p8130_FinalProject

Jiaqi Li

12/9/2017

```
library(tidyverse)
library(janitor)
library(psych)
library(knitr)
```

```
ghc <- readxl::read_excel("GHProject_Dataset.xlsx") %>%
  clean_names()
```

Each VisitID represents a unique visit. However, it is possible that a patient visited the hospital more than once. Summarize the number of visits per patient and if multiple visits per patient, select the first visit (by date) recorded.

```
ghc %>%
  group_by(patientid) %>%
  summarize(n = n())  ## summarize the number of visits per patient
```

```
## # A tibble: 3,612 x 2
##    patientid     n
##        <dbl> <int>
##  1       185     1
##  2       748     1
##  3      1572     1
##  4      1837     1
##  5      2513     1
##  6      4121     1
##  7      4956     1
##  8      5514     1
##  9      6445     1
## 10      6674     1
## # ... with 3,602 more rows
```

```
ghc <- ghc[!duplicated(ghc$patientid),]  ## if multiple visits per patient, select the first visit (by d

factor_var <- c("is30dayreadmit", "mews", "cindex", "icu_flag")
ghc[, factor_var] <- lapply(ghc[, factor_var], factor) ## convert to factor variables
```

```
## descriptive summary for continuous data
ghc %>%
  describe() %>%
  .[-c(1, 2, 5:8, 10, 12:18 ), -c(1, 2, 6, 7, 11, 12, 13)] %>%
  kable()
```

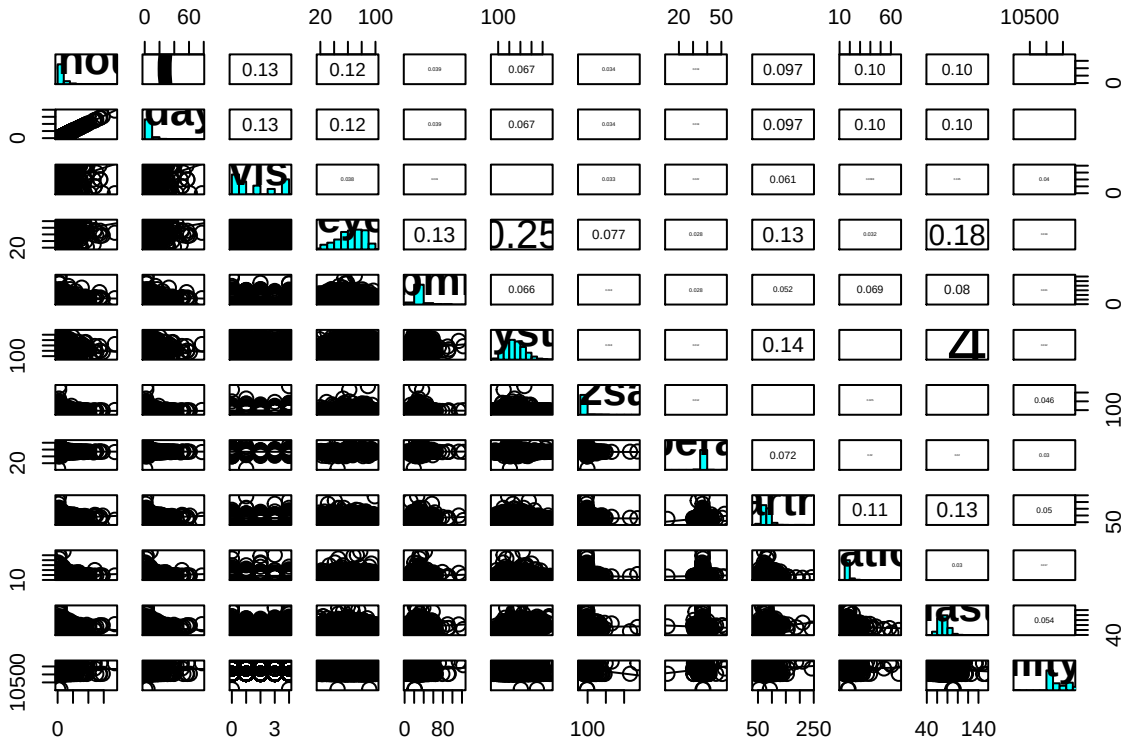
	mean	sd	median	min	max	range
loshours	131.067829	142.0840262	92.000000	1.000000e+00	2111.00000	2110.00000
losdays2	5.461160	5.9201678	3.833333	4.166670e-02	87.95833	87.91667
evisit	1.753599	1.5766841	1.000000	0.000000e+00	4.00000	4.00000
ageyear	65.687154	18.6905338	68.000000	1.800000e+01	105.00000	87.00000
bmi	28.350708	7.9909843	27.100000	3.100000e+00	122.65000	119.55000
bpsystolic	130.551176	16.7150773	129.217391	8.878261e+01	193.96296	105.18035

	mean	sd	median	min	max	range
o2sat	97.861452	4.9084193	97.585366	8.000000e+01	236.52632	156.52632
temperature	36.729473	0.8989632	36.728571	1.185000e+01	52.27500	40.42500
heartrate	80.069760	12.9998580	79.200000	3.758333e+01	242.58333	205.00000
respirationrate	18.195561	2.6330404	17.760000	1.200000e+01	67.71795	55.71795
bpdiastolic	72.520989	9.7982812	71.846154	2.956349e+01	154.40000	124.83651
facilityzip	11317.998339	311.0451673	11375.000000	1.007500e+04	11803.00000	1728.00000

```
colnames(ghc)[c(5:8, 10, 12:18 )] ## These are indicator variables
```

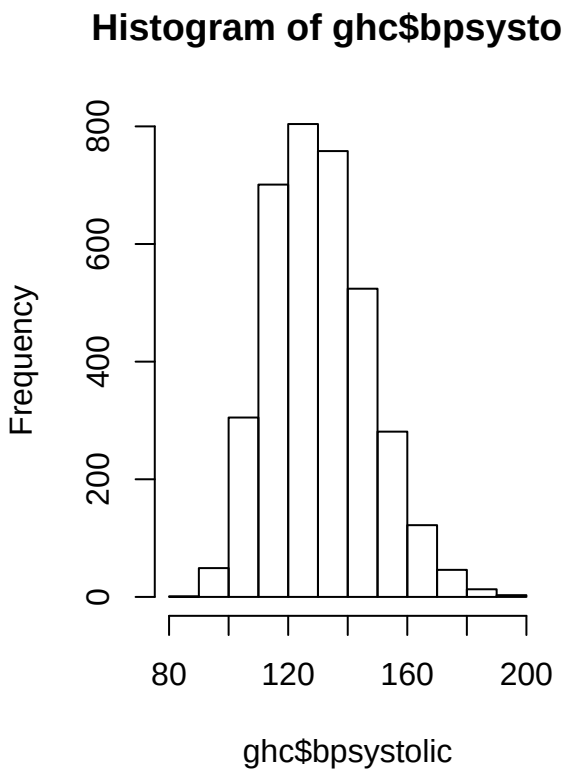
```
##  [1] "is30dayreadmit" "admitdtm"       "mews"           "cindex"
##  [5] "icu_flag"       "postalcode"     "gender"         "race"
##  [9] "religion"       "maritalstatus"  "facilityname"   "insurancetype"
```

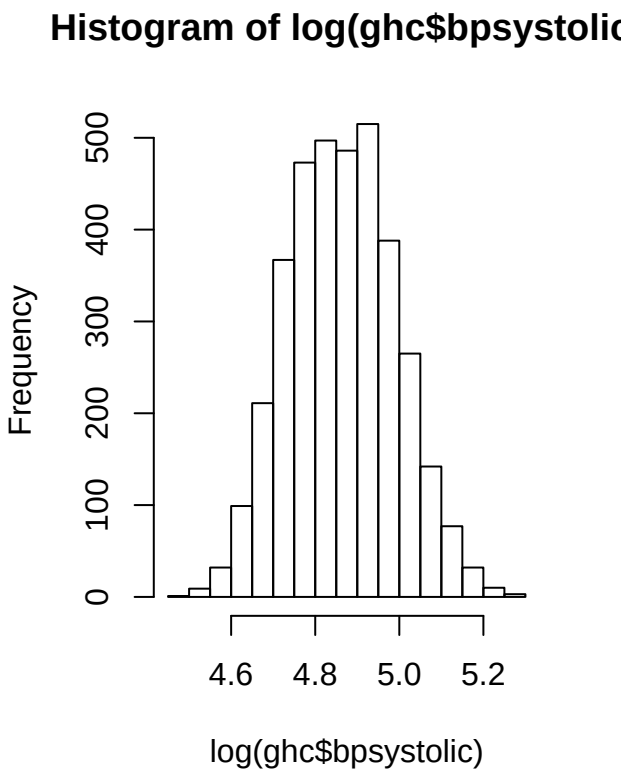
```
ghc_conti <- ghc[, -c(1, 2, 5:8, 10, 12:18 )] %>%
  drop_na()

panel.hist <- function(x, ...){
  usr <-par("usr"); on.exit(par(usr))
  par(usr =c(usr[1:2], 0, 1.5) )
  h <-hist(x, plot = FALSE, 8)
  breaks <- h$breaks; nB <-length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...){
  usr <-par("usr");on.exit(par(usr))
  par(usr =c(0, 1, 0, 1))
  r <-abs(cor(x, y))
  txt <-format(c(r, 0.123456789), digits = digits)[1]
  txt <-paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 1.2/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor*r*5)
}

my_line <- function(x,y,...){
  points(x,y,...)
  abline(a =lm(y~x)$coefficients[1] , b =lm(y~x)$coefficients[2] , ...)
}

pairs(ghc_conti, lower.panel = my_line,
      upper.panel = panel.cor,
      cex = 1.5, pch = 1, bg = "light blue",
      diag.panel = panel.hist, cex.labels = 2, font.labels = 2)
```



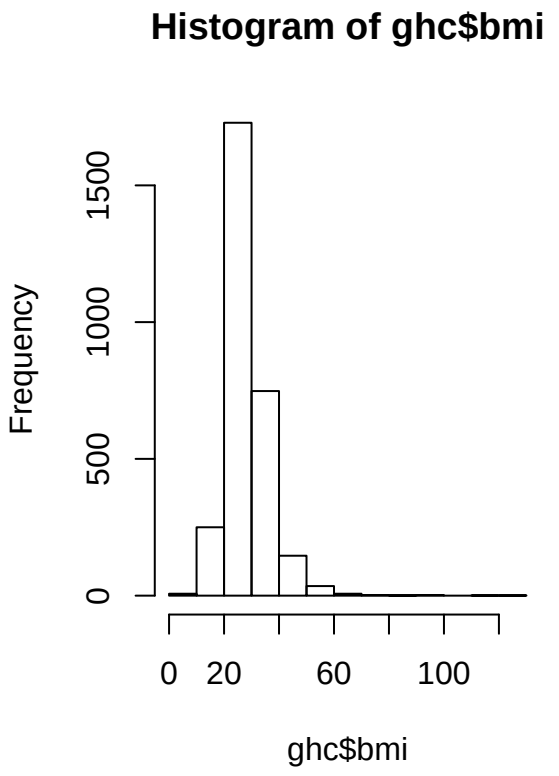
```
## We can see from the panel that the correlation between loshours and losdays2 is 1. It means they are

ghc_conti <- ghc_conti[, -2]
```

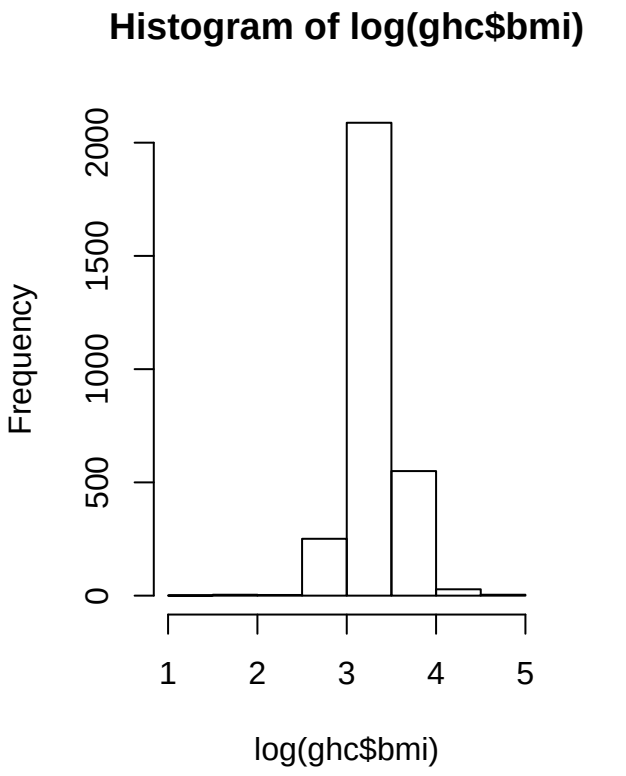
```
par(mfrow = c(1, 2))

hist(ghc$bpsystolic)
hist(log(ghc$bpsystolic))
```



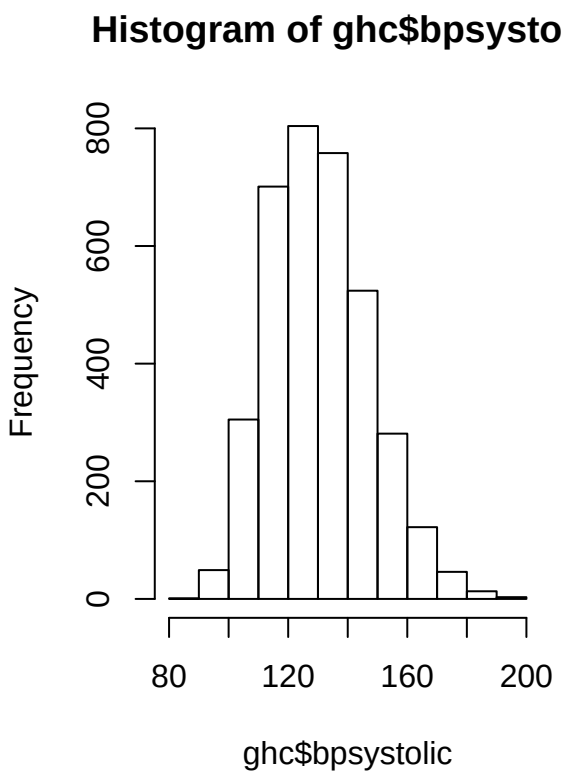


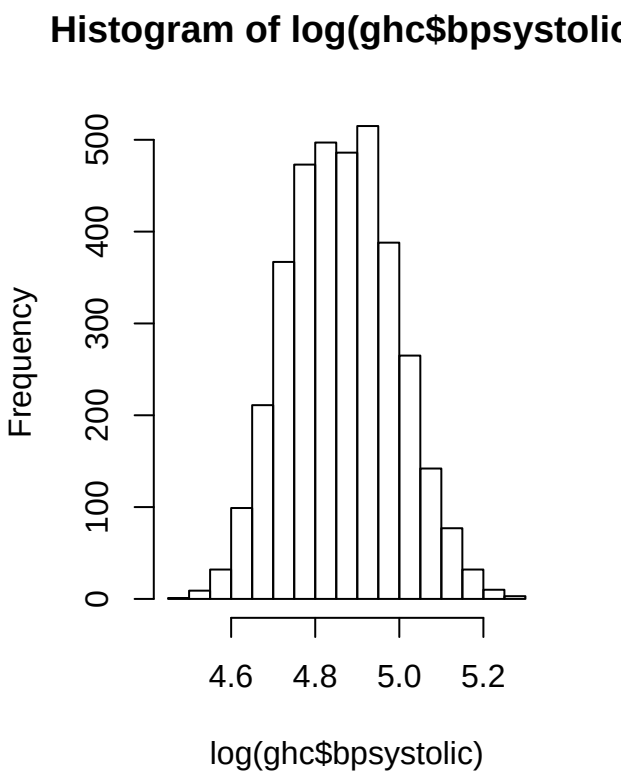

```
hist(ghc$bmi)
hist(log(ghc$bmi))
```



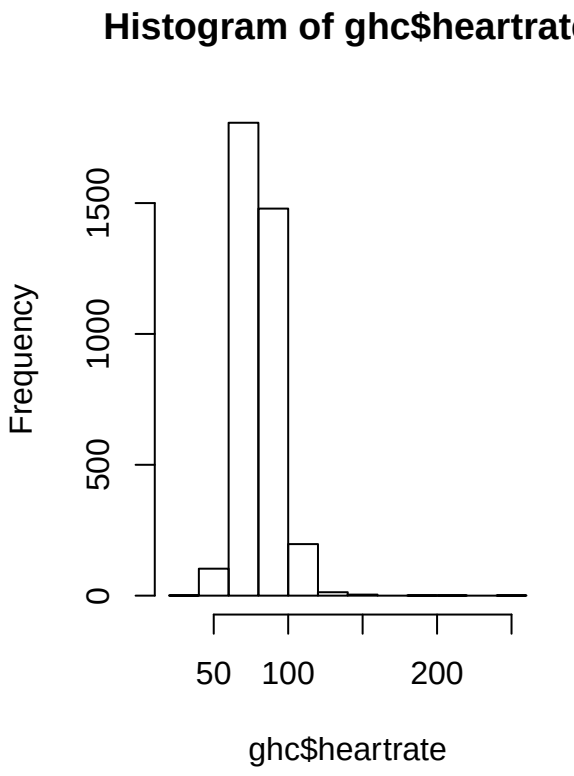


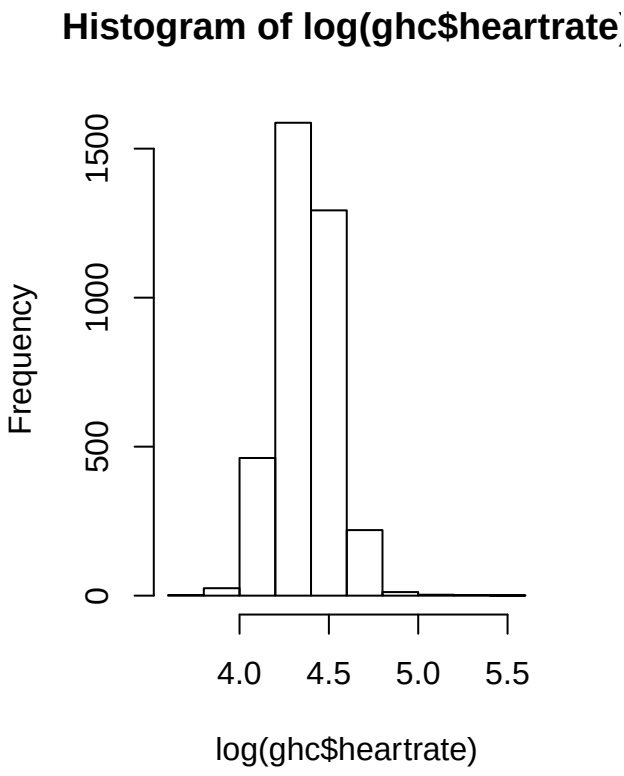

```
hist(ghc$bpsystolic)
hist(log(ghc$bpsystolic))
```



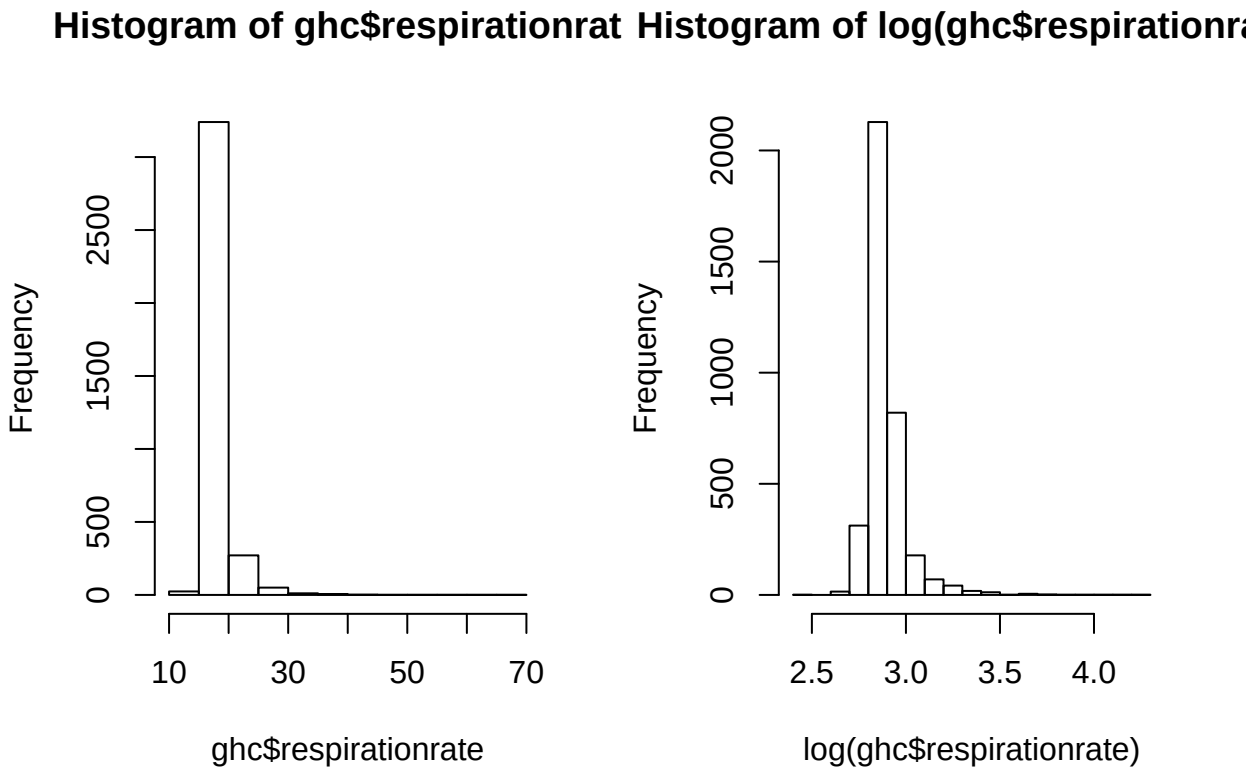



```
hist(ghc$heartrate)
hist(log(ghc$heartrate))
```

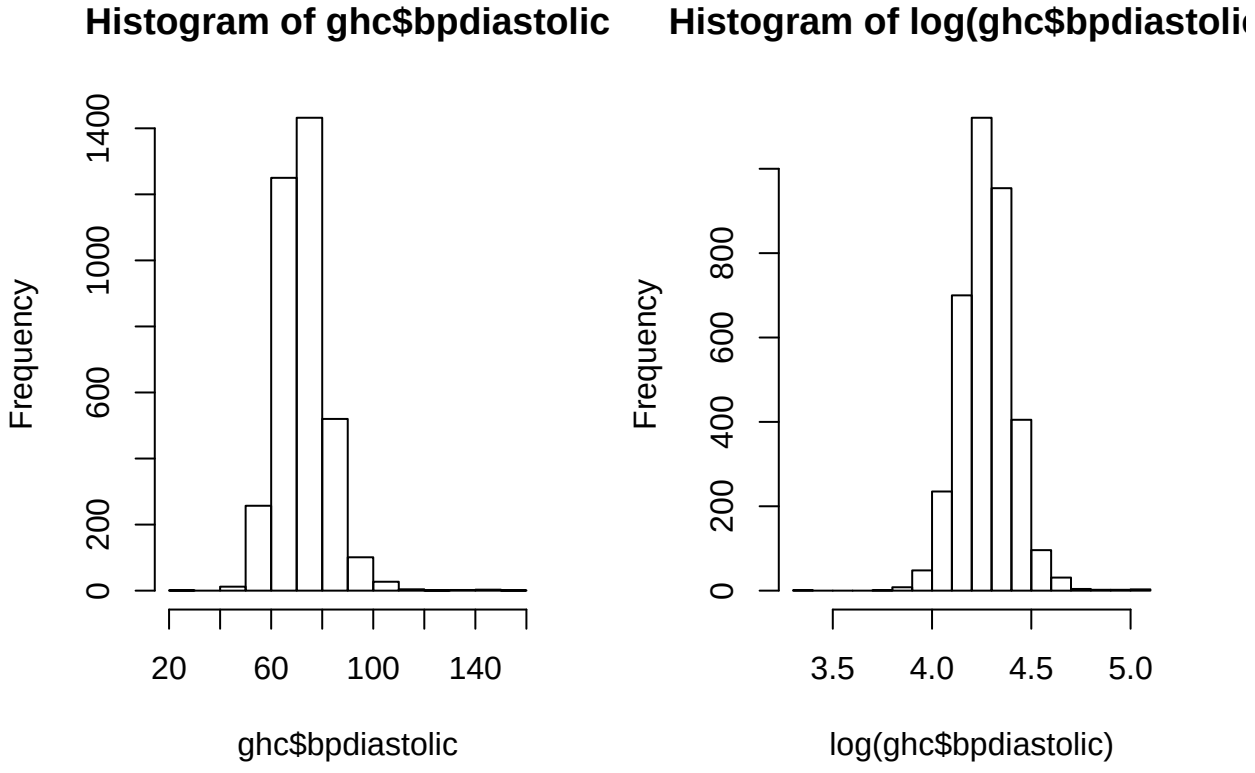





```
hist(ghc$respirationrate)
hist(log(ghc$respirationrate)) ## still skewed after transformation
```




```
hist(ghc$bpdiastolic)
hist(log(ghc$bpdiastolic)) ## transformation may not need?
```



Besides these, I also tried several transformations with *evisit*, but none of them looks good.