Project 1

Isabel Renteria and Zuzu Trottier

#filtering data

```
# read in data
raw_1516 <- read.csv("~/stat_proj_1/data/Y1516.csv", header = T)
raw_1617 <- read.csv("~/stat_proj_1/data/Y1617.csv", header = T)
raw_1718 <- read.csv("~/stat_proj_1/data/Y1718.csv", header = T)
raw_1819 <- read.csv("~/stat_proj_1/data/Y1819.csv", header = T)
raw_2122 <- read.csv("~/stat_proj_1/data/Y2122.csv", header = T)
raw_2223 <- read.csv("~/stat_proj_1/data/Y2223.csv", header = T)

# filter data for the attendance and suspension columns
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
manova_1516 <- raw_1516 %>%
  select(contains("Attendance") |
         contains("Suspensions") |
         "Primary_Category") %>%
  select(ends_with("2_Pct")) %>%
  select(!contains("Avg")) %>%
  select(!contains("Lbl")) %>%
  select(!contains("PSAT")) %>%
  na.omit()

manova_1617 <- raw_1617 %>%
  select(contains("Attendance") |
         contains("Suspensions") |
         "Primary_Category") %>%
  select(ends_with("2_Pct")) %>%
  select(!contains("Avg")) %>%
  select(!contains("Lbl")) %>%
  select(!contains("PSAT")) %>%
  na.omit()
```

```
manova_1718 <- raw_1718 %>%
  select(contains("Attendance") |
         contains("Suspensions") |
         "Primary_Category") %>%
  select(ends_with("2_Pct")) %>%
  select(!contains("Avg")) %>%
  select(!contains("Lbl")) %>%
  select(!contains("PSAT")) %>%
  na.omit()

manova_1819 <- raw_1819 %>%
  select(contains("Attendance") |
         contains("Suspensions") |
         "Primary_Category") %>%
  select(ends_with("2_Pct")) %>%
  select(!contains("Avg")) %>%
  select(!contains("Lbl")) %>%
  select(!contains("PSAT")) %>%
  na.omit()

manova_2122 <- raw_2122 %>%
  select(contains("Attendance") |
         contains("Suspensions") |
         "Primary_Category") %>%
  select(ends_with("2_Pct")) %>%
  select(!contains("Avg")) %>%
  select(!contains("Lbl")) %>%
  select(!contains("PSAT")) %>%
  na.omit()

manova_2223 <- raw_2223 %>%
  select(contains("Attendance") |
         contains("Suspensions") |
         "Primary_Category") %>%
  select(ends_with("2_Pct")) %>%
  select(!contains("Avg")) %>%
  select(!contains("Lbl")) %>%
  select(!contains("PSAT")) %>%
  na.omit()
```

our different considerations of groupings

```
pre_covid_mano <- rbind(manova_1516,manova_1617,manova_1718,manova_1819)
post_covid_mano <- rbind(manova_2122,manova_2223)



pre_covid_mano <- mutate(pre_covid_mano, across(everything(), ~ ifelse(.x == 0, 0.01, .x)))
post_covid_mano <- mutate(post_covid_mano, across(everything(), ~ ifelse(.x == 0, 0.01, .x)))
```

```r
all_data <- rbind(pre_covid_mano,post_covid_mano)
```

Independent?

```r
cor(pre_covid_mano)
```

```
##                                          Student_Attendance_Year_2_Pct
## Student_Attendance_Year_2_Pct                              1.00000000
## Teacher_Attendance_Year_2_Pct                              0.09082425
## Suspensions_Per_100_Students_Year_2_Pct                   -0.61485895
## Misconducts_To_Suspensions_Year_2_Pct                     -0.14450671
##                                          Teacher_Attendance_Year_2_Pct
## Student_Attendance_Year_2_Pct                              0.09082425
## Teacher_Attendance_Year_2_Pct                              1.00000000
## Suspensions_Per_100_Students_Year_2_Pct                   -0.06036965
## Misconducts_To_Suspensions_Year_2_Pct                     -0.03519533
##                                          Suspensions_Per_100_Students_Year_2_Pct
## Student_Attendance_Year_2_Pct                                         -0.61485895
## Teacher_Attendance_Year_2_Pct                                         -0.06036965
## Suspensions_Per_100_Students_Year_2_Pct                                1.00000000
## Misconducts_To_Suspensions_Year_2_Pct                                  0.20385934
##                                          Misconducts_To_Suspensions_Year_2_Pct
## Student_Attendance_Year_2_Pct                                       -0.14450671
## Teacher_Attendance_Year_2_Pct                                       -0.03519533
## Suspensions_Per_100_Students_Year_2_Pct                              0.20385934
## Misconducts_To_Suspensions_Year_2_Pct                                1.00000000
```

```r
cor(post_covid_mano)
```

```
##                                          Student_Attendance_Year_2_Pct
## Student_Attendance_Year_2_Pct                              1.00000000
## Teacher_Attendance_Year_2_Pct                              0.28228203
## Suspensions_Per_100_Students_Year_2_Pct                   -0.64915992
## Misconducts_To_Suspensions_Year_2_Pct                     -0.09092687
##                                          Teacher_Attendance_Year_2_Pct
## Student_Attendance_Year_2_Pct                               0.2822820
## Teacher_Attendance_Year_2_Pct                               1.0000000
## Suspensions_Per_100_Students_Year_2_Pct                    -0.0706949
## Misconducts_To_Suspensions_Year_2_Pct                      -0.1175303
##                                          Suspensions_Per_100_Students_Year_2_Pct
## Student_Attendance_Year_2_Pct                                         -0.6491599
## Teacher_Attendance_Year_2_Pct                                         -0.0706949
## Suspensions_Per_100_Students_Year_2_Pct                                1.0000000
## Misconducts_To_Suspensions_Year_2_Pct                                  0.1379702
##                                          Misconducts_To_Suspensions_Year_2_Pct
## Student_Attendance_Year_2_Pct                                       -0.09092687
## Teacher_Attendance_Year_2_Pct                                       -0.11753033
## Suspensions_Per_100_Students_Year_2_Pct                              0.13797017
## Misconducts_To_Suspensions_Year_2_Pct                                1.00000000
```

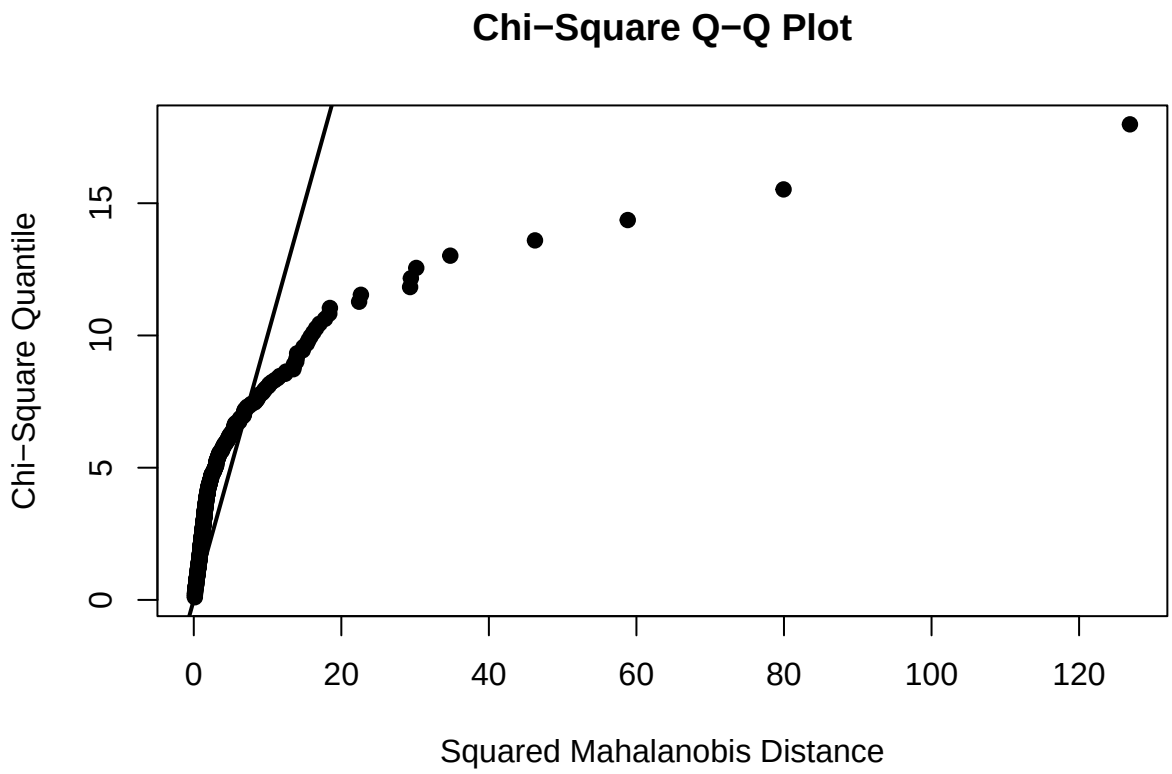
```r
cor(all_data)
```

```
##                                          Student_Attendance_Year_2_Pct
## Student_Attendance_Year_2_Pct                                1.0000000
## Teacher_Attendance_Year_2_Pct                                0.1472220
## Suspensions_Per_100_Students_Year_2_Pct                     -0.4805800
## Misconducts_To_Suspensions_Year_2_Pct                       -0.1045475
##                                          Teacher_Attendance_Year_2_Pct
## Student_Attendance_Year_2_Pct                               0.14722202
## Teacher_Attendance_Year_2_Pct                               1.00000000
## Suspensions_Per_100_Students_Year_2_Pct                    -0.05647993
## Misconducts_To_Suspensions_Year_2_Pct                      -0.04478753
##                                          Suspensions_Per_100_Students_Year_2_Pct
## Student_Attendance_Year_2_Pct                                        -0.48057999
## Teacher_Attendance_Year_2_Pct                                        -0.05647993
## Suspensions_Per_100_Students_Year_2_Pct                               1.00000000
## Misconducts_To_Suspensions_Year_2_Pct                                0.19411630
##                                          Misconducts_To_Suspensions_Year_2_Pct
## Student_Attendance_Year_2_Pct                                       -0.10454753
## Teacher_Attendance_Year_2_Pct                                       -0.04478753
## Suspensions_Per_100_Students_Year_2_Pct                              0.19411630
## Misconducts_To_Suspensions_Year_2_Pct                               1.00000000
```

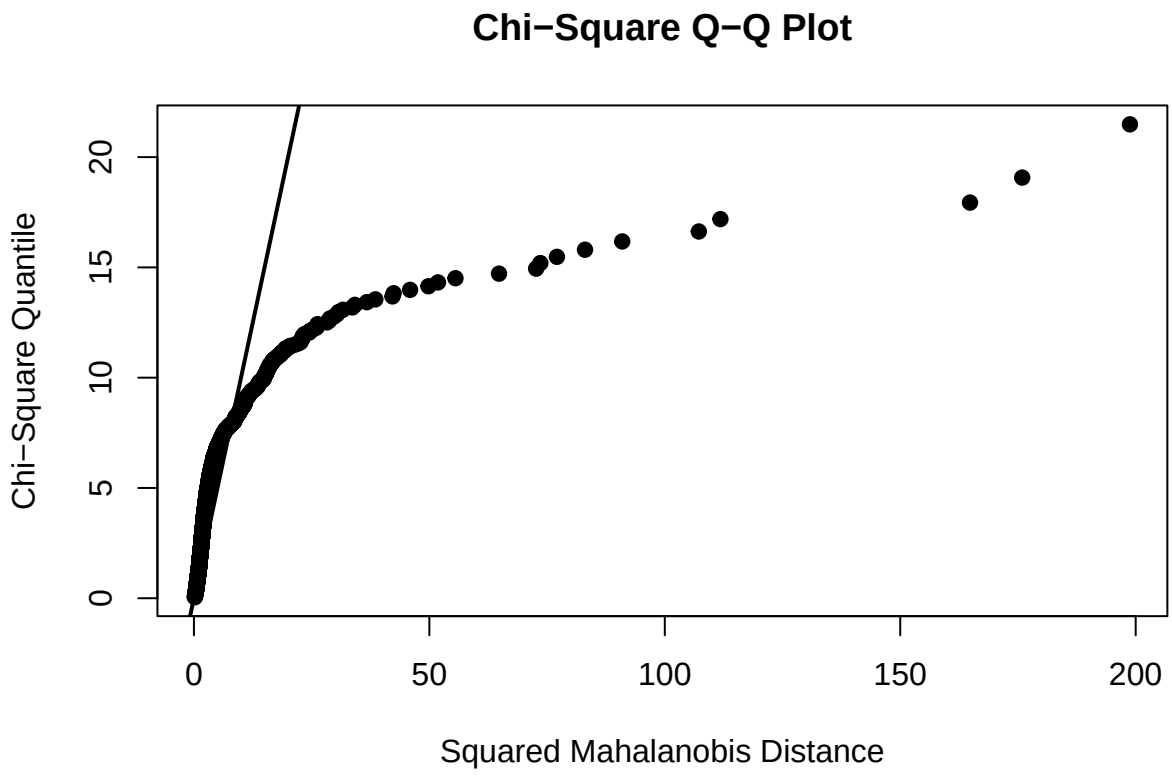
```r
# generally independent, with some higher correlation values for (suspensions and student attendance) a
```

Normality?

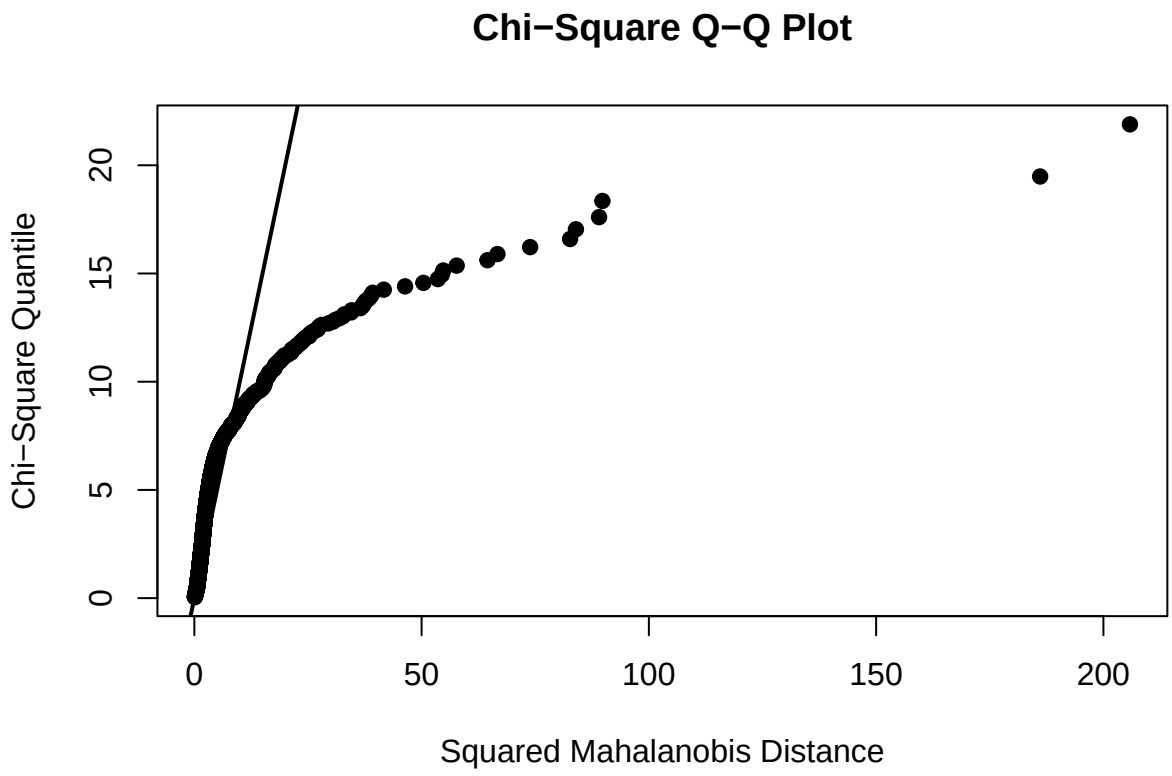
```r
library(MVN)
mvn_pre <- mvn(post_covid_mano, mvnTest="hz", multivariatePlot = "qq") #no
```




```
mvn_post <- mvn(pre_covid_mano, mvnTest="hz" , multivariatePlot = "qq") #no
```




```
mvn_all <- mvn(all_data, mvnTest="hz" , multivariatePlot = "qq") #no
```



#MANOVA (IGNORE THIS MANOVA)

```
mano_pre <- manova( cbind(Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_2_Pct,Miscondu
summary(mano_pre)
```

```
##                               Df  Pillai approx F num Df den Df    Pr(>F)
## Student_Attendance_Year_2_Pct    1 0.38128   404.05      3   1967 < 2.2e-16 ***
## Residuals                     1969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mano_post <- manova( cbind(Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_2_Pct,Miscond
summary(mano_post)
```

```
##                               Df  Pillai approx F num Df den Df    Pr(>F)
## Student_Attendance_Year_2_Pct   1 0.47818   121.57      3    398 < 2.2e-16 ***
## Residuals                     400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mano_all <- manova( cbind(Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_2_Pct,Miscondu
summary(mano_all, test = "Wilks")
```

```
##                               Df   Wilks approx F num Df den Df    Pr(>F)
```

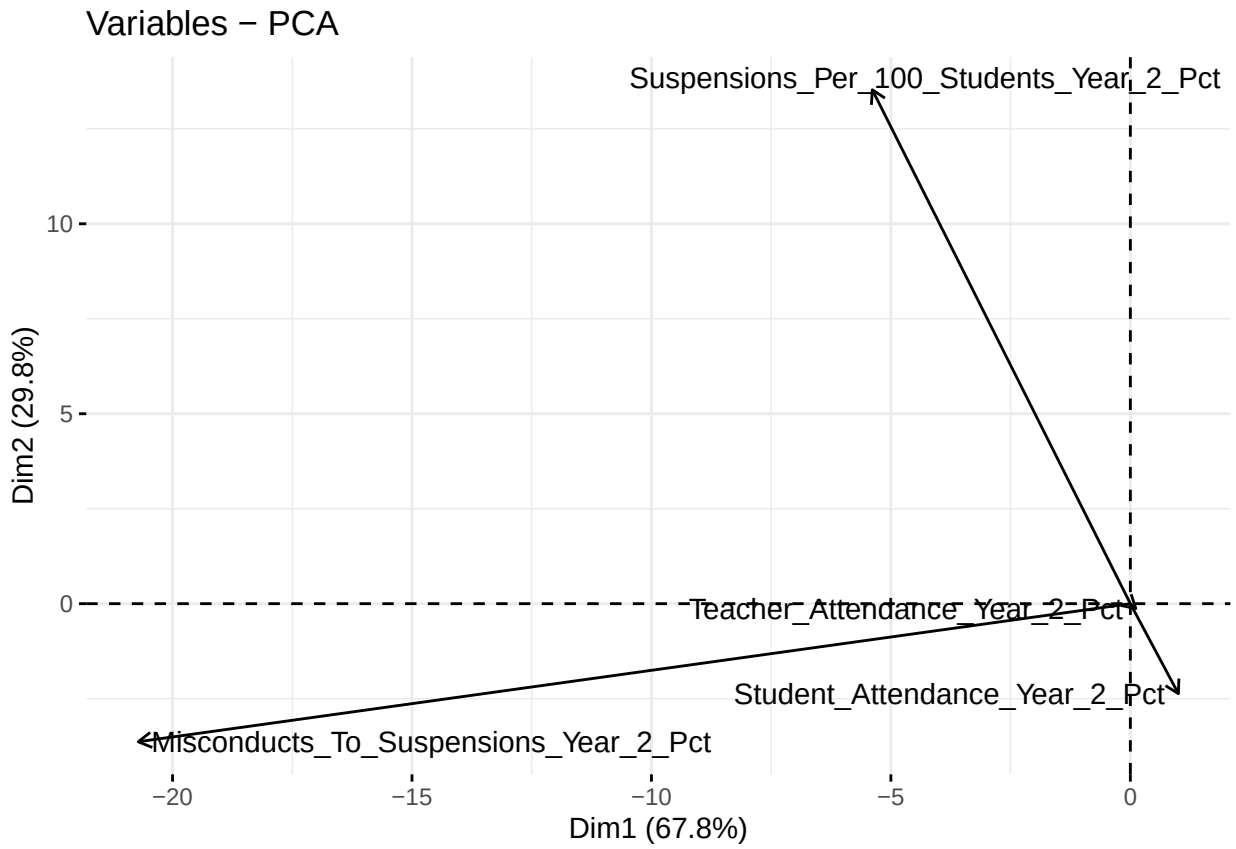
```
## Student_Attendance_Year_2_Pct    1 0.75452   256.91      3   2369 < 2.2e-16 ***
## Residuals                     2371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#PCA

```
library(factoextra)
```

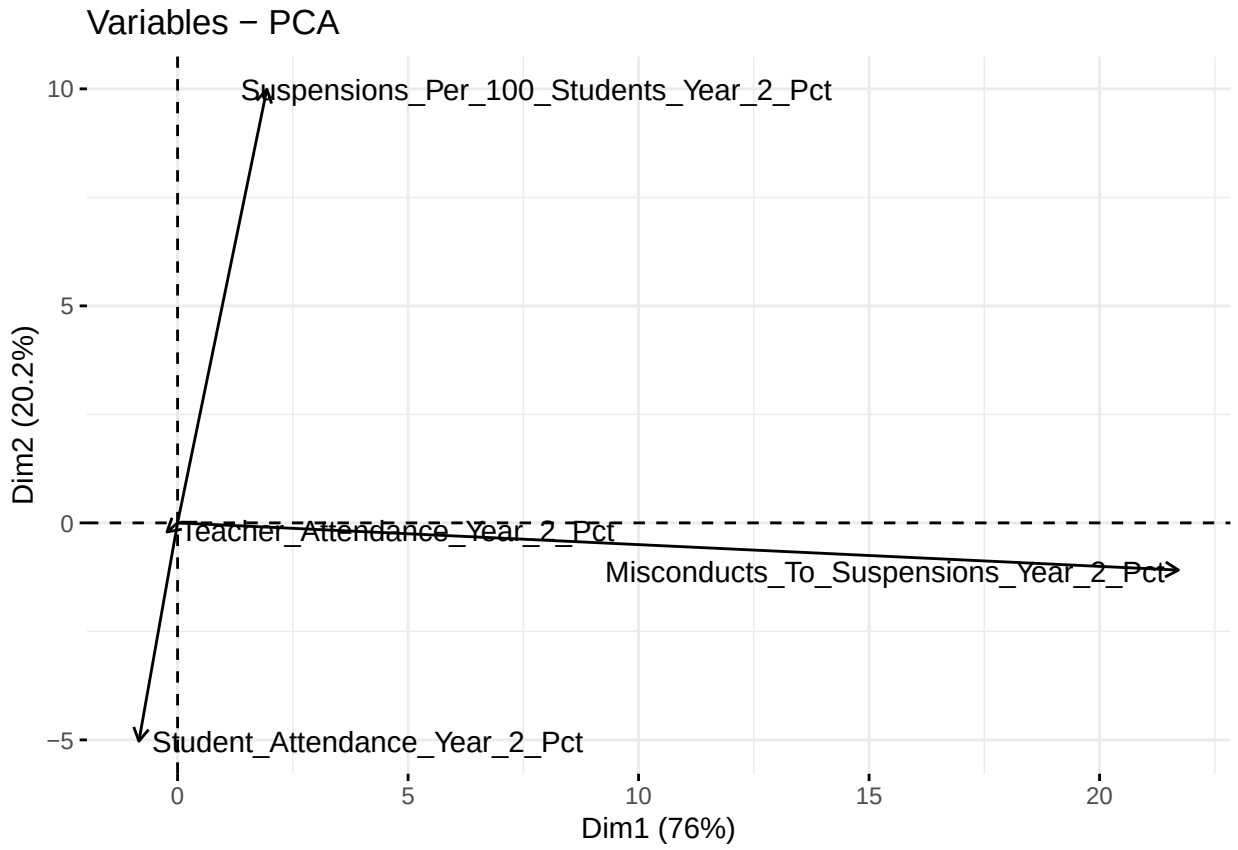
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(broom)
PCA_pre <- prcomp(pre_covid_mano)
prop_pre <- sum(PCA_pre$sdev[1:2])/sum(PCA_pre$sdev) # 86.4%
fviz_pca_var(PCA_pre,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```

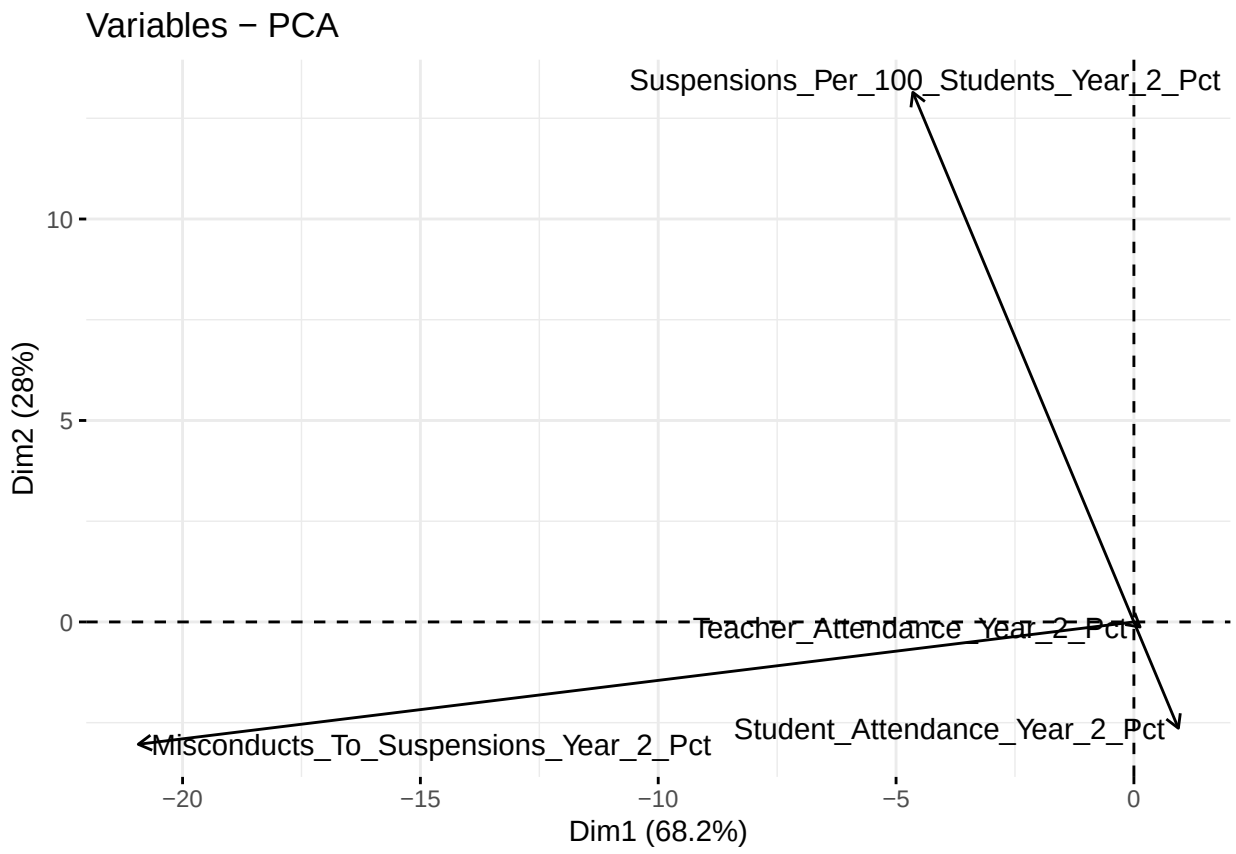



```
PCA_post <- prcomp(post_covid_mano)
prop_post <- sum(PCA_post$sdev[1:2])/sum(PCA_post$sdev) # 84.0%
fviz_pca_var(PCA_post,
             col.ind = "cos2", # Color by the quality of representation
```

```
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```

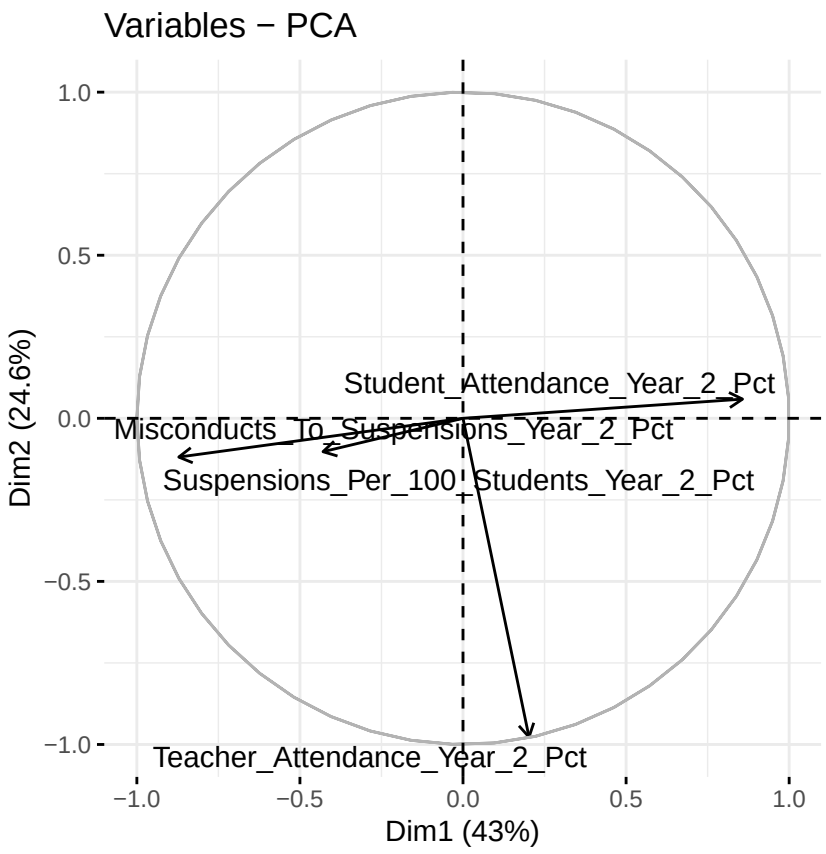



```
PCA_all <- prcomp(all_data)
prop_all <- sum(PCA_all$sdev[1:2])/sum(PCA_all$sdev) # 83.6%
fviz_pca_var(PCA_all,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```

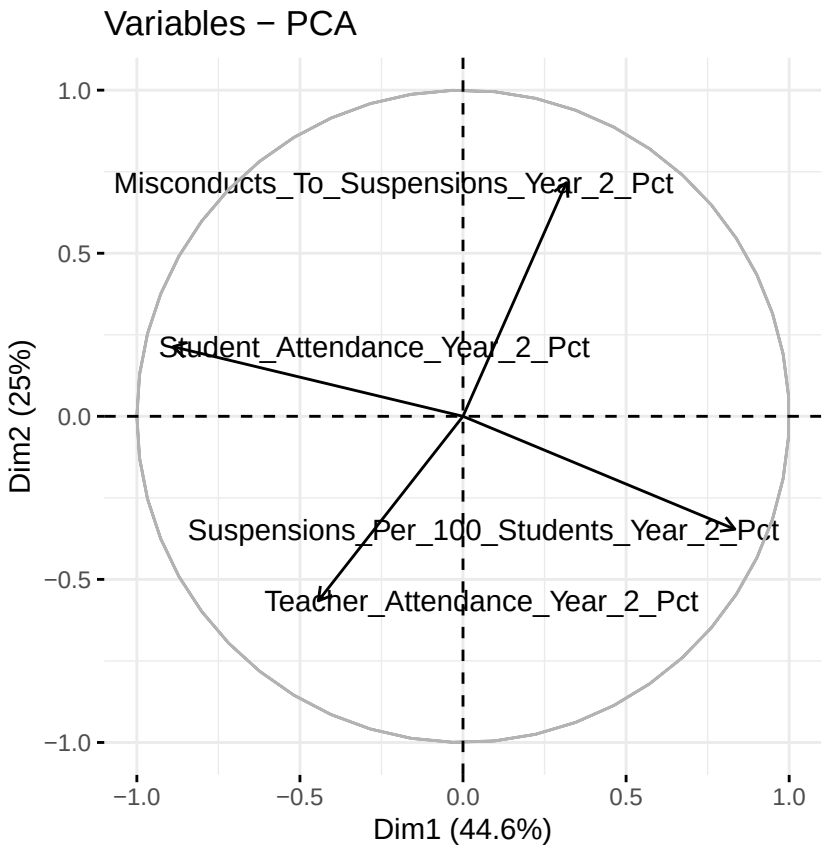


#PCA if scaled by sd

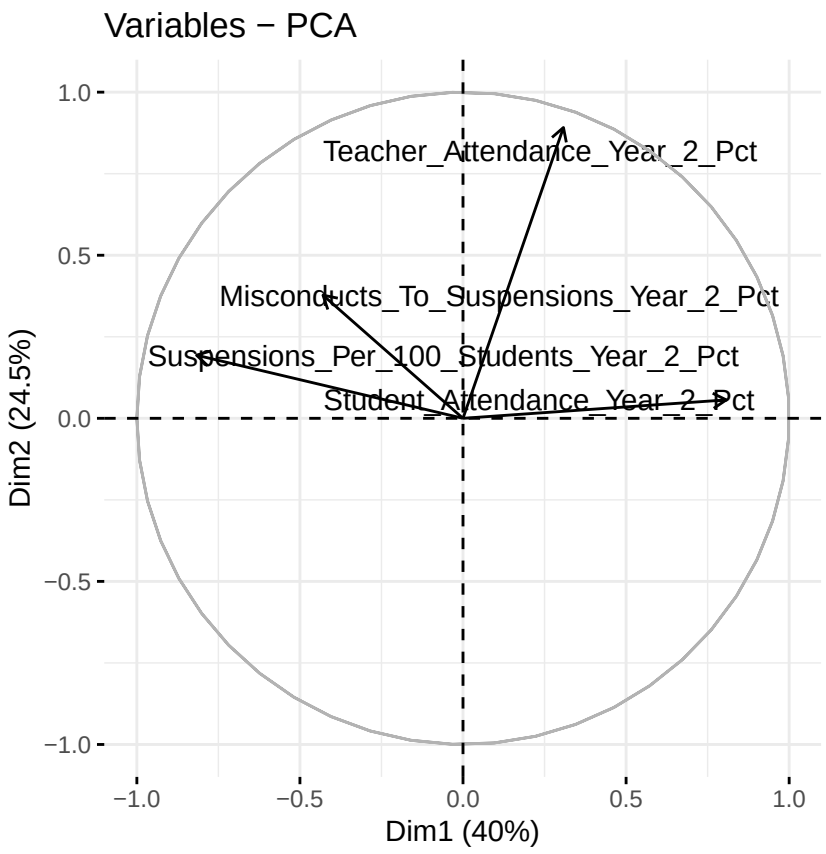
```
PCA_pre <- prcomp(pre_covid_mano, scale = sapply(pre_covid_mano,sd))
prop_pre <- sum(PCA_pre$sdev[1:3])/sum(PCA_pre$sdev) # 84.1%
fviz_pca_var(PCA_pre,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```




```
PCA_post <- prcomp(post_covid_mano, scale = sapply(post_covid_mano,sd))
prop_post <- sum(PCA_post$sdev[1:3])/sum(PCA_post$sdev) # 85.5%
fviz_pca_var(PCA_post,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```




```
PCA_all <- prcomp(all_data, scale = sapply(all_data,sd))
prop_all <- sum(PCA_all$sdev[1:3])/sum(PCA_all$sdev) # 81.4%
fviz_pca_var(PCA_all,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```




```
# need to use another PC to get the same level of varability explained when not scaled by sd
```

extra analysis on the difference between elementary school and high school

```
extra_1516 <- raw_1516 %>%
  select(Student_Attendance_Year_2_Pct,Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_
  na.omit()

extra_1617 <- raw_1617 %>%
  select(Student_Attendance_Year_2_Pct,Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_
  na.omit()

extra_1718 <- raw_1718 %>%
  select(Student_Attendance_Year_2_Pct,Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_
  na.omit()

extra_1819 <- raw_1819 %>%
  select(Student_Attendance_Year_2_Pct,Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_
  na.omit()

extra_2122 <- raw_2122 %>%
```

```
  select(Student_Attendance_Year_2_Pct,Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_
  na.omit()

extra_2223 <- raw_2223 %>%
  select(Student_Attendance_Year_2_Pct,Teacher_Attendance_Year_2_Pct,Suspensions_Per_100_Students_Year_
  na.omit()

box_pre <- rbind(extra_1516,extra_1617,extra_1718,extra_1819)
box_post <- rbind(extra_2122,extra_2223)
```

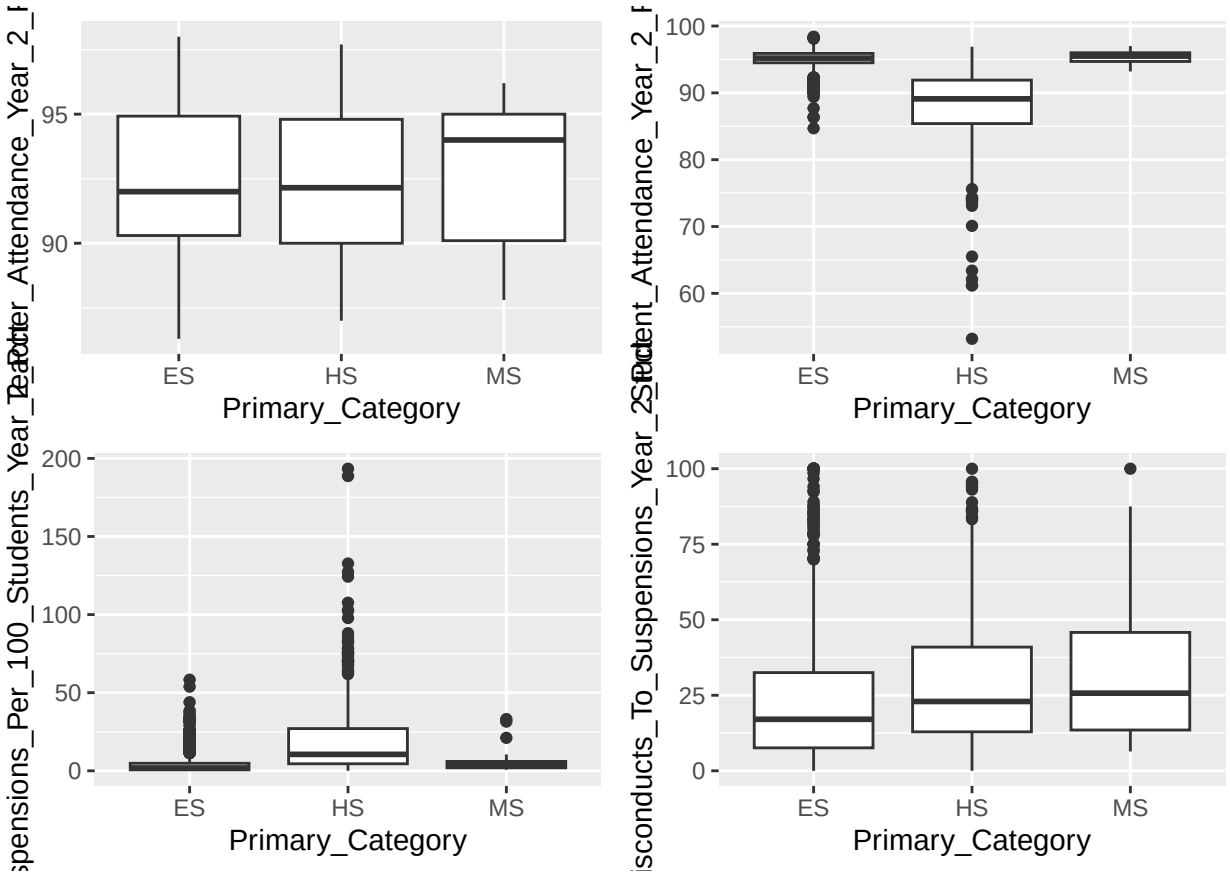
#box plot

```
library(gridExtra)
```

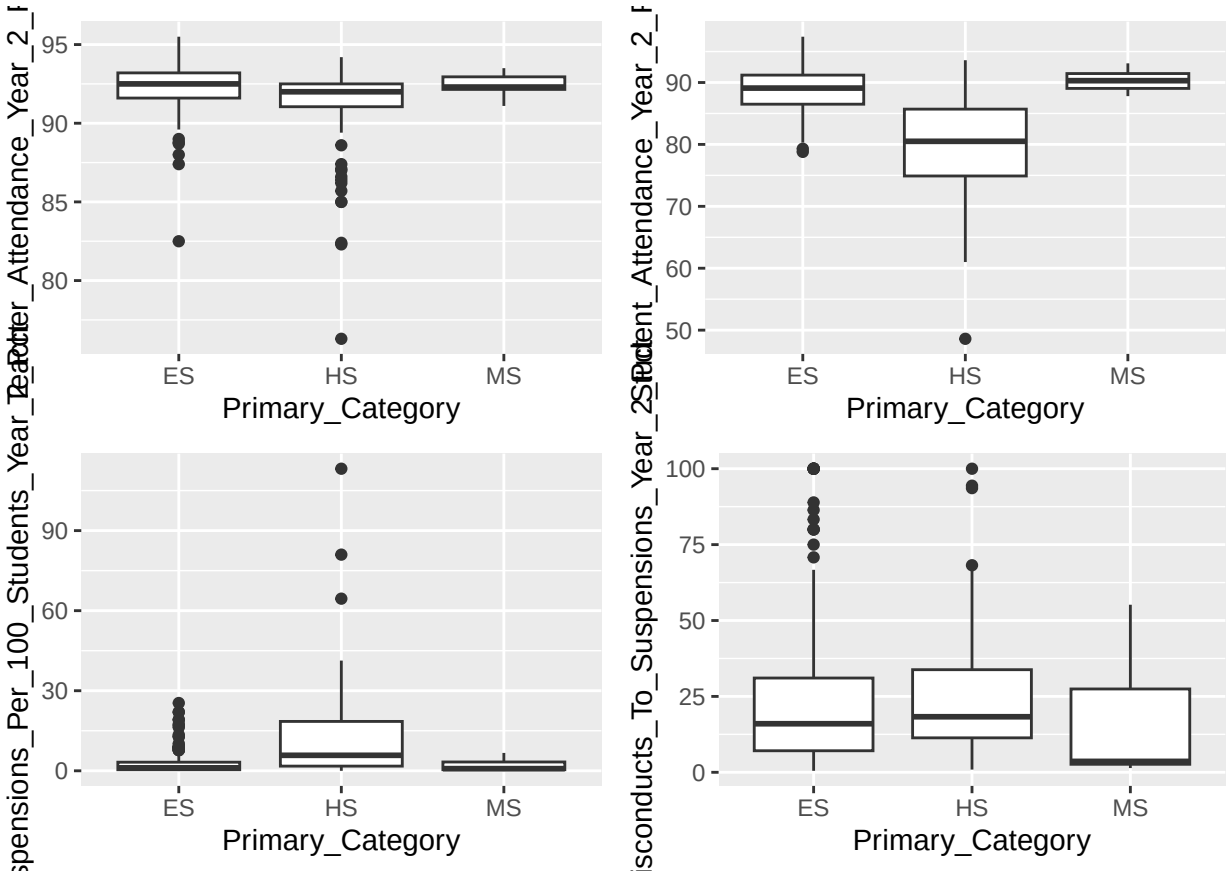
```
##
## Attaching package: 'gridExtra'

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

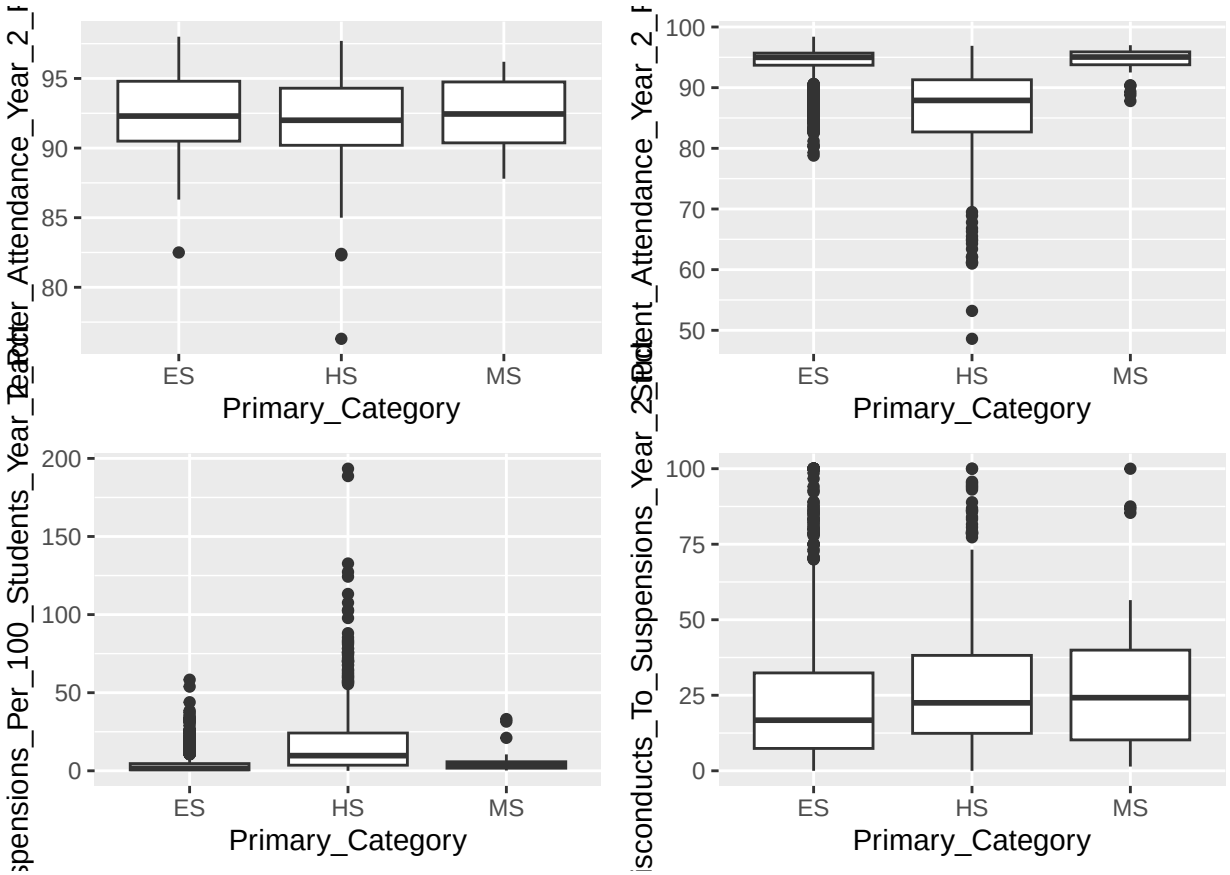
```
library(ggplot2)
#pre box plot
box_ta_pre <- ggplot(box_pre, aes(x = Primary_Category, y = Teacher_Attendance_Year_2_Pct)) +
  geom_boxplot()
box_sa_pre <- ggplot(box_pre, aes(x = Primary_Category, y = Student_Attendance_Year_2_Pct)) +
  geom_boxplot()
box_sus_pre <- ggplot(box_pre, aes(x = Primary_Category, y = Suspensions_Per_100_Students_Year_2_Pct)) +
  geom_boxplot()
box_mis_pre <- ggplot(box_pre, aes(x = Primary_Category, y = Misconducts_To_Suspensions_Year_2_Pct)) +
  geom_boxplot()
grid.arrange(box_ta_pre, box_sa_pre, box_sus_pre, box_mis_pre, ncol = 2, nrow = 2)
```




```
#post box plot
box_ta_post <- ggplot(box_post, aes(x = Primary_Category, y = Teacher_Attendance_Year_2_Pct)) +
  geom_boxplot()
box_sa_post <- ggplot(box_post, aes(x = Primary_Category, y = Student_Attendance_Year_2_Pct)) +
  geom_boxplot()
box_sus_post <- ggplot(box_post, aes(x = Primary_Category, y = Suspensions_Per_100_Students_Year_2_Pct))
  geom_boxplot()
box_mis_post <- ggplot(box_post, aes(x = Primary_Category, y = Misconducts_To_Suspensions_Year_2_Pct)) +
  geom_boxplot()
grid.arrange(box_ta_post, box_sa_post, box_sus_post, box_mis_post, ncol = 2, nrow = 2)
```




```
#all box plot
box_ta_all <- ggplot(rbind(box_pre,box_post), aes(x = Primary_Category, y = Teacher_Attendance_Year_2_P
  geom_boxplot()
box_sa_all <- ggplot(rbind(box_pre,box_post), aes(x = Primary_Category, y = Student_Attendance_Year_2_P
  geom_boxplot()
box_sus_all <- ggplot(rbind(box_pre,box_post), aes(x = Primary_Category, y = Suspensions_Per_100_Studen
  geom_boxplot()
box_mis_all <- ggplot(rbind(box_pre,box_post), aes(x = Primary_Category, y = Misconducts_To_Suspensions_
  geom_boxplot()
grid.arrange(box_ta_all, box_sa_all, box_sus_all, box_mis_all, ncol = 2, nrow = 2)
```



IGNORE

```
mano_pre_attend <- manova( cbind(Suspensions_Per_100_Students_Year_2_Pct,Misconducts_To_Suspensions_Yea
summary(mano_pre_attend)
```

```
##                                Df  Pillai approx F num Df den Df    Pr(>F)
## Teacher_Attendance_Year_2_Pct   1 0.00642     6.35      2   1967  0.001774 **
## Student_Attendance_Year_2_Pct   1 0.37614   592.98      2   1967 < 2.2e-16 ***
## Residuals                    1968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mano_post_attend <- manova( cbind(Suspensions_Per_100_Students_Year_2_Pct,Misconducts_To_Suspensions_Yea
summary(mano_post_attend)
```

```
##                                Df  Pillai approx F num Df den Df  Pr(>F)
## Teacher_Attendance_Year_2_Pct   1 0.02009     4.08      2    398 0.01761 *
## Student_Attendance_Year_2_Pct   1 0.43300   151.97      2    398 < 2e-16 ***
## Residuals                    399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##THIS IS THE ONLY MANOVA WE CARE ABOUT

```r
mano_pre_PC <- manova( cbind(Suspensions_Per_100_Students_Year_2_Pct,Misconducts_To_Suspensions_Year_2_
summary(mano_pre_PC)
```

```
##                     Df  Pillai approx F num Df den Df    Pr(>F)
## Primary_Category     2 0.50769   167.21      8   3932 < 2.2e-16 ***
## Residuals         1968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mano_post_PC <- manova( cbind(Suspensions_Per_100_Students_Year_2_Pct,Misconducts_To_Suspensions_Year_2_
summary(mano_post_PC)
```

```
##                     Df  Pillai approx F num Df den Df    Pr(>F)
## Primary_Category     2 0.41585   26.054      8    794 < 2.2e-16 ***
## Residuals         399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#pre
summary.aov(mano_pre_PC)
```

```
##  Response Suspensions_Per_100_Students_Year_2_Pct :
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## Primary_Category     2  93365   46682  282.76 < 2.2e-16 ***
## Residuals         1968 324906     165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Misconducts_To_Suspensions_Year_2_Pct :
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## Primary_Category     2  17546  8773.0  20.219 2.033e-09 ***
## Residuals         1968 853930   433.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Teacher_Attendance_Year_2_Pct :
##                     Df  Sum Sq Mean Sq F value Pr(>F)
## Primary_Category     2    10.3  5.1285  0.8069 0.4464
## Residuals         1968 12508.1  6.3558
##
##  Response Student_Attendance_Year_2_Pct :
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## Primary_Category     2  15691  7845.6   980.6 < 2.2e-16 ***
## Residuals         1968  15746     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#post
summary.aov(mano_post_PC)
```

```
##  Response Suspensions_Per_100_Students_Year_2_Pct :
```

```
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## Primary_Category   2   7769  3884.4  43.747 < 2.2e-16 ***
## Residuals        399  35428    88.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Misconducts_To_Suspensions_Year_2_Pct :
##                   Df Sum Sq Mean Sq F value Pr(>F)
## Primary_Category   2    532  266.19  0.5621 0.5705
## Residuals        399 188951  473.56
##
##  Response Teacher_Attendance_Year_2_Pct :
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## Primary_Category   2  110.43  55.216  17.142 7.213e-08 ***
## Residuals        399 1285.23   3.221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Student_Attendance_Year_2_Pct :
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## Primary_Category   2  6768.1  3384.0  133.43 < 2.2e-16 ***
## Residuals        399 10119.7    25.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```