Project 1

Annika Gandhi

10/20/2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(readr)
epilepsy <- read_csv("Downloads/epilepsy.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   treatment = col_character(),
##   base = col_double(),
##   age = col_double(),
##   seizure.rate = col_double(),
##   period = col_double(),
##   subject = col_double()
## )
```

```
library(readr)
seizure <- read_csv("Downloads/seizure.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   y1 = col_double(),
##   y2 = col_double(),
##   y3 = col_double(),
##   y4 = col_double(),
##   trt = col_double(),
##   base = col_double(),
##   age = col_double()
## )
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────────
──────────────────────────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.2.1          ✔ purrr   0.3.2
## ✔ tibble  2.1.3          ✔ dplyr   0.8.3
## ✔ tidyr   1.0.0.9000     ✔ stringr 1.4.0
## ✔ ggplot2 3.2.1          ✔ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
summary(epilepsy)
```

```
##        X1         treatment              base             age
##  Min.   :  1.0   Length:236         Min.   :  6.00   Min.   :18.00
##  1st Qu.: 60.5   Class :character   1st Qu.: 12.00   1st Qu.:23.00
##  Median :232.5   Mode  :character   Median : 22.00   Median :28.00
##  Mean   :251.9                      Mean   : 31.22   Mean   :28.34
##  3rd Qu.:413.2                      3rd Qu.: 41.00   3rd Qu.:32.00
##  Max.   :593.0                      Max.   :151.00   Max.   :42.00
##   seizure.rate         period        subject
##  Min.   :  0.000   Min.   :1.00   Min.   : 1
##  1st Qu.:  2.750   1st Qu.:1.75   1st Qu.:15
##  Median :  4.000   Median :2.50   Median :30
##  Mean   :  8.263   Mean   :2.50   Mean   :30
##  3rd Qu.:  9.000   3rd Qu.:3.25   3rd Qu.:45
##  Max.   :102.000   Max.   :4.00   Max.   :59
```

```
summary(seizure)
```

```
##        X1             y1               y2              y3
##  Min.   : 1.0   Min.   :  0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:15.5   1st Qu.:  2.000   1st Qu.: 3.000   1st Qu.: 2.000
##  Median :30.0   Median :  4.000   Median : 5.000   Median : 4.000
##  Mean   :30.0   Mean   :  8.949   Mean   : 8.356   Mean   : 8.441
##  3rd Qu.:44.5   3rd Qu.: 10.500   3rd Qu.:11.500   3rd Qu.: 8.000
##  Max.   :59.0   Max.   :102.000   Max.   :65.000   Max.   :76.000
##        y4              trt              base             age
##  Min.   : 0.000   Min.   :0.0000   Min.   :  6.00   Min.   :18.00
##  1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 12.00   1st Qu.:22.50
##  Median : 5.000   Median :1.0000   Median : 22.00   Median :28.00
##  Mean   : 7.339   Mean   :0.5254   Mean   : 31.24   Mean   :28.85
##  3rd Qu.: 8.000   3rd Qu.:1.0000   3rd Qu.: 41.50   3rd Qu.:33.50
##  Max.   :63.000   Max.   :1.0000   Max.   :151.00   Max.   :57.00
```

```
nrow(epilepsy)
```

```
## [1] 236
```

```
nrow(seizure)
```

```
## [1] 59
```

##The epilepsy and seizure datasets are r datasets that detail the results of a randomized clinical trial for patients with epileptic seizures. The clinical trial tested the effects that being on a anti-epileptic drug, Progabide, or a placebo had on the patients' seizure rates. There were 59 patients that participated in the trial. The dataset seizure has a row for each patient with their age, treatment type (0 or 1), an ID variable, and a count of the number of baseline seizures that occured in an 8 week period before the start of the trial.

##The epilepsy dataset contains four rows for each of the 59 patients. It seperates each patient's 8 weeks during the trial into 4 periods of two weeks. The "period" column of this dataset indicates which two week period of the trial that row accounts for and the corresponding "seizure.rate"" column shows how many seizures that individual had during each 2 week period. This dataset also had age, baseline, and identifying columns as well as a categorical treatment column.

##With these two datasets, I wanted to see how much of a difference there was in seizure rates for those taking the drug and those not taking the drug compared to their baseline value. I also wanted to see how the seizure rate differed from period 1-4 (throughout the length of the trial) and by age. This could have a relation to my future career as a physician if I am involved in clinical trials for my patients. This data is already tidy and ready to be joined.

```
library(dplyr)
epilepsy %>% left_join(seizure, by= c("subject"="X1", "age", "base")) ->seizurejoined
head(seizurejoined)
```

```
## # A tibble: 6 x 12
##      X1 treatment  base   age seizure.rate period subject    y1    y2    y3
##   <dbl> <chr>     <dbl> <dbl>        <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1     1 placebo      11    31            5      1       1     5     3     3
## 2   110 placebo      11    31            3      2       1     5     3     3
## 3   112 placebo      11    31            3      3       1     5     3     3
## 4   114 placebo      11    31            3      4       1     5     3     3
## 5     2 placebo      11    30            3      1       2     3     5     3
## 6   210 placebo      11    30            5      2       2     3     5     3
## # … with 2 more variables: y4 <dbl>, trt <dbl>
```

```
glimpse(seizurejoined)
```

```
## Observations: 236
## Variables: 12
## $ X1           <dbl> 1, 110, 112, 114, 2, 210, 212, 214, 3, 310, 312, 31…
## $ treatment    <chr> "placebo", "placebo", "placebo", "placebo", "placeb…
## $ base         <dbl> 11, 11, 11, 11, 11, 11, 11, 11, 6, 6, 6, 6, 8, 8, 8…
## $ age          <dbl> 31, 31, 31, 31, 30, 30, 30, 30, 25, 25, 25, 25, 36,…
## $ seizure.rate <dbl> 5, 3, 3, 3, 3, 5, 3, 3, 2, 4, 0, 5, 4, 4, 1, 4, 7, …
## $ period       <dbl> 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, …
## $ subject      <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, …
## $ y1           <dbl> 5, 5, 5, 5, 3, 3, 3, 3, 2, 2, 2, 2, 4, 4, 4, 4, 7, …
## $ y2           <dbl> 3, 3, 3, 3, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 18,…
## $ y3           <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 0, 0, 0, 0, 1, 1, 1, 1, 9, …
## $ y4           <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 5, 5, 5, 5, 4, 4, 4, 4, 21,…
## $ trt          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
```

#To join these to datasets, I chose to use the left join function because I wanted to ke ep all the columns from both datasets and all of the rows from the epilepsy dataset, whi ch has more observations because of the 4 periods for each subject. Performing this join took all the rows from the epilepsy data set and matched them with rows in the seizure d ata with matching values for age, base, and the identifying column X1. No observations w ere dropped.

```
seizurejoined %>% pivot_wider(names_from="treatment", values_from= "trt") ->epilepsywide
glimpse(epilepsywide)
```

```
## Observations: 236
## Variables: 12
## $ X1           <dbl> 1, 110, 112, 114, 2, 210, 212, 214, 3, 310, 312, 31…
## $ base         <dbl> 11, 11, 11, 11, 11, 11, 11, 11, 6, 6, 6, 6, 8, 8, 8…
## $ age          <dbl> 31, 31, 31, 31, 30, 30, 30, 30, 25, 25, 25, 25, 36,…
## $ seizure.rate <dbl> 5, 3, 3, 3, 3, 5, 3, 3, 2, 4, 0, 5, 4, 4, 1, 4, 7, …
## $ period       <dbl> 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, …
## $ subject      <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, …
## $ y1           <dbl> 5, 5, 5, 5, 3, 3, 3, 3, 2, 2, 2, 2, 4, 4, 4, 4, 7, …
## $ y2           <dbl> 3, 3, 3, 3, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 18,…
## $ y3           <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 0, 0, 0, 0, 1, 1, 1, 1, 9, …
## $ y4           <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 5, 5, 5, 5, 4, 4, 4, 4, 21,…
## $ placebo      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ Progabide    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
```

```
head(epilepsywide)
```

```
## # A tibble: 6 x 12
##      X1  base   age seizure.rate period subject    y1    y2    y3    y4
##   <dbl> <dbl> <dbl>        <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1    11    31            5      1       1     5     3     3     3
## 2   110    11    31            3      2       1     5     3     3     3
## 3   112    11    31            3      3       1     5     3     3     3
## 4   114    11    31            3      4       1     5     3     3     3
## 5     2    11    30            3      1       2     3     5     3     3
## 6   210    11    30            5      2       2     3     5     3     3
## # … with 2 more variables: placebo <dbl>, Progabide <dbl>
```

```
epilepsywide %>% pivot_longer(c(placebo, Progabide), names_to = "treatment", values_to =
"trt") ->epilepsyshort
na.omit(epilepsyshort)->epilepsyshort
glimpse(epilepsyshort)
```

```
## Observations: 220
## Variables: 12
## $ X1           <dbl> 1, 110, 112, 114, 2, 210, 212, 214, 3, 310, 312, 31…
## $ base         <dbl> 11, 11, 11, 11, 11, 11, 11, 11, 6, 6, 6, 6, 8, 8, 8…
## $ age          <dbl> 31, 31, 31, 31, 30, 30, 30, 30, 25, 25, 25, 25, 36,…
## $ seizure.rate <dbl> 5, 3, 3, 3, 3, 5, 3, 3, 2, 4, 0, 5, 4, 4, 1, 4, 7, …
## $ period       <dbl> 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, …
## $ subject      <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, …
## $ y1           <dbl> 5, 5, 5, 5, 3, 3, 3, 3, 2, 2, 2, 2, 4, 4, 4, 4, 7, …
## $ y2           <dbl> 3, 3, 3, 3, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 18,…
## $ y3           <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 0, 0, 0, 0, 1, 1, 1, 1, 9, …
## $ y4           <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 5, 5, 5, 5, 4, 4, 4, 4, 21,…
## $ treatment    <chr> "placebo", "placebo", "placebo", "placebo", "placeb…
## $ trt          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
```

##Although this data is already tidy, I demonstrated my knowledge of these functions by creating new columns with the treatment names and treatment numerical identifier for values. However, because half of the values had NAs in either column, I had to delete the NAs causing some rows to be deleted for individuals that did not have matching data in the epilepsy dataset.

```
seizurejoined %>% group_by(treatment) %>% summarize_if(is.numeric, mean, na.rm = T) ->df
1
head(df1)
```

```
## # A tibble: 2 x 12
##   treatment    X1  base   age seizure.rate period subject    y1    y2    y3
##   <chr>     <dbl> <dbl> <dbl>        <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 placebo    151.  30.8  29           8.60    2.5    14.5  9.36  8.29  8.79
## 2 Progabide  343.  31.6  27.7          7.96    2.5    44    8.96  8.56  8.19
## # … with 2 more variables: y4 <dbl>, trt <dbl>
```

```
## A tibble: 2 x 12
#treatment     X     base   age    seizure.rate period subject y1   y2    y3    y4   trt
#<fct>     <dbl> <dbl> <dbl>        <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 placebo    151.  30.8  29           8.60    2.5    14.5  9.36  8.29  8.79  8
 0
# 2 Progabide  343.  31.6  27.7         7.96    2.5    44    8.96  8.56  8.19  6.67
 1
```

```
seizurejoined %>% group_by(treatment) %>% summarize_if(is.numeric, sd, na.rm = T) ->df2
head(df2)
```

```
## # A tibble: 2 x 12
##   treatment    X1  base   age seizure.rate period subject    y1    y2    y3
##   <chr>     <dbl> <dbl> <dbl>        <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 placebo    118.  25.7  5.92         10.4   1.12    8.11 10.00  8.05  14.5
## 2 Progabide  190.  27.6  6.52         13.9   1.12    8.98 19.2  12.5   14.2
## # … with 2 more variables: y4 <dbl>, trt <dbl>
```

```
#### A tibble: 2 x 12
#treatment     X    base   age seizure.rate  period subject    y1    y2    y3    y4   tr
t
#<fct>     <dbl> <dbl> <dbl>        <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 placebo    118.  25.7  5.92         10.4   1.12    8.11 10.00  8.05  14.5  7.51
 0
 #2 Progabide  190.  27.6  6.52         13.9   1.12    8.98 19.2  12.5   14.2 11.8
 0
```

```
seizurejoined %>% filter(period == 1 & trt == 0) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      9.36
```

```
#  mean_rate
#1  9.357143
```

```
seizurejoined %>% filter(period == 2 & trt == 0) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      8.29
```

```
#  mean_rate
#1  8.285714

seizurejoined %>% filter(period == 3 & trt == 0) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      8.79
```

```
#mean_rate
#1  8.785714

seizurejoined %>% filter(period == 4 & trt == 0) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      7.96
```

```
#mean_rate
#1  7.964286

seizurejoined %>% filter(period == 1 & trt == 1) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      8.96
```

```
# mean_rate
#1  8.962963

seizurejoined %>% filter(period == 2 & trt == 1) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      8.56
```

```
#  mean_rate
#1  8.555556

seizurejoined %>% filter(period == 3 & trt == 1) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      8.19
```

```
#  mean_rate
#1  8.185185

seizurejoined %>% filter(period == 4 & trt == 1) %>% summarise(mean_rate = mean(seizure.
rate, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_rate
##       <dbl>
## 1      6.63
```

```
# mean_rate
#1   6.62963

seizurejoined %>% select(age) %>% arrange(age)
```

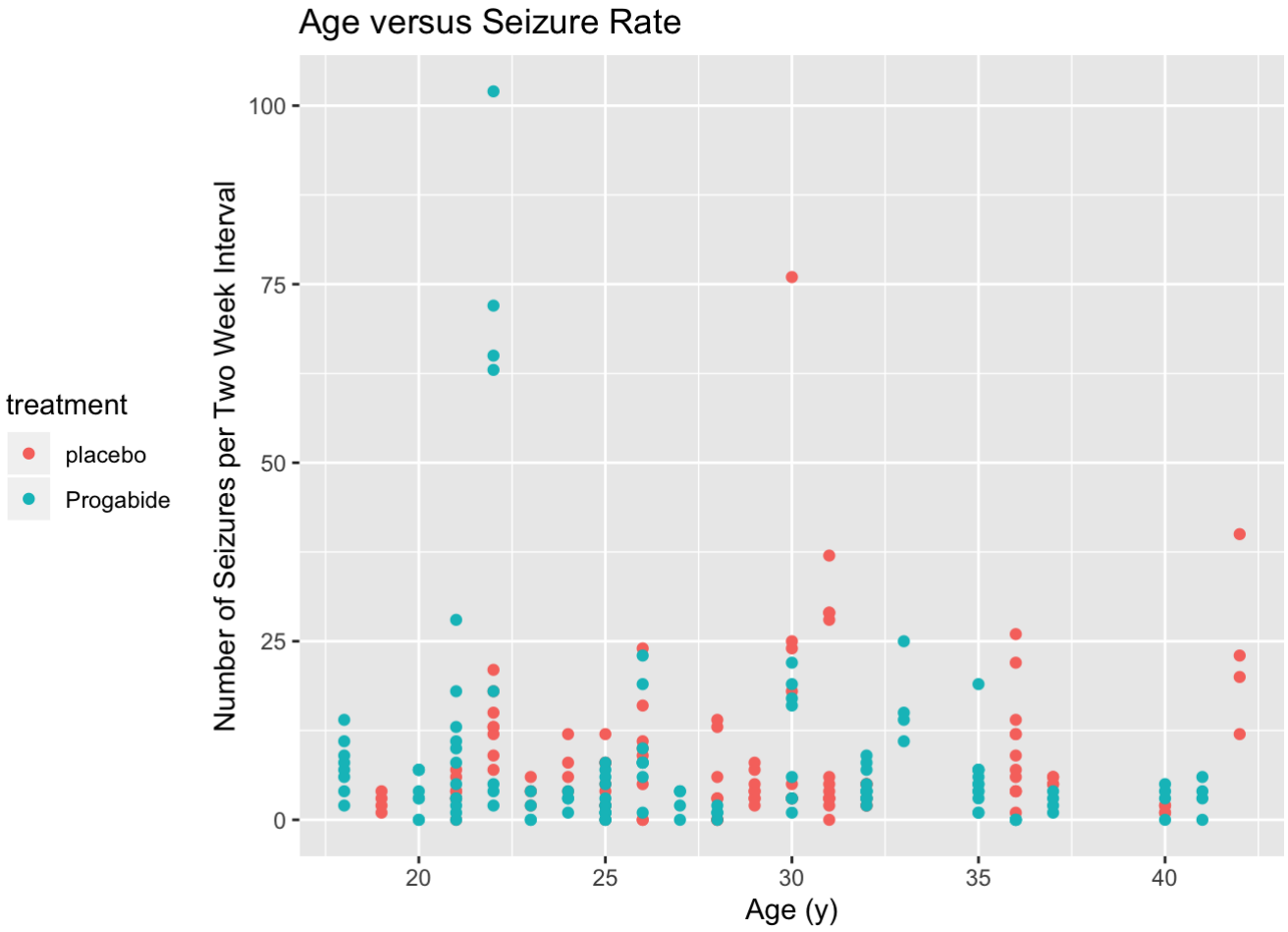
```
## # A tibble: 236 x 1
##      age
##    <dbl>
##  1    18
##  2    18
##  3    18
##  4    18
##  5    18
##  6    18
##  7    18
##  8    18
##  9    19
## 10    19
## # … with 226 more rows
```

```
seizurejoined %>% mutate(ratevsbaseline = base/4) -> seizurejoined
```

##The first two tibbles show the means and standard deviations for each of the numeric v ariables for individuals that were taking the placebo pill and individuals taking the an ti-seizure medication. What is particularly important here is the seizure rate column as that is how we can see what affect taking the drug had assuming that both populations ha d fairly similar seizure rates. Those taking the placebo had an average of only .64 more seizures, so it doesn't seem very conclusive that the anti-seizure medication was effect ive in reducing seizure rates at first glance. I then further investigated this by findi ng the mean seizure rates per period and per treatment to see whether this mean number c hanged over time for each treatment. The mean seizure rate over time from period 1 to 4 does on average go down, from 9.357143 to 7.964286 for individuals on the placebo over the course of 8 weeks and from 8.962963 to 6.62963 for individuals on Progabide over the course of 8 weeks. This could be a promising sign of some effect of the trial.

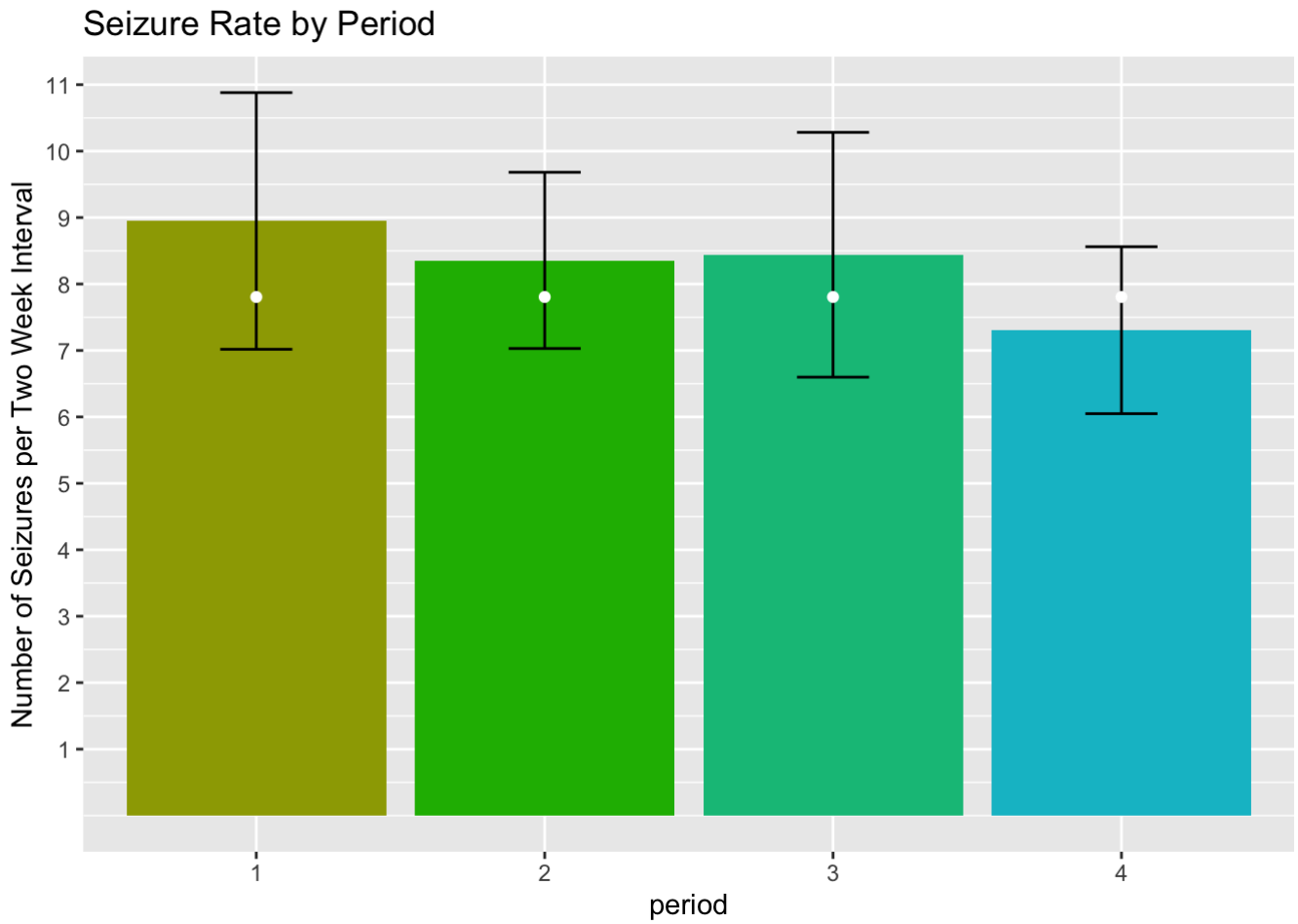
##I then selected for age and arranged ascendingly to see what the minimum and maximum a ges were easily. The minimum age for the trial was 18 and the maximum age was 42. Next, I made a new variable to make the baseline number of seizures over 8 weeks an average o f each of the two week periods during the baseline time. I did this by dividing the 8 we ek baseline value by 4. This was done so I could easily compare the baseline seizure rat e to the seizure rates during the course of the trial.

```
library(ggplot2)

ggplot(data = seizurejoined, aes(x= age, y= seizure.rate, color= treatment)) +
  geom_point() +  labs(title="Age versus Seizure Rate ", x="Age (y)", y = "Number of Sei
zures per Two Week Interval") +
  theme(legend.position = "left") + scale_fill_hue(h=c(0,90))
```



##This plot shows the number of seizures per two week period versus the age of the individual color-coded by treatment type. We can see that the number of seizures does not differ significantly in patients of different age. There does to be a slightly lower number of seizures for those in the blue that were on the anti-seizure medication, however there is not a drastic enough trend here that we can make any definitive claims on its effectiveness. We can also see that there are more high outliers for those on the drug, showing a possible negative side effect or reaction. I changed the color and moved the legend for the treatment variable to the left-hand-side of the plot.

```
seizurejoined$period <- as.character(as.numeric(seizurejoined$period))

ggplot(data=seizurejoined, aes(x=period, y=seizure.rate)) +
  geom_bar(aes(y=seizure.rate, fill = period), stat = "summary", fun.y= "mean") +
  scale_y_continuous(name = "Number of Seizures per Two Week Interval", breaks = c(1,2,3
,4,5,6,7,8,9,10,11,12)) +
  geom_errorbar(fun.data='mean_se', stat = "summary", width = .25) +
  ggtitle("Seizure Rate by Period") + scale_fill_hue(h=c(90,200)) +  theme(legend.positi
on = "none") +
  geom_point(aes(y=ratevsbaseline), stat = "summary", fun.y= "mean", color = "white")
```



##This plot shows the mean number of seizures per two week interval for all individuals. I also included an error bar that showed the amount of variability in the form of the st andard error of the mean. I also overlayed the mean value of the new baseline per two we ek variable to see how seizure rates differed over time compared to the baseline time pe riod. Unfortunately, there does not seem to be a steady decline in seizure rates over ti me during the trial, due to periods 2 and 3 being quite similar. However, there is a not icable decrease in seizure rate from period 1 to period 4.Compared to the baseline numbe r of seizures, the average is greater than baseline for weeks 1-6 and slightly lower tha n the mean baseline for weeks 6-8 of the trial. This could possibly be due to the effect s of the drug over time, but this is not conclusive. For this plot, I changed the color to greens and blues and removed the legend. I also created custom tick marks for the y axis.

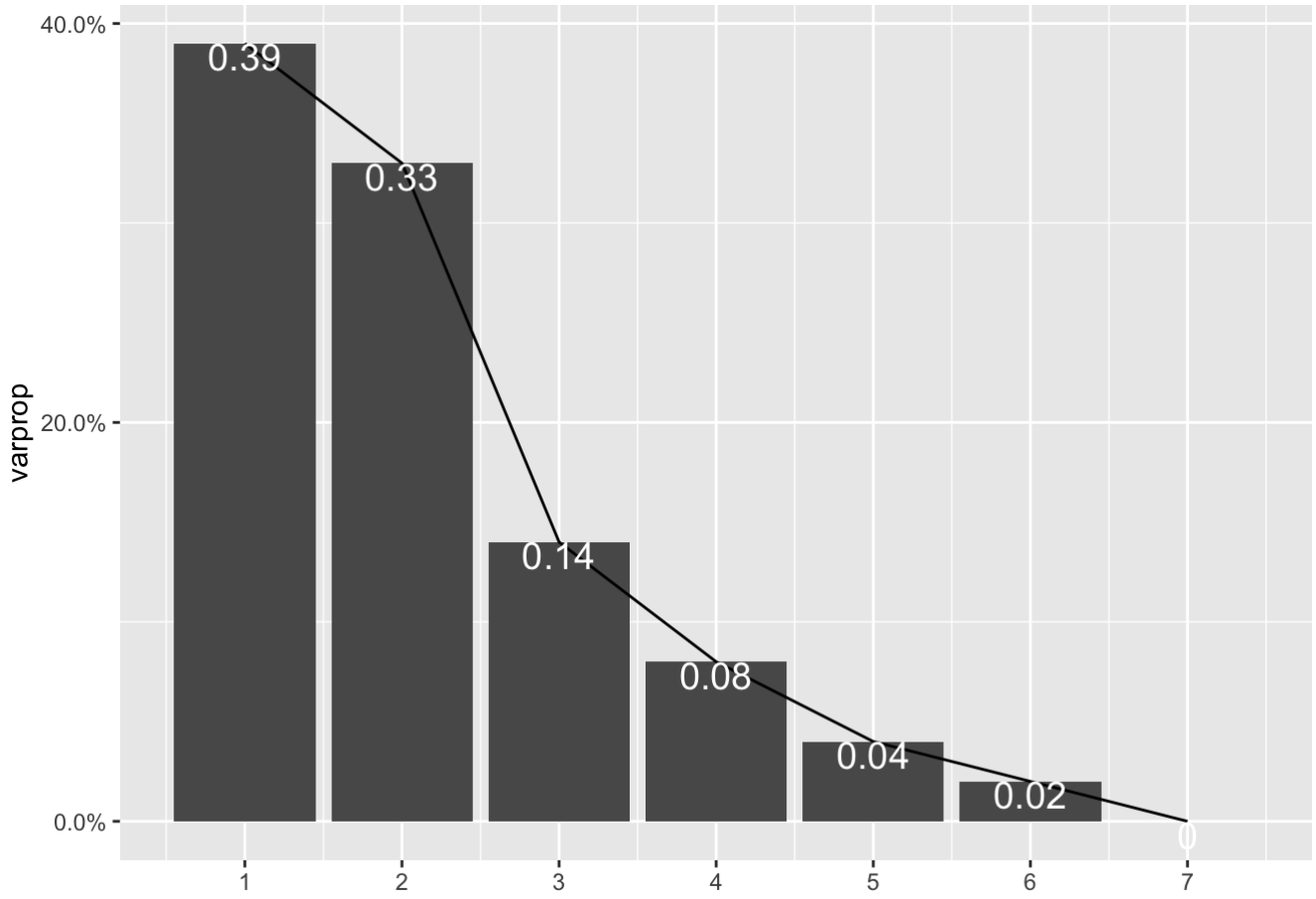
```
sj_nums<-seizurejoined%>%select(-y1,-y2,-y3,-y4,-period,-treatment)%>%scale
rownames(sj_nums)<-seizurejoined$treatment
sj_pca<-princomp(na.omit(sj_nums), cor = TRUE)
names(sj_pca)
```

```
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
## [7] "call"
```

```
summary(sj_pca, loadings=T)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
## Standard deviation     1.6537859 1.5249663 0.9836591 0.73988297 0.5466253
## Proportion of Variance 0.3907154 0.3322174 0.1382265 0.07820383 0.0426856
## Cumulative Proportion  0.3907154 0.7229328 0.8611593 0.93936313 0.9820487
##                            Comp.6 Comp.7
## Standard deviation     0.35448400      0
## Proportion of Variance 0.01795127      0
## Cumulative Proportion  1.00000000      1
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## X1                     0.502         0.858
## base            0.588                       0.384        -0.707
## age            -0.158         0.972         0.150
## seizure.rate    0.522         0.202        -0.818  0.104
## subject                0.608        -0.317        -0.715
## trt                    0.601        -0.400         0.685
## ratevsbaseline  0.588                       0.384         0.707
```

```
##Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4    Comp.5     Comp.6 C
omp.7
##Standard deviation     1.6537859 1.5249663 0.9836591 0.73988297 0.5466253 0.35448400
0
##Proportion of Variance 0.3907154 0.3322174 0.1382265 0.07820383 0.0426856 0.01795127
0
##Cumulative Proportion  0.3907154 0.7229328 0.8611593 0.93936313 0.9820487 1.00000000
1

eigval<-sj_pca$sdev^2
varprop=round(eigval/sum(eigval),2)
ggplot()+geom_bar(aes(y=varprop,x=1:7),stat="identity")+xlab("")+geom_path(aes(y=varpro
p,x=1:7))+
  geom_text(aes(x=1:7,y=varprop,label=round(varprop,2)),vjust=1,col="white",size=5)+
  scale_y_continuous(breaks=seq(0,.6,.2),labels = scales::percent)+
  scale_x_continuous(breaks=1:10)
```




```
round(cumsum(eigval)/sum(eigval),2)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
##   0.39   0.72   0.86   0.94   0.98   1.00   1.00
```

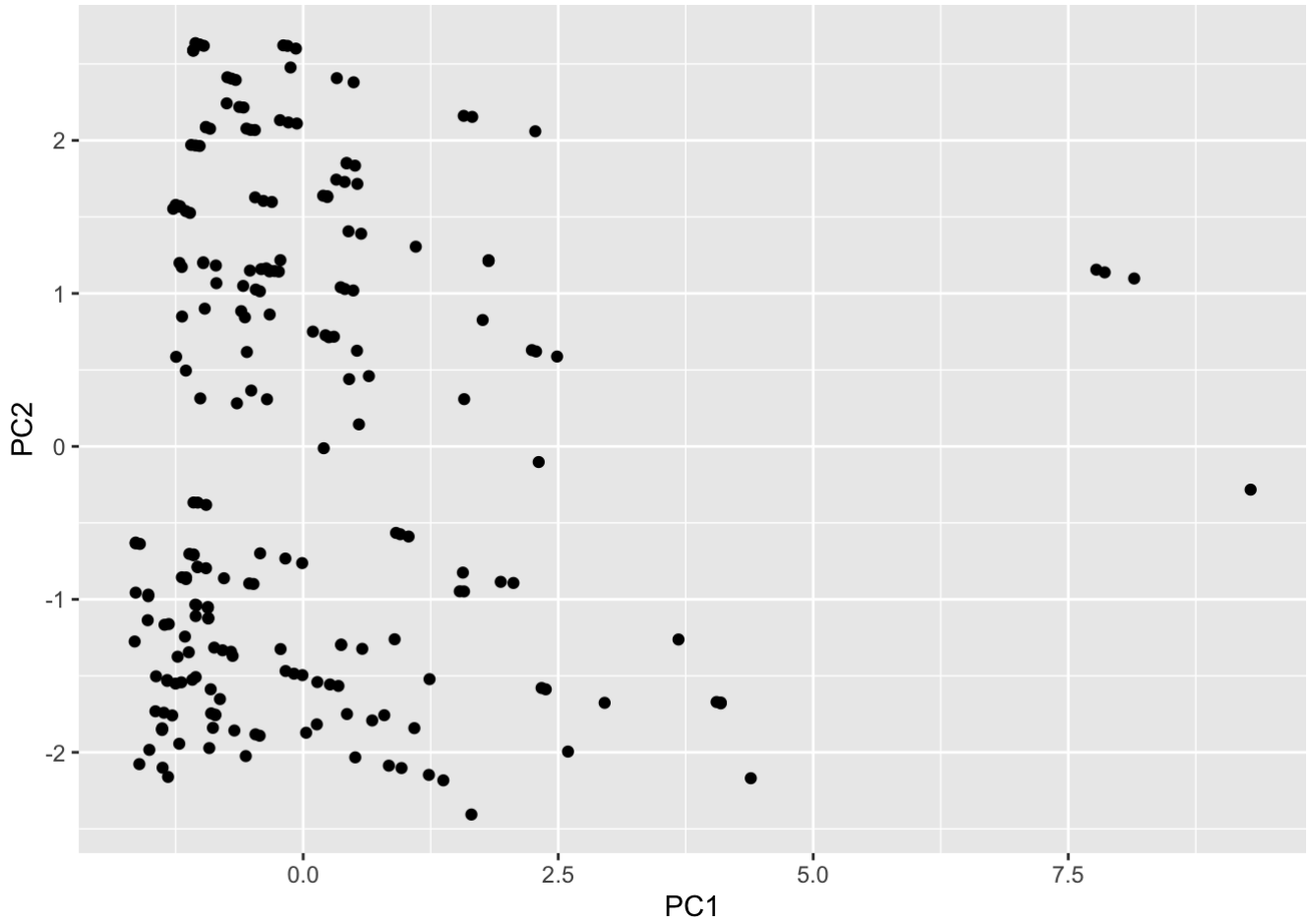
```
eigval
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 2.7350078 2.3255221 0.9675852 0.5474268 0.2987992 0.1256589 0.0000000
```

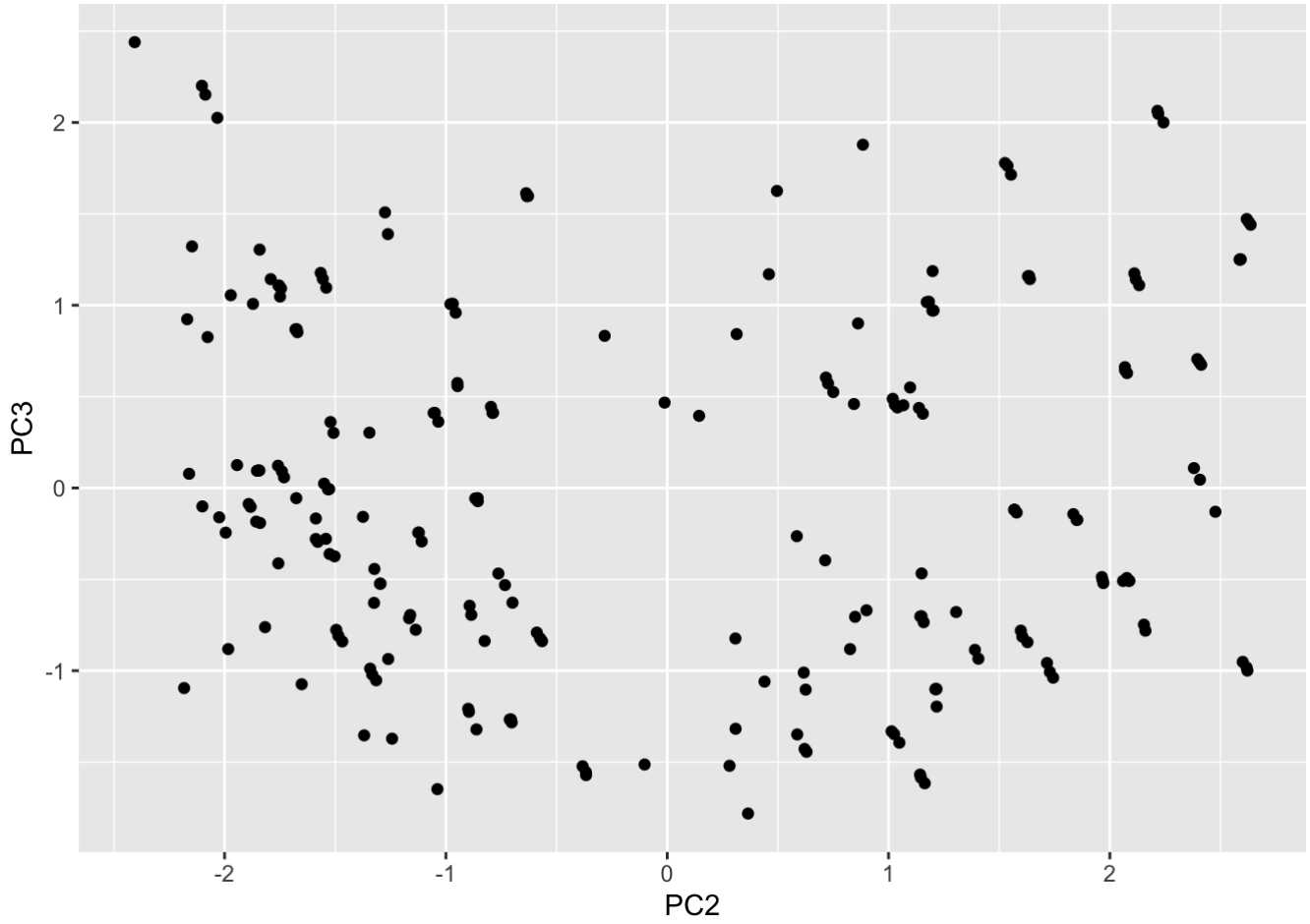
```
summary(sj_pca, loadings=T)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
## Standard deviation     1.6537859 1.5249663 0.9836591 0.73988297 0.5466253
## Proportion of Variance 0.3907154 0.3322174 0.1382265 0.07820383 0.0426856
## Cumulative Proportion  0.3907154 0.7229328 0.8611593 0.93936313 0.9820487
##                           Comp.6 Comp.7
## Standard deviation     0.35448400      0
## Proportion of Variance 0.01795127      0
## Cumulative Proportion  1.00000000      1
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## X1                     0.502         0.858
## base            0.588                       0.384        -0.707
## age            -0.158         0.972         0.150
## seizure.rate    0.522         0.202        -0.818  0.104
## subject                0.608        -0.317        -0.715
## trt                    0.601        -0.400         0.685
## ratevsbaseline  0.588                       0.384         0.707
```

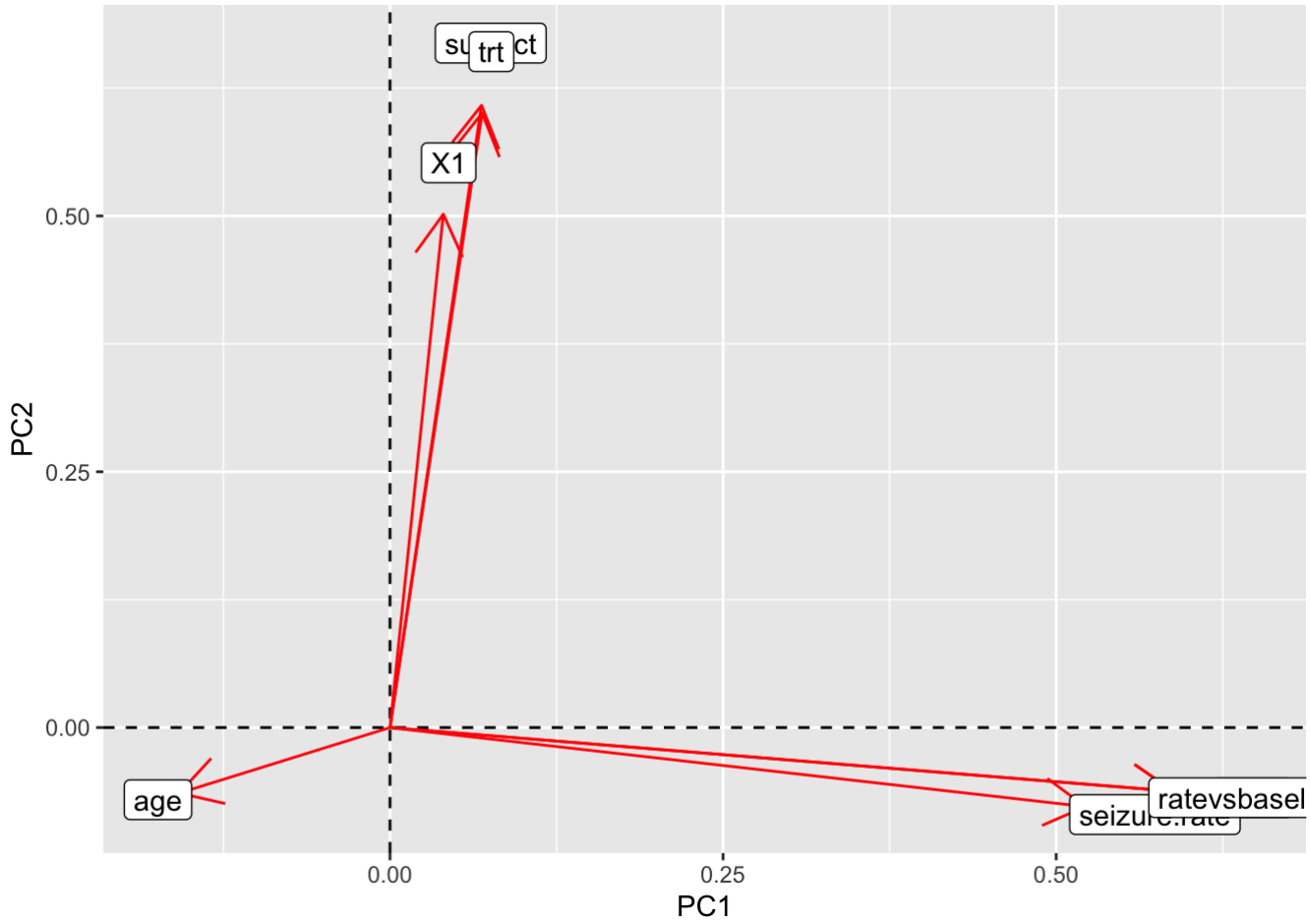
```
sjdf<-data.frame(PC1=sj_pca$scores[,1], PC2=sj_pca$scores[,2])
ggplot(sjdf,aes(PC1, PC2))+geom_point()
```



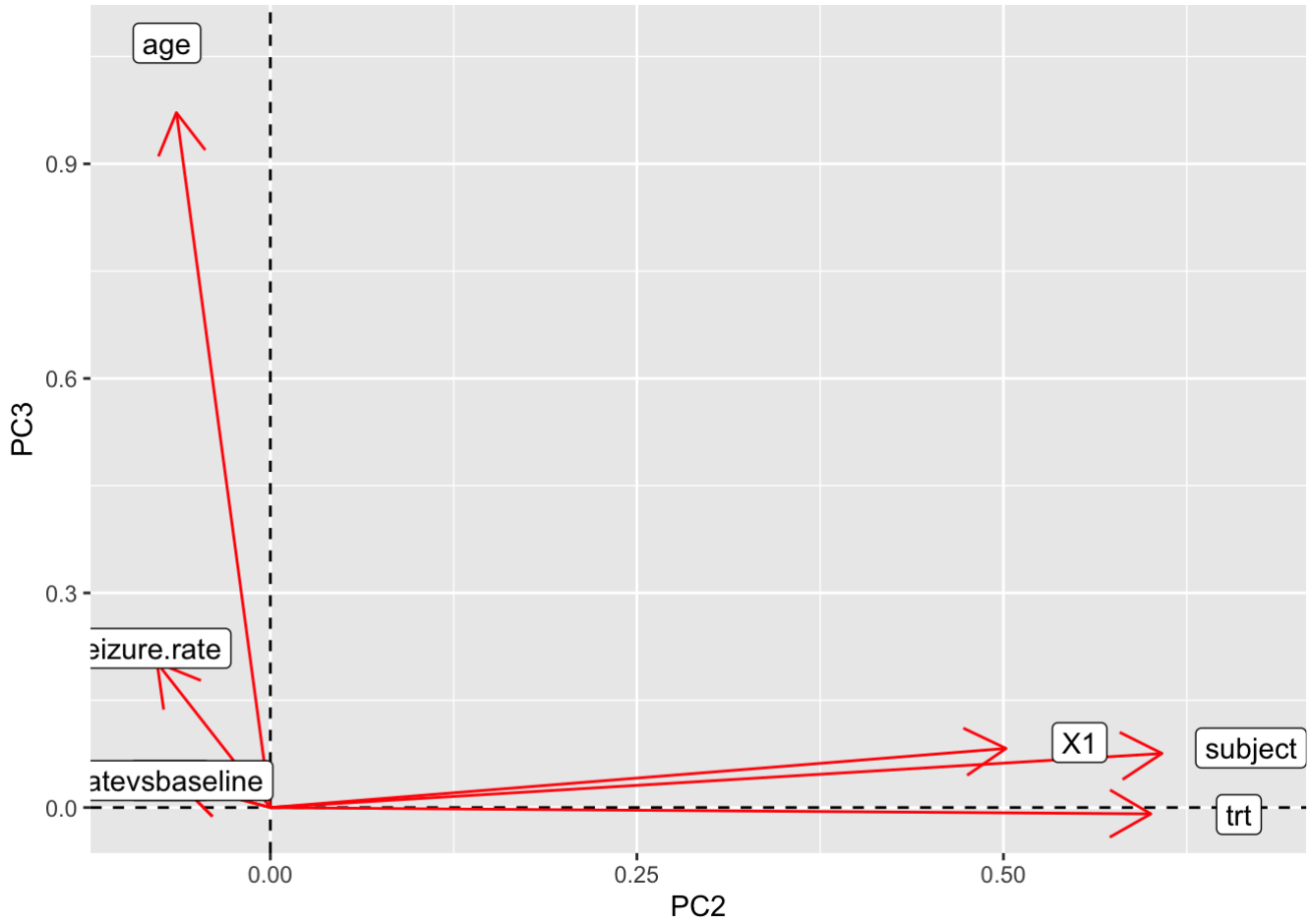
```
sjdf2<-data.frame(PC2=sj_pca$scores[,2], PC3=sj_pca$scores[,3])
ggplot(sjdf2,aes(PC2, PC3))+geom_point()
```




```
sj_pca$loadings[1:7,1:2]%>%as.data.frame%>%rownames_to_column%>%
ggplot()+geom_hline(aes(yintercept=0),lty=2)+
 geom_vline(aes(xintercept=0),lty=2)+ylab("PC2")+xlab("PC1")+
 geom_segment(aes(x=0,y=0,xend=Comp.1,yend=Comp.2),arrow=arrow(),col="red")+
 geom_label(aes(x=Comp.1*1.1,y=Comp.2*1.1,label=rowname))
```




```
sj_pca$loadings[1:7,2:3]%>%as.data.frame%>%rownames_to_column%>%
 ggplot()+geom_hline(aes(yintercept=0),lty=2)+
 geom_vline(aes(xintercept=0),lty=2)+ylab("PC3")+xlab("PC2")+
 geom_segment(aes(x=0,y=0,xend=Comp.2,yend=Comp.3),arrow=arrow(),col="red")+
 geom_label(aes(x=Comp.2*1.1,y=Comp.3*1.1,label=rowname))
```



From the scree plot and its cumulative sums of variance, we can see that we should keep three principal components.Although PC 1 doesn't have a correlation value for each variable, it does load based on base, age, and seizure rate which are the non-ID variables we would like to compare. we can see from the scatterplot of PC2 vs PC1, PC1 explains most of the variance in the two main groups shown on the plot. PC2 also divides the variance into two main halves, upper and lower, of the plot. When looking at the plot of loadings we can see that the seizure rate and rate versus baseline are redundant and strongly influence PC1 while the age is slightly negatively correlated and also slightly affects PC1. The variables X, trt, and subject do not really affect PC1 but do affect PC2. PC3 is influenced by age.