Homework 8

STAT 5014

Adeline Guthrie

11/6/2019

Problem 3

```
#download("http://databank.worldbank.org/data/download/Edstats_csv.zip",
#          dest = "Edstats_csv.zip")
#unzip("Edstats_csv.zip", exdir = "C:/Users/apgut/Documents/STAT_5014")

EdStatsData <- read_csv("EdStatsData.csv")

edu <- EdStatsData %>%
  rename("Country" = "Country Code") %>%
  rename("Indicator" = "Indicator Code") %>%
  select(-c(1, 3, 68))

NAs_full <- sum(apply(edu[, 3:67], 2, is.na))
total_cells_full <- 886930 * 66
data_points_full <- total_cells_full - NAs_full

edu2 <- melt(edu, c(1, 2), c(3:67), variable.name = "year", na.rm = TRUE)

swe_fin <- filter(edu2, Country == "FIN" | Country == "SWE")
swe_fin2 <- swe_fin %>%
  group_by(Indicator, year) %>%
  summarize(Ind_size = n()) %>%
  filter(Ind_size == 2) %>%
  select(-3)

swe_fin3 <- left_join(swe_fin2, swe_fin, by = c("Indicator" = "Indicator", "year" = "year"))

swe_fin3 <- swe_fin3 %>%
  transform(year = as.numeric(year)) %>%
  transform(year = year + 1969)

swe_fin3 <- filter(swe_fin3, year < 2015)

swe <- swe_fin3 %>%
  filter(Country == "SWE")

fin <- swe_fin3 %>%
  filter(Country == "FIN")

swe_years <- length(levels(unique(filter(swe_fin, Country == "SWE")$year)))
fin_years <- length(levels(unique(filter(swe_fin, Country == "FIN")$year)))
swe_inds <- length(unique(filter(swe_fin, Country == "SWE")$Indicator))
fin_inds <- length(unique(filter(swe_fin, Country == "FIN")$Indicator))
```

```
swe_fin_summary <- kable(data.frame("Number of Years" = c(swe_years, fin_years),
                                    "Number of Indicators" = c(swe_inds, fin_inds),
                                    row.names = c("Sweden", "Finland")))
```

In the complete data set there was a total of 58537380 possible cells for data, only 5080738 actually contained data, the rest were NA.

After tidying the data, there were still 58537380 before removing NAs and then 5080738 after removing the NAs.

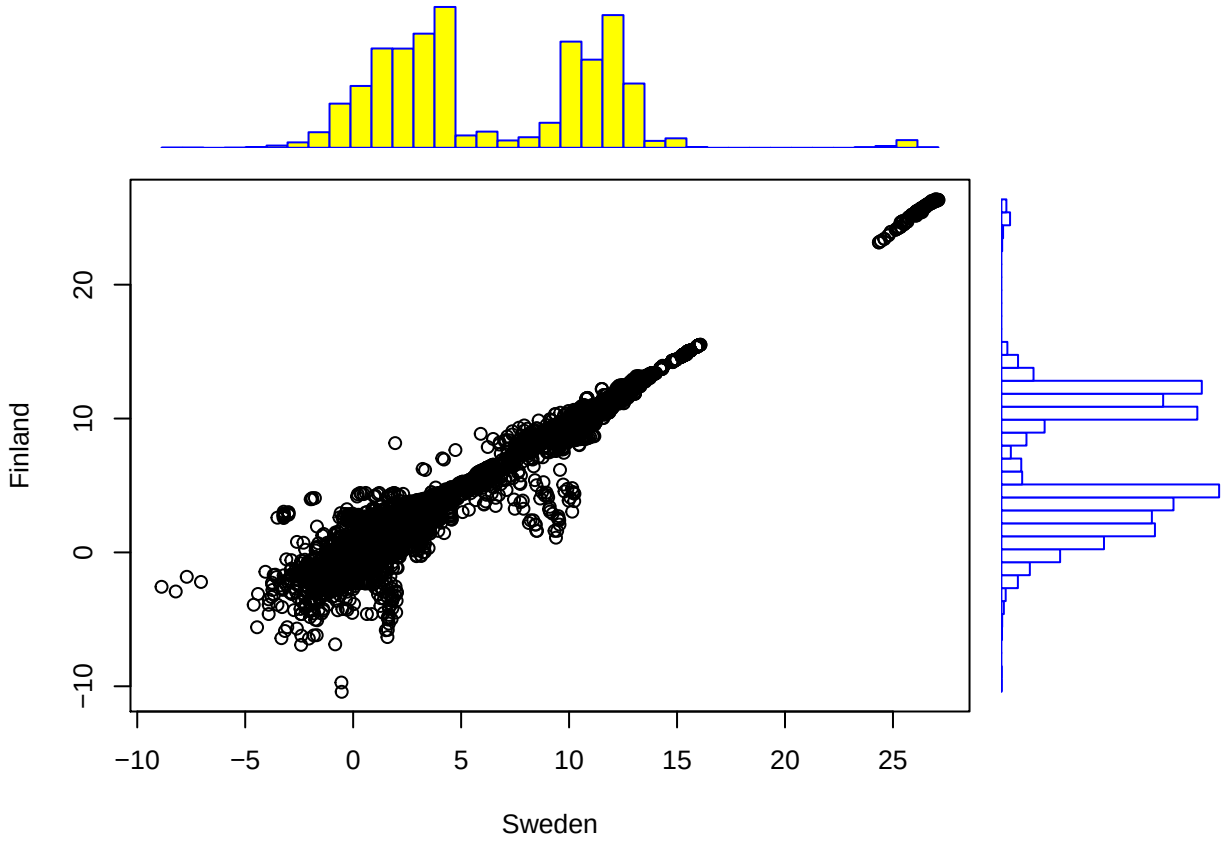
Below is a table summarizing the number of years each country has data for and the number of unique identifiers for each country.

	Number.of.Years	Number.of.Indicators
Sweden	65	1964
Finland	65	2040

Problem 4

```
scatterhist <- function(x, y, xlab = "", ylab = ""){
  zones <- matrix(c(2, 0, 1, 3), ncol = 2, byrow = TRUE)
  layout(zones, widths = c(4/5, 1/5), heights = c(1/5, 4/5))
  xhist <- hist(x, breaks = 50, plot = FALSE)
  yhist <- hist(y, breaks = 50, plot = FALSE)
  top <- max(c(xhist$counts, yhist$counts))
  par(mar = c(4, 4, 1, 1))
  plot(x, y, xlab = xlab, ylab = ylab)
  par(mar = c(0, 4, 1, 1))
  barplot(xhist$counts, axes = FALSE, ylim = c(0, top), space = 0,  col = "yellow", border = "blue")
  par(mar = c(4, 0, 1, 1))
  barplot(yhist$counts, axes = FALSE, xlim = c(0, top), space = 0, horiz = TRUE, col = "white",
          border = "blue")
}

scatterhist(log(swe$value), log(fin$value), xlab = "Sweden", ylab = "Finland")
```

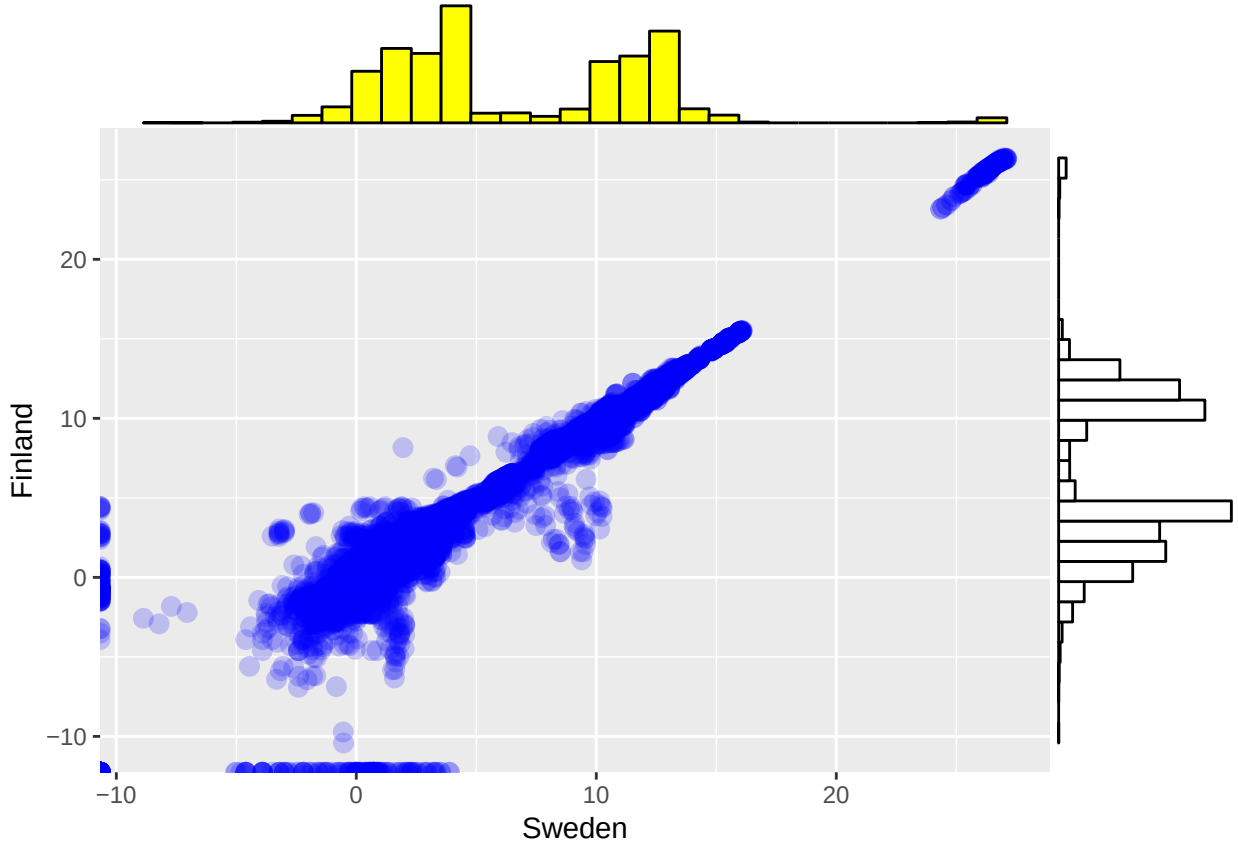


I took the log of the value for both Sweden and Finland because there were both very small and very large values in the data.

Problem 5

```
swe_vals <- swe$value
fin_vals <- fin$value
vals <- data.frame(Sweden = log(swe_vals), Finland = log(fin_vals))

g <- ggplot(vals) + geom_point(aes(Sweden, Finland), color = "Blue", size = 3, alpha = .2)
g2 <- ggMarginal(g, type = "histogram", xparams = list(fill = "Yellow"),
                 yparams = list(fill = "White"))
g2
```



References

Code for creating the plot for Problem 3 was adapted from the site below:

```
https://www.r-bloggers.com/example-8-41-scatterplot-with-marginal-histograms/
```