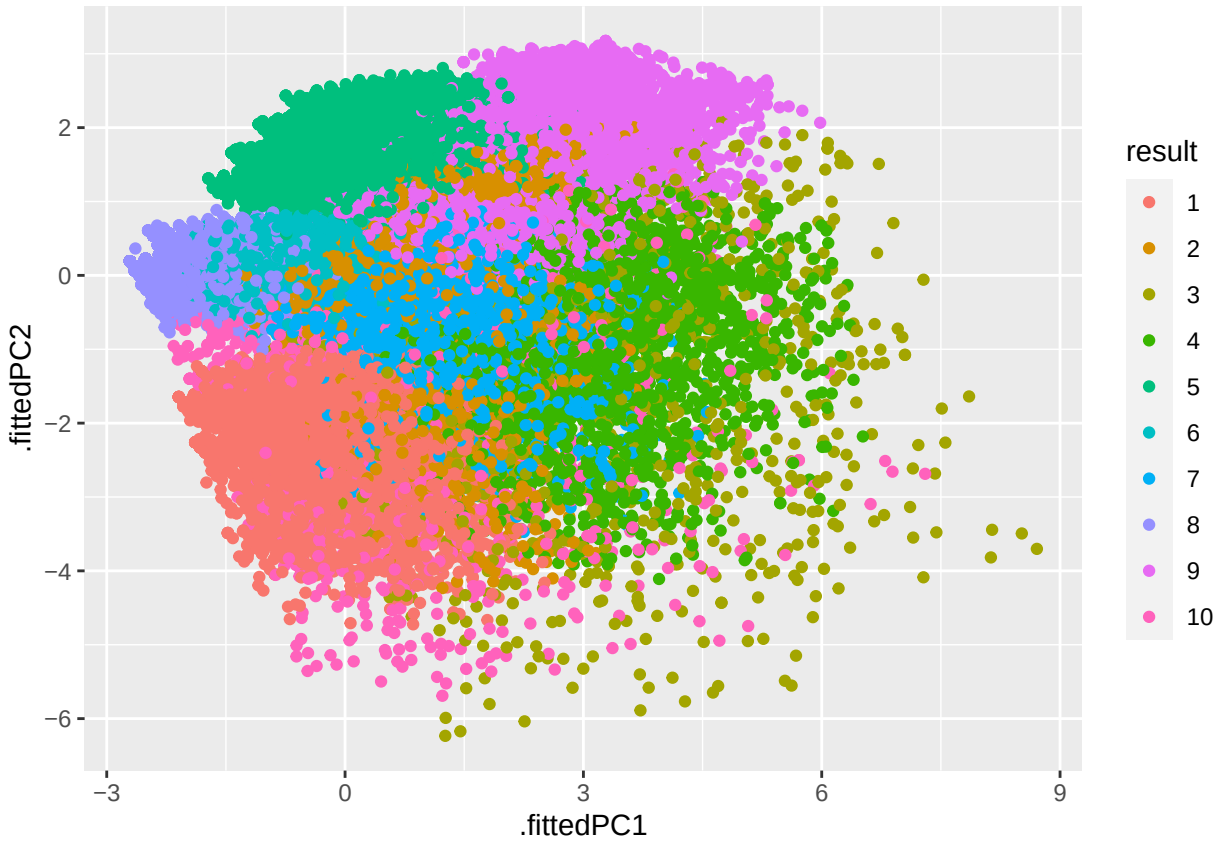



```
ggplot(Bully_cluster, aes(x = result, fill = factor(Sex, levels = c(0,1), labels = c("Female", "Male")))
```

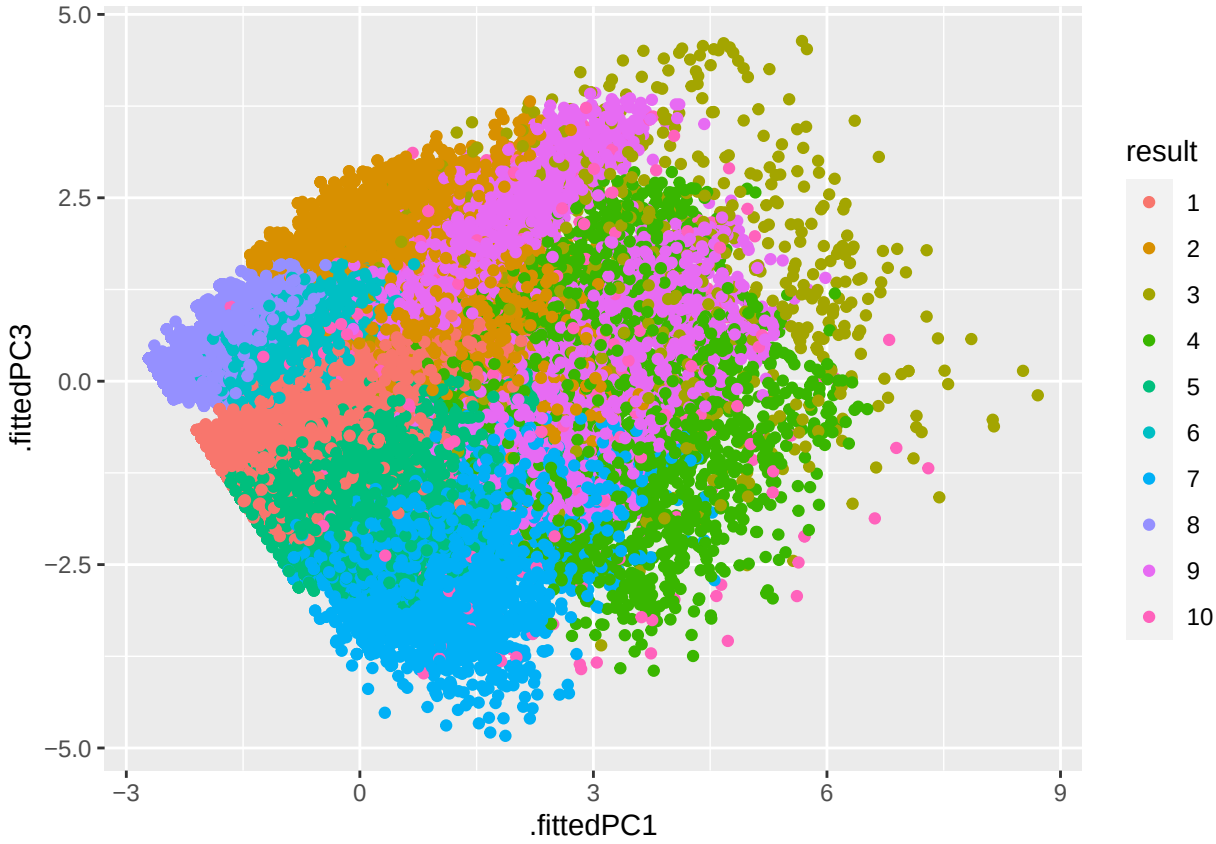


PCA Graphs after the clusters

```
ggplot(Bully_augmented) + geom_point(aes(x = .fittedPC1, y = .fittedPC2 , col = result))
```




```
ggplot(Bully_augmented) + geom_point(aes(x = .fittedPC1, y = .fittedPC3 , col = result))
```


5.0
2.5
0.0
−2.5
−5.0
.fittedPC3
−3
0
3
6
9
.fittedPC1
result
1
2
3
4
5
6
7
8
9
10