STAT 8416 Homework 1

Brian Detweiler

Wednesday, September 24, 2016

1. (a) What is data science? Data science is the act of gathering, cleaning, analyzing, and delivering a data product. It requires in-depth domain knowledge, hacking skills (curiosity and problem solving), and strong mathematical and statistical knowledge.

(b) Explain with an example what you mean by data product. A data product is the end result of a data transformation that contains some value to your customer. Data products can come in many forms, including reports, presentations, production-grade software decision support systems, or even academic publications.

(c) Carefully read the Cleveland's paper shown in lecture 2 and discuss what he suggested about the field of statistics and data science. *TODO*

(d) Explain in a short paragraph how data science is different from computer science. Computer Science is the study of computing; the infrastructure and mechanisms that process data. Data Science is the study of the data itself and deriving meaning from it. Incidentally, it happens to be very convenient to use computers for the analysis. It should be noted that Computer Science can and has existed without Data Science, but it would be very difficult (in all practicality, impossible) for Data Science to exist without Computer Science.

```
getSquare <- function(x) {
  if (x > 100) {
    return("Big number")
  } else {
    return(x^2)
  }
}

getSquare(5)
```

```
## [1] 25
```

```
getSquare(500)
```

```
## [1] "Big number"
```

```
getSquare <- function(x) {

  z <- lapply(x, function(y) {
    if (y > 100) {
      return("Big number")
    } else {
      return(y^2)
    }
  })
  return(z)
  # TODO: Get this unlisted and keep numerics numeric
```

```
}

getSquare(5)
```

```
## [[1]]
## [1] 25
```

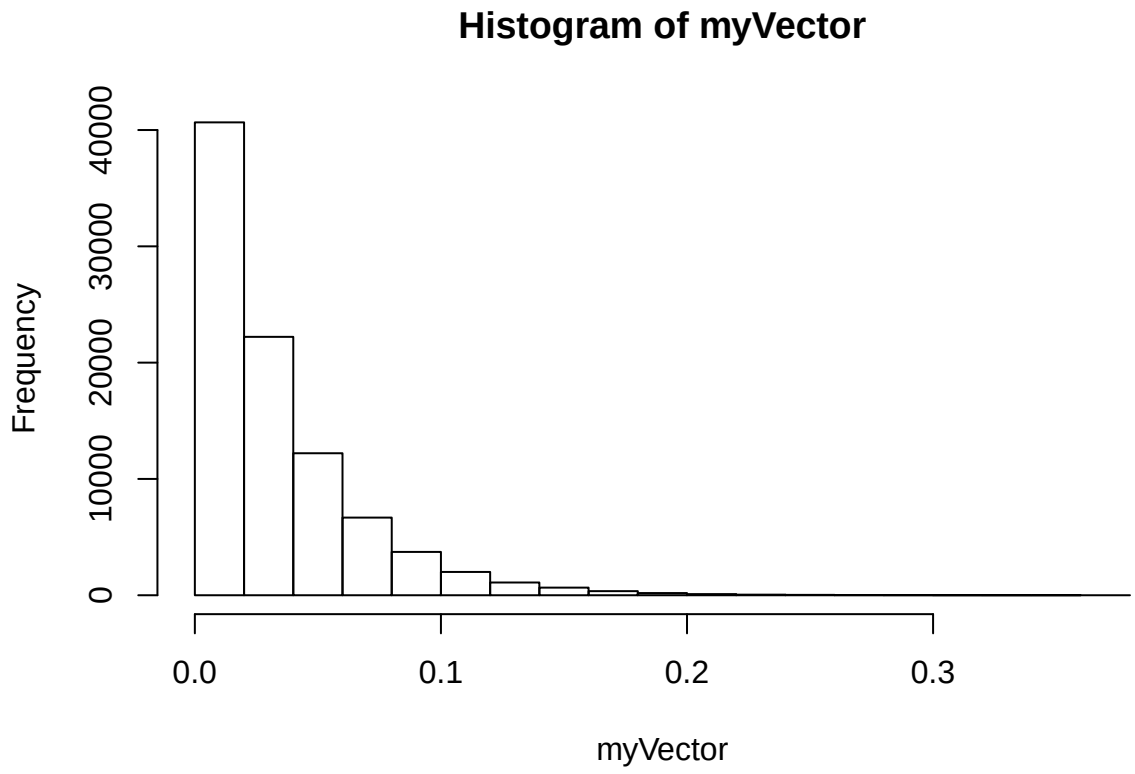
```
getSquare(500)
```

```
## [[1]]
## [1] "Big number"
```

```
x <- c(25, 200)
getSquare(x)
```

```
## [[1]]
## [1] 625
##
## [[2]]
## [1] "Big number"
```

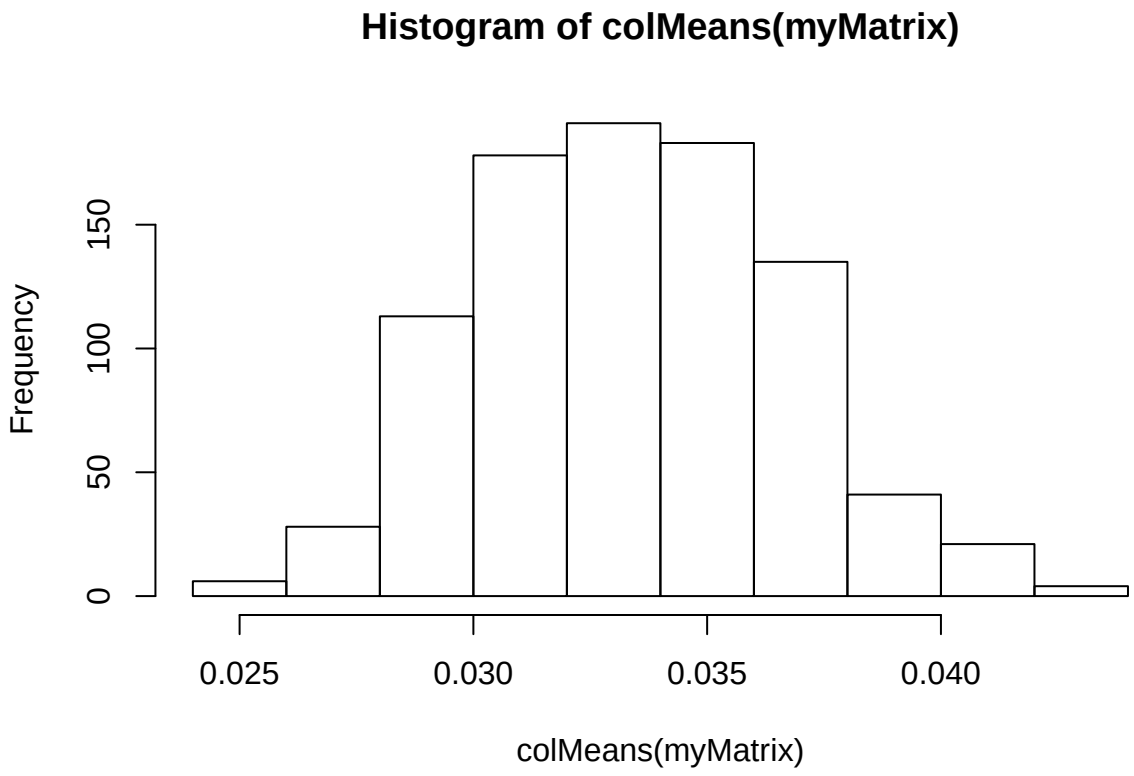
```
myVector <- rexp(90000, r=30)
hist(myVector)
```



```
myMatrix <- matrix(myVector, ncol=900)
dim(myMatrix)
```

```
## [1] 100 900
```

```
hist(colMeans(myMatrix))
```



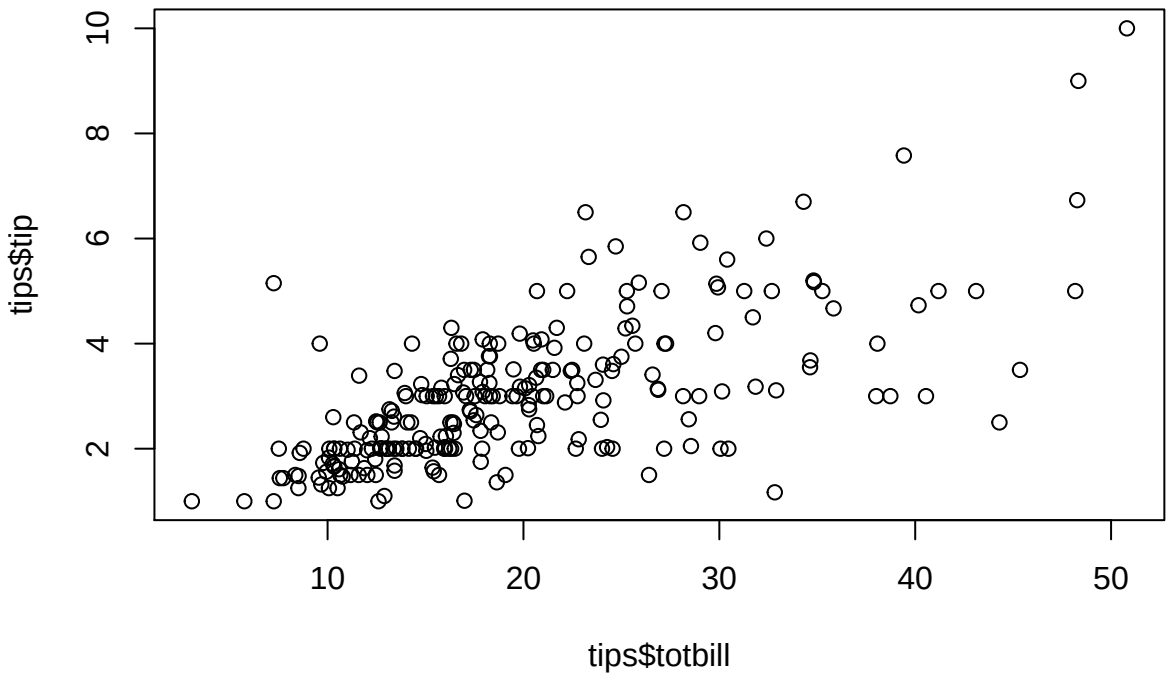
Explain why the two histograms you have created in questions XXX and XXXY are different in shapes.

The two histograms are different shapes because the first follows the exponential distribution, as we would expect. The second is due to the Central Limit Theorem, which states that the sampling distribution of the sample mean approximates a normal distribution, regardless of underlying distribution.

What are the very first few steps one should do once data is loaded onto **R**? Demonstrate that by loading tips data from http://www.ggobi.org/book/data/tips.csv

```
# Note: On my work PC, I have to use a proxy.
if(Sys.info()[1] == "Windows") {
  setInternet2(TRUE)
}

tips <- read.csv('http://www.ggobi.org/book/data/tips.csv')
plot(tips$totbill, tips$tip)
```


tips$tip
2
4
6
8
10
10
20
30
40
50
tips$totbill