MBTA

Hanh Nguyen

6/4/2017

The Massachusetts Bay Transportation Authority ("MBTA" or just "the T" for short) manages America's oldest subway, as well as Greater Boston's commuter rail, ferry, and bus systems.
The dataset is stored as an Excel spreadsheet called mbta.xlsx, which is a set of MBTA ridership data. The first row is a title, so it needs to be skipped.

```
library(readxl)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("/Users/user/GitHub/data-vis")
mbta = read_excel("mbta.xlsx",skip=1)
```

Using *str(mbta), head(mbta), summary(mbta)*, we explore the dataset and notice some unnecessary rows and columns. All of the NA values are stored in the All Modes by Qtr row. This row is a quarterly average of weekday MBTA ridership and since this dataset tracks monthly average ridership, it can be removed. Similarly, the 7th row (Pct Chg / Yr) and the 11th row (TOTAL) are not really observations and will be removed. The first column also needs to be removed because it's just listing the row numbers.

Besides, this dataset is stored as a tibble which is just a specific type of data frame.

```
mbta = mbta[-c(1, 7, 11), ]
mbta = mbta[, -1]
```

The different modes of transportation (commuter rail, bus, subway, ferry, etc.) are variables, providing information about each month's average ridership. The months themselves are observations. The variables are stored in rows instead of columns and since we actually want to represent variables in columns rather than rows, we use the **gather()** and **separate()** functions from the tidyr package. Further illustration of them is in **Cleaning_yymmdd.Rmd.**
Also, we change the average weekday ridership column, thou_riders, into numeric values rather than character strings.

```
mbta2 = mbta %>%
  gather(month, thou_riders, -mode)
mbta2$thou_riders = as.numeric(mbta2$thou_riders)
mbta2 = mbta2 %>%
  spread(mode,thou_riders) %>%
  separate(month, into=c("year","month"),sep="-")
```

By running *summary(mbta2), hist(mbta2$Boat)*, we see that every value of the Boat column clustered around 4 and one loner out around 40. Every month, average weekday commuter boat ridership was on either side of four thousand. Then, one month it jumped to 40 thousand without warning? This value is likely an error as being accidentally typed 40 instead of 4. Therefore, we'll locate the incorrect value and change it to 4.

```
i = which(mbta2$Boat > 30)
mbta2$Boat[i] = 4
summary(mbta2)
```

```
##      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 :4.268   Median :359.9
##                                        Mean   :4.447   Mean   :358.6
##                                        3rd Qu.:5.178   3rd Qu.:372.2
##                                        Max.   :6.733   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
```

Source:
https://www.datacamp.com/courses/importing-cleaning-data-in-r-case-studies