Final Project

2024-06-03

#Data Collection This data set is downloaded from Kaggle. https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

```r
library(survival)
library(cluster)
library(ggplot2)
library(dplyr)
library(broom)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:survival':
##
##     cluster
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(stats)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
set.seed(333)

data <- read.csv("AIDS_Classification.csv")
```

Before we begin, we check if there happen to be missing values for this dataset

```
cat("Number of missing values per column:\n")
```

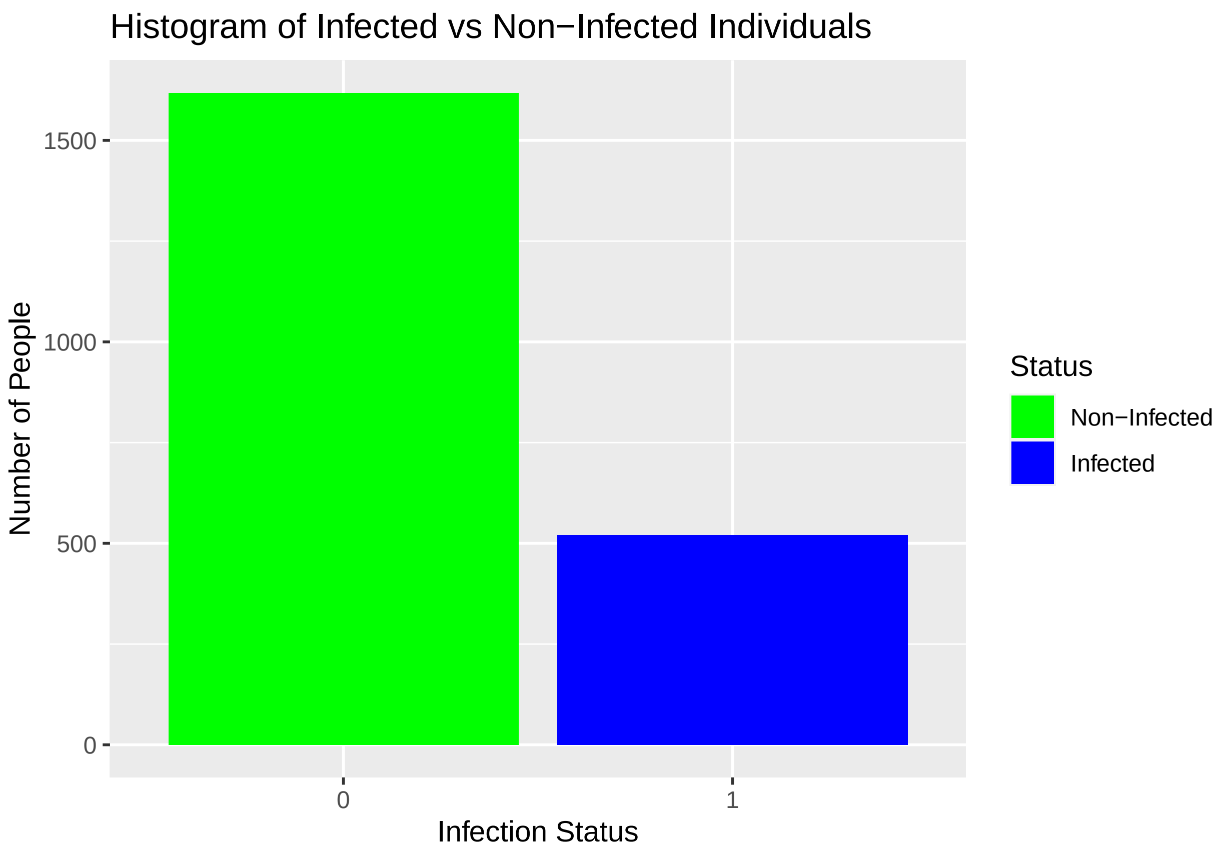
```
## Number of missing values per column:
```

```
sapply(data, function(x) sum(is.na(x)))
```

```
##     time      trt      age     wtkg     hemo     homo    drugs   karnof 
##        0        0        0        0        0        0        0        0 
##   oprior      z30  preanti     race   gender     str2    strat  symptom 
##        0        0        0        0        0        0        0        0 
##    treat   offtrt     cd40    cd420     cd80    cd820 infected 
##        0        0        0        0        0        0        0
```

#Exploratory Data Analysis There doesn't seem to be any missing data, now we can look at the frequencies of the classes to look at the distribution of the data

```
#raw number of infect and noninfected
ggplot(data, aes(x=factor(infected), fill=factor(infected))) +
  geom_bar() +
  scale_fill_manual(values=c("green", "blue"), labels=c("Non-Infected", "Infected")) +
  labs(x="Infection Status", y="Number of People", fill="Status") +
  ggtitle("Histogram of Infected vs Non-Infected Individuals")
```



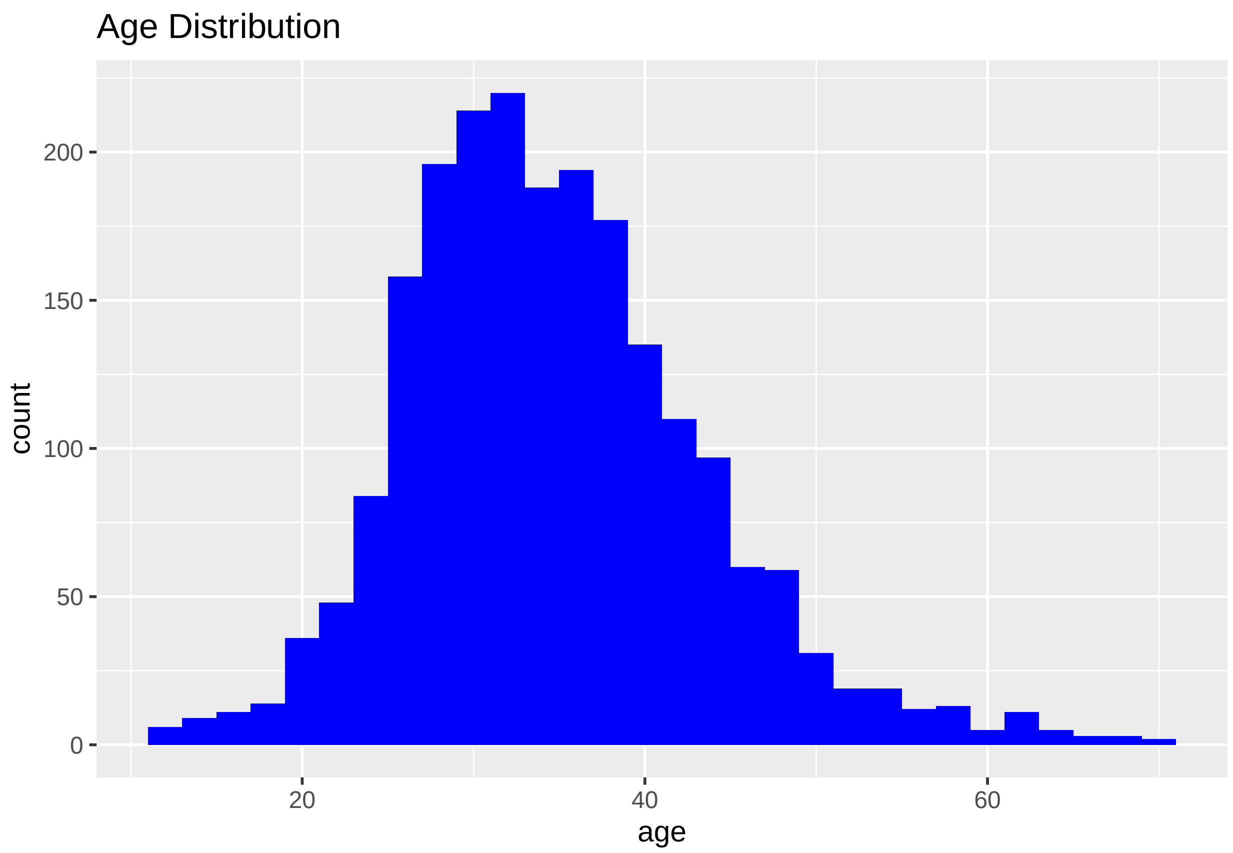
The data is not horribly imbalanced but still shows slight signs of imbalance. Checking the specific distributive percentages of the classes

```
data_percentage <- data %>%
  group_by(infected) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = Count / sum(Count) * 100)

print(data_percentage)
```

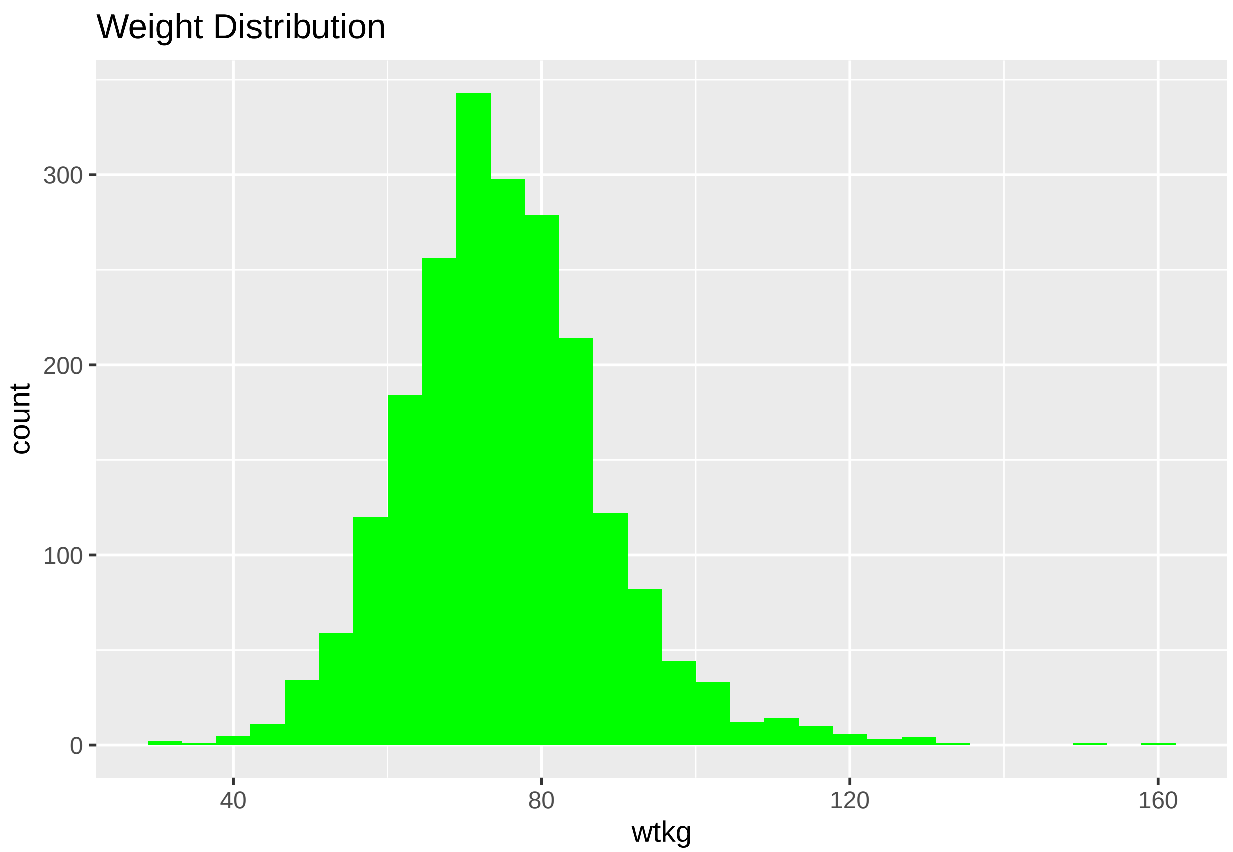
```
## # A tibble: 2 x 3
##   infected Count Percentage
##      <int> <int>      <dbl>
## 1        0  1618       75.6
## 2        1   521       24.4
```

##Visualizing key variables

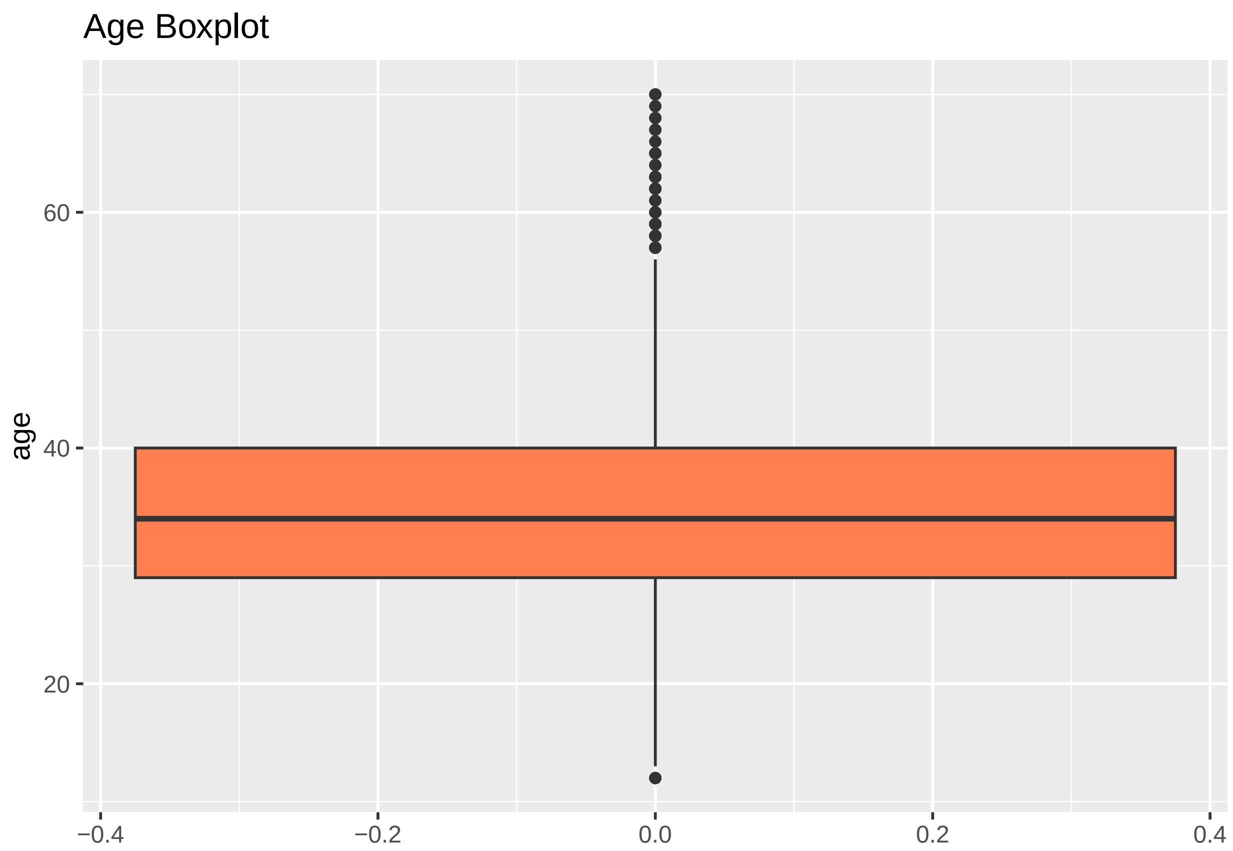
```
# visualize distributions of key variables
ggplot(data, aes(x=age)) + geom_histogram(bins=30, fill="blue") + ggtitle("Age Distribution")
```



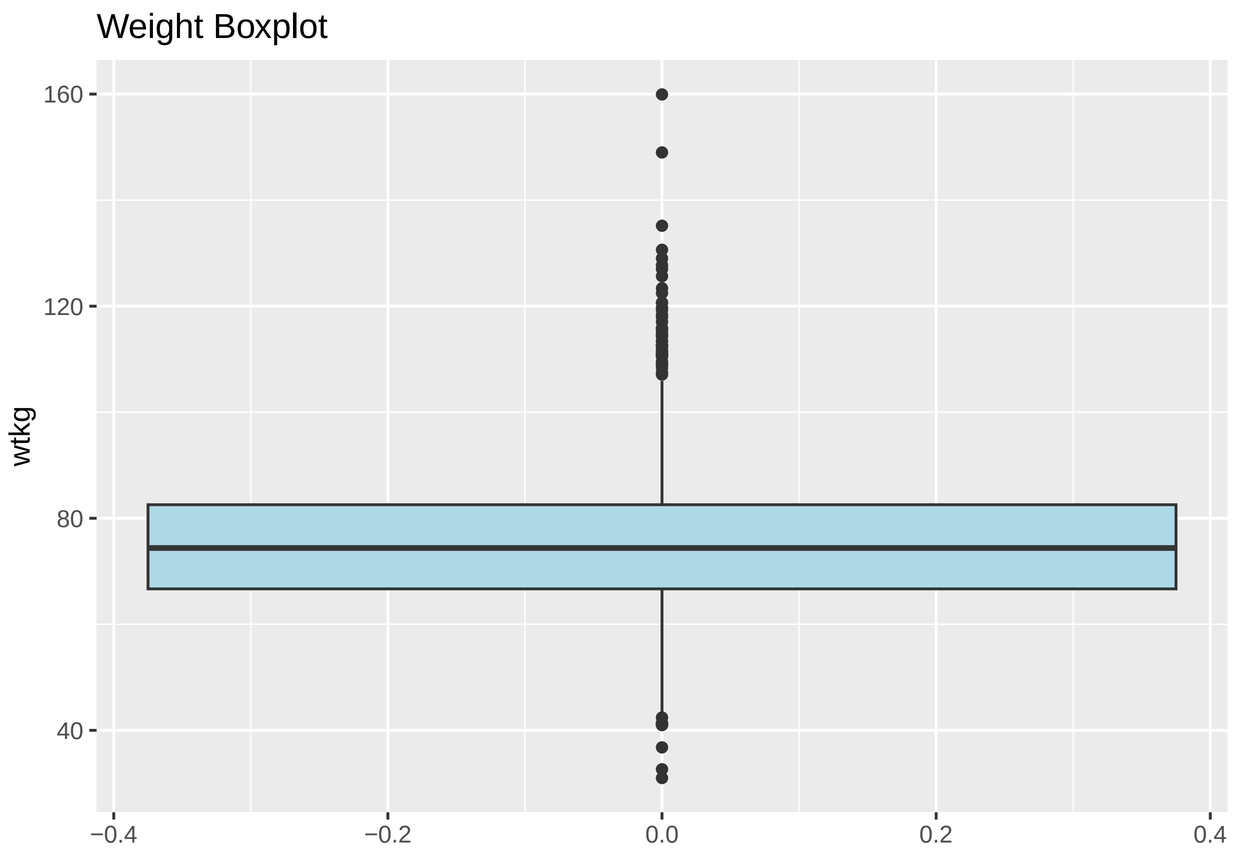
```r
ggplot(data, aes(x=wtkg)) + geom_histogram(bins=30, fill="green") + ggtitle("Weight Distribution")
```



```
# check for outliers
ggplot(data, aes(y=age)) + geom_boxplot(fill="coral") + ggtitle("Age Boxplot")
```

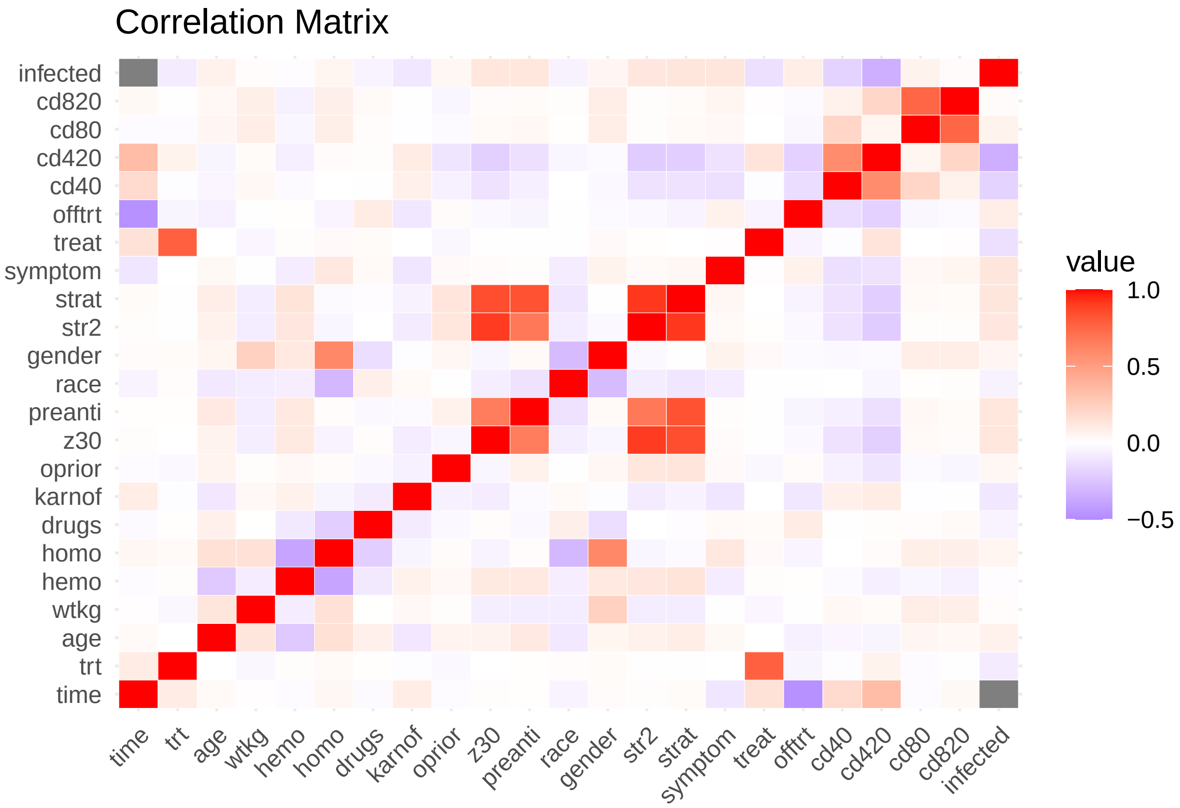


```
ggplot(data, aes(y=wtkg)) + geom_boxplot(fill="lightblue") + ggtitle("Weight Boxplot")
```



##Correlation Matrix

```
data$infected <- as.numeric(data$infected)
numerical_data <- select_if(data, is.numeric)

correlation_matrix <- cor(numerical_data)

melted_correlation_matrix <- melt(correlation_matrix)

ggplot(melted_correlation_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(limit = c(-0.5, 1), mid = "white", high = "red", low = "blue", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  labs(title = "Correlation Matrix", x = "", y = "")
```



The vast number of variables make the correlation matrix hard to analyze.

##Feature Selection Using statistical techniques to identify the most important features

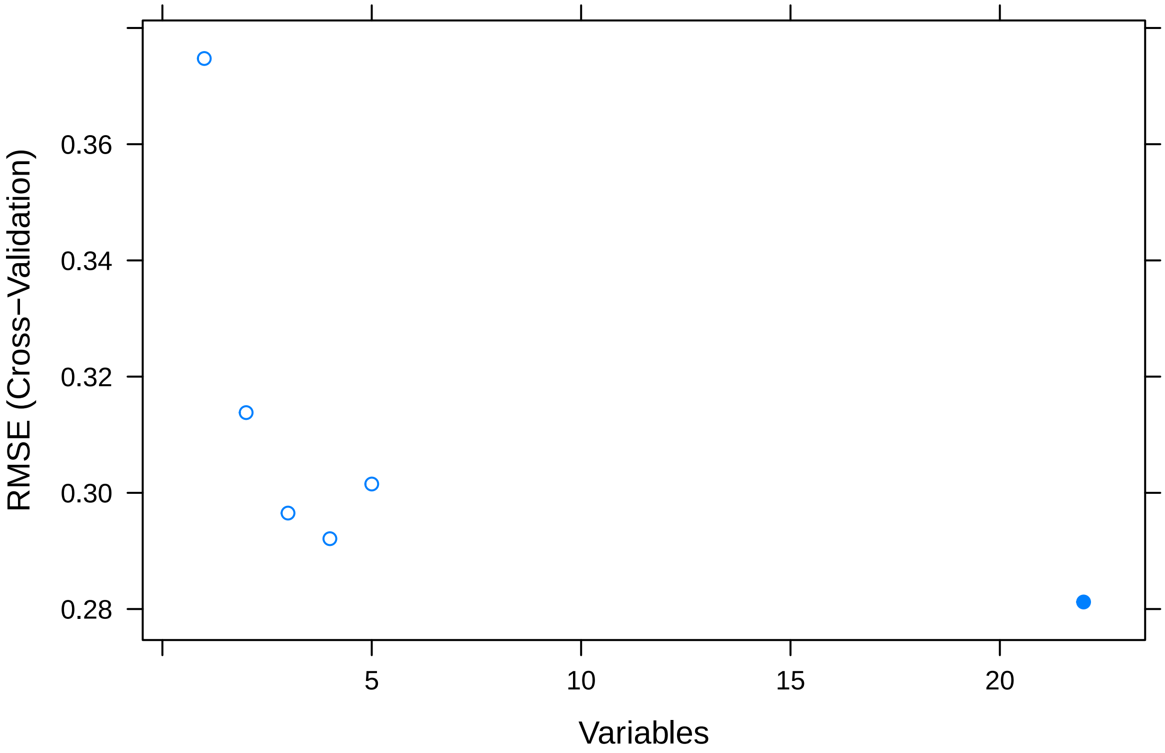
```
# using recursive feature elimination
control <- rfeControl(functions=rfFuncs, method="cv", number=10)
results <- rfe(data[, -ncol(data)], data[, ncol(data)], sizes=c(1:5), rfeControl=control)
```

```
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
## Warning in randomForest.default(x, y, importance = TRUE, ...): The response has
## five or fewer unique values.  Are you sure you want to do regression?
```

```r
print(results)
```

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold)
##
## Resampling performance over subset size:
##
##  Variables   RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
##          1 0.3748   0.2956 0.2168 0.02777    0.06483 0.02188
##          2 0.3138   0.4943 0.2313 0.02408    0.06806 0.01845
##          3 0.2965   0.5521 0.2125 0.02613    0.06667 0.01911
##          4 0.2921   0.5712 0.2100 0.02565    0.06430 0.01770
##          5 0.3015   0.5580 0.2261 0.02743    0.06109 0.02387
##         22 0.2812   0.5738 0.1755 0.02069    0.05274 0.01550        *
##
## The top 5 variables (out of 22):
##    time, offtrt, cd420, preanti, age
```

```r
# Plotting feature importance
plot(results)
```

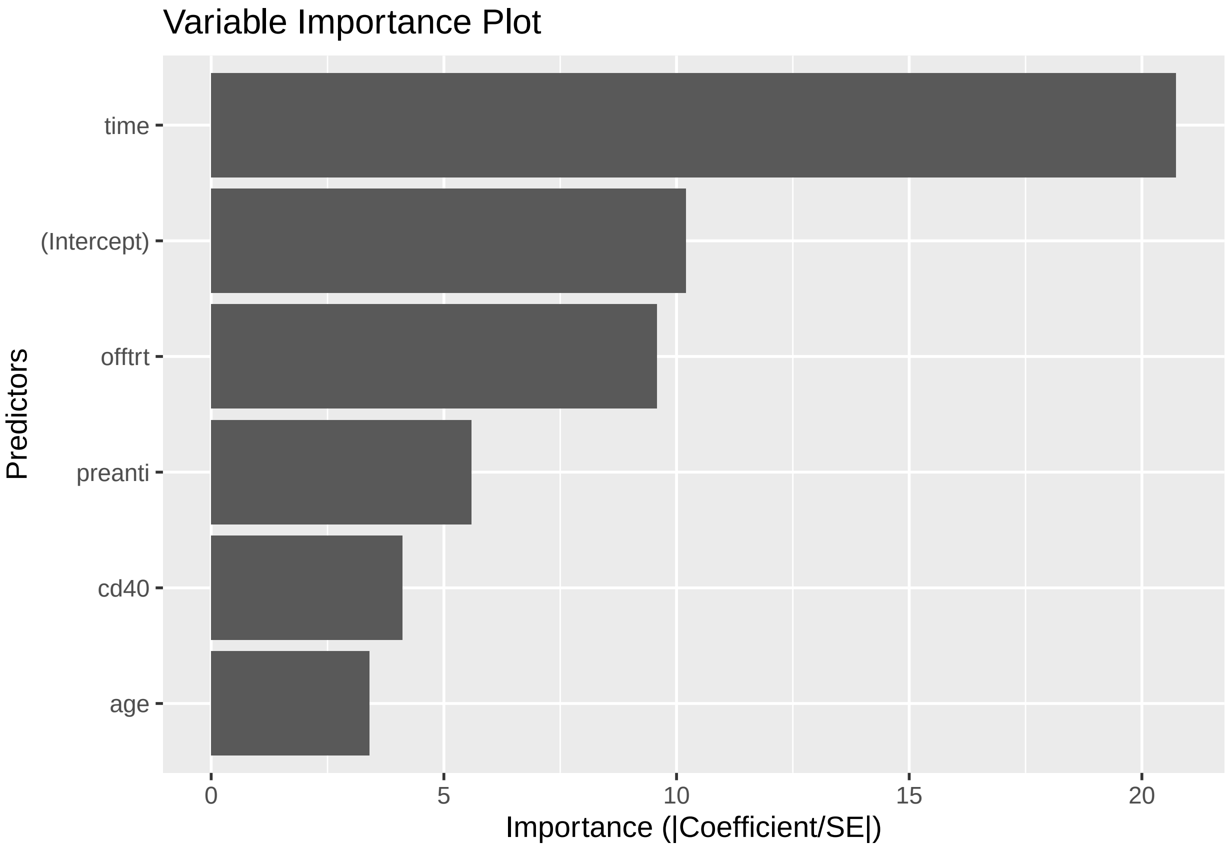


##model biulding

```r
# Logistic regression with glm
model <- glm(infected ~ age + time + cd40 + offtrt + preanti, data = data, family = binomial())
summary(model)
```

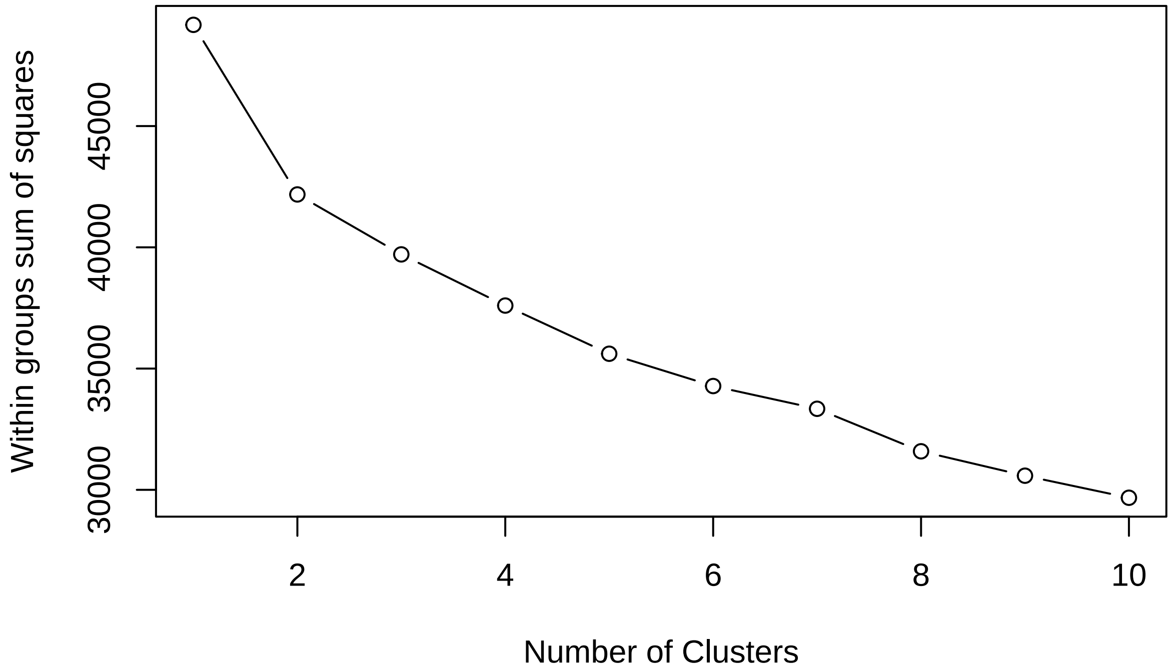
```
##
## Call:
## glm(formula = infected ~ age + time + cd40 + offtrt + preanti,
##     family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4444  -0.5025  -0.3479  -0.1373   2.8969
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.4339395  0.4345856  10.203  < 2e-16 ***
## age          0.0251593  0.0074097   3.395 0.000685 ***
## time        -0.0064937  0.0003132 -20.736  < 2e-16 ***
## cd40        -0.0024482  0.0005961  -4.107 4.01e-05 ***
## offtrt      -1.8331326  0.1914483  -9.575  < 2e-16 ***
## preanti      0.0007763  0.0001389   5.588 2.30e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2375.0  on 2138  degrees of freedom
## Residual deviance: 1507.8  on 2133  degrees of freedom
## AIC: 1519.8
##
## Number of Fisher Scoring iterations: 5
```

```r
# extracting model coefficients
tidy_model <- tidy(model)
tidy_model$importance <- abs(tidy_model$estimate / tidy_model$std.error)

# plotting variable importance
ggplot(tidy_model, aes(x = reorder(term, importance), y = importance)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Variable Importance Plot", x = "Predictors", y = "Importance (|Coefficient/SE|)")
```



##clustering Lets begin by determining the #of clusters we want to use for k

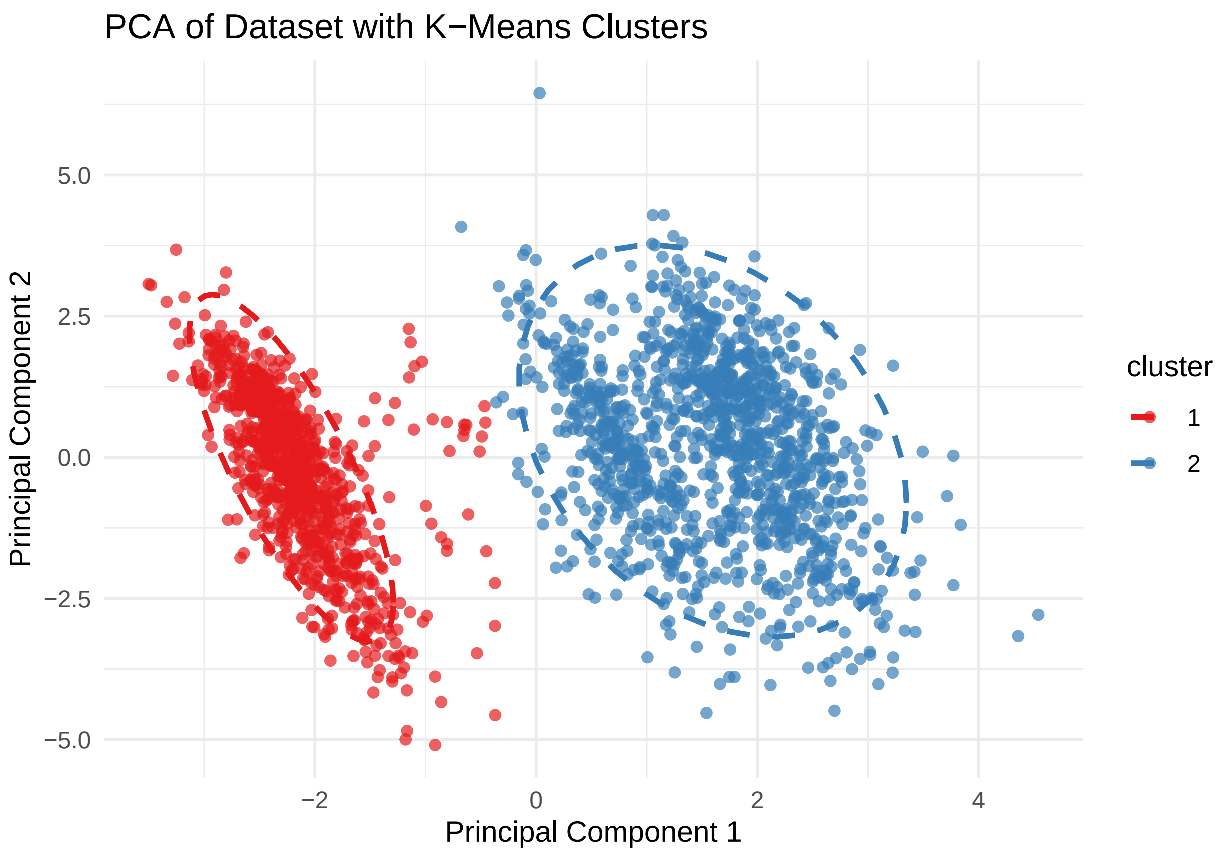
```
numeric_columns <- sapply(data, is.numeric)
data[numeric_columns] <- scale(data[numeric_columns])
new<- data[numeric_columns]
#Elbow method
wss <- sapply(1:10, function(k) sum(kmeans(data[numeric_columns], centers=k, nstart=10)$withinss))
plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
```



4 seems to be a reasonable choice

```r
# K-means clustering
kmeans_result <- kmeans(data[numeric_columns], centers=2, nstart=25)
data$cluster <- kmeans_result$cluster

# visualizing in pca
pca_results <- prcomp(data[numeric_columns])
pca_data <- data.frame(PC1 = pca_results$x[,1], PC2 = pca_results$x[,2], cluster = as.factor(kmeans_resu

# Assuming 'pca_data' contains the PCA results and cluster assignments
ggplot(pca_data, aes(x = PC1, y = PC2, color = cluster)) +
  geom_point(alpha=0.7) +
  stat_ellipse(type = "t", linetype = 2, size = 1, level = 0.95) +  # Adds ellipses
  labs(title = "PCA of Dataset with K-Means Clusters", x = "Principal Component 1", y = "Principal Compo
  scale_color_brewer(type = "qual", palette = "Set1") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



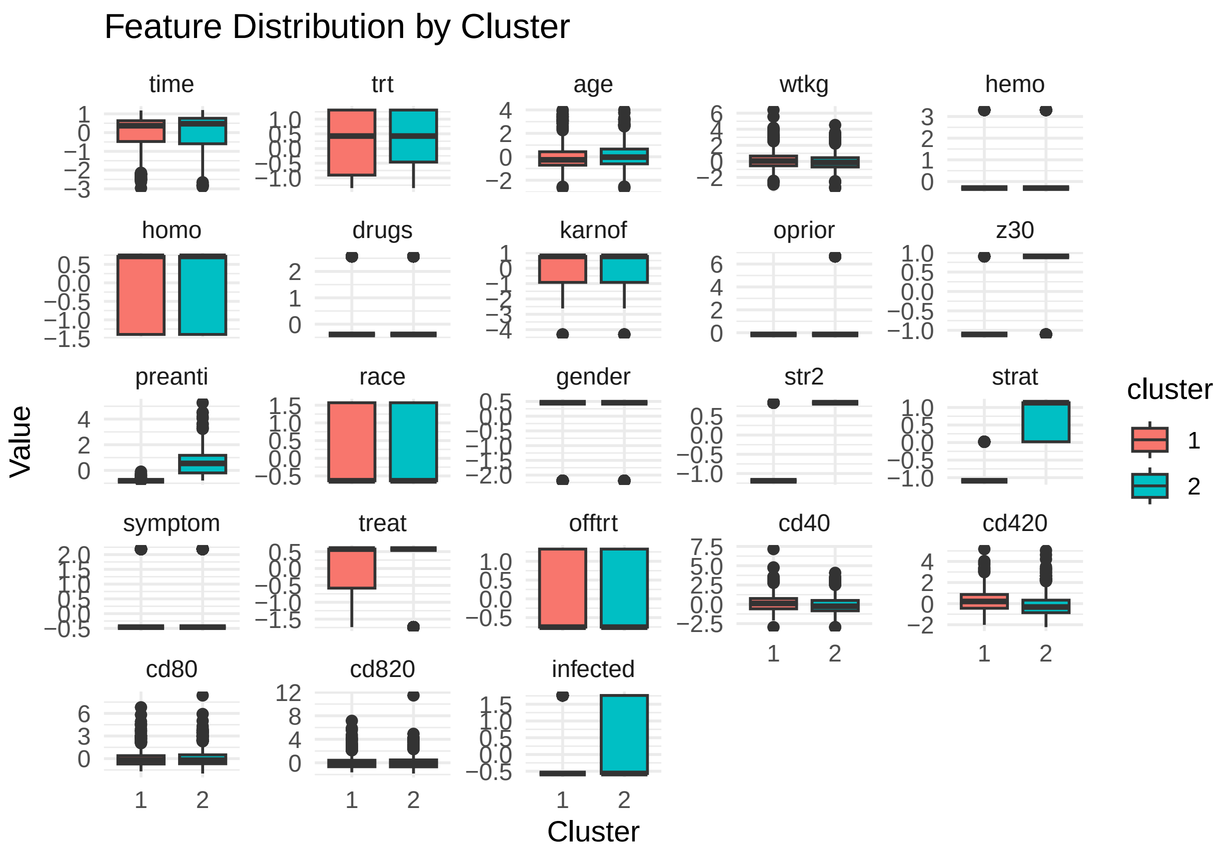
```
#looking at loadings
loadings <- pca_results$rotation[, 1:2]
print(loadings)
```

```
##                   PC1           PC2
## time     -0.058424067  4.686767e-01
## trt      -0.015856340  2.176077e-01
## age       0.058258503  2.597159e-02
## wtkg     -0.054118723  2.473253e-02
## hemo      0.088231796  7.257587e-05
## homo     -0.018046634  7.763264e-02
## drugs    -0.006741498 -3.506161e-02
## karnof   -0.063304753  1.104698e-01
## oprior    0.068884005 -5.711919e-02
## z30       0.455978279  1.098262e-01
## preanti   0.413414090  1.269287e-01
## race     -0.066075558 -7.073092e-02
## gender    0.002944844  6.266751e-02
## str2      0.473974112  1.055538e-01
## strat     0.482369514  1.223308e-01
## symptom   0.043514229 -1.229823e-01
## treat    -0.019856301  2.551445e-01
## offtrt    0.012547015 -3.278950e-01
## cd40     -0.134861770  3.221517e-01
## cd420    -0.193065703  4.081232e-01
## cd80      0.002992836  1.679659e-01
## cd820    -0.007981432  1.938717e-01
## infected  0.138410197 -3.454319e-01
## cluster   0.235441469  5.064225e-02
```

```r
# Add cluster assignments to original data
data$cluster <- as.factor(kmeans_result$cluster)
summary_stats <- aggregate(. ~ cluster, data, mean)
print(summary_stats)
```

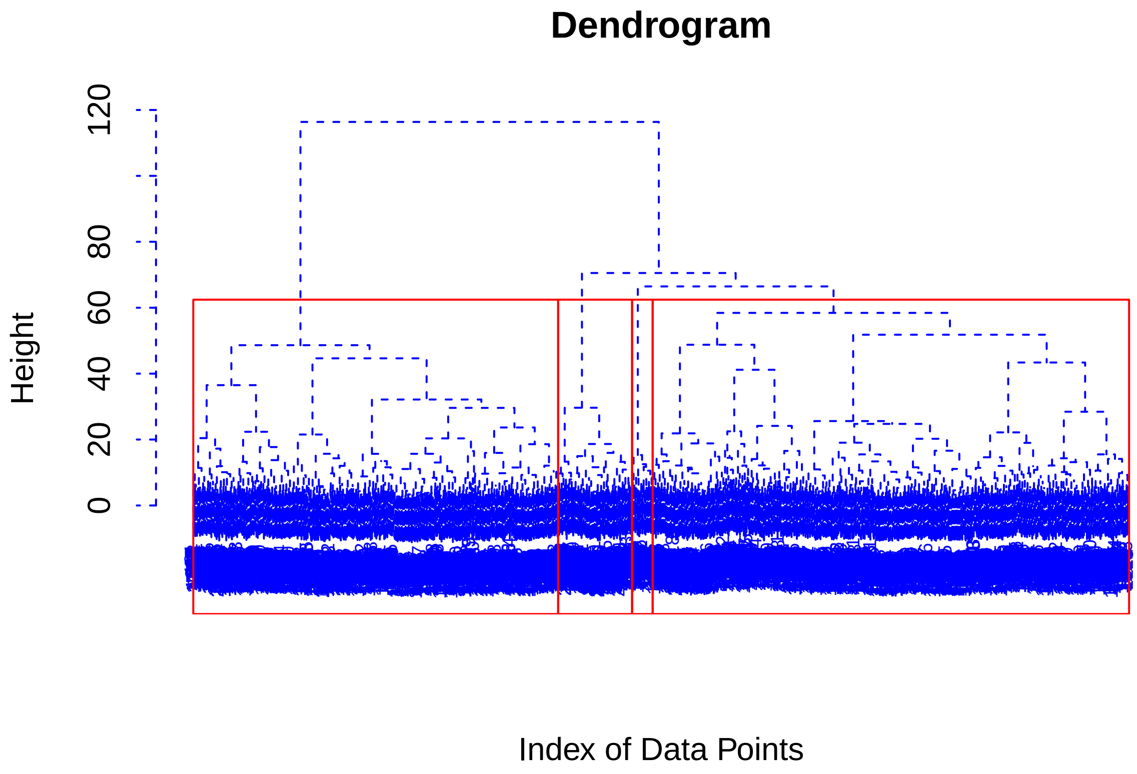
```
##   cluster         time          trt         age        wtkg       hemo
## 1       1 -0.008419867  0.004526576 -0.08813169  0.09760078 -0.1521721
## 2       2  0.006198717 -0.003332471  0.06488267 -0.07185382  0.1120293
##          homo        drugs      karnof      oprior        z30    preanti
## 1  0.04988409  0.002765750  0.1009028 -0.1498534 -1.0748366 -0.8032242
## 2 -0.03672473 -0.002036149 -0.0742848  0.1103223  0.7912961  0.5913347
##          race      gender       str2       strat     symptom        treat
## 1  0.09093190  0.03225197 -1.1419398 -1.0641691 -0.03174006 -0.003610993
## 2 -0.06694418 -0.02374394  0.8406976  0.7834427  0.02336707  0.002658418
##        offtrt       cd40      cd420        cd80       cd820   infected
## 1  0.03657292  0.1570617  0.2609956 -0.01608132 -0.01811826 -0.1538731
## 2 -0.02692503 -0.1156291 -0.1921453  0.01183909  0.01333869  0.1132816
```

```r
# Melting data for easier plotting
data_melted <- reshape2::melt(data, id.vars = "cluster")
ggplot(data_melted, aes(x = cluster, y = value, fill = cluster)) +
  geom_boxplot() +
  facet_wrap(~ variable, scales = "free_y") +
  theme_minimal() +
  labs(title = "Feature Distribution by Cluster", y = "Value", x = "Cluster")
```



The axes (PC1 and PC2) represent the principal components that account for the most variance in the data ##Hierarchical clustering

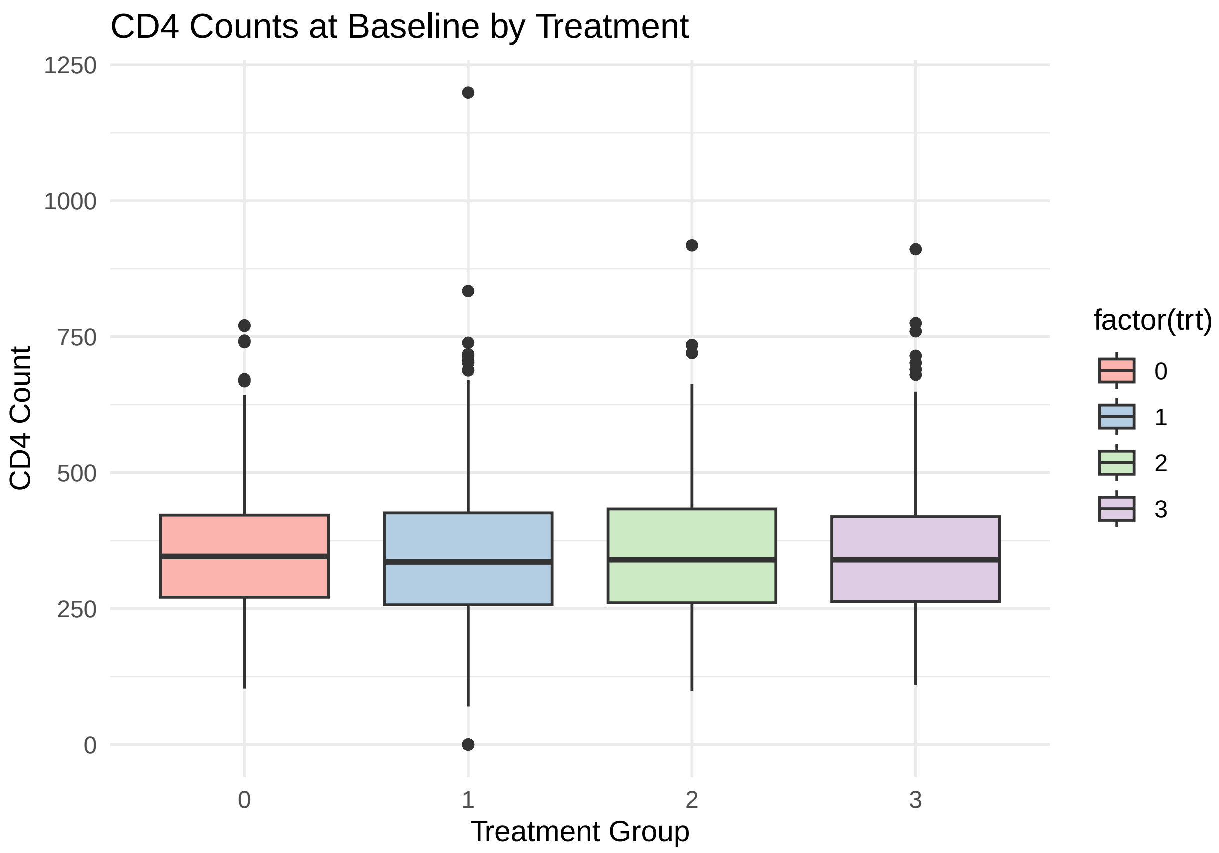
```
hc_result <- hclust(dist(data[numeric_columns]), method="ward.D2")
plot(hc_result, main = "Dendrogram", xlab = "Index of Data Points", ylab = "Height",
     lty = 2, col = "blue", sub = "", cex = 0.6)
rect.hclust(hc_result, k = 4, border = "red")
```



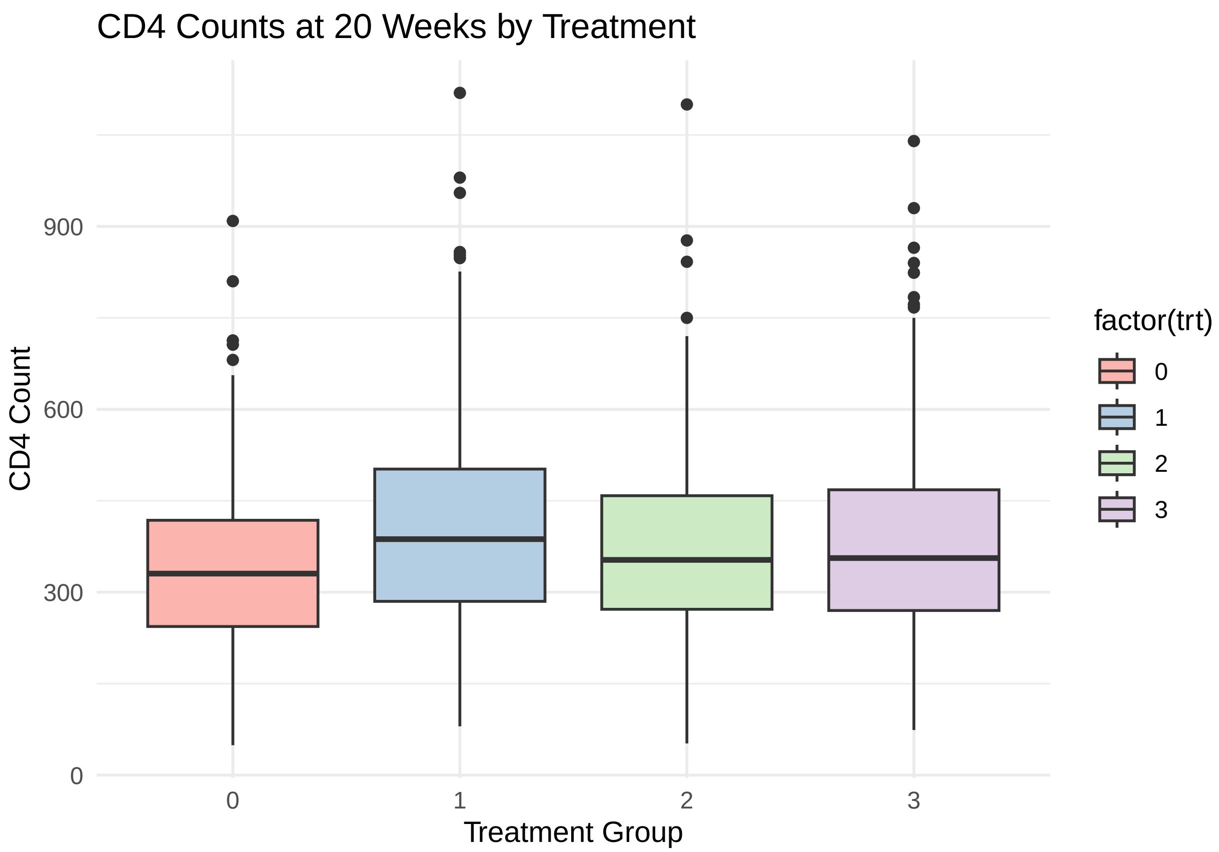
##Do AIDS patients exhibit different patterns of treatment response based on the type of treatment received? How do CD4/CD8 counts change over time for patients under different treatment regimens?

Box plots

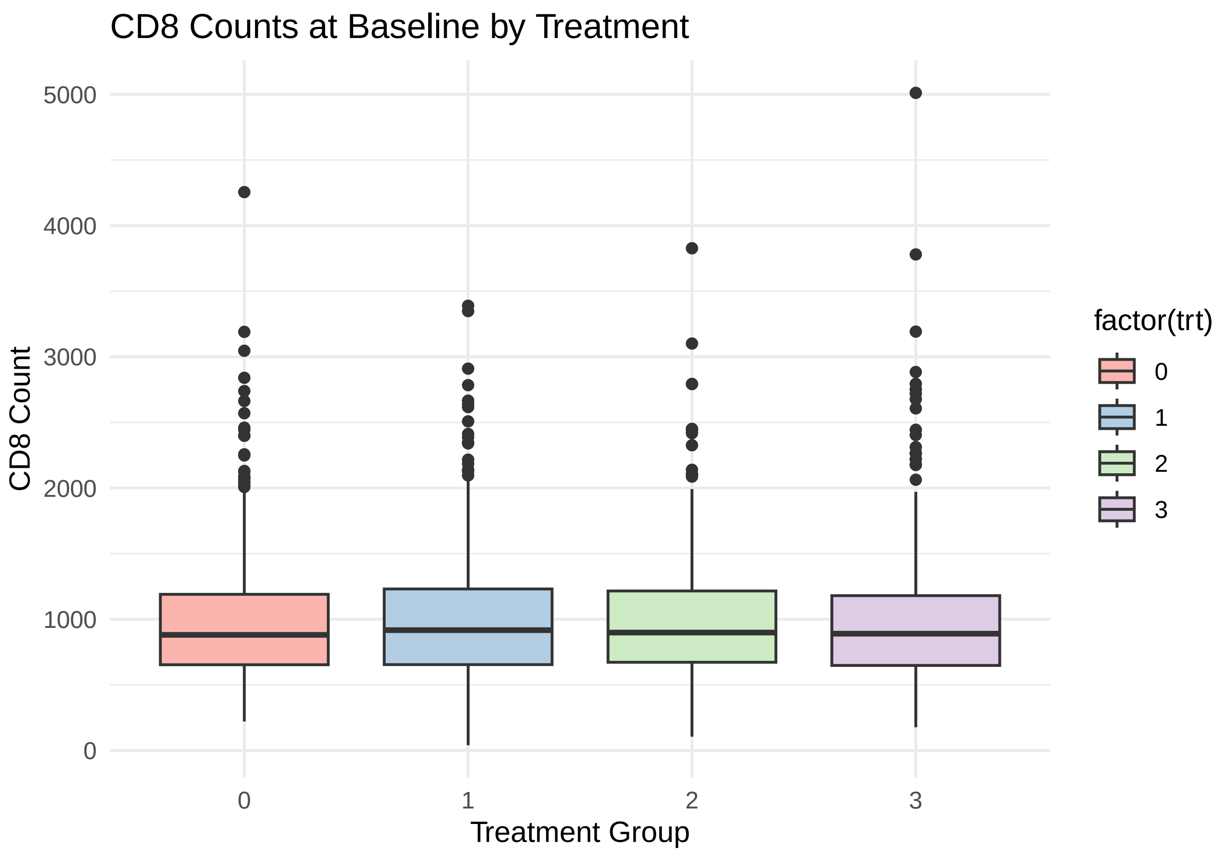
```
data <- read.csv("AIDS_Classification.csv")
data$trt <- as.factor(data$trt)

#CD4 counts at baseline
ggplot(data, aes(x = factor(trt), y = cd40, fill = factor(trt))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "CD4 Counts at Baseline by Treatment", x = "Treatment Group", y = "CD4 Count") +
  theme_minimal()
```



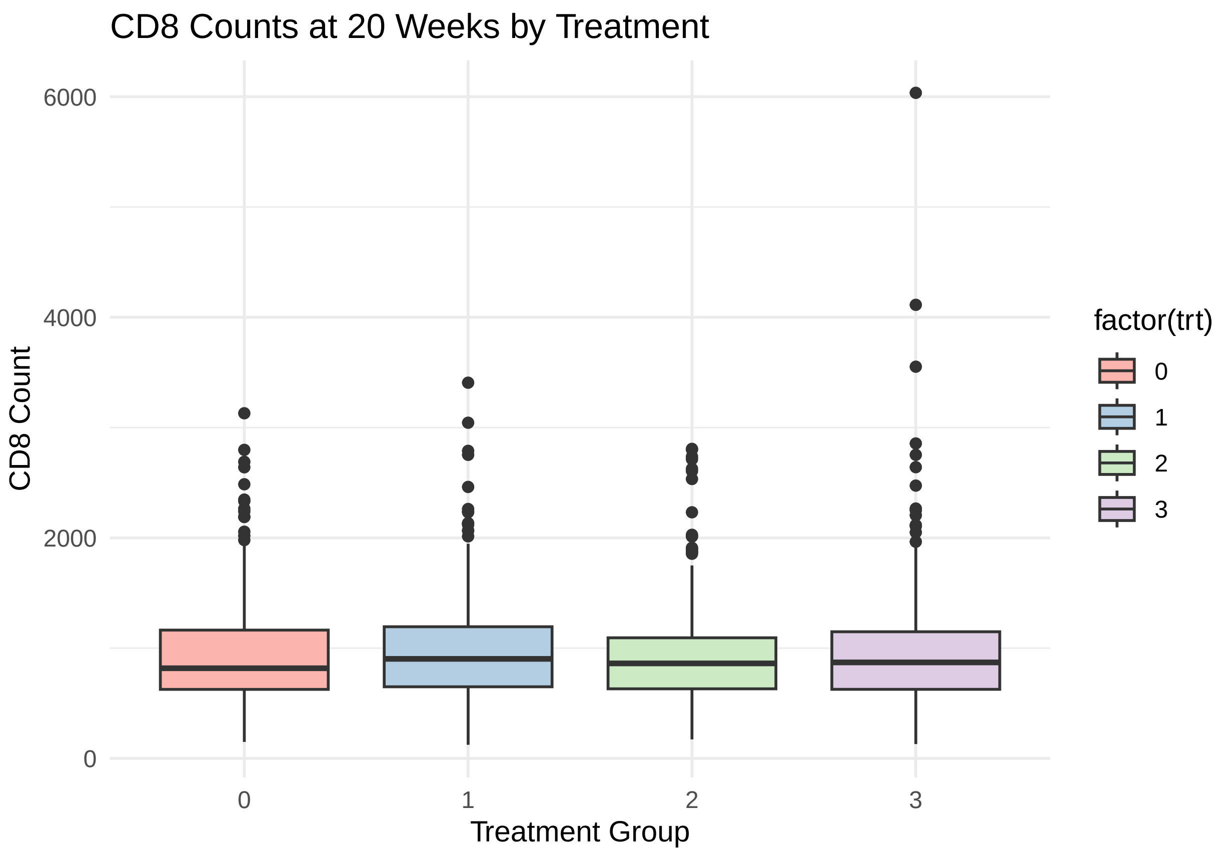
```
#D4 counts at 20 week
ggplot(data, aes(x = factor(trt), y = cd420, fill = factor(trt))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "CD4 Counts at 20 Weeks by Treatment", x = "Treatment Group", y = "CD4 Count") +
  theme_minimal()
```



```
# CD8 counts at baseline
ggplot(data, aes(x = factor(trt), y = cd80, fill = factor(trt))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "CD8 Counts at Baseline by Treatment", x = "Treatment Group", y = "CD8 Count") +
  theme_minimal()
```

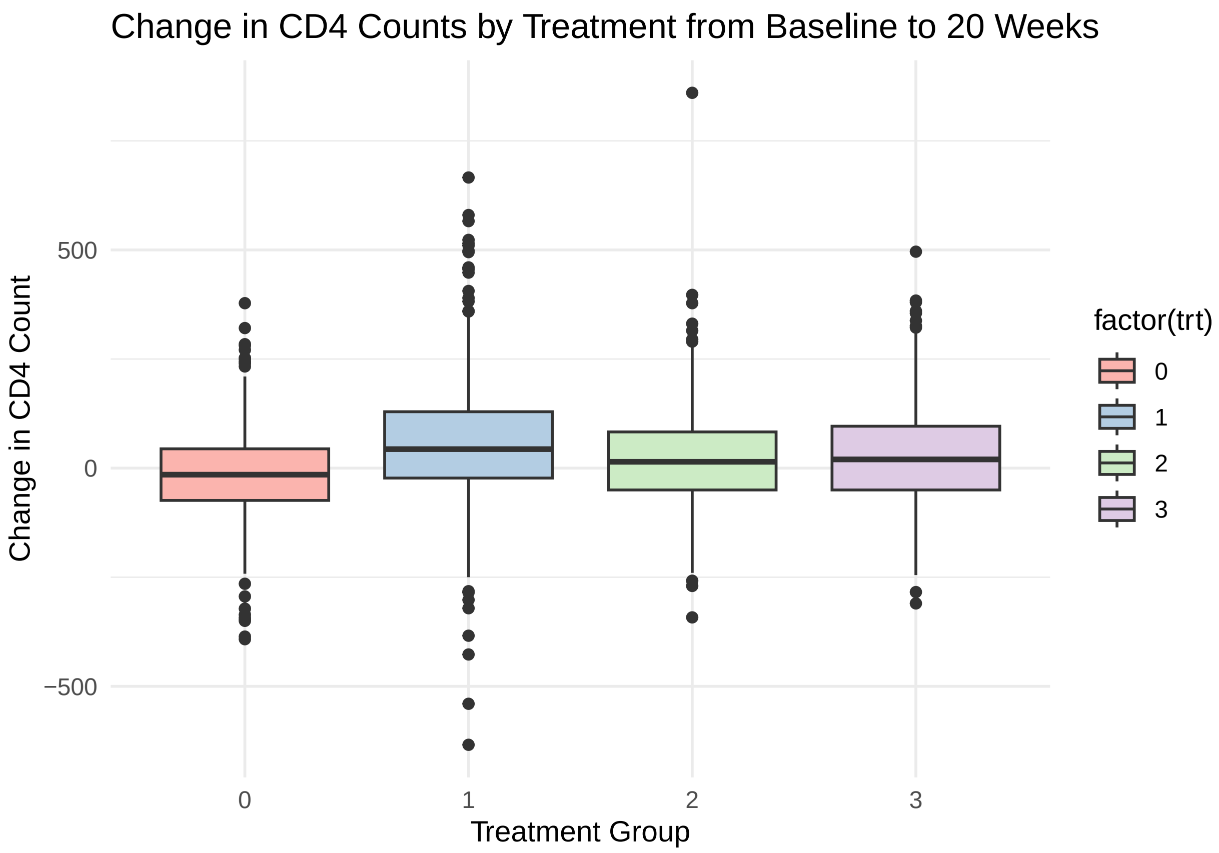


```
#CD8 counts at 20 week
ggplot(data, aes(x = factor(trt), y = cd820, fill = factor(trt))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "CD8 Counts at 20 Weeks by Treatment", x = "Treatment Group", y = "CD8 Count") +
  theme_minimal()
```

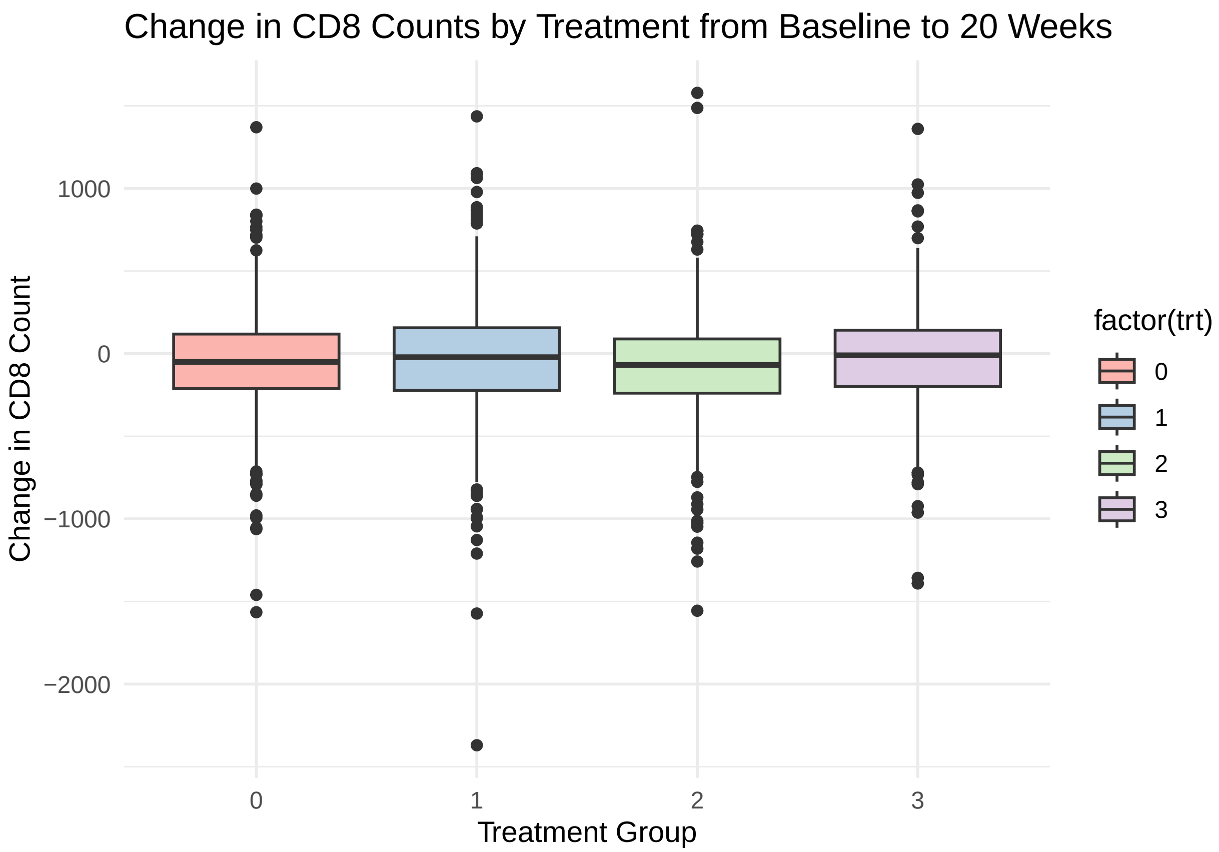


Change plots

```
data <- data %>%
  mutate(cd4_change = cd420 - cd40,
         cd8_change = cd820 - cd80)

# Change in CD4 and CD8 counts
ggplot(data, aes(x = factor(trt), y = cd4_change,fill = factor(trt))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Change in CD4 Counts by Treatment from Baseline to 20 Weeks", x = "Treatment Group", y =
  theme_minimal()
```

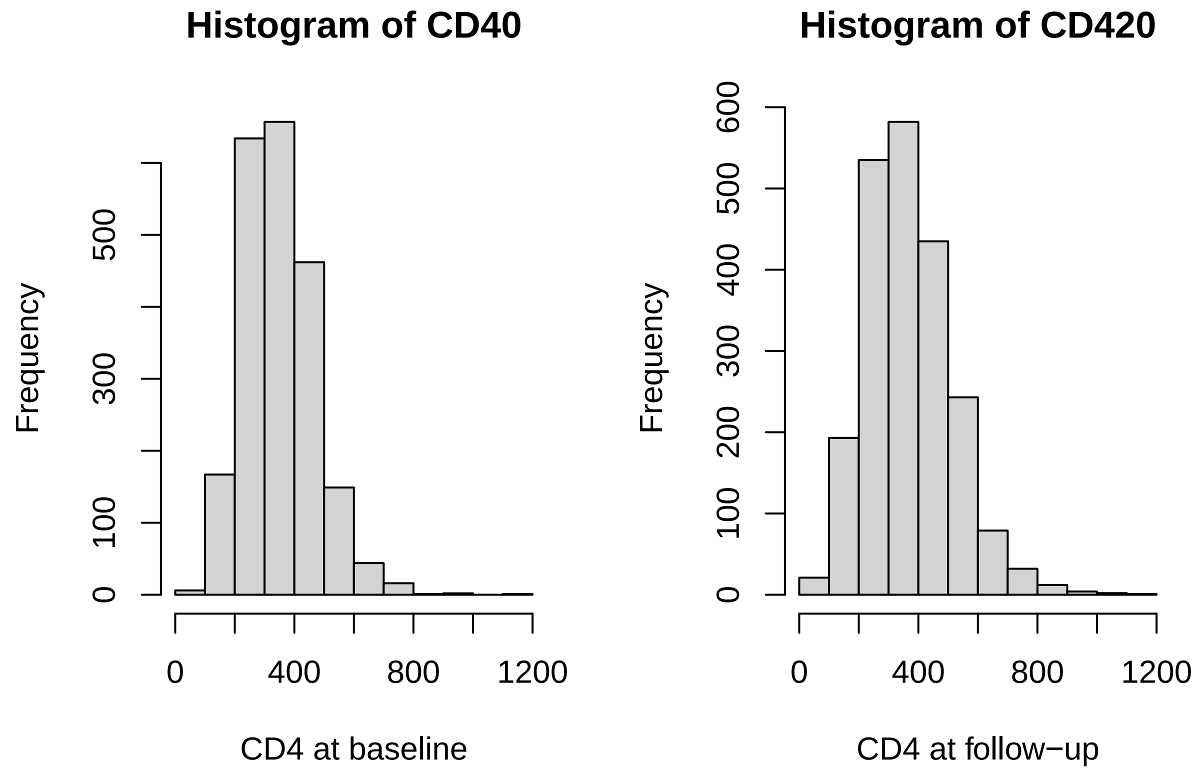


```r
ggplot(data, aes(x = factor(trt), y = cd8_change,fill = factor(trt))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Change in CD8 Counts by Treatment from Baseline to 20 Weeks", x = "Treatment Group", y =
  theme_minimal()
```

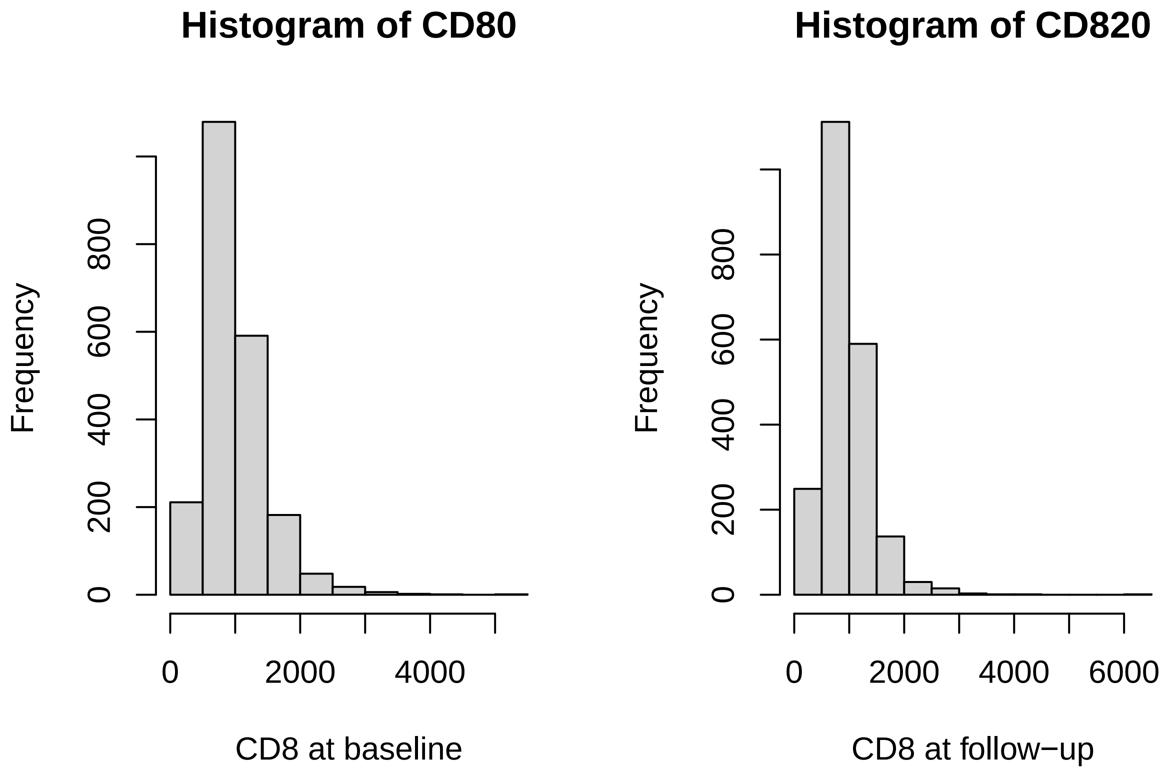

Change in CD8 Counts by Treatment from Baseline to 20 Weeks

Histogram

```
#histogram for cd4-cd4 20 weeks
par(mfrow=c(1,2))
hist(data$cd40, main="Histogram of CD40", xlab="CD4 at baseline")
hist(data$cd420, main="Histogram of CD420", xlab="CD4 at follow-up")
```



```r
par(mfrow=c(1,1))

#histogram for cd8-cd8 20 weeks
par(mfrow=c(1,2))
hist(data$cd80, main="Histogram of CD80", xlab="CD8 at baseline")
hist(data$cd820, main="Histogram of CD820", xlab="CD8 at follow-up")
```



```
par(mfrow=c(1,1))
```

ANOVA tables

```
# ANOVA for CD4 counts from baseline to follow-up
anova_cd4 <- aov(cd40 ~ cd420 + trt, data = data)
summary(anova_cd4)
```

```
##               Df   Sum Sq  Mean Sq F value   Pr(>F)
## cd420          1 10237263 10237263  1123.5  < 2e-16 ***
## trt            3   377202   125734    13.8 6.51e-09 ***
## Residuals   2134 19445303     9112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA for CD8 counts from baseline to follow-up
anova_cd8 <- aov(cd80 ~ cd820 + trt, data = data)
summary(anova_cd8)
```

```
##               Df    Sum Sq   Mean Sq  F value Pr(>F)
## cd820          1 281930714 281930714 2859.166 <2e-16 ***
## trt            3    645363    215121    2.182 0.0882 .
## Residuals   2134 210425085     98606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA to test for differences in changes in CD4 counts across treatment groups
anova_result <- aov(cd4_change ~ factor(trt), data = data)
summary(anova_result)
```

```
##                Df   Sum Sq Mean Sq F value Pr(>F)    
## factor(trt)     3  1375624  458541   31.98 <2e-16 ***
## Residuals    2135 30613936   14339                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```