Datacamp_Importing & Cleaning Data in R: Case Studies___MBTA Ridership Data

dizhen

2019/4/5

Using readxl

```
# Load readxl
library(readxl)

# Import mbta.xlsx and skip first row: mbta
mbta<-read_excel("data/mbta.xlsx",skip = 1)
```

```
## New names:
## * `` -> ...1
```

Examining the data

```
# View the structure of mbta
str(mbta)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    11 obs. of  60 variables:
##  $ ...1   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ mode   : chr  "All Modes by Qtr" "Boat" "Bus" "Commuter Rail" ...
##  $ 2007-01: chr  "NA" "4" "335.819" "142.2" ...
##  $ 2007-02: chr  "NA" "3.6" "338.675" "138.5" ...
##  $ 2007-03: num  1188 40 340 138 459 ...
##  $ 2007-04: chr  "NA" "4.3" "352.162" "139.5" ...
##  $ 2007-05: chr  "NA" "4.9" "354.367" "139" ...
##  $ 2007-06: num  1246 5.8 350.5 143 477 ...
##  $ 2007-07: chr  "NA" "6.521" "357.519" "142.391" ...
##  $ 2007-08: chr  "NA" "6.572" "355.479" "142.364" ...
##  $ 2007-09: num  1256.57 5.47 372.6 143.05 499.57 ...
##  $ 2007-10: chr  "NA" "5.145" "368.847" "146.542" ...
##  $ 2007-11: chr  "NA" "3.763" "330.826" "145.089" ...
##  $ 2007-12: num  1216.89 2.98 312.92 141.59 448.27 ...
##  $ 2008-01: chr  "NA" "3.175" "340.324" "142.145" ...
##  $ 2008-02: chr  "NA" "3.111" "352.905" "142.607" ...
##  $ 2008-03: num  1253.52 3.51 361.15 137.45 494.05 ...
##  $ 2008-04: chr  "NA" "4.164" "368.189" "140.389" ...
##  $ 2008-05: chr  "NA" "4.015" "363.903" "142.585" ...
##  $ 2008-06: num  1314.82 5.19 362.96 142.06 518.35 ...
##  $ 2008-07: chr  "NA" "6.016" "370.921" "145.731" ...
##  $ 2008-08: chr  "NA" "5.8" "361.057" "144.565" ...
##  $ 2008-09: num  1307.04 4.59 389.54 141.91 517.32 ...
##  $ 2008-10: chr  "NA" "4.285" "357.974" "151.957" ...
##  $ 2008-11: chr  "NA" "3.488" "345.423" "152.952" ...
##  $ 2008-12: num  1232.65 3.01 325.77 140.81 446.74 ...
##  $ 2009-01: chr  "NA" "3.014" "338.532" "141.448" ...
##  $ 2009-02: chr  "NA" "3.196" "360.412" "143.529" ...
##  $ 2009-03: num  1209.79 3.33 353.69 142.89 467.22 ...
##  $ 2009-04: chr  "NA" "4.049" "359.38" "142.34" ...
##  $ 2009-05: chr  "NA" "4.119" "354.75" "144.225" ...
##  $ 2009-06: num  1233.1 4.9 347.9 142 473.1 ...
##  $ 2009-07: chr  "NA" "6.444" "339.477" "137.691" ...
##  $ 2009-08: chr  "NA" "5.903" "332.661" "139.158" ...
##  $ 2009-09: num  1230.5 4.7 374.3 139.1 500.4 ...
##  $ 2009-10: chr  "NA" "4.212" "385.868" "137.104" ...
##  $ 2009-11: chr  "NA" "3.576" "366.98" "129.343" ...
##  $ 2009-12: num  1207.85 3.11 332.39 126.07 440.93 ...
##  $ 2010-01: chr  "NA" "3.207" "362.226" "130.91" ...
##  $ 2010-02: chr  "NA" "3.195" "361.138" "131.918" ...
##  $ 2010-03: num  1208.86 3.48 373.44 131.25 483.4 ...
##  $ 2010-04: chr  "NA" "4.452" "378.611" "131.722" ...
##  $ 2010-05: chr  "NA" "4.415" "380.171" "128.8" ...
##  $ 2010-06: num  1244.41 5.41 363.27 129.14 490.26 ...
##  $ 2010-07: chr  "NA" "6.513" "353.04" "122.935" ...
##  $ 2010-08: chr  "NA" "6.269" "343.688" "129.732" ...
##  $ 2010-09: num  1225.5 4.7 381.6 132.9 521.1 ...
##  $ 2010-10: chr  "NA" "4.402" "384.987" "131.033" ...
##  $ 2010-11: chr  "NA" "3.731" "367.955" "130.889" ...
##  $ 2010-12: num  1216.26 3.16 326.34 121.42 450.43 ...
##  $ 2011-01: chr  "NA" "3.14" "334.958" "128.396" ...
##  $ 2011-02: chr  "NA" "3.284" "346.234" "125.463" ...
##  $ 2011-03: num  1223.45 3.67 380.4 134.37 516.73 ...
##  $ 2011-04: chr  "NA" "4.251" "380.446" "134.169" ...
##  $ 2011-05: chr  "NA" "4.431" "385.289" "136.14" ...
##  $ 2011-06: num  1302.41 5.47 376.32 135.58 529.53 ...
##  $ 2011-07: chr  "NA" "6.581" "361.585" "132.41" ...
##  $ 2011-08: chr  "NA" "6.733" "353.793" "130.616" ...
##  $ 2011-09: num  1291 5 388 137 550 ...
##  $ 2011-10: chr  "NA" "4.484" "398.456" "128.72" ...
```

```r
# View the first 6 rows of mbta
head(mbta, n=6)
```

```
## # A tibble: 6 x 60
##    ...1 mode  `2007-01` `2007-02` `2007-03` `2007-04` `2007-05` `2007-06`
##   <dbl> <chr> <chr>     <chr>         <dbl> <chr>     <chr>         <dbl>
## 1     1 All ~ NA        NA            1188. NA        NA            1246.
## 2     2 Boat  4         3.6             40  4.3       4.9              5.8
## 3     3 Bus   335.819   338.675        340. 352.162   354.367        351.
## 4     4 Comm~ 142.2     138.5          138. 139.5     139            143
## 5     5 Heav~ 435.294   448.271        459. 472.201   474.579        477.
## 6     6 Ligh~ 227.231   240.262        241. 255.557   248.262        246.
## # ... with 52 more variables: `2007-07` <chr>, `2007-08` <chr>,
## #   `2007-09` <dbl>, `2007-10` <chr>, `2007-11` <chr>, `2007-12` <dbl>,
## #   `2008-01` <chr>, `2008-02` <chr>, `2008-03` <dbl>, `2008-04` <chr>,
## #   `2008-05` <chr>, `2008-06` <dbl>, `2008-07` <chr>, `2008-08` <chr>,
## #   `2008-09` <dbl>, `2008-10` <chr>, `2008-11` <chr>, `2008-12` <dbl>,
## #   `2009-01` <chr>, `2009-02` <chr>, `2009-03` <dbl>, `2009-04` <chr>,
## #   `2009-05` <chr>, `2009-06` <dbl>, `2009-07` <chr>, `2009-08` <chr>,
## #   `2009-09` <dbl>, `2009-10` <chr>, `2009-11` <chr>, `2009-12` <dbl>,
## #   `2010-01` <chr>, `2010-02` <chr>, `2010-03` <dbl>, `2010-04` <chr>,
## #   `2010-05` <chr>, `2010-06` <dbl>, `2010-07` <chr>, `2010-08` <chr>,
## #   `2010-09` <dbl>, `2010-10` <chr>, `2010-11` <chr>, `2010-12` <dbl>,
## #   `2011-01` <chr>, `2011-02` <chr>, `2011-03` <dbl>, `2011-04` <chr>,
## #   `2011-05` <chr>, `2011-06` <dbl>, `2011-07` <chr>, `2011-08` <chr>,
## #   `2011-09` <dbl>, `2011-10` <chr>
```

```
# View a summary of mbta
summary(mbta)
```

```
##       ...1          mode             2007-01            2007-02         
##  Min.   : 1.0   Length:11          Length:11          Length:11         
##  1st Qu.: 3.5   Class :character   Class :character   Class :character  
##  Median : 6.0   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 6.0                                                           
##  3rd Qu.: 8.5                                                           
##  Max.   :11.0                                                           
##     2007-03           2007-04            2007-05         
##  Min.   :   0.114   Length:11          Length:11         
##  1st Qu.:   9.278   Class :character   Class :character  
##  Median : 137.700   Mode  :character   Mode  :character  
##  Mean   : 330.293                                        
##  3rd Qu.: 399.225                                        
##  Max.   :1204.725                                        
##     2007-06           2007-07            2007-08         
##  Min.   :   0.096   Length:11          Length:11         
##  1st Qu.:   5.700   Class :character   Class :character  
##  Median : 143.000   Mode  :character   Mode  :character  
##  Mean   : 339.846                                        
##  3rd Qu.: 413.788                                        
##  Max.   :1246.129                                        
##     2007-09           2007-10            2007-11         
##  Min.   :  -0.007   Length:11          Length:11         
##  1st Qu.:   5.539   Class :character   Class :character  
##  Median : 143.051   Mode  :character   Mode  :character  
##  Mean   : 352.554                                        
##  3rd Qu.: 436.082                                        
##  Max.   :1310.764                                        
##     2007-12           2008-01            2008-02         
##  Min.   :  -0.060   Length:11          Length:11         
##  1st Qu.:   4.385   Class :character   Class :character  
##  Median : 141.585   Mode  :character   Mode  :character  
##  Mean   : 321.588                                        
##  3rd Qu.: 380.594                                        
##  Max.   :1216.890                                        
##     2008-03           2008-04            2008-05         
##  Min.   :   0.058   Length:11          Length:11         
##  1st Qu.:   5.170   Class :character   Class :character  
##  Median : 137.453   Mode  :character   Mode  :character  
##  Mean   : 345.604                                        
##  3rd Qu.: 427.601                                        
##  Max.   :1274.031                                        
##     2008-06           2008-07            2008-08         
##  Min.   :   0.060   Length:11          Length:11         
##  1st Qu.:   5.742   Class :character   Class :character
##  Median : 142.057   Mode  :character   Mode  :character
##  Mean   : 359.667
##  3rd Qu.: 440.656
##  Max.   :1320.728
##     2008-09            2008-10             2008-11
##  Min.   :   0.021   Length:11          Length:11
##  1st Qu.:   5.691   Class :character   Class :character
##  Median : 141.907   Mode  :character   Mode  :character
##  Mean   : 362.099
##  3rd Qu.: 453.430
##  Max.   :1338.015
##     2008-12            2009-01             2009-02
##  Min.   :  -0.015   Length:11          Length:11
##  1st Qu.:   4.689   Class :character   Class :character
##  Median : 140.810   Mode  :character   Mode  :character
##  Mean   : 319.882
##  3rd Qu.: 386.255
##  Max.   :1232.655
##     2009-03            2009-04             2009-05
##  Min.   :  -0.050   Length:11          Length:11
##  1st Qu.:   5.003   Class :character   Class :character
##  Median : 142.893   Mode  :character   Mode  :character
##  Mean   : 330.142
##  3rd Qu.: 410.455
##  Max.   :1210.912
##     2009-06            2009-07             2009-08
##  Min.   :  -0.079   Length:11          Length:11
##  1st Qu.:   5.845   Class :character   Class :character
##  Median : 142.006   Mode  :character   Mode  :character
##  Mean   : 333.194
##  3rd Qu.: 410.482
##  Max.   :1233.085
##     2009-09            2009-10             2009-11
##  Min.   :  -0.035   Length:11          Length:11
##  1st Qu.:   5.693   Class :character   Class :character
##  Median : 139.087   Mode  :character   Mode  :character
##  Mean   : 346.687
##  3rd Qu.: 437.332
##  Max.   :1291.564
##     2009-12            2010-01             2010-02
##  Min.   :  -0.022   Length:11          Length:11
##  1st Qu.:   4.784   Class :character   Class :character
##  Median : 126.066   Mode  :character   Mode  :character
##  Mean   : 312.962
##  3rd Qu.: 386.659
##  Max.   :1207.845
##     2010-03            2010-04             2010-05
##  Min.   :   0.012   Length:11          Length:11
##  1st Qu.:   5.274   Class :character   Class :character
##  Median : 131.252   Mode  :character   Mode  :character
##  Mean   : 332.726
##  3rd Qu.: 428.420
##  Max.   :1225.556
##     2010-06           2010-07            2010-08
##  Min.   :   0.008   Length:11          Length:11
##  1st Qu.:   6.436   Class :character   Class :character
##  Median : 129.144   Mode  :character   Mode  :character
##  Mean   : 335.964
##  3rd Qu.: 426.769
##  Max.   :1244.409
##     2010-09           2010-10            2010-11
##  Min.   :   0.001   Length:11          Length:11
##  1st Qu.:   5.567   Class :character   Class :character
##  Median : 132.892   Mode  :character   Mode  :character
##  Mean   : 346.524
##  3rd Qu.: 451.361
##  Max.   :1293.117
##     2010-12           2011-01            2011-02
##  Min.   :  -0.004   Length:11          Length:11
##  1st Qu.:   4.466   Class :character   Class :character
##  Median : 121.422   Mode  :character   Mode  :character
##  Mean   : 312.917
##  3rd Qu.: 388.385
##  Max.   :1216.262
##     2011-03          2011-04            2011-05
##  Min.   :   0.05   Length:11          Length:11
##  1st Qu.:   6.03   Class :character   Class :character
##  Median : 134.37   Mode  :character   Mode  :character
##  Mean   : 345.17
##  3rd Qu.: 448.56
##  Max.   :1286.66
##     2011-06           2011-07            2011-08
##  Min.   :   0.054   Length:11          Length:11
##  1st Qu.:   6.926   Class :character   Class :character
##  Median : 135.581   Mode  :character   Mode  :character
##  Mean   : 353.331
##  3rd Qu.: 452.923
##  Max.   :1302.414
##     2011-09           2011-10
##  Min.   :   0.043   Length:11
##  1st Qu.:   6.660   Class :character
##  Median : 136.901   Mode  :character
##  Mean   : 362.555
##  3rd Qu.: 469.204
##  Max.   :1348.754
```

Removing unnecessary rows and columns

```r
# Remove rows 1, 7, and 11 of mbta: mbta2
mbta2 <- mbta[-c(1,7,11),]

# Remove the first column of mbta2: mbta3
mbta3 <- mbta2[,-1]
```

Observations are stored in columns

```
# Load tidyr
library(tidyr)

# Gather columns of mbta3: mbta4
mbta4 <- gather(mbta3, month,thou_riders,-mode)

# View the head of mbta4
head(mbta4)
```

```
## # A tibble: 6 x 3
##   mode           month   thou_riders
##   <chr>          <chr>   <chr>
## 1 Boat           2007-01 4
## 2 Bus            2007-01 335.819
## 3 Commuter Rail  2007-01 142.2
## 4 Heavy Rail     2007-01 435.294
## 5 Light Rail     2007-01 227.231
## 6 Private Bus    2007-01 4.772
```

Type conversions

```
# Coerce thou_riders to numeric
mbta4$thou_riders <- as.numeric(mbta4$thou_riders)
```

Variables are stored in both rows and columns

```
# Spread the contents of mbta4: mbta5
mbta5 <- spread(mbta4, mode,thou_riders)

# View the head of mbta5
head(mbta5)
```

```
## # A tibble: 6 x 9
##   month  Boat   Bus `Commuter Rail` `Heavy Rail` `Light Rail` `Private Bus`
##   <chr> <dbl> <dbl>           <dbl>        <dbl>        <dbl>         <dbl>
## 1 2007~   4    336.             142.         435.         227.           4.77
## 2 2007~   3.6  339.             138.         448.         240.           4.42
## 3 2007~  40    340.             138.         459.         241.           4.57
## 4 2007~   4.3  352.             140.         472.         256.           4.54
## 5 2007~   4.9  354.             139          475.         248.           4.77
## 6 2007~   5.8  351.             143          477.         246.           4.72
## # ... with 2 more variables: RIDE <dbl>, `Trackless Trolley` <dbl>
```

Separating columns

```
# View the head of mbta5
head(mbta5)
```

```
## # A tibble: 6 x 9
##   month  Boat   Bus `Commuter Rail` `Heavy Rail` `Light Rail` `Private Bus`
##   <chr> <dbl> <dbl>           <dbl>        <dbl>        <dbl>         <dbl>
## 1 2007~   4    336.            142.         435.         227.          4.77
## 2 2007~   3.6  339.            138.         448.         240.          4.42
## 3 2007~  40    340.            138.         459.         241.          4.57
## 4 2007~   4.3  352.            140.         472.         256.          4.54
## 5 2007~   4.9  354.            139          475.         248.          4.77
## 6 2007~   5.8  351.            143          477.         246.          4.72
## # ... with 2 more variables: RIDE <dbl>, `Trackless Trolley` <dbl>
```

```
# Split month column into month and year: mbta6
mbta6 <- separate(mbta5,month,c("year","month"))

# View the head of mbta6
head(mbta6)
```

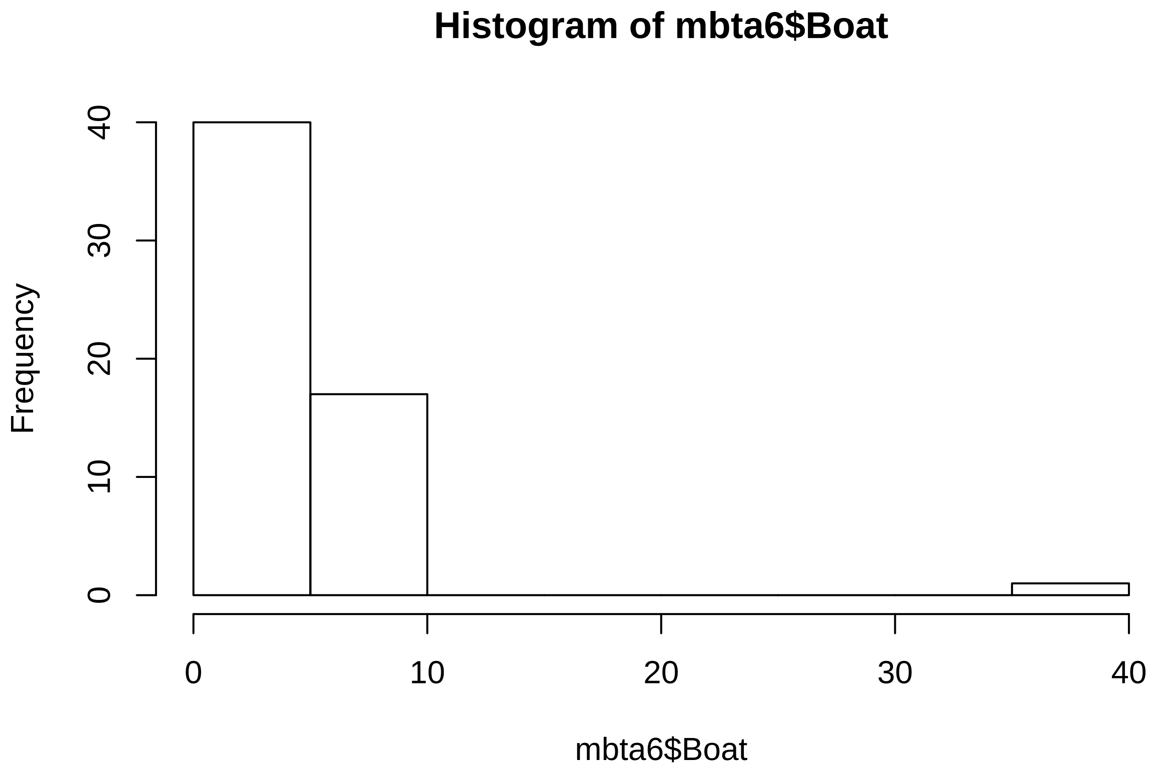
```
## # A tibble: 6 x 10
##   year  month  Boat   Bus `Commuter Rail` `Heavy Rail` `Light Rail`
##   <chr> <chr> <dbl> <dbl>           <dbl>        <dbl>        <dbl>
## 1 2007  01      4    336.             142.         435.         227.
## 2 2007  02      3.6  339.             138.         448.         240.
## 3 2007  03     40    340.             138.         459.         241.
## 4 2007  04      4.3  352.             140.         472.         256.
## 5 2007  05      4.9  354.             139          475.         248.
## 6 2007  06      5.8  351.             143          477.         246.
## # ... with 3 more variables: `Private Bus` <dbl>, RIDE <dbl>, `Trackless
## #   Trolley` <dbl>
```

Do your values seem reasonable?

```
# View a summary of mbta6
summary(mbta6)
```

```
##      year               month                 Boat             Bus
##  Length:58          Length:58          Min.   : 2.985   Min.   :312.9
##  Class :character   Class :character   1st Qu.: 3.494   1st Qu.:345.6
##  Mode  :character   Mode  :character   Median : 4.293   Median :359.9
##                                        Mean   : 5.068   Mean   :358.6
##                                        3rd Qu.: 5.356   3rd Qu.:372.2
##                                        Max.   :40.000   Max.   :398.5
##  Commuter Rail     Heavy Rail      Light Rail     Private Bus
##  Min.   :121.4   Min.   :435.3   Min.   :194.4   Min.   :2.213
##  1st Qu.:131.4   1st Qu.:471.1   1st Qu.:220.6   1st Qu.:2.641
##  Median :138.8   Median :487.3   Median :231.9   Median :2.820
##  Mean   :137.4   Mean   :489.3   Mean   :233.0   Mean   :3.352
##  3rd Qu.:142.4   3rd Qu.:511.3   3rd Qu.:244.5   3rd Qu.:4.167
##  Max.   :153.0   Max.   :554.9   Max.   :271.1   Max.   :4.878
##       RIDE       Trackless Trolley
##  Min.   :4.900   Min.   : 5.777
##  1st Qu.:5.965   1st Qu.:11.679
##  Median :6.615   Median :12.598
##  Mean   :6.604   Mean   :12.125
##  3rd Qu.:7.149   3rd Qu.:13.320
##  Max.   :8.598   Max.   :15.109
```

```
# Generate a histogram of Boat column
hist(mbta6$Boat)
```

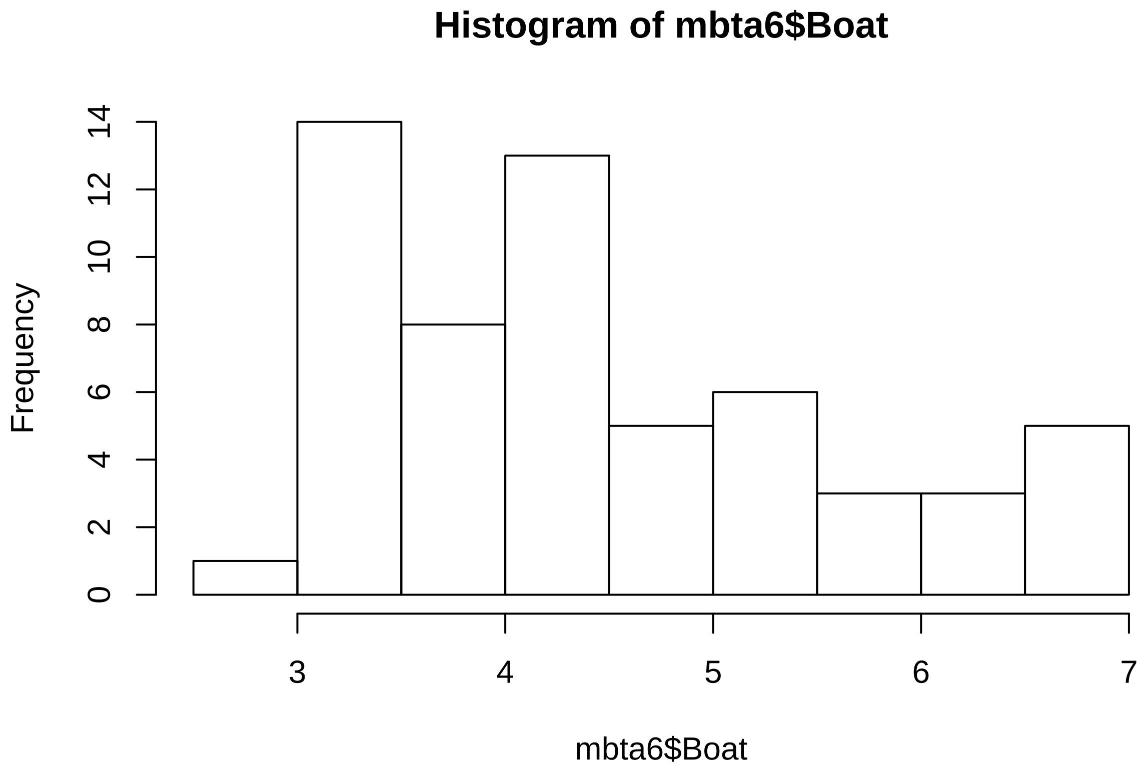


Dealing with entry error

```
# Find the row number of the incorrect value: i
i <- which(mbta6$Boat > 15)

# Replace the incorrect value with 4
mbta6$Boat[i] <- 4

# Generate a histogram of Boat column
hist(mbta6$Boat)
```



```
# # Look at Boat and Trackless Trolley ridership over time (don't change)
# ggplot(mbta_boat, aes(x = month, y = thou_riders, col = mode)) +  geom_point() +
#   scale_x_discrete(name = "Month", breaks = c(200701, 200801, 200901, 201001, 201101)) +
#   scale_y_continuous(name = "Avg Weekday Ridership (thousands)")
#
# # Look at all T ridership over time (don't change)
# ggplot(mbta_all, aes(x = month, y = thou_riders, col = mode)) + geom_point() +
#   scale_x_discrete(name = "Month", breaks = c(200701, 200801, 200901, 201001, 201101)) +
#   scale_y_continuous(name = "Avg Weekday Ridership (thousands)")
```