hw8

Shuangshuang Xu

2019/11/4

```
#unzip("C:/Users/44653/Desktop/gitfile/Edstats_csv.zip", exdir = "C:/Users/44653/Desktop/gitfile")
data <- read.csv("C:/Users/44653/Desktop/gitfile/EdStatsData.csv", header = TRUE, skip = 1, stringsAsFa
TF <- !is.na(data)
# delete colums are all NA
data1 <- data[,apply(as.matrix(TF), 2, sum) != 0]
# delete row with all NA except first four columns.
data1 <- data1[apply(as.matrix(TF), 1, sum) != 4, ]
```

There are 886929 observations in the complete dataset.

In my cleaned dataset, there are 357405 observations.

```
t1 <- summary(data1[data1$Arab.World=="Vietnam",-1:-4])
kable(t1, caption = "summary for Vietnam")
```

X	X.1	X.2	X.3	X.4	X.5
Min. : 0	Min. : 2	Min. : 2	Min. : 2	Min. : 2	Min. : 0
1st Qu.: 1	1st Qu.: 59	1st Qu.: 59	1st Qu.: 58	1st Qu.: 58	1st Qu.: 1
Median : 4	Median : 1560026	Median : 1622813	Median : 1678224	Median : 1726474	Median : 5
Mean : 547383	Mean : 4608829	Mean : 4734003	Mean : 4857771	Mean : 4979666	Mean : 622219
3rd Qu.: 41	3rd Qu.: 3564034	3rd Qu.: 3829140	3rd Qu.: 4028418	3rd Qu.: 4087553	3rd Qu.: 41
Max. :42729000	Max. :43725000	Max. :44758000	Max. :45825000	Max. :46918000	Max. :48030000
NA's :1755	NA's :2115	NA's :2115	NA's :2115	NA's :2115	NA's :1755

```
t1 <- summary(data1[data1$Arab.World=="Vietnam",-1:-4])
kable(t1, caption = "summary for Vietnam")
```

X	X.1	X.2	X.3	X.4	X.5
Min. : 0	Min. : 2	Min. : 2	Min. : 2	Min. : 2	Min. : 0
1st Qu.: 1	1st Qu.: 59	1st Qu.: 59	1st Qu.: 58	1st Qu.: 58	1st Qu.: 1
Median : 4	Median : 1560026	Median : 1622813	Median : 1678224	Median : 1726474	Median : 5
Mean : 547383	Mean : 4608829	Mean : 4734003	Mean : 4857771	Mean : 4979666	Mean : 622219
3rd Qu.: 41	3rd Qu.: 3564034	3rd Qu.: 3829140	3rd Qu.: 4028418	3rd Qu.: 4087553	3rd Qu.: 41
Max. :42729000	Max. :43725000	Max. :44758000	Max. :45825000	Max. :46918000	Max. :48030000
NA's :1755	NA's :2115	NA's :2115	NA's :2115	NA's :2115	NA's :1755

```
# use the data from Norway, plot "X" and "X.1", delete rows with NA
data_plot <- data1[data1$Arab.World=="Norway", 5:6]
data_plot <- data_plot[apply(is.na(data_plot), 1, sum) == 0, ]
data_plot$X <- log(data_plot$X)
data_plot$X.1 <- log(data_plot$X.1)
```

```
#take log transformation, since data are too sparse

# create plot with histgram
par(fig=c(0,0.8,0,0.8))
plot(data_plot$X, data_plot$X.1, xlab="X",
  ylab="X.1")
par(fig=c(0,0.8,0.55,1), new=TRUE)
hist(data_plot$X, axes=FALSE, main = "marginal dist of X", xlab = NULL, ylab = NULL)
par(fig=c(0.65,1,0,0.8),new=TRUE)
hist(data_plot$X.1, axes=FALSE, main = "marginal dist of X.1", xlab = NULL, ylab = NULL)
```

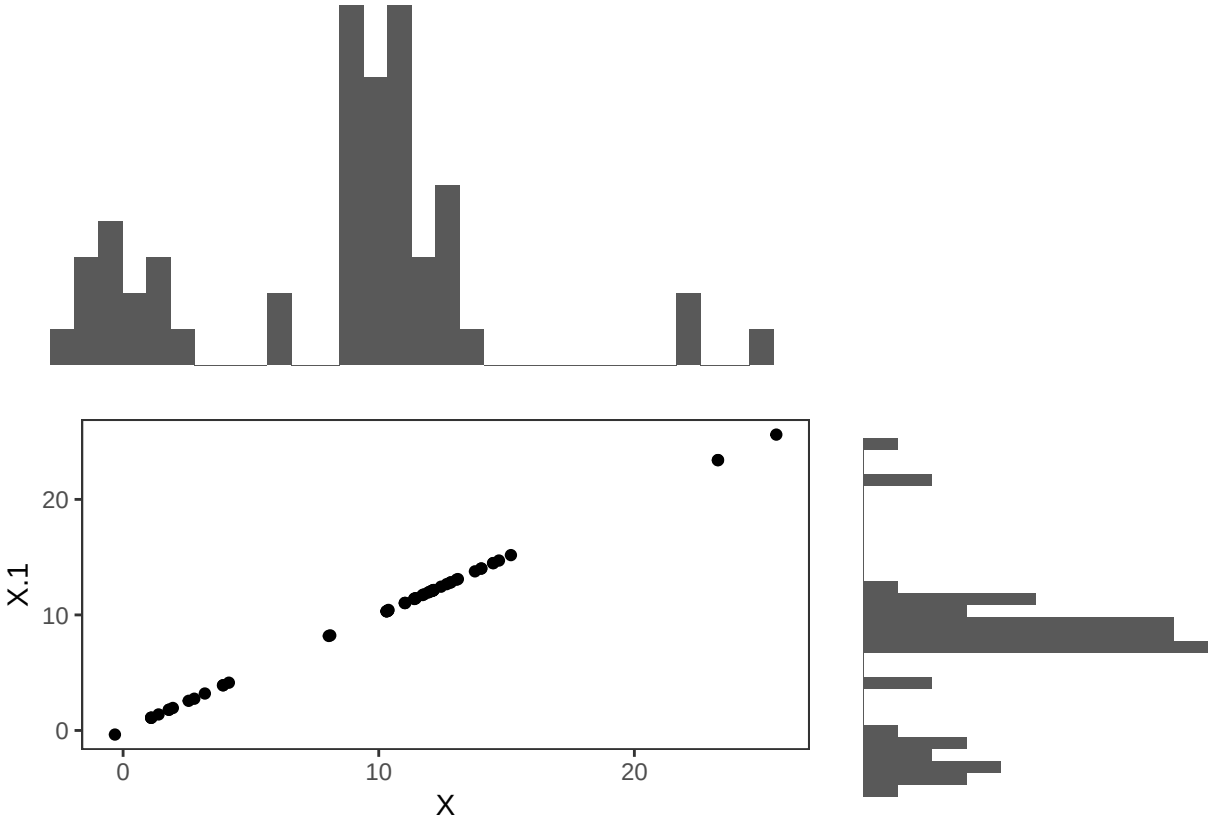

```
par(new=TRUE)
```

```
p1 <-ggplot(data_plot, aes(X, X.1))+geom_point()+ theme_bw() + theme(panel.grid.major = element_blank()
p2 <-qplot(X, data = data_plot) + theme_bw() + theme(panel.grid.major = element_blank(), panel.grid.min

p3 <-qplot(X.1, data = data_plot) + coord_flip() + theme_bw() + theme(panel.grid.major = element_blank()

grid.arrange(
  grobs = list(p2, p1, p3),
  widths = c(1, 0.5),
  layout_matrix = rbind(c(1, NA),
                        c(2, 3))
)
```

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


X.1
20
10
0
0
10
20
X