Project 2 - MC DATA 101

Juan J. Nunez

April 14, 2018

First I'm going to start by turning an SPSS file (i.e., .sav) into an RMarkdown file (i.e., .Rda). The data I will be using is from the World Bank. For more information on how to download World Bank data, please visit https://data.worldbank.org

Once the data is downloaded to a .SAV file, it is easy to use it using R by bringing up the 'foreign' package.

```
setwd("C:\\Users\\Juan Nunez\\Desktop\\MC_DATA_101\\ASSIG_2_DATA101")
```

I can look at the data using 'ggplot'. I download and bring up the package. First I get the data.

```
##install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

Now I look at the dimensions of the ASSIG2_DATA data frame.

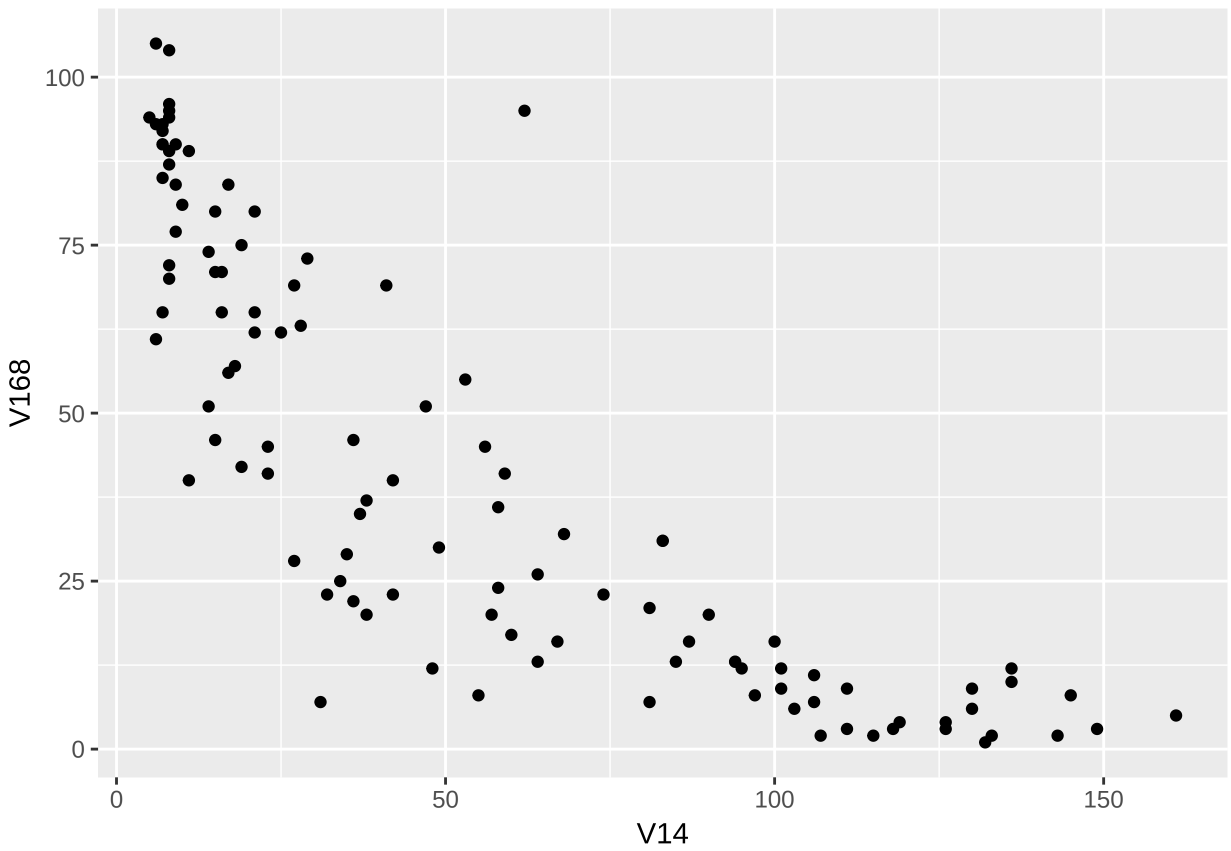
```
library(foreign)
NEWLES <- read.spss("C:\\Users\\Juan Nunez\\Desktop\\MC_DATA_101\\ASSIG_2_DATA101\\LES101.SAV", use.valu

ASSIG2_DATA <- NEWLES

save(ASSIG2_DATA, file = "ASSIG2_DATA.Rda")

dim(ASSIG2_DATA)
```

```
## [1] 148 356
```

I see that there are 148 rows and 356 columns. Let's look at a basic scatterplot first.

```
ggplot(ASSIG2_DATA, aes(V14, V168)) +
  geom_point()
```

```
## Warning: Removed 38 rows containing missing values (geom_point).
```



What we can see here is that there is a trend where as V14 increases, V168 decreases. To understand this better, we use the codebook below to understand what our variables of interest mean.

V1 COUNTRY NUMBER ; V2 ABBREVIATED COUNTRY NAME ; V3 COUNTRY NAME ; V5 % ADULT FEMALE ILLITERACY 1990 ; V12 ENERGY CONSUMPTION/CAPITA 1991 ; V14 INFANT MORTALITY RATE 1991 ; V168 FEMALE SECODARY SCHOOL ENROLLMENT GROSS 1980 ; V133 CIVIL LIBERTIES 1991 ; V188 WORLD AS 5 REGIONS ;

```
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
```

```
SHORT_ASSIG2_DATA <- select(ASSIG2_DATA, V1, V2, V3, V5, V12, V14, V168, V133)
```

There is another issue that we have to take care of before we move on. Are there any missing values in V14 or V168?

```
is.na(ASSIG2_DATA$V168)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE
##  [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
##  [23] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##  [34] FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
##  [45]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
##  [67] FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE
##  [78] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
##  [89]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
## [111] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
## [122]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
## [133] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## [144]  TRUE FALSE FALSE FALSE FALSE
```

```
summary(ASSIG2_DATA$V14)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    5.00   16.00   40.50   53.70   89.25  161.00      10
```

```
summary(ASSIG2_DATA$V168)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    1.00   12.00   36.00   42.11   71.00  105.00      31
```

```
is.na(ASSIG2_DATA$V14)
```

```
##   [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
##  [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [45]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
##  [67] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##  [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [144]  TRUE FALSE  TRUE  TRUE  TRUE
```

Apparently there are some missing values. So we should take a look at some statistics. Let's look at the variance, standard deviation, and inner quartile range.

```
class(SHORT_ASSIG2_DATA)
```

```
## [1] "data.frame"
```

```
class(SHORT_ASSIG2_DATA$V14)
```

```
## [1] "numeric"
```

```
var(SHORT_ASSIG2_DATA$V14, na.rm = TRUE)
```

```
## [1] 1855.002
```

```
sd(SHORT_ASSIG2_DATA$V14, na.rm = TRUE)
```

```
## [1] 43.06973
```

```r
IQR(SHORT_ASSIG2_DATA$V14, na.rm = TRUE)
```

```
## [1] 73.25
```

```r
var(ASSIG2_DATA$V168, na.rm = TRUE)
```

```
## [1] 1047.117
```

```r
sd(ASSIG2_DATA$V168, na.rm = TRUE)
```

```
## [1] 32.35919
```

```r
IQR(ASSIG2_DATA$V168, na.rm = TRUE)
```

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

What about the mode?

```r
vect2 <- na.omit(SHORT_ASSIG2_DATA$V168)

getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

v <- vect2


result <- getmode(v)
print(result)
```

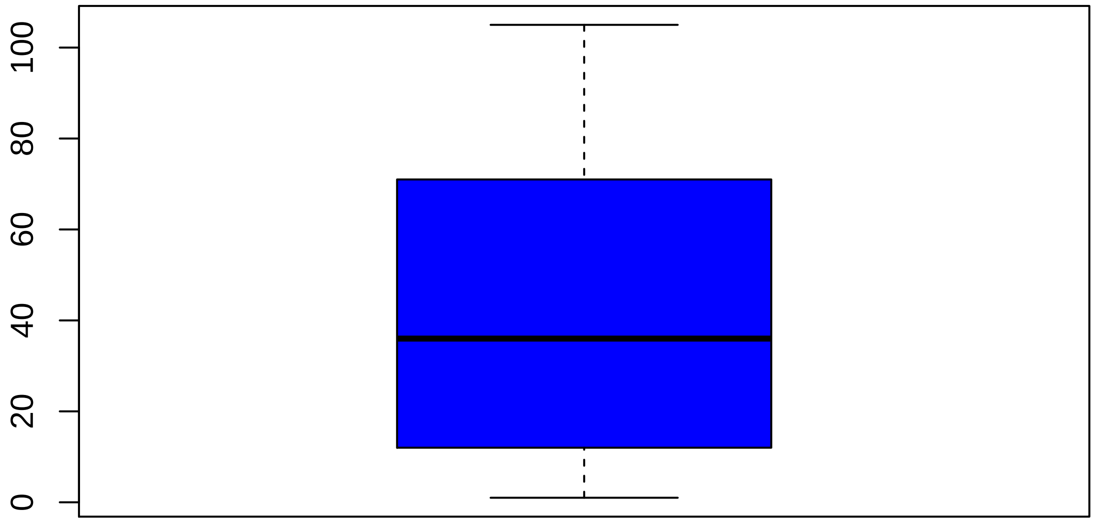
```
## [1] 2
```

```r
vect3 <- na.omit(SHORT_ASSIG2_DATA$V14)

getmode2 <- function(z) {
   uniqz <- unique(z)
   uniqz[which.max(tabulate(match(z, uniqv)))]
}

z <- vect3

result2 <- getmode(z)
print(result2)
```

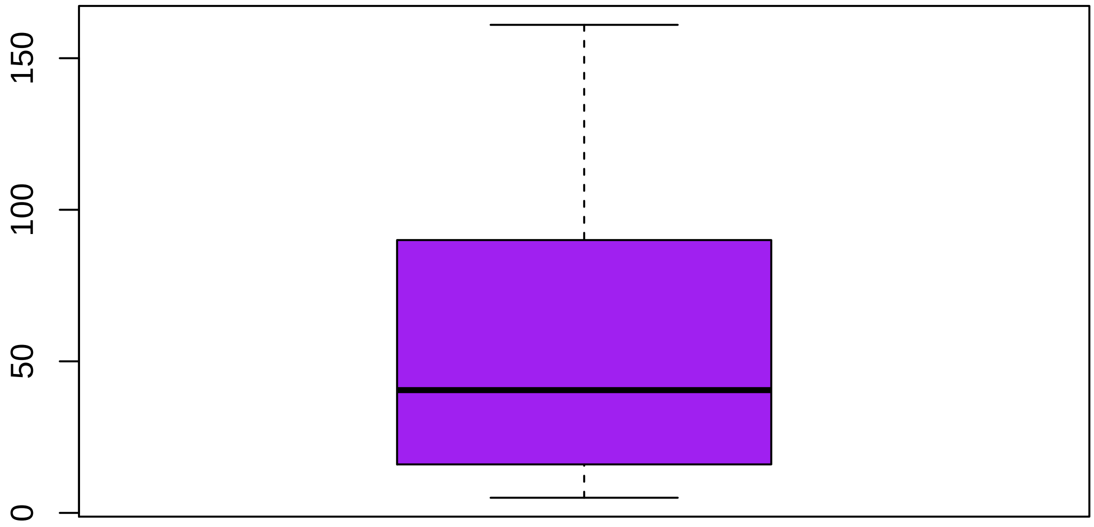
```
## [1] 8
```

Now that we have seen the statistics, we can start to look at the variables graphically.

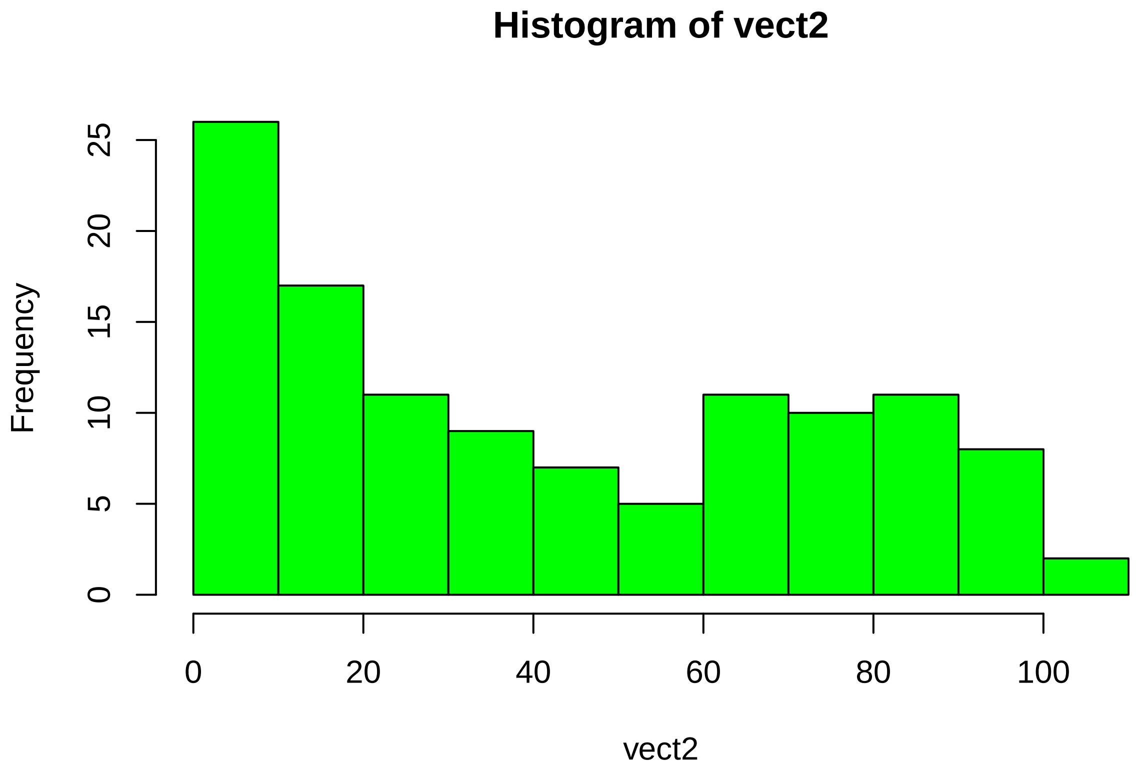
```r
 boxplot(vect2, col = "blue")
```



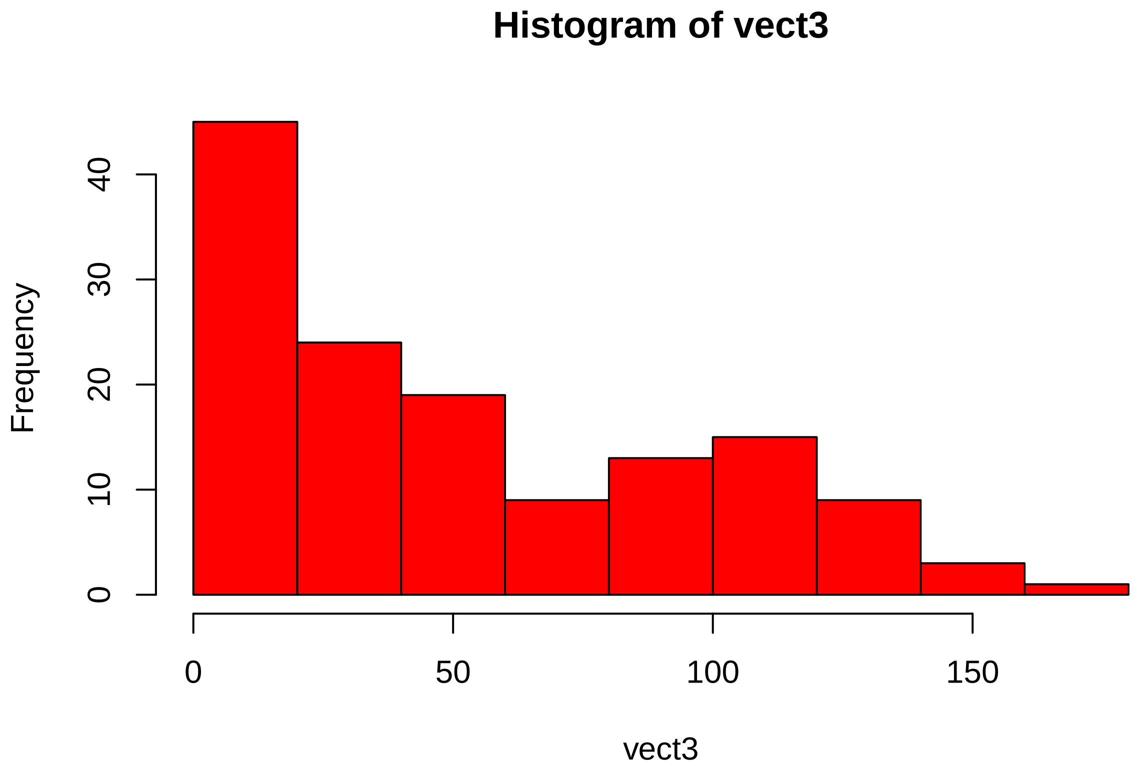
```
boxplot(vect3, col = "purple")
```



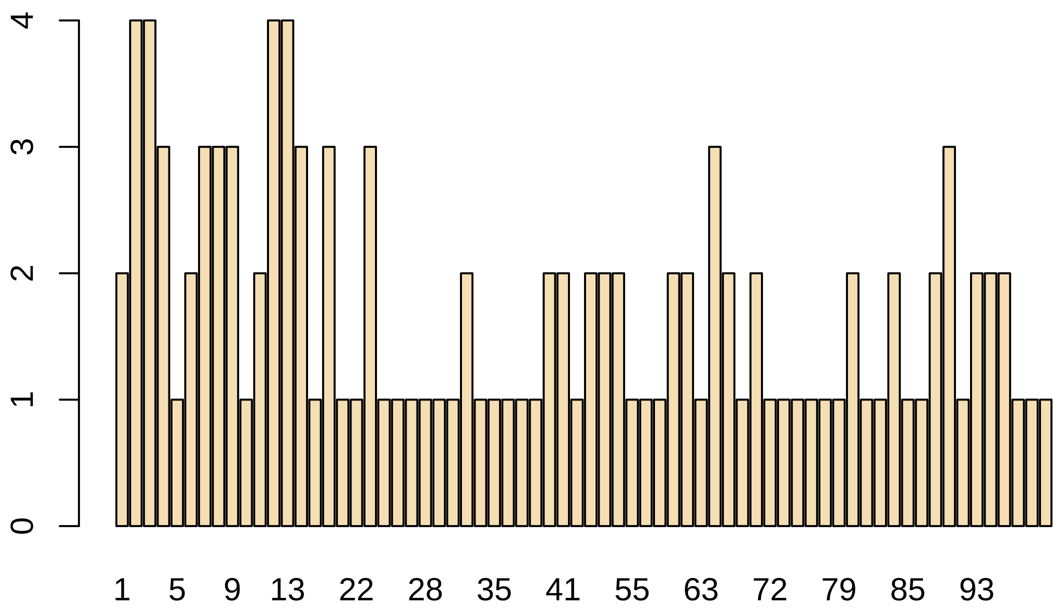
```
hist(vect2, col = "green")
```



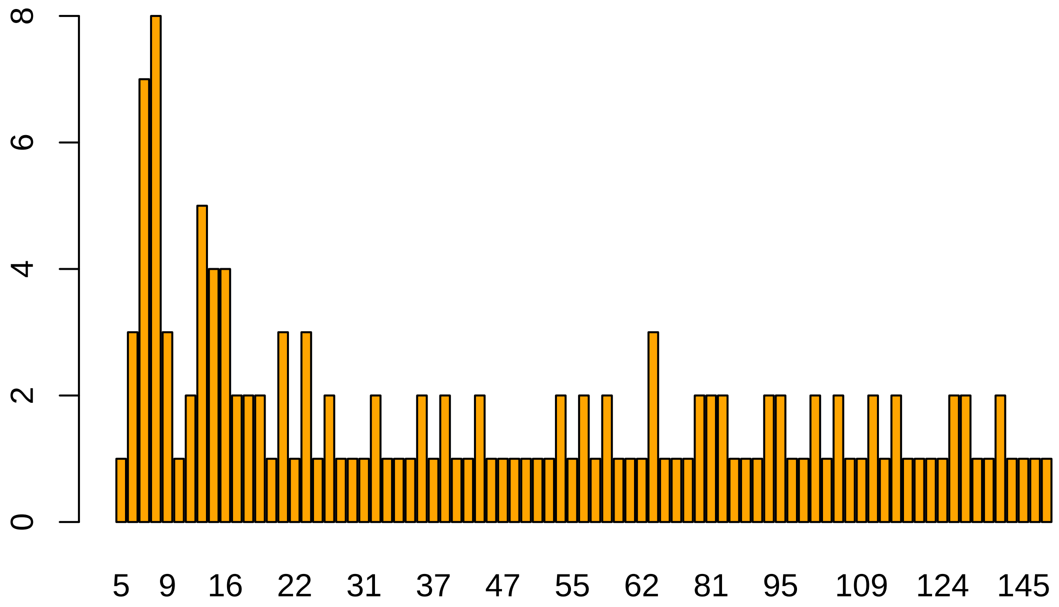
```r
hist(vect3,col = "red")
```



```
table(vect2) %>% barplot(col = "wheat")
```



```r
table(vect3) %>% barplot(col = "orange")
```

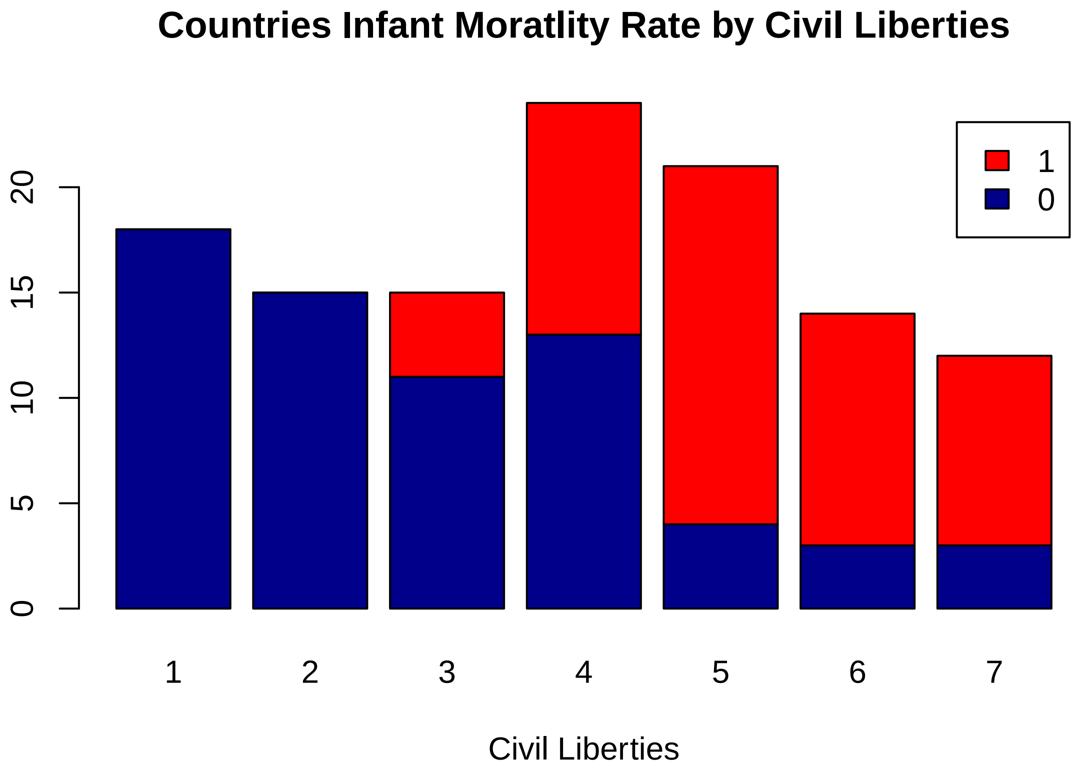


Let's try stacked barplots. We are going to use the variable civil liberties. We have to make sure to use listwise deletion for this to work.

```
SMALL_DATA <- select(SHORT_ASSIG2_DATA, V14, V133)

SMALL_DATA <- na.omit(SMALL_DATA)

SMALL_DATA <- mutate(SMALL_DATA, INF_DUM = as.numeric(SMALL_DATA$V14 >= mean(SMALL_DATA$V14)))

head(SMALL_DATA)
```

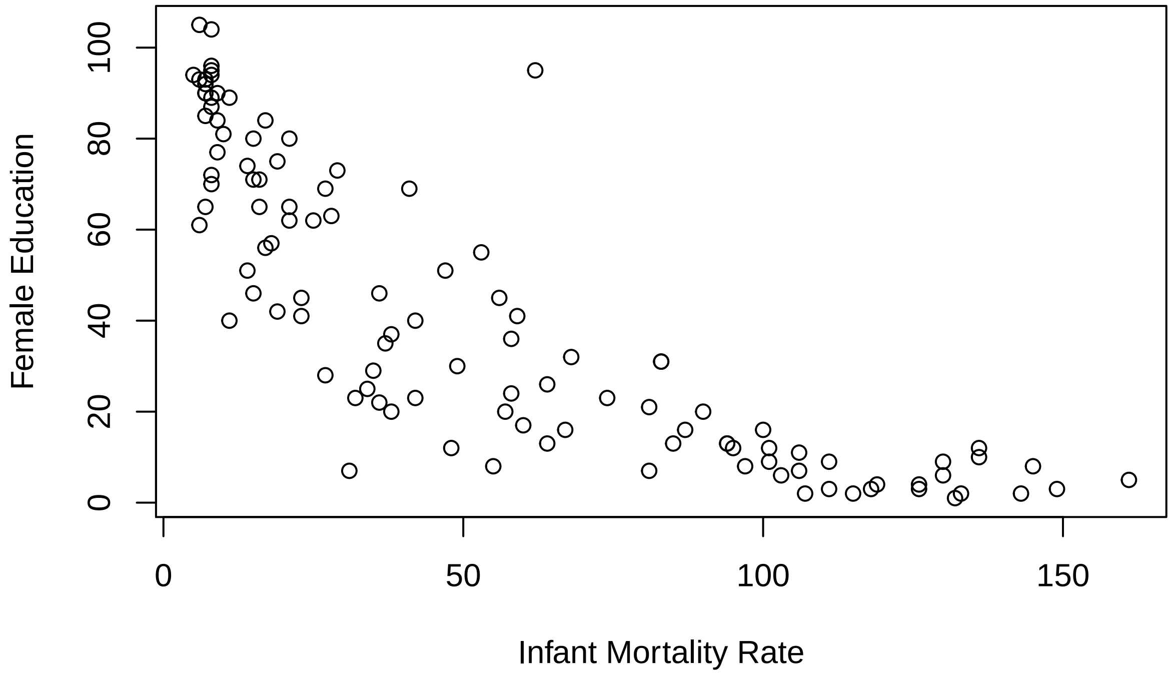
```
##   V14 V133 INF_DUM
## 1  28    6       0
## 2  64    4       1
## 3 130    7       1
## 4  25    3       0
## 5   8    1       0
## 6   8    1       0
```

```
counts <- table(SMALL_DATA$INF_DUM, SMALL_DATA$V133)
barplot(counts, main="Countries Infant Moratlity Rate by Civil Liberties",
  xlab="Civil Liberties", col=c("darkblue","red"),
  legend = rownames(counts))
```



These graphics are very helpful to identify trends in the data. Let's take a look at a scatterplot now.

```
SCATDAT <- select(SHORT_ASSIG2_DATA, V14, V168)
SCATDAT <- na.omit(SCATDAT)


 with(SCATDAT, plot(V14, V168, xlab = "Infant Mortality Rate", ylab = "Female Education"))
```



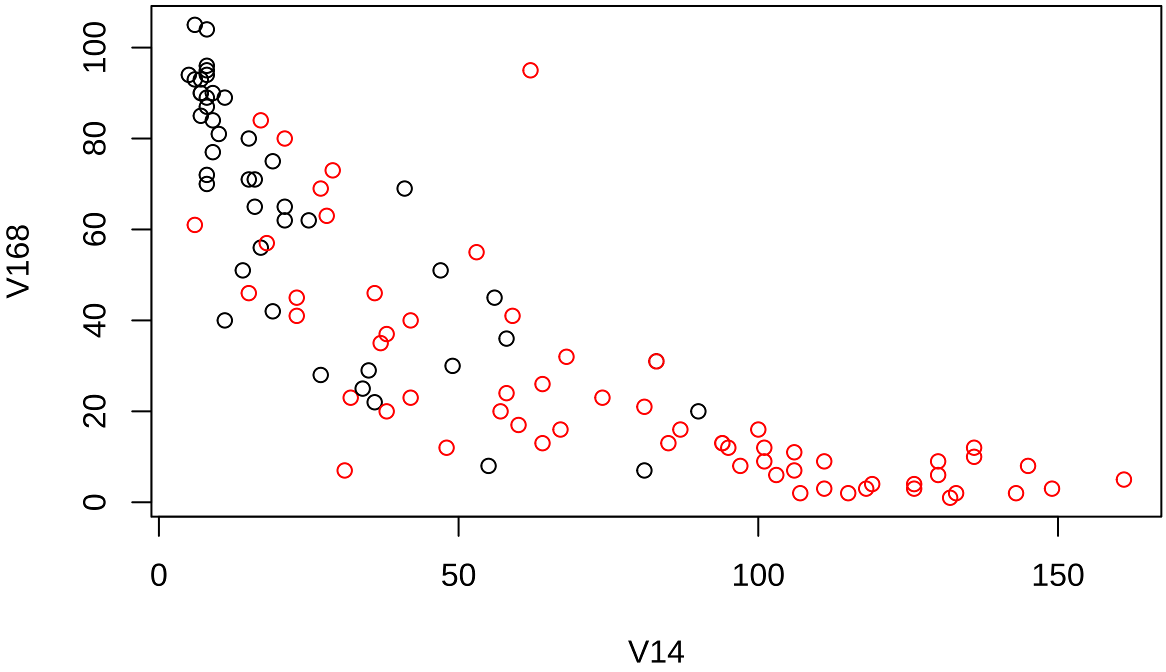
We can confirm the association that we observed at the beginning of this analysis. Sometimes statistical analyses are not linear.

Now let's see if we can incorporate civil liberties.

```
SCATDAT2 <- select(SHORT_ASSIG2_DATA, V14, V168, V133)
SCATDAT2 <- na.omit(SCATDAT2)
SCATDAT2 <- mutate(SCATDAT2, CLIBDUM = as.numeric(SCATDAT2$V133 > 3))

head(SCATDAT2)
```

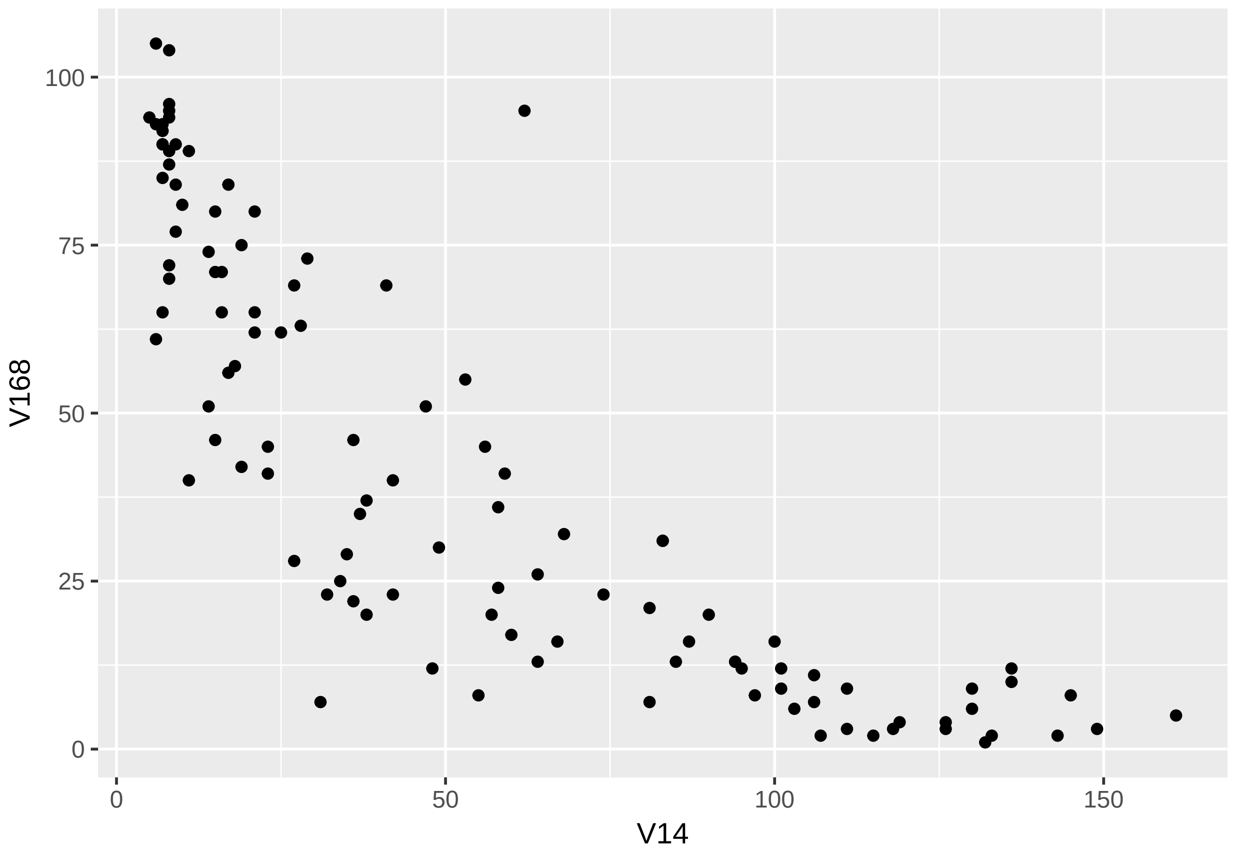
```
##   V14 V168 V133 CLIBDUM
## 1  28   63    6       1
## 2  64   26    4       1
## 3 130    9    7       1
## 4  25   62    3       0
## 5   8   72    1       0
## 6   8   87    1       0
```

```
 with(SCATDAT2, plot(V14, V168, col = as.factor(CLIBDUM) ) )
```



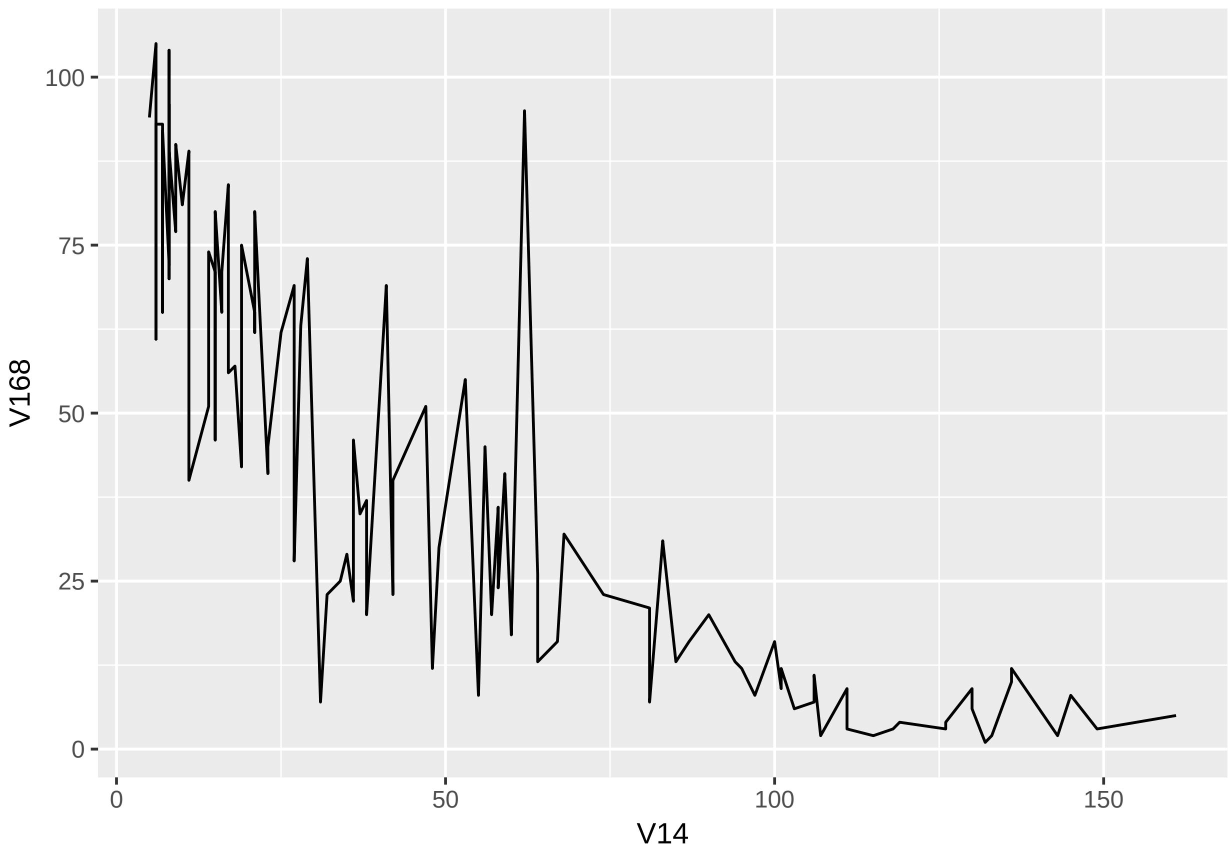
Finally, we can see how those with more than 3 in civil liberties are red. Now we'll use ggplot again.

```
ggplot(SCATDAT) +
  aes(x = V14, y = V168) +
  geom_point()
```



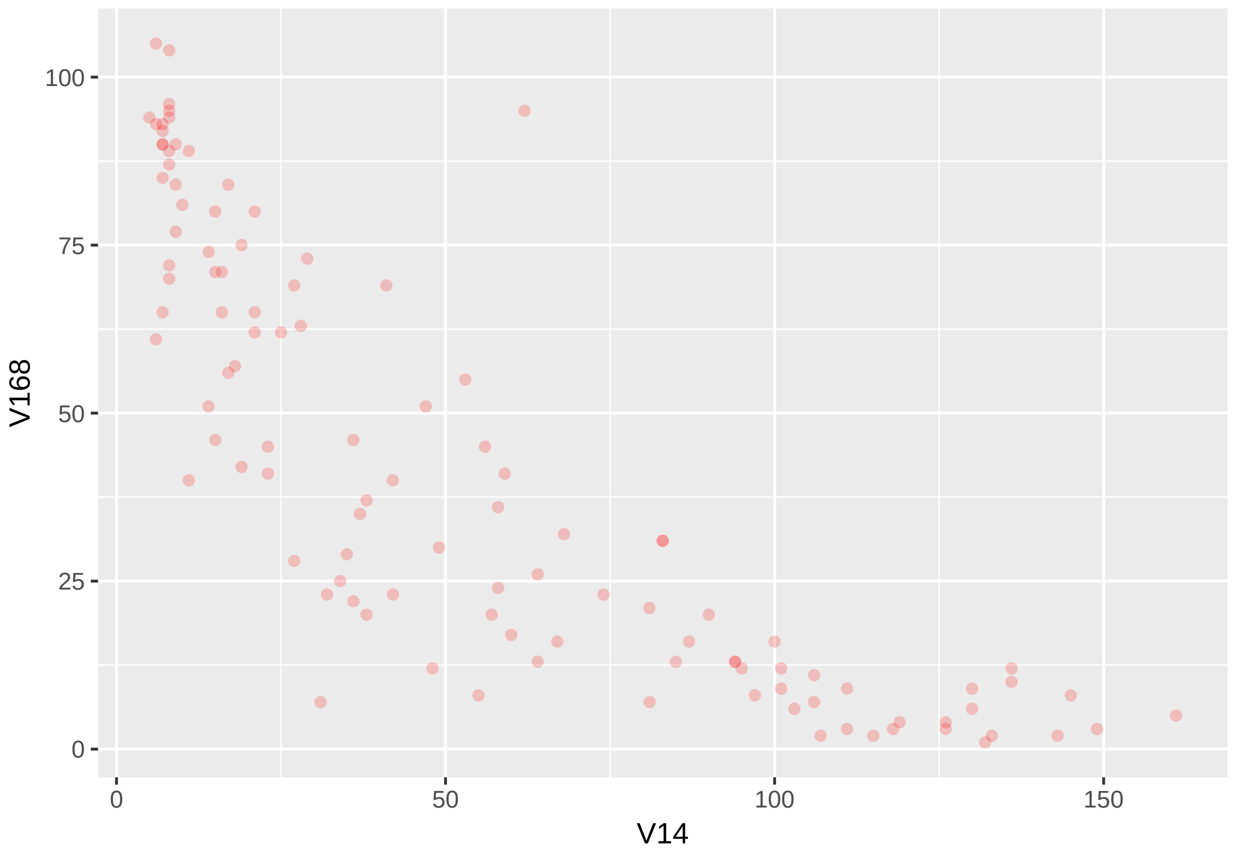
aes stands for aesthetics. Now we are going to get a line instead of dots.

```
ggplot(SCATDAT) +
  aes(x = V14, y = V168) +
  geom_line()
```



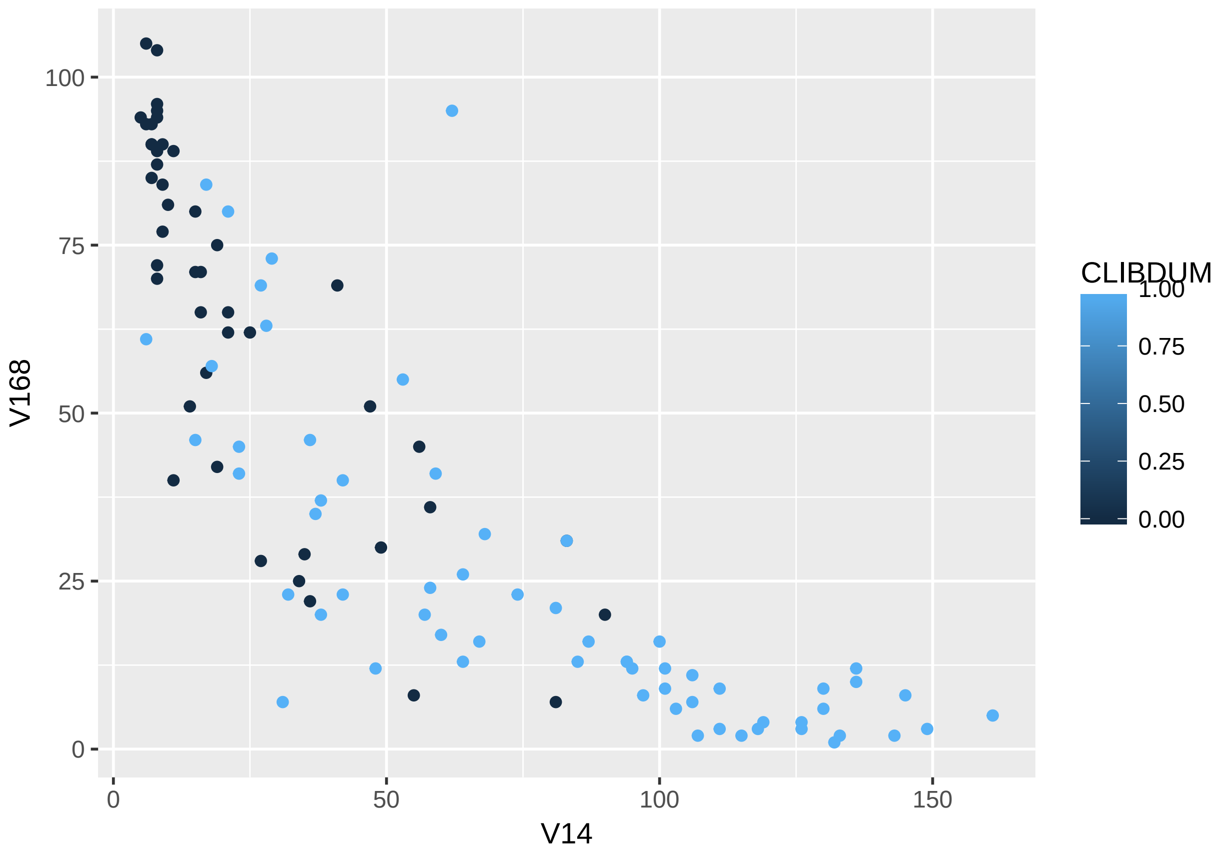
Let’s add some layers to our previous graphs.

```r
ggplot(SCATDAT) +
  aes(x = V14, y = V168) +
  geom_point(colour = 'red', alpha = 0.2)
```



Now let's look at the three variables together.

```
ggplot(SCATDAT2) +
  aes(x = V14, y = V168, colour = CLIBDUM) +
  geom_point()
```

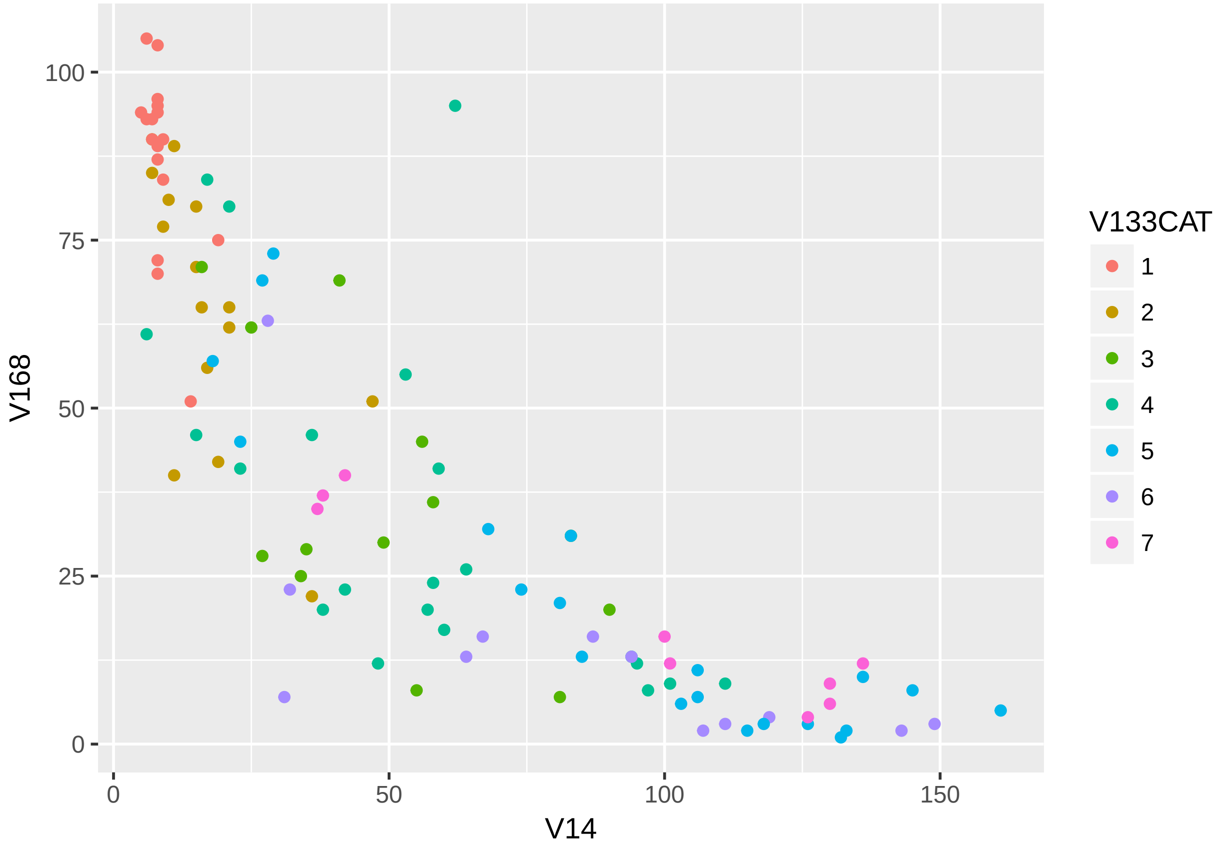


We can use a variable with more categories.

```
SCATDAT3 <- select(SHORT_ASSIG2_DATA, V14, V168, V133)
SCATDAT3 <- na.omit(SCATDAT3)
SCATDAT3 <- mutate(SCATDAT3, V133CAT = as.factor(SCATDAT2$V133))

head(SCATDAT3)
```

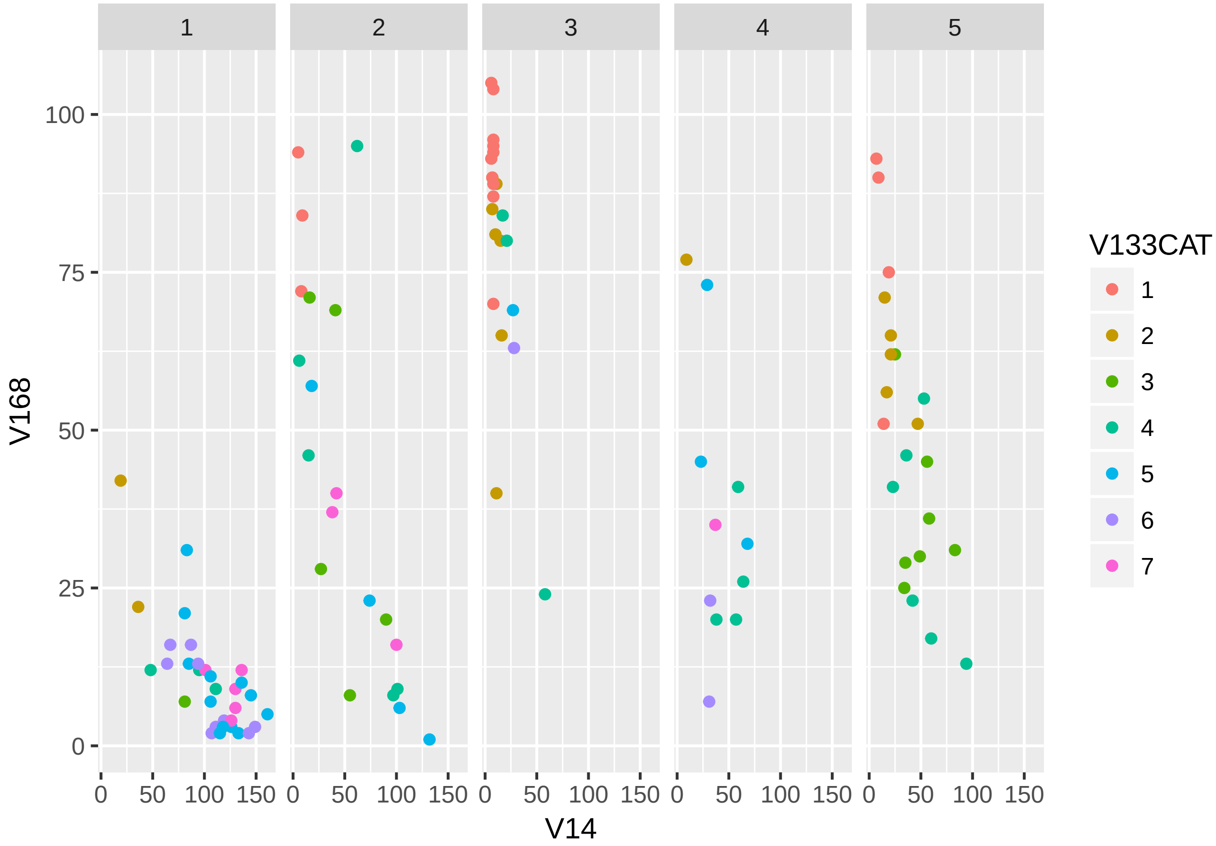
```
##   V14 V168 V133 V133CAT
## 1  28   63    6       6
## 2  64   26    4       4
## 3 130    9    7       7
## 4  25   62    3       3
## 5   8   72    1       1
## 6   8   87    1       1
```

```
ggplot(SCATDAT3) +
  aes(x = V14, y = V168, colour = V133CAT) +
  geom_point()
```



We can even do what is called a facet. We are going to use region of the world.

```
SCATDAT4 <- select(ASSIG2_DATA, V14, V168, V133, V188)
SCATDAT4 <- na.omit(SCATDAT4)
SCATDAT4 <- mutate(SCATDAT4, V133CAT = as.factor(SCATDAT4$V133))
SCATDAT4 <- mutate(SCATDAT4, V188CAT = as.factor(SCATDAT4$V188))

ggplot(SCATDAT4) +
  aes(x = V14, y = V168, colour = V133CAT) +
  geom_point() +
  facet_grid(~ V188CAT)
```

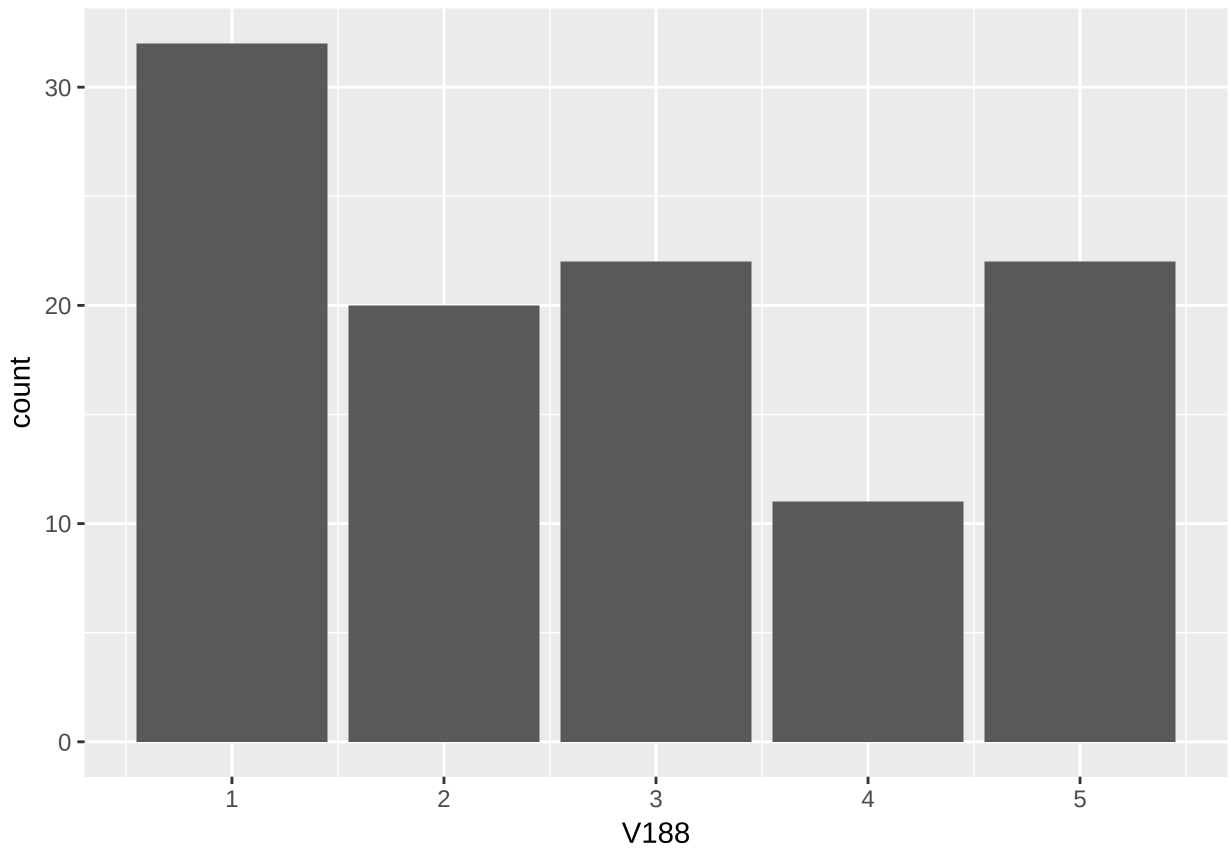


Now we have a plot for each world region. The codebook of V188 states that the regions are:

1=Sub-Saharan Africa; 2=South Asia, East Asia, and Pacific; 3=Europe/Central Asia; 4=Middle East and North Africa; 5=Americas;

Now let’s see what ggplot can do.

```
ggplot(SCATDAT4) +
  aes(x = V188) +
  geom_bar()
```

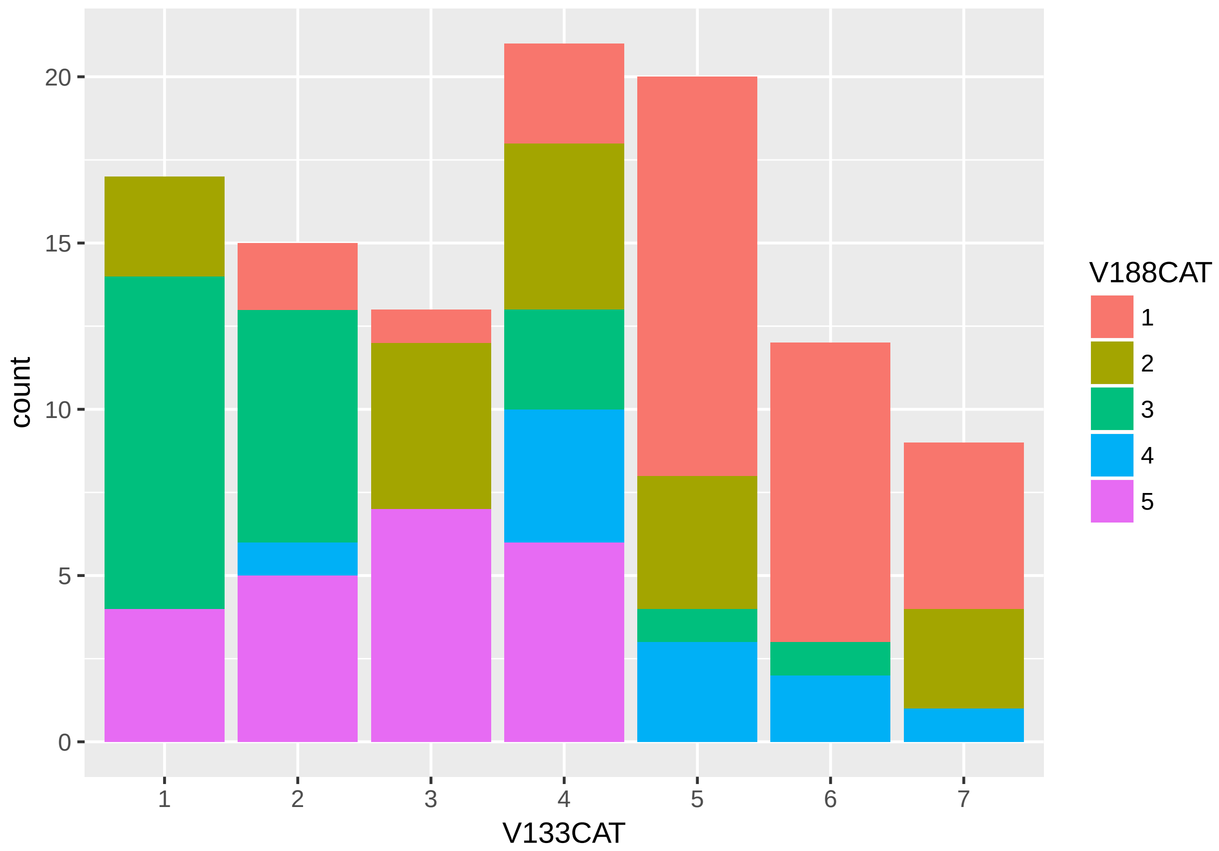


Now we create a dummy variable for civil liberties where those with more than 3 in the scale are coded as 1.

```
SCATDAT4 <- mutate(SCATDAT4, CLIBDUM = as.numeric(SCATDAT2$V133 > 3))
head(SCATDAT4)
```

```
##   V14 V168 V133 V188 V133CAT V188CAT CLIBDUM
## 1  28   63    6    3       6       3       1
## 2  64   26    4    4       4       4       1
## 3 130    9    7    1       7       1       1
## 4  25   62    3    5       3       5       0
## 5   8   72    1    2       1       2       0
## 6   8   87    1    3       1       3       0
```

```
SCATDAT4 %>%
ggplot() +
  aes(x = V133CAT, fill = V188CAT) +
  geom_bar()
```

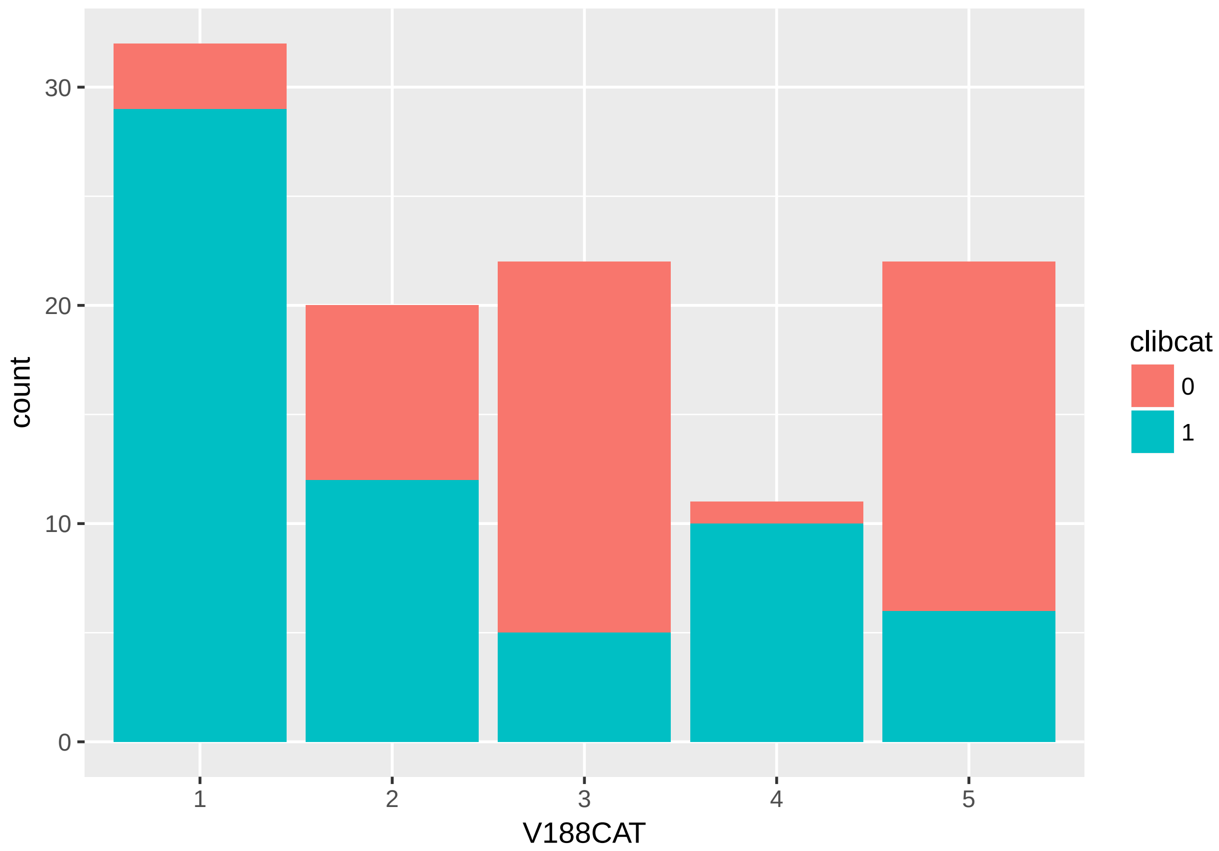


This is a bit messy, so we try to ammeliorate it.

```
SCATDAT4 <- mutate(SCATDAT4, CLIBDUM = as.numeric(SCATDAT2$V133 > 3))
SCATDAT4 <- mutate(SCATDAT4, clibcat = as.factor(SCATDAT4$CLIBDUM))
head(SCATDAT4)
```

```
##   V14 V168 V133 V188 V133CAT V188CAT CLIBDUM clibcat
## 1  28   63    6    3       6       3       1       1
## 2  64   26    4    4       4       4       1       1
## 3 130    9    7    1       7       1       1       1
## 4  25   62    3    5       3       5       0       0
## 5   8   72    1    2       1       2       0       0
## 6   8   87    1    3       1       3       0       0
```

```
SCATDAT4 %>%
ggplot() +
  aes(x = V188CAT, fill = clibcat) +
  geom_bar()
```

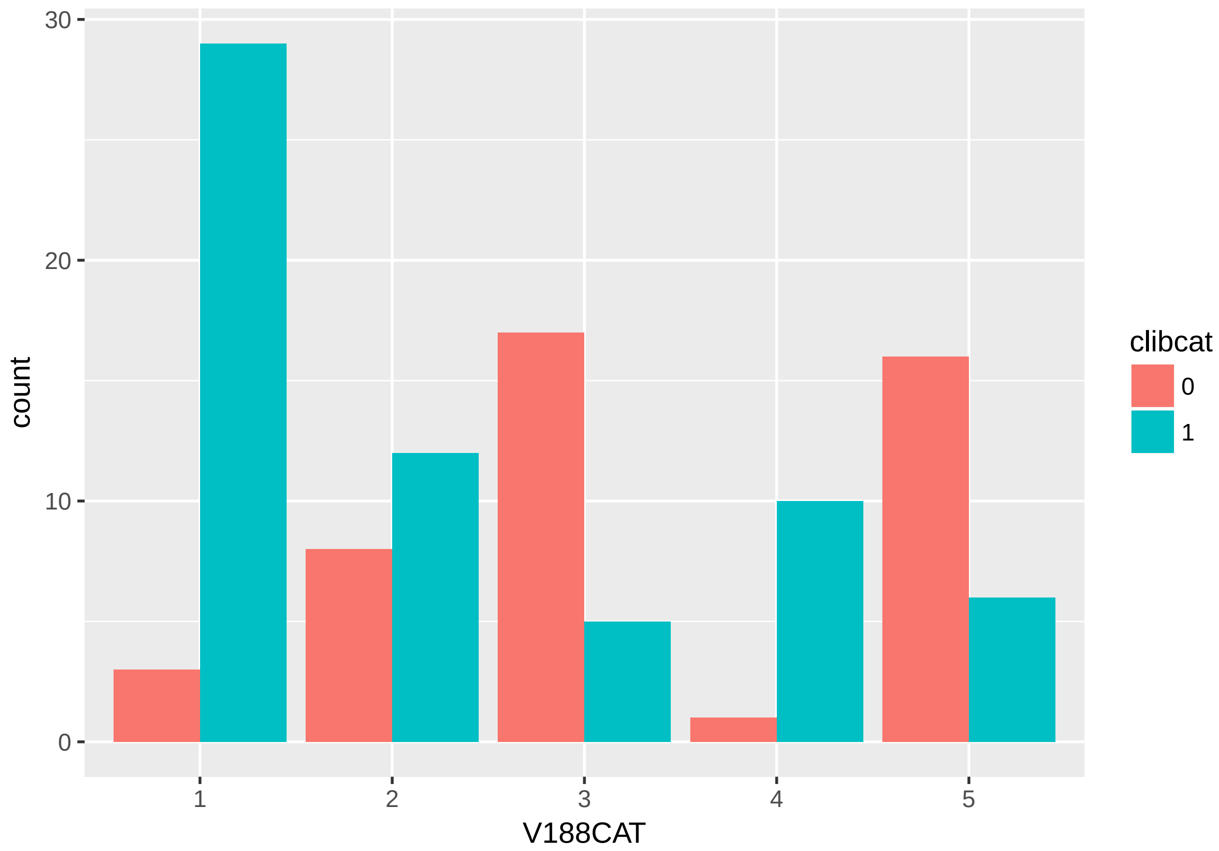


Much better. We can make it even better by changing the positions of the bars.

```r
SCATDAT4 <- mutate(SCATDAT4, CLIBDUM = as.numeric(SCATDAT2$V133 > 3))
SCATDAT4 <- mutate(SCATDAT4, clibcat = as.factor(SCATDAT4$CLIBDUM))
head(SCATDAT4)
```

```
##   V14 V168 V133 V188 V133CAT V188CAT CLIBDUM clibcat
## 1  28   63    6    3       6       3       1       1
## 2  64   26    4    4       4       4       1       1
## 3 130    9    7    1       7       1       1       1
## 4  25   62    3    5       3       5       0       0
## 5   8   72    1    2       1       2       0       0
## 6   8   87    1    3       1       3       0       0
```

```r
SCATDAT4 %>%
ggplot() +
  aes(x = V188CAT, fill = clibcat) +
  geom_bar(position = 'dodge')
```

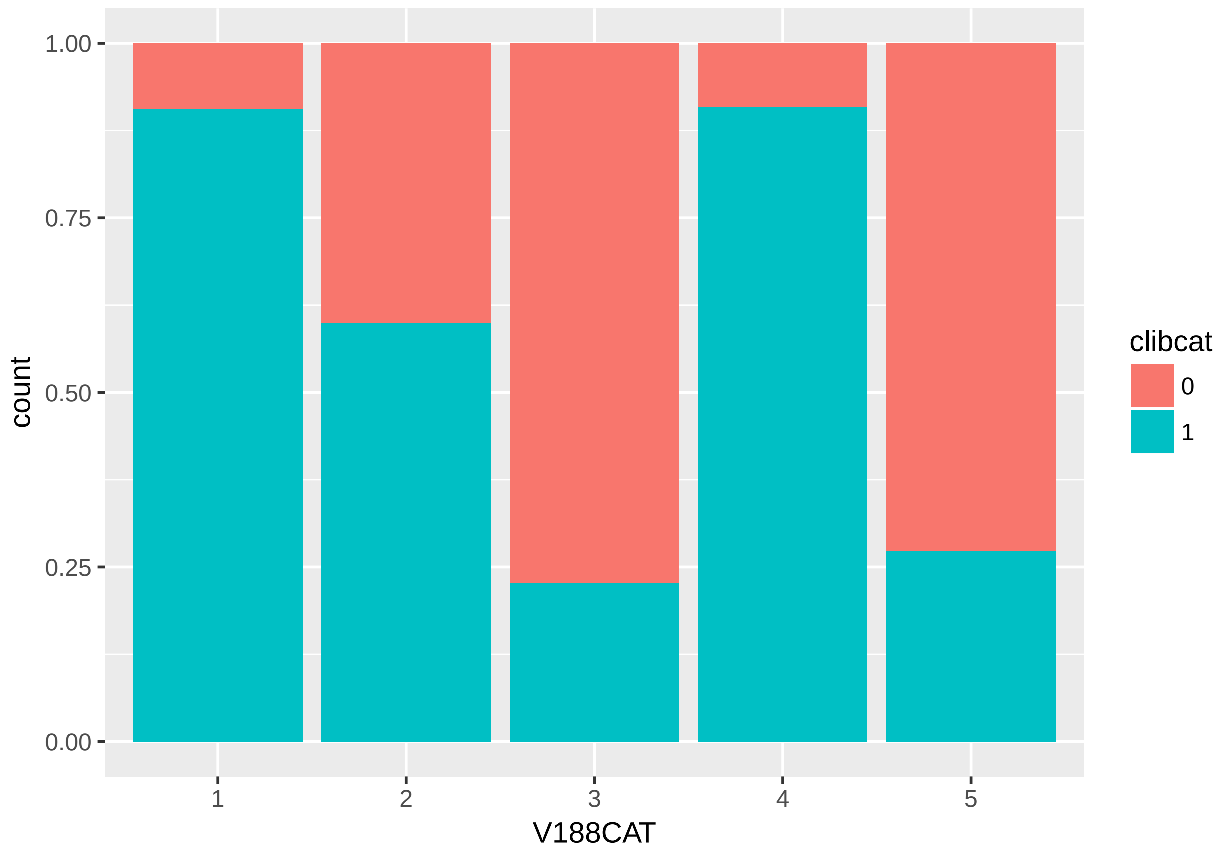


We can look at proportions of the categorical value also.

```
SCATDAT4 <- mutate(SCATDAT4, CLIBDUM = as.numeric(SCATDAT2$V133 > 3))
SCATDAT4 <- mutate(SCATDAT4, clibcat = as.factor(SCATDAT4$CLIBDUM))
head(SCATDAT4)
```

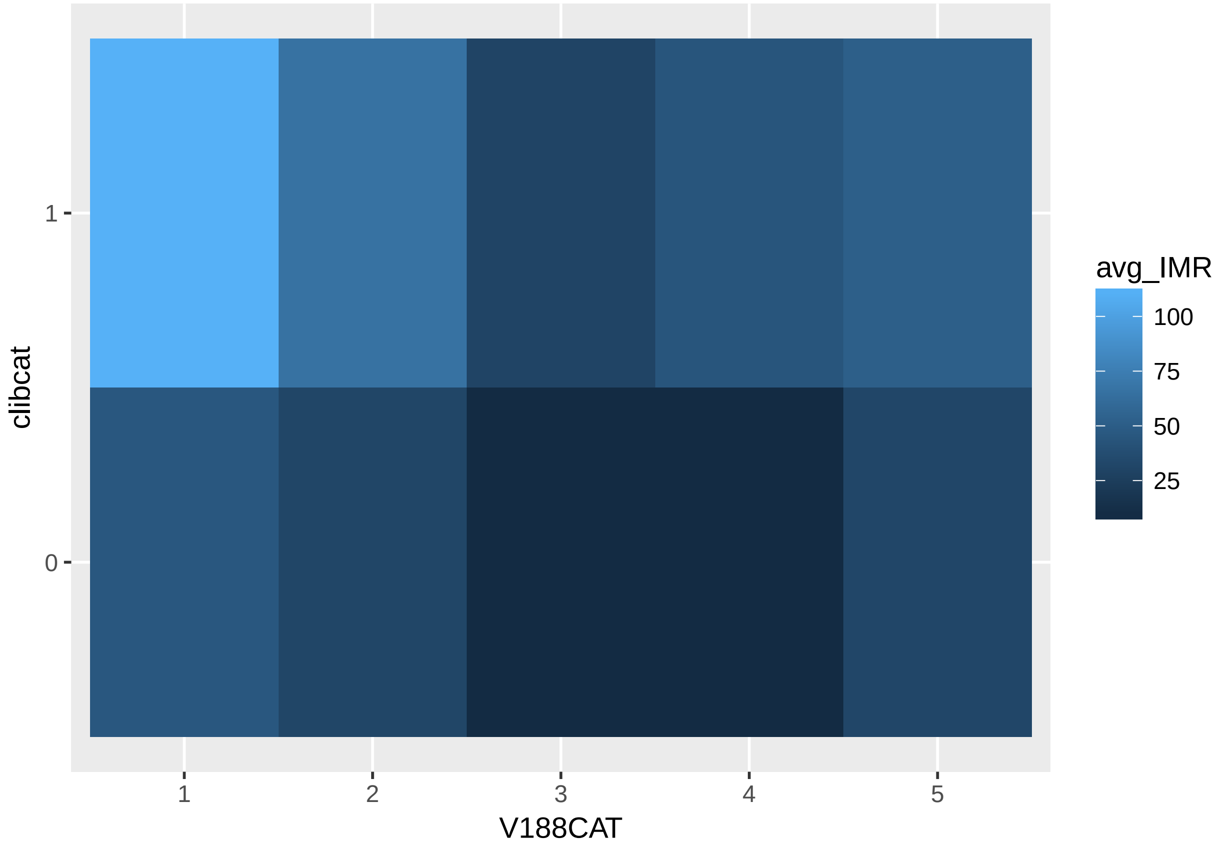
```
##   V14 V168 V133 V188 V133CAT V188CAT CLIBDUM clibcat
## 1  28   63    6    3       6       3       1       1
## 2  64   26    4    4       4       4       1       1
## 3 130    9    7    1       7       1       1       1
## 4  25   62    3    5       3       5       0       0
## 5   8   72    1    2       1       2       0       0
## 6   8   87    1    3       1       3       0       0
```

```
SCATDAT4 %>%
ggplot() +
  aes(x = V188CAT, fill = clibcat) +
  geom_bar(position = 'fill')
```



We can see how civil liberties take over proportions in different regions. Now we are going to look at heat plot. This is a useful tool to look for interactions.

```
HEATDAT <-

SCATDAT4 %>%
  group_by(clibcat, V188CAT) %>%
  summarise(avg_IMR = mean(V14)) %>%
  ungroup()

ggplot(HEATDAT) +
  aes(y = clibcat, x = V188CAT, fill = avg_IMR) +
  geom_raster()
```



```
HEATDAT
```

```
## # A tibble: 10 x 3
##    clibcat V188CAT avg_IMR
##    <fct>   <fct>     <dbl>
##  1 0       1         45.3
##  2 0       2         31.4
##  3 0       3          8.94
##  4 0       4          9.00
##  5 0       5         31.9
##  6 1       1        111.
##  7 1       2         65.7
##  8 1       3         30.2
##  9 1       4         43.8
## 10 1       5         51.3
```

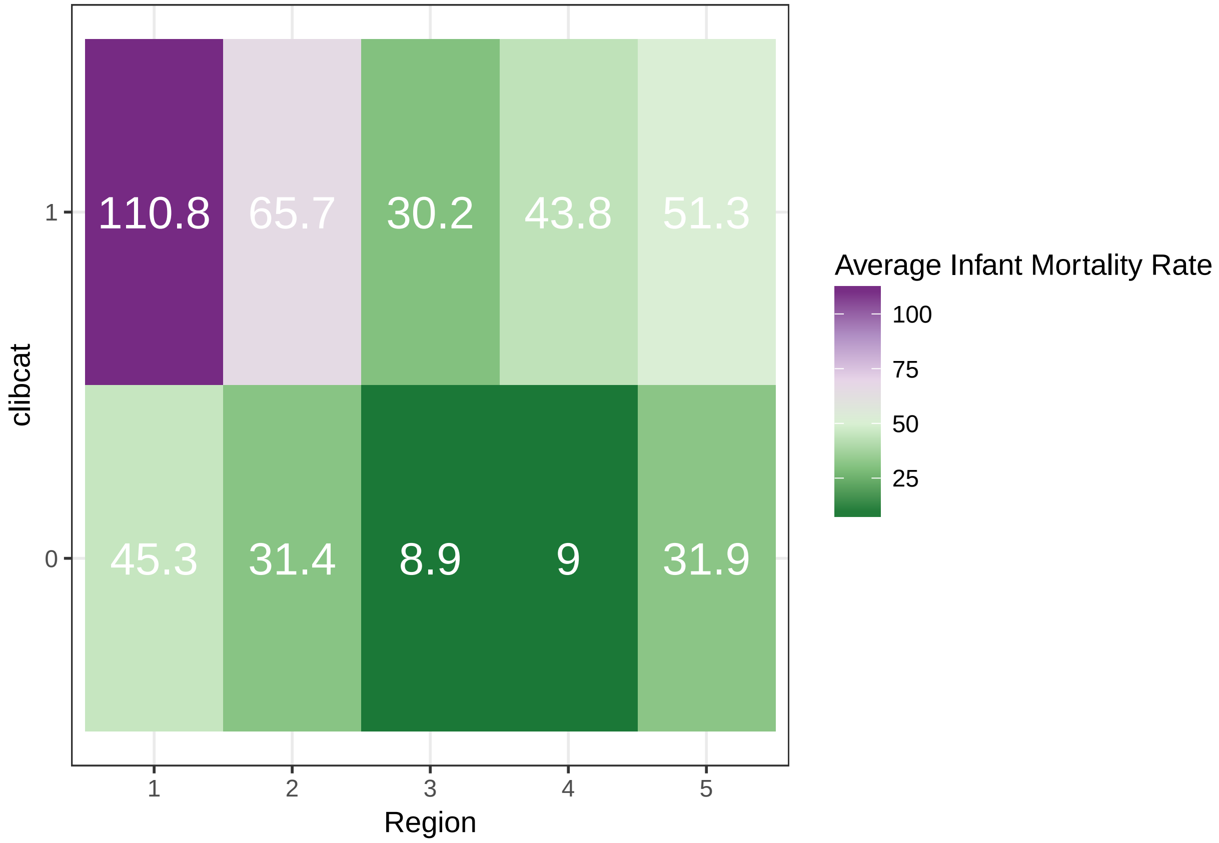
We can see in table all of the combinations of civil liberty and region and their corresponding average infant mortality rate. Let's put the mean infant mortality in the graph.

```
HEATDAT$label <- HEATDAT$avg_IMR %>% round(1) %>% as.character
HEATDAT
```

```
## # A tibble: 10 x 4
##    clibcat V188CAT avg_IMR label
##    <fct>   <fct>     <dbl> <chr>
##  1 0       1         45.3  45.3
##  2 0       2         31.4  31.4
##  3 0       3          8.94 8.9
##   4 0       4          9.00 9
##   5 0       5         31.9  31.9
##   6 1       1        111.   110.8
##   7 1       2         65.7  65.7
##   8 1       3         30.2  30.2
##   9 1       4         43.8  43.8
##  10 1       5         51.3  51.3
```

We can also add color schemes. We need to install and download RColorBrewer.

```
##install.packages('RColorBrewer')
library(RColorBrewer)

ggplot(HEATDAT) +
  aes(y = clibcat, x = V188CAT, fill = avg_IMR, label = label) +
  geom_raster() +
  geom_text(colour = 'white',
            size = 6) +
  scale_fill_distiller(name = 'Average Infant Mortality Rate',
                       type = 'div',
                       palette = 3) +
  xlab('Region') +
  ylab('clibcat') +
  theme_bw()
```


clibcat
Region
Average Infant Mortality Rate