BM_HW4

Coco

11/9/2018

Problem 2

Here is the **code chunk** to load the data file

```
heartdisease_df <- read_csv(file = "./data/HeartDisease.csv") %>%
  mutate(gender = as.numeric(gender))
```

a

There are in total 788 observations and 10 variables. The main outcome is 'total cost'. The main predictor is the 'number of emergency room (ER) visits', which is 'ERvisits' in the data set. The other important covariants including 'age', 'gender', 'number of complications' that arose during treatment, and 'duration of treatment condition', which in the data are indicated as 'age', 'gender', 'complications' and 'duration' repectively.

The minimum, first quatile, median, mean, third quatile and maximum value of age, number of complications and duration of treatment conditions are shown below repectively.

```
summary(heartdisease_df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.00   55.00   60.00   58.72   64.00   70.00
```

```
summary(heartdisease_df$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   41.75  165.50  164.03  281.00  372.00
```

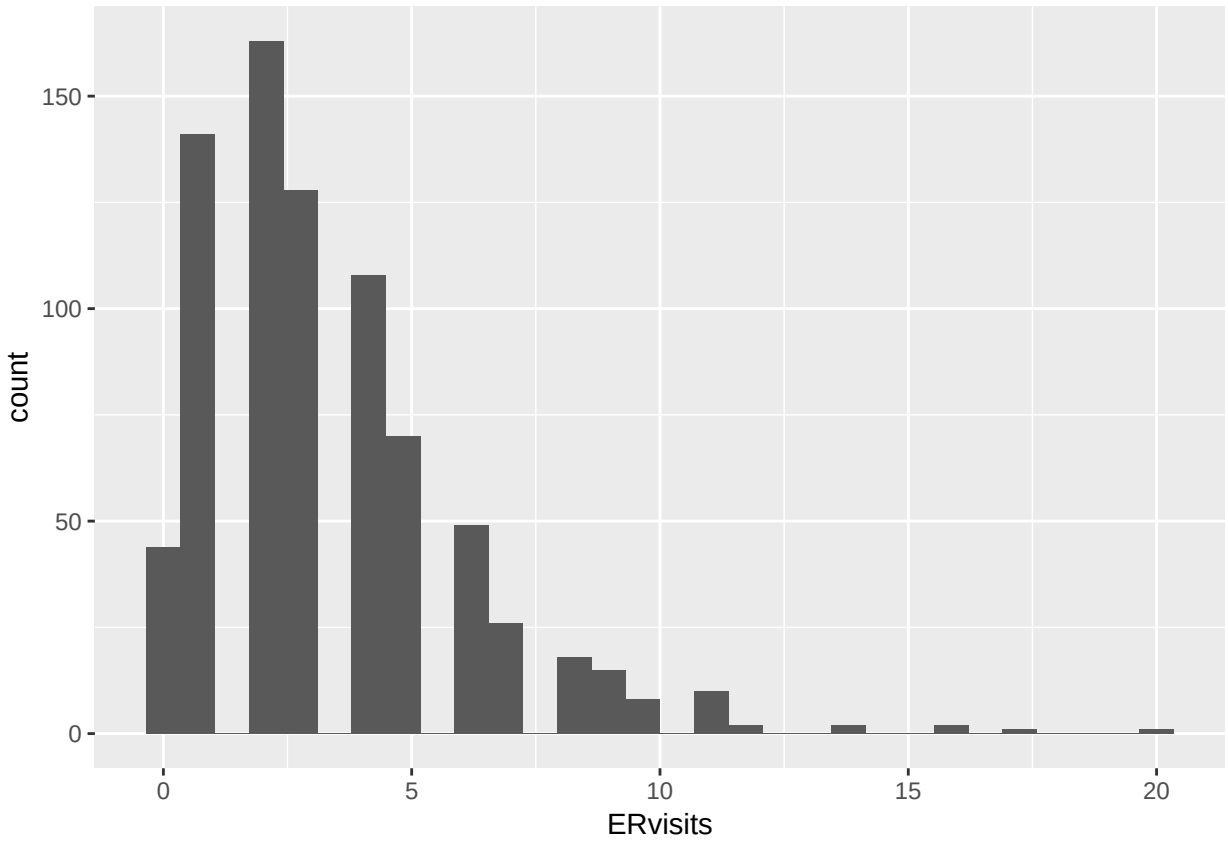
In the 788 observations, there are xx male and xx female. Also, there are

b

Investigate the shape of the distribution for variable 'total cost' and try different transformations, if needed.

```
heartdisease_df %>%
  ggplot(aes(x = ERvisits))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



c

already done?

d