Assignment-1.R

purba

Thu Sep 28 16:33:22 2017

```
## Q1
 I am a graduate student with specialization in machine learning and natural
language processing. Pursuing this course could help me to understand and ana
lyze the data well

 Completing the sequel of this course with 521 would help me to gain MS minor
in Statistics

I have used R before but this class could help me to learn and use R more
efficiently

## Q2
Random Sampling is a sampling technique where each sample has an equal
probability of getting selected from population. This process is completely
based on chance

Random assignment is an experimental process which allows the experimental
participants to be assigned to treatment or control group using random proced
ure
These two procedures ensures that experiment is fair, avoids any kind of
unfairness and lurking variables that can influence the outcome of the experi
ment


## Q3-Ex-1.16
library(Sleuth3)

## Warning: package 'Sleuth3' was built under R version 3.4.1

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.1

boxplot(ex0116$PerCapitaGDP, main="Per Capita GDP",
        ylab="Gross DOmestic Product Per Capita in $U.S.")
```
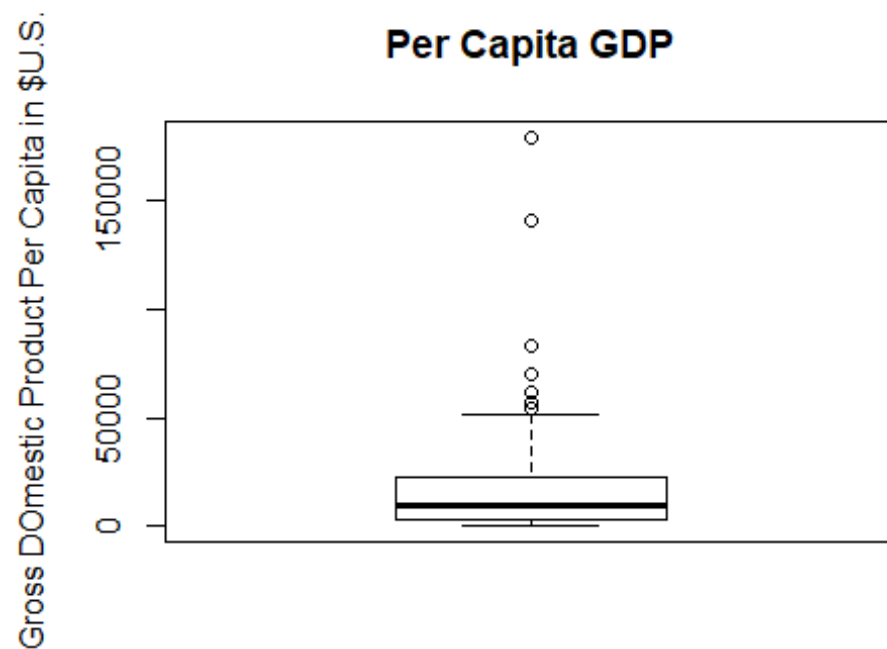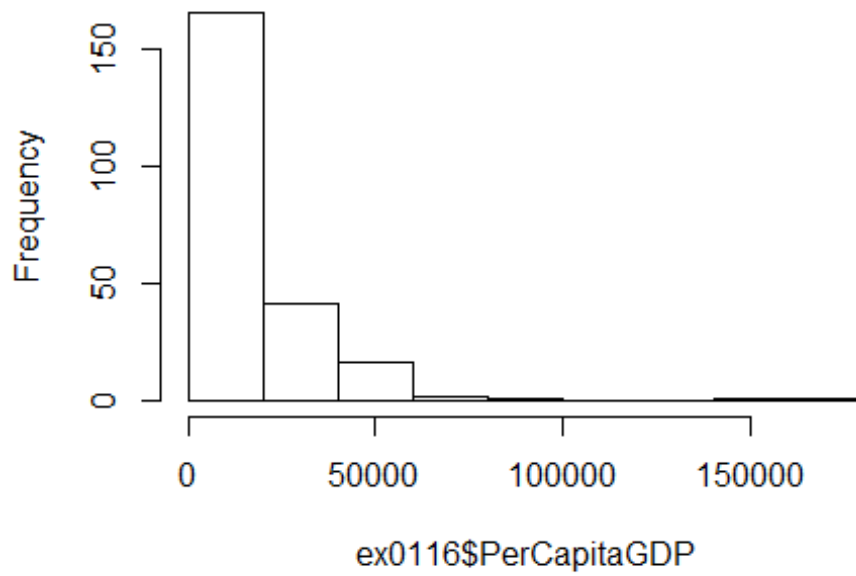
Per Capita GDP

Gross DOmestic Product Per Capita in $U.S.

The display is not very different from the Display 1.11. The slight difference is there in the extreme points, where two of them overlaps after the largest value whereas in Display 1.11, the extreme points are well separated.

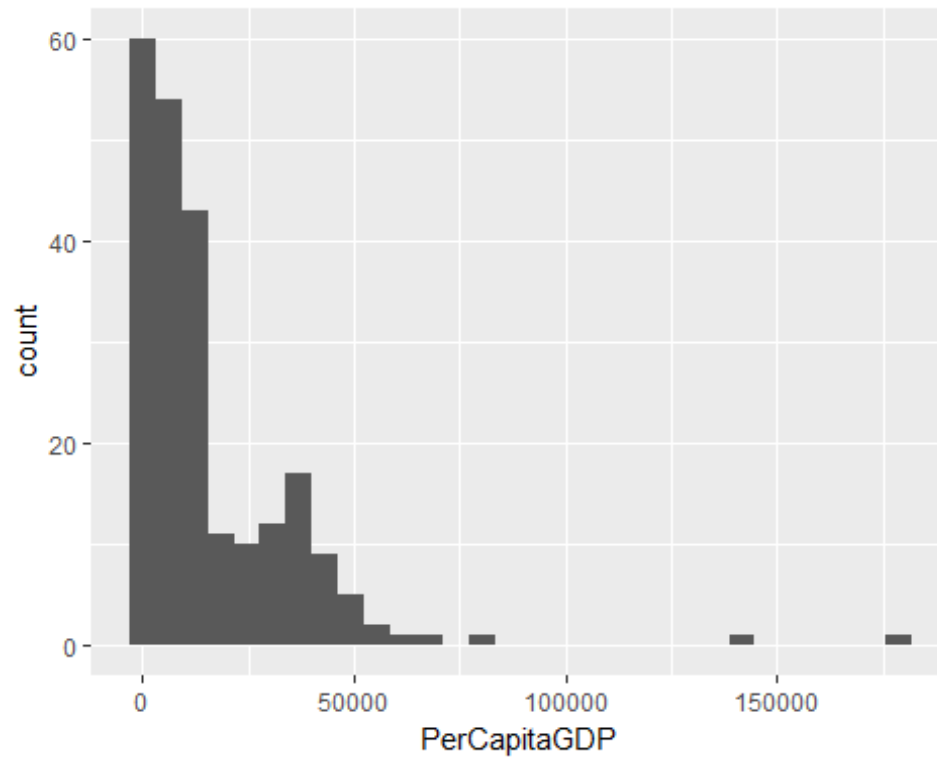```
hist(ex0116$PerCapitaGDP)
```

## Histogram of ex0116$PerCapitaGDP



```
Bin width is $20,000 (9 breaks)

ggplot(data=ex0116, aes(PerCapitaGDP)) + geom_histogram()

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
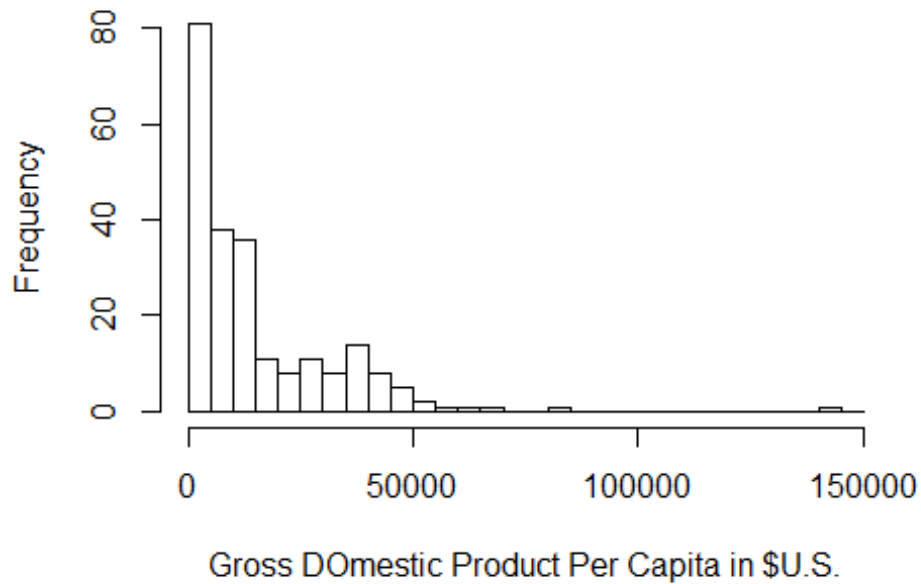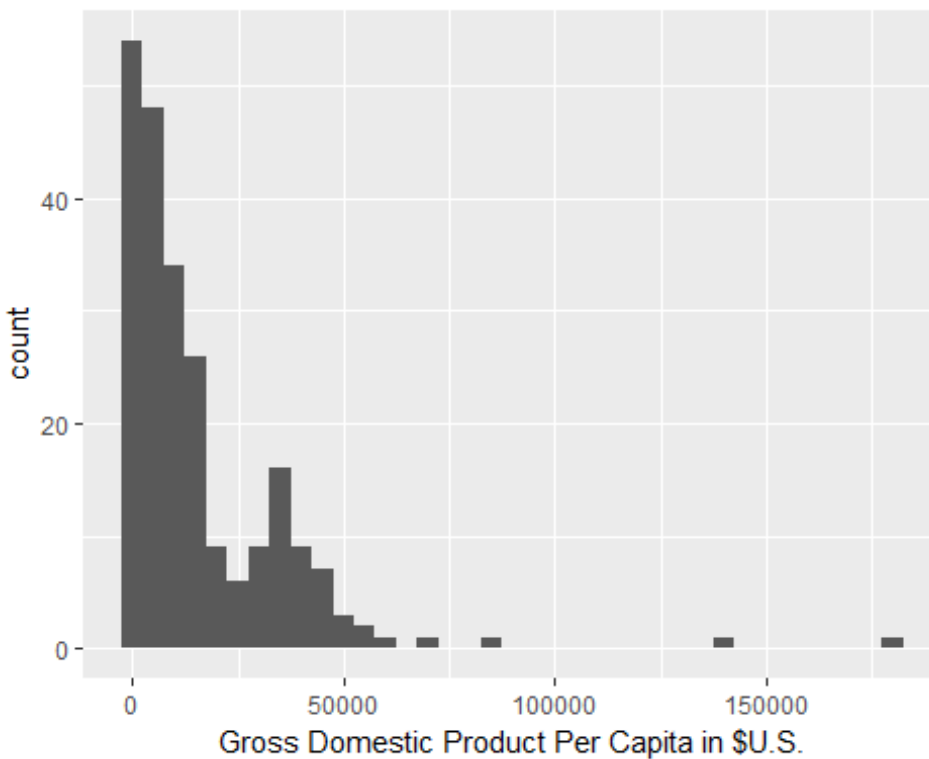
```
##In this case the bin width is $5956.67

hist(ex0116$PerCapitaGDP[ex0116$PerCapitaGDP >= 0 & ex0116$PerCapitaGDP < 150
000],
    xlab="Gross DOmestic Product Per Capita in $U.S.",
    main="Per Capita GDP",breaks=seq(0,150000,by=5000))
```

# Per Capita GDP



Gross DOmestic Product Per Capita in $U.S.

```
##Another method to get the bin width of $5000
p <- ggplot(data=ex0116, aes(PerCapitaGDP)) + geom_histogram(binwidth = 5000)
p + xlab("Gross Domestic Product Per Capita in $U.S.")
```
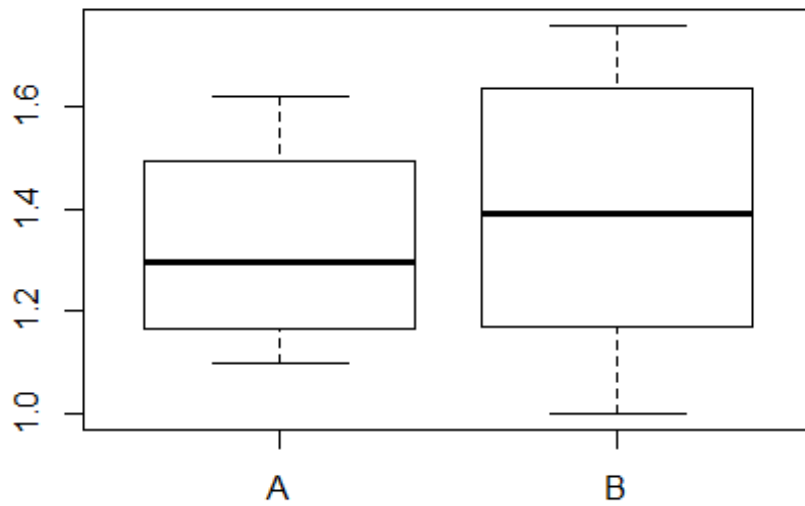
```
Histogram with 30 breaks
hist(ex0116$PerCapitaGDP, xlab="Gross DOmestic Product Per Capita in $U.S.",
     main="Per Capita GDP", breaks=30)
```
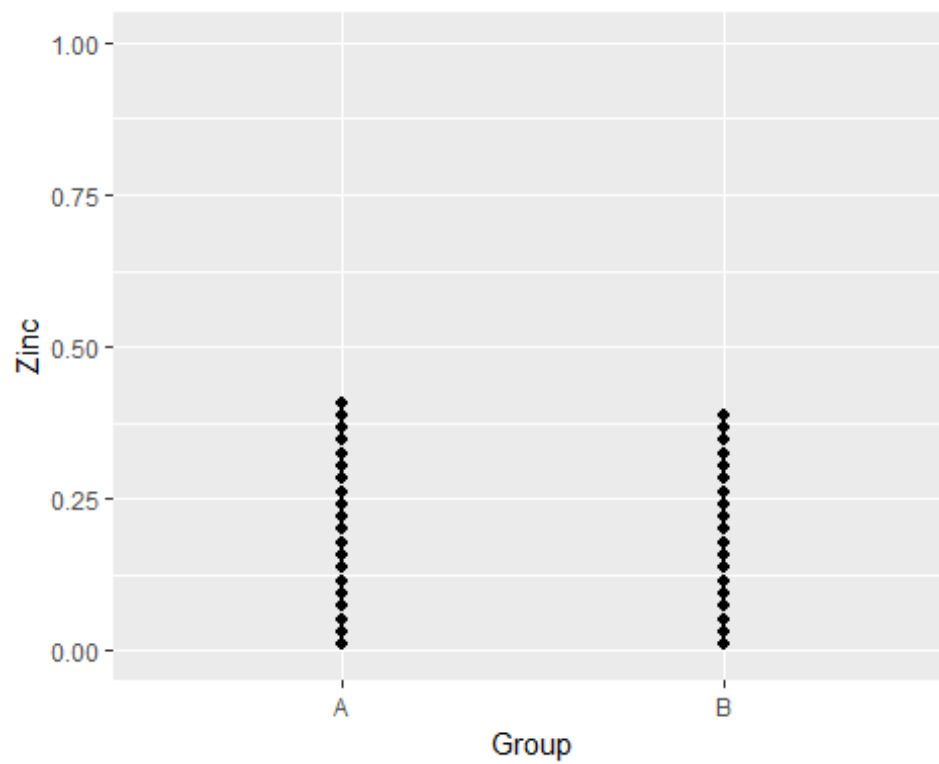
**Per Capita GDP**



Gross DOmestic Product Per Capita in $U.S.
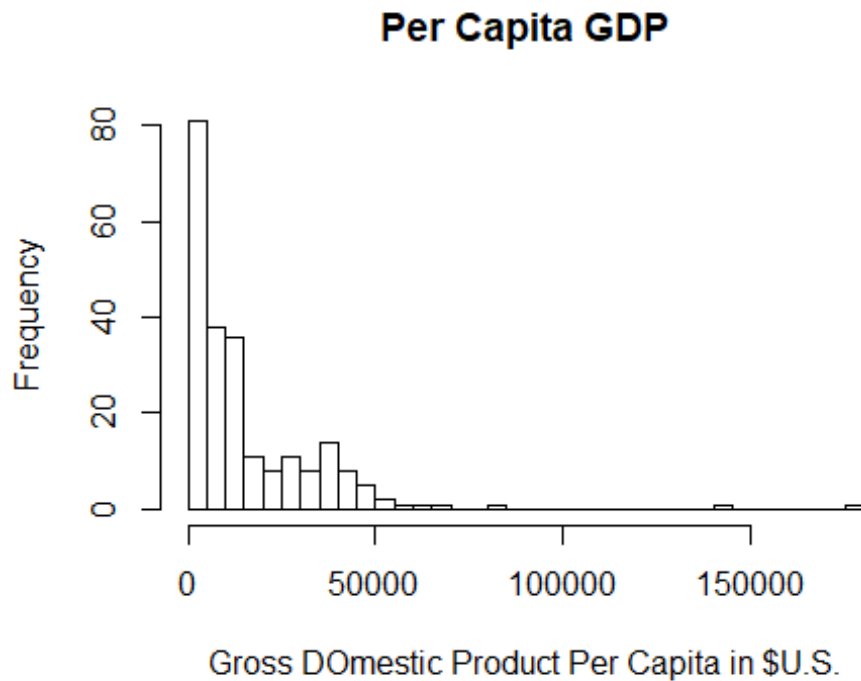
```
## Q4-Ex-1.25

boxplot(ex0125$Zinc~ex0125$Group)
```

```
ggplot(data=ex0125, aes(x=Group, y=Zinc)) + geom_dotplot()
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```
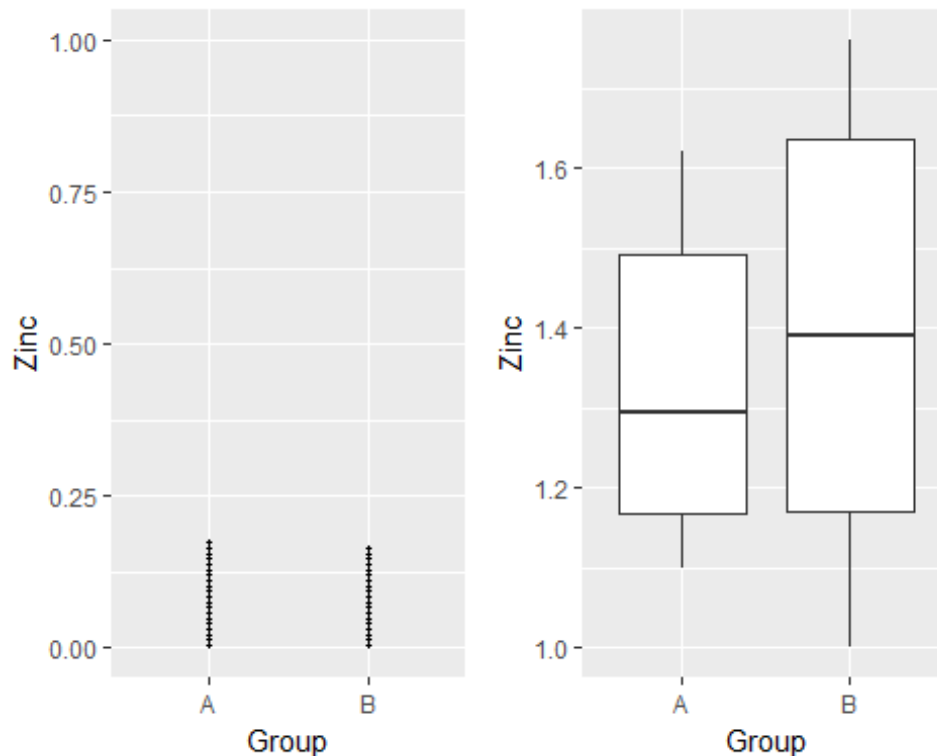
```
require(gridExtra)

## Loading required package: gridExtra

## Warning: package 'gridExtra' was built under R version 3.4.1

plot1 <- ggplot(data=ex0125, aes(x=Group, y=Zinc)) + geom_boxplot()
plot2 <- ggplot(data=ex0125, aes(x=Group, y=Zinc)) + geom_dotplot()
grid.arrange(plot2, plot1, ncol=2)

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


Boxplot is better because it displays the distribution of data values using median, quartiles and extremes. This helps in identifying outliers and compa-ring distribution whereas dotplot are used for continuous and univariate data . It counts the frequency within group, making it difficult to make interpret ation.