Anagha Uppal - MW Section - Homework 1

R Crash Course

by Sept 1, 2015 3:30pm

Question 1: Define a variable `x` to be 2. Translate the following expressions into R and evaluate them.

a. $$0.5(9-3)+3^2 \times \frac{6}{2x+3}$$

```
#Replace with your R code
x <- 2
0.5 * (9-3)+3^2 * ({6}/{2*x+3})
## [1] 10.71429
#10.71429
```

b. $$3x(x-5+2x^4)$$

```
#Replace with your R code
3*x * (x-5+2*x^4)
## [1] 174
#174
```

c. $$\frac{5-2e^{-x/2}}{5+\sqrt{(}\ln x+9)}$$

```
#Replace with your R code
{5-2*exp(-x/2)}/{5+sqrt(log(x+9))} # = 0.6511769
## [1] 0.6511769
```

d. $$\frac{x+7}{3} \times \log_{10}(0.5x^2 - x/3+1)$$

```
#Replace with your R code
{{x+7}/{3}} * log10(0.5*x^2-x/3+1) # = 1.10393
## [1] 1.10393
```

e. $$(1+2x^2)^{4/3}$$

```
#Replace with your R code
(1+2*x^2)^{4/3} # = 18.72075
## [1] 18.72075
```

Question 2: Download HW1-clickthru.xls. This is a small part of a kaggle.com competition concerned with predicting the clickthru rate of ads displayed on mobile devices. Open this file with Excel and save it as a comma-delimited file. Then, use `read.csv` to load this file into R, naming it . The column names are mostly self-explanatory. Variables starting with the word Site deal with the particular website on which the ad was displayed, variables starting with App deal with the App used to view the ad (Safari, Chrome, through a game, etc.), and variables starting with x are anonymized categorical variables (so trade secrets are not revealed).

```
#Note:  remove the eval=FALSE on the line above ONLY after you have created the .c
sv file
CLICK <- read.csv("HW1-clickthru.csv")
```

a. How many rows and columns are in the data frame?

```
#Insert R code if you used any (can get answer without running a command)
nrow(CLICK) # = 13594
## [1] 13594
ncol(CLICK) # = 15
## [1] 15
```

Response:13594 rows and 15 columns

b. Show the first eight lines of the data using `head` .

```
#Replace with your R code
head(CLICK, 8)
##   Click BannerPosition SiteID SiteDomain SiteCategory AppDomain AppCategory Dev
iceModel
## 1   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
D11
## 2   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
D6
## 3   Yes           Pos1     S6        SD5        SCat5       AD2         AC1
D12
## 4    No           Pos2     S6        SD5        SCat5       AD2         AC1
D17
## 5    No           Pos2     S6        SD5        SCat5       AD2         AC1
D12
## 6    No           Pos2     S6        SD5        SCat5       AD2         AC1
D5
## 7   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
D11
## 8   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
D14
##     x1 x2 x3   x4    x5     x6 x7
## 1 1005  J  b val2 type1 class2 CC
## 2 1005  K  b val2 type1 class3 CC
## 3 1005  E  a val1 type3 class2 EE
## 4 1005  K  b val2 type1 class1 CC
## 5 1005  H  a val1 type3 class3 EE
## 6 1005  K  b val2 type1 class1 CC
## 7 1005  K  b val2 type1 class1 CC
## 8 1005  J  b val2 type1 class1 CC
#   Click BannerPosition SiteID SiteDomain SiteCategory AppDomain AppCategory
# 1   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
# 2   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
# 3   Yes           Pos1     S6        SD5        SCat5       AD2         AC1
# 4    No           Pos2     S6        SD5        SCat5       AD2         AC1
# 5    No           Pos2     S6        SD5        SCat5       AD2         AC1
# 6    No           Pos2     S6        SD5        SCat5       AD2         AC1
# 7   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
# 8   Yes           Pos2     S6        SD5        SCat5       AD2         AC1
#   DeviceModel   x1 x2 x3   x4    x5     x6 x7
# 1         D11 1005  J  b val2 type1 class2 CC
# 2          D6 1005  K  b val2 type1 class3 CC
# 3         D12 1005  E  a val1 type3 class2 EE
# 4         D17 1005  K  b val2 type1 class1 CC
# 5         D12 1005  H  a val1 type3 class3 EE
# 6          D5 1005  K  b val2 type1 class1 CC
# 7         D11 1005  K  b val2 type1 class1 CC
# 8         D14 1005  J  b val2 type1 class1 CC
```

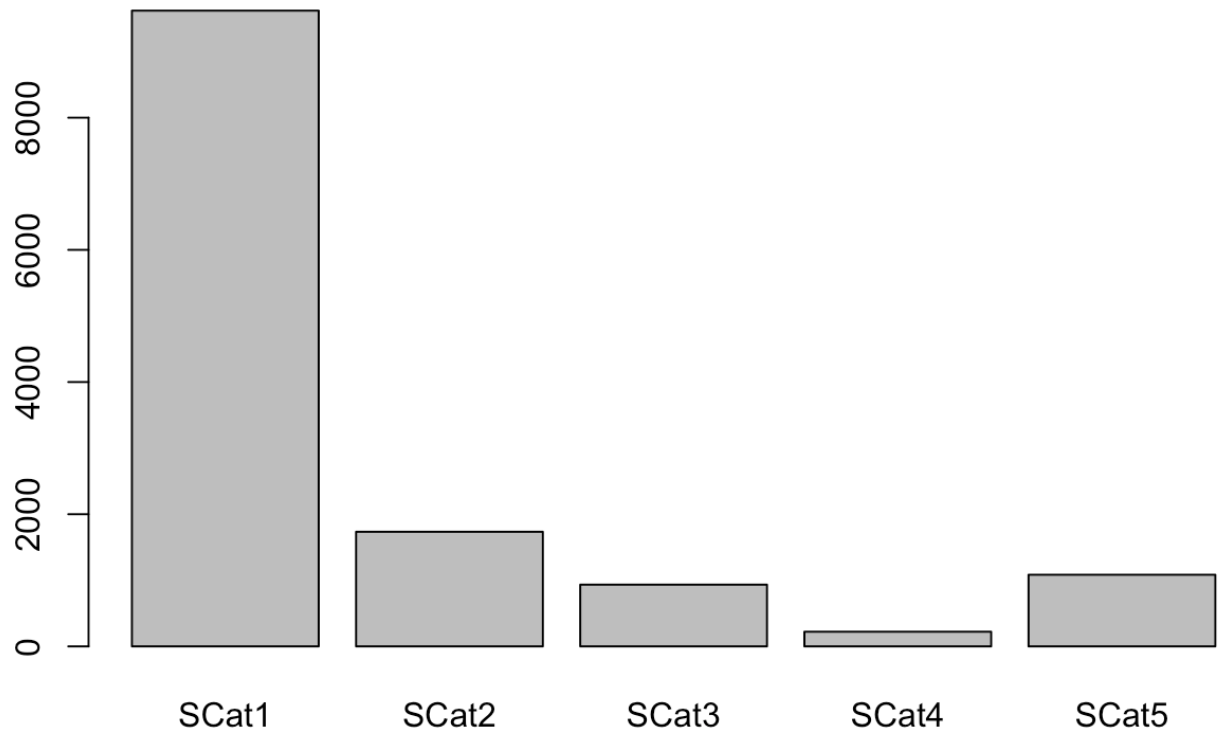
c. Show rows 2015 and 2016. Did the people click the ad? Note: write your response beneath the chunk delimiting the R code.

```
#Replace with your R code
CLICK[2015:2016,]
##      Click BannerPosition SiteID SiteDomain SiteCategory AppDomain AppCategory
## 2015    No           Pos2     S5         SD6         SCat3       AD2          AC2
## 2016    No           Pos2     S5         SD6         SCat3       AD2          AC2
##      DeviceModel   x1 x2 x3   x4    x5     x6 x7
## 2015         D18 1010  R  f val3 type3 class1 AA
## 2016         D18 1010  R  f val3 type3 class1 AA
#      Click BannerPosition SiteID SiteDomain SiteCategory AppDomain AppCategory
# 2015    No           Pos2     S5         SD6         SCat3       AD2          AC2
# 2016    No           Pos2     S5         SD6         SCat3       AD2          AC2
#      DeviceModel   x1 x2 x3   x4    x5     x6 x7
# 2015         D18 1010  R  f val3 type3 class1 AA
# 2016         D18 1010  R  f val3 type3 class1 AA
```

Response: Nope.

d. Make a bar chart of the column `SiteCategory`. What is the most common and least common categories?

```
#Replace with your R code
plot(CLICK$SiteCategory)
```



Response: Most=SCat1; Least=SCat4

e. Make a *relative* frequency table of the `x3` column. What percentage of cases had the level `d`?

```
#Replace with your R code
table(CLICK$x3)/length(CLICK$x3)
##
##          a          b          c          d          e          f
## 0.72598205 0.03869354 0.04023834 0.13792850 0.02037664 0.03678093
```

Response: 13.793%

f. Create two subsets of the data called `SUB.D11` and `SUB.D12` which contain only rows where `DeviceModel` is `D11` and `D12`, respectively. Calculate the fraction of users who clicked the ad (hint: look at a `summary` of the `Click` columns in each subset), and comment on whether there looks to be a difference in click-thru rate between these two devices.

```
#Replace with your R code
SUB.D11 <- subset(CLICK, DeviceModel=="D11")
SUB.D12 <- subset(CLICK, DeviceModel=="D12")
summary(SUB.D11)
##  Click    BannerPosition     SiteID      SiteDomain  SiteCategory AppDomain App
Category
##  No :96   Pos1:60        S6     :123   SD5    :123   SCat1:  0    AD1:  1   AC
1:155
##  Yes:60   Pos2:96        S8     : 22   SD3    : 22   SCat2:  2    AD2:155   AC
2:  1
##                          S1     :  8   SD1    :  8   SCat3:  1    AD3:  0
##                          S4     :  2   SD7    :  2   SCat4:  8
##                          S5     :  1   SD6    :  1   SCat5:145
##                          S2     :  0   SD2    :  0
##                          (Other):  0   (Other):  0
##   DeviceModel        x1             x2     x3        x4         x5          x6
##  D11    :156   Min.   :1005   K      :47   a:26   val1:63   type1:93   class1:1
28
##  D1     :  0   1st Qu.:1005   J      :46   b:93   val2:93   type2: 0   class2:
13
##  D10    :  0   Median :1005   P      :22   c:36   val3: 0   type3:62   class3:
15
##  D12    :  0   Mean   :1005   L      :14   d: 1              type4: 1   class4:
0
##  D13    :  0   3rd Qu.:1005   D      : 7   e: 0
##  D14    :  0   Max.   :1005   B      : 5   f: 0
##  (Other):  0                  (Other):15
##   x7
##  AA: 0
##  BB:37
##  CC:93
##  DD: 0
##  EE:26
##
```

```
##
summary(SUB.D12)
##  Click      BannerPosition     SiteID       SiteDomain   SiteCategory AppDomain
##  No :2356   Pos1:2502      S2     :2297   SD8    :2297   SCat1:2327   AD1:  13
##  Yes: 323   Pos2: 177      S6     : 154   SD5    : 154   SCat2:   8   AD2:2666
##                            S8     :  85   SD3    :  85   SCat3:  13   AD3:   0
##                            S1     :  59   SD1    :  59   SCat4:  59
##                            S7     :  33   SD4    :  33   SCat5: 272
##                            S3     :  30   SD2    :  30
##                            (Other):  21   (Other):  21
##  AppCategory  DeviceModel         x1             x2      x3          x4
##  AC1:2666    D12    :2679   Min.   :1005   A      :300   a:2498   val1:2670
##  AC2:  13    D1     :   0   1st Qu.:1005   B      :300   b:   0   val2:   0
##              D10    :   0   Median :1005   E      :283   c: 148   val3:   9
##              D11    :   0   Mean   :1005   H      :279   d:  24
##              D13    :   0   3rd Qu.:1005   I      :278   e:   9
##              D14    :   0   Max.   :1005   F      :269   f:   0
##              (Other):   0                  (Other):970
##      x5            x6        x7
##  type1:   0   class1:1899   AA:   0
##  type2:   9   class2: 201   BB: 172
##  type3:2646   class3: 569   CC:   0
##  type4:  24   class4:  10   DD:   9
##                             EE:2498
##
##
```

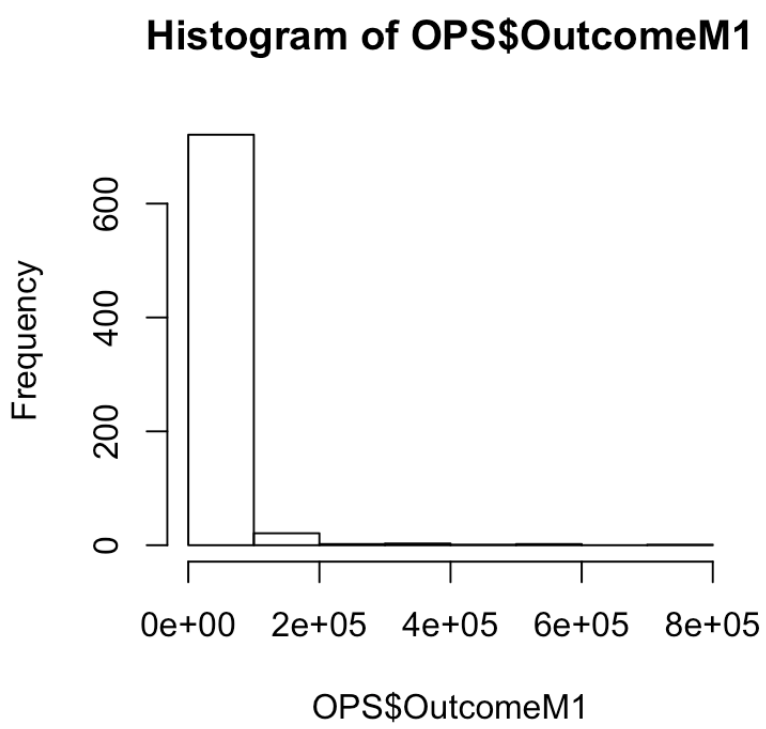
Response: D11:38.46%; D12:12.057%. That for D11 is far higher.

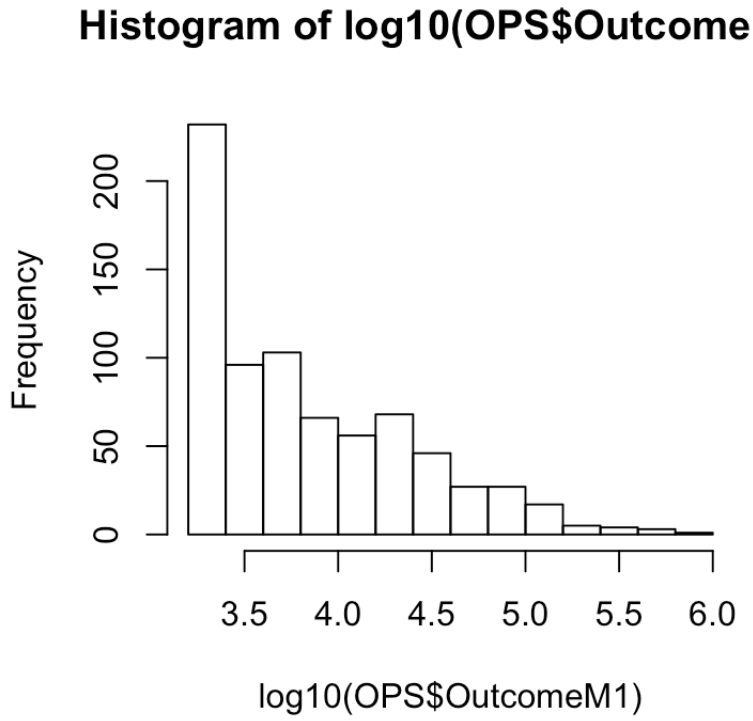
Question 3: Download and read in HW1-OPS.dat using `read.csv`, calling it `OPS`. This is from the kaggle.com Online Product Sales competition that is interested in predicting the amount of sales (for each of twelve consecutive months) after release of a new product (variables OutcomeM1, OutcomeM2, etc.) from various anonymized predictors (Quan1, Quant2, Cat1, Cat2, etc.).

```
OPS <- read.csv("HW1-OPS.dat")
```

a. Make a histogram of the values in `OutcomeM1` and another of the logarithm (base 10) of these values (note: you do not need to add a `breaks` argument; the default is fine). In your R code, separate the commands by a semi-colon so the plots appear side-by-side. Which is more useful for visualizing the distribution and why?

```
#Replace with your R code
hist(OPS$OutcomeM1); hist(log10(OPS$OutcomeM1))
```





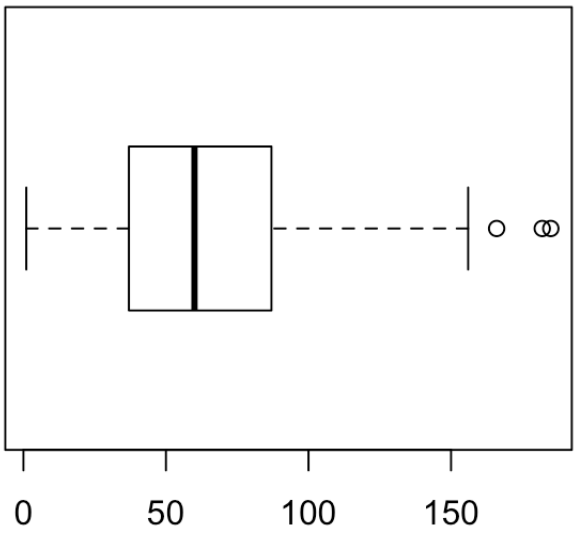
Response: The barchart provides more visually useful data as the x-values that hold most of the data become more spread out.

b. Using the preferred histogram, comment on how many peaks you believe the distribution to have.

Response: Using the logarithmic histogram, the chart is highly skewed but really only has one peak/mode. The same can be seen from the other graph.

c. Make a (horizontal) boxplot of the column `Quan3`. Between what two values (at least approximately) do the central 50% of values fall?

```
#Replace with your R code
boxplot(OPS$Quan3, horizontal = TRUE)
```

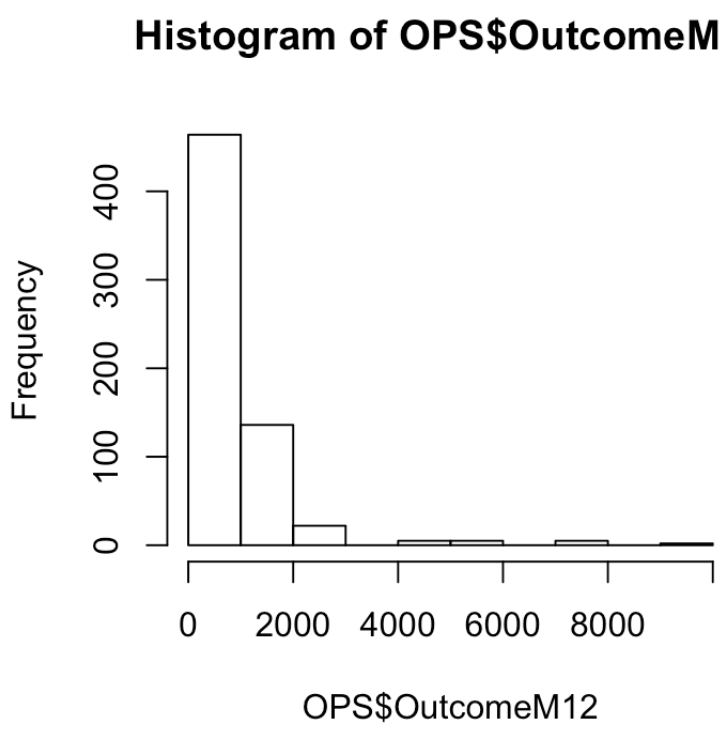


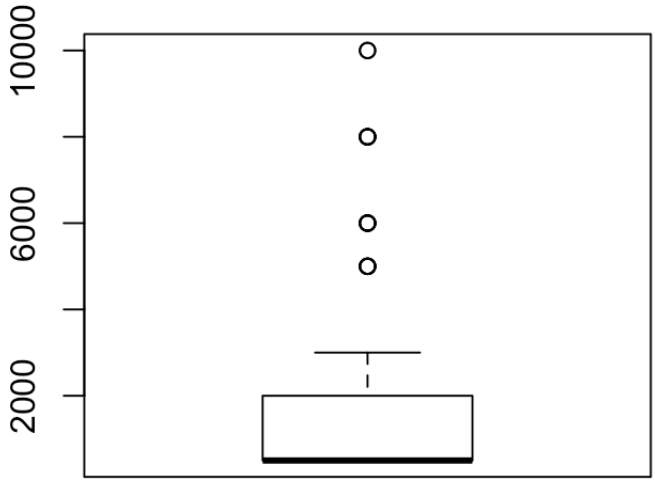

```
summary(OPS$Quan3)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   37.00   60.00   62.17   87.00  185.00       5
```

Reponse: Between Q1(37) and Q3(87)

d. What is a typical value for `OutcomeM12`? Note: you'll have to look at a histogram (or boxplot) of values and the results of `summary` to decide the most appropriate number. Note: to make a dollar sign, you need to type `\$`

```
#Replace with your R code
hist(OPS$OutcomeM12); boxplot(OPS$OutcomeM12)
```






```
#Highly skewed - Using median

summary(OPS$OutcomeM12) # med = 500
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     500     500     500    1072    2000   10000     112
```

Reponse: Median = 500

e. Of interest are products that have values of `Quan1` less than one million. Make a subset containing rows that meet this condition, then find the average value of `OutcomeM1` for these cases.

```
#Replace with your R code
CHEAP <- subset(OPS, Quan1<1000000)
summary(CHEAP$Quan1)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31600  281900  613700  569700  887900  994700
mean(CHEAP$OutcomeM1) # = 14057.14
## [1] 14057.14
```