Untitled

L. Austin Hadamuscin

3/10/2022

Packages

```
library(dplyr)
```

Data

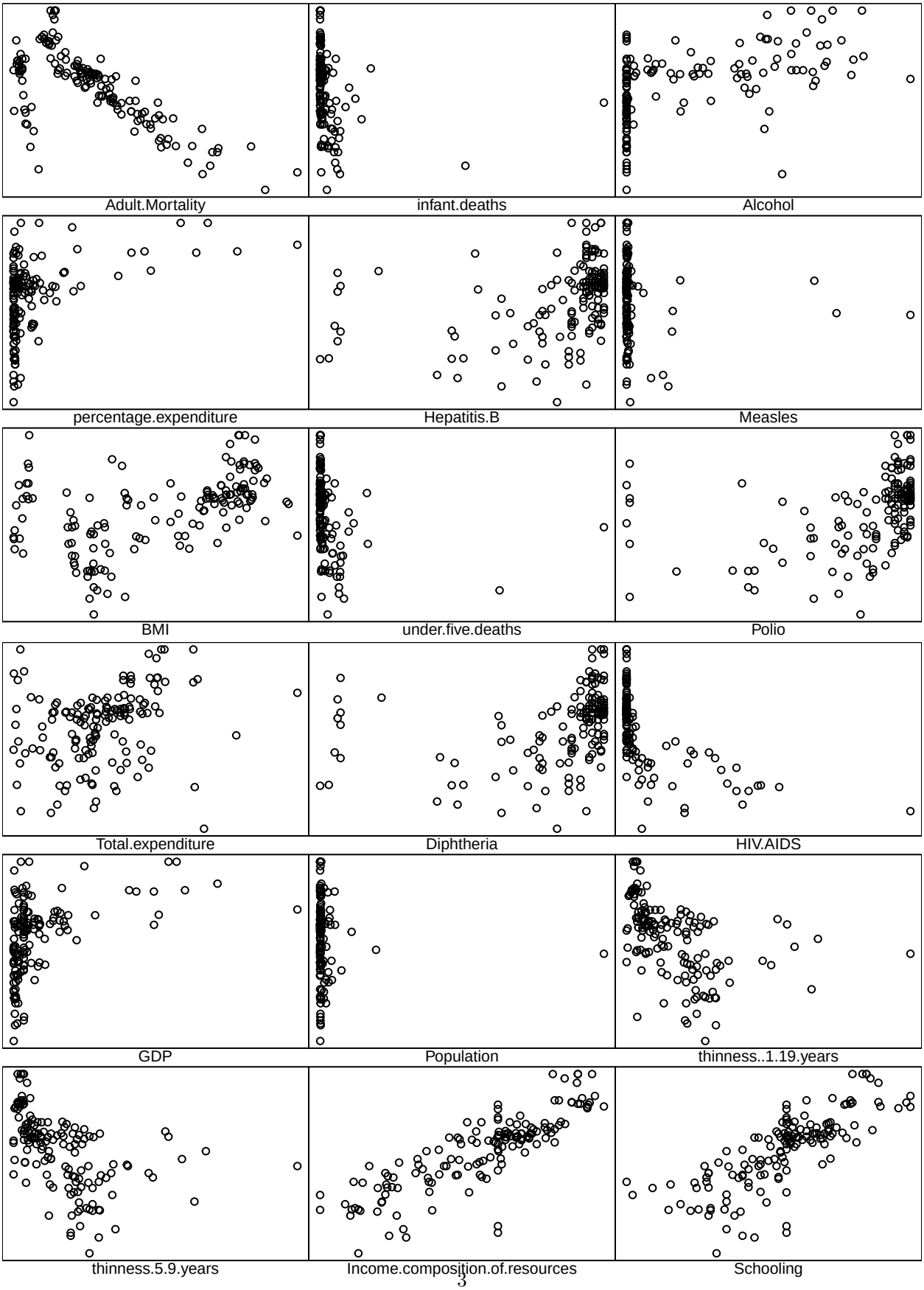
```
dat <- read.csv("Life Expectancy Data.csv") |>
         filter(Year==2014) |>
         select(-Year)
```

Partitioning The Data

```
RNGkind(sample.kind = "Rounding")
set.seed(0)
# Combined Training and validation data
trainval_ind <- sample(1:nrow(dat), size = ceiling(0.80*nrow(dat))) %>% sort()

trainval_dat <- mutate_if(dat[trainval_ind,],
                          is.numeric,
                          function(x) ifelse(is.na(x), median(x, na.rm = T), x))


# Training Data
train_ind <- sample(trainval_ind, size = ceiling(0.60*nrow(dat))) %>% sort()

train_dat <- dat[train_ind,]


# Validation Data
val_dat <- dat[-train_ind,]


# Testing Data
test_dat <- cbind(dat[-trainval_ind,1:2],

                  sapply(1:ncol(dat[,-c(1,2)]), function(i){
```

```
                    ifelse(dat[-trainval_ind,i+2] %>% is.na(),
                           median(dat[trainval_ind,i+2], na.rm = T),
                           dat[-trainval_ind,i+2])

                    }) %>% as.data.frame()

                )
names(test_dat) <- names(trainval_dat)
```

Plotting Life Expectancy vs Explainatory Variables. (linearity check)



Adjusting the data

```
dat_trans <- cbind(dat[,c(1,2)],
                   data.frame(Life.expectancy      = dat$Life.expectancy,
                        Adult.Mortality_sqrt       = dat$Adult.Mortality^0.5,
                        infant.deaths_p0.5_log     = log(dat$infant.deaths+0.5),
                        Alcohol_sqrt               = dat$Alcohol^.5,
                        percentage.expenditure_p0.5_log = log(dat$percentage.expenditure+0.5),
                        Hepatitis.B_sqrt           = dat$Hepatitis.B^5,
                        Measles_p0.5_log           = log(dat$Measles+0.5),
                        BMI                        = dat$BMI,
                        under.five.deaths_p0.5_log      = log(dat$under.five.deaths+0.5),
                        Polio_5                    = dat$Polio^5,
                        Total.expenditure_sqrt     = dat$Total.expenditure^.5,
                        Diphtheria_5               = dat$Diphtheria^5,
                        HIV.AIDS_log               = log(dat$HIV.AIDS),
                        GDP_log                    = log(dat$GDP),
                        Population_log             = log(dat$Population),
                        thinness..1.19.years_log   = log(dat$thinness..1.19.years),
                        thinness.5.9.years_sqrt    = sqrt(dat$thinness.5.9.years),
                        Income.composition.of.resources = dat$Income.composition.of.resources,
                        Schooling                  = dat$Schooling)
)
```

Partitioning the Transformed Data

```
trainval_dat_trans <- dat_trans[trainval_ind,]
train_dat_trans <- dat_trans[train_ind,]
val_dat_trans <- dat_trans[-train_ind,]
test_dat_trans <- dat_trans[-trainval_ind,]
```

Plotting Life Expectancy vs Transformed Explainatory Variables (linearity check)