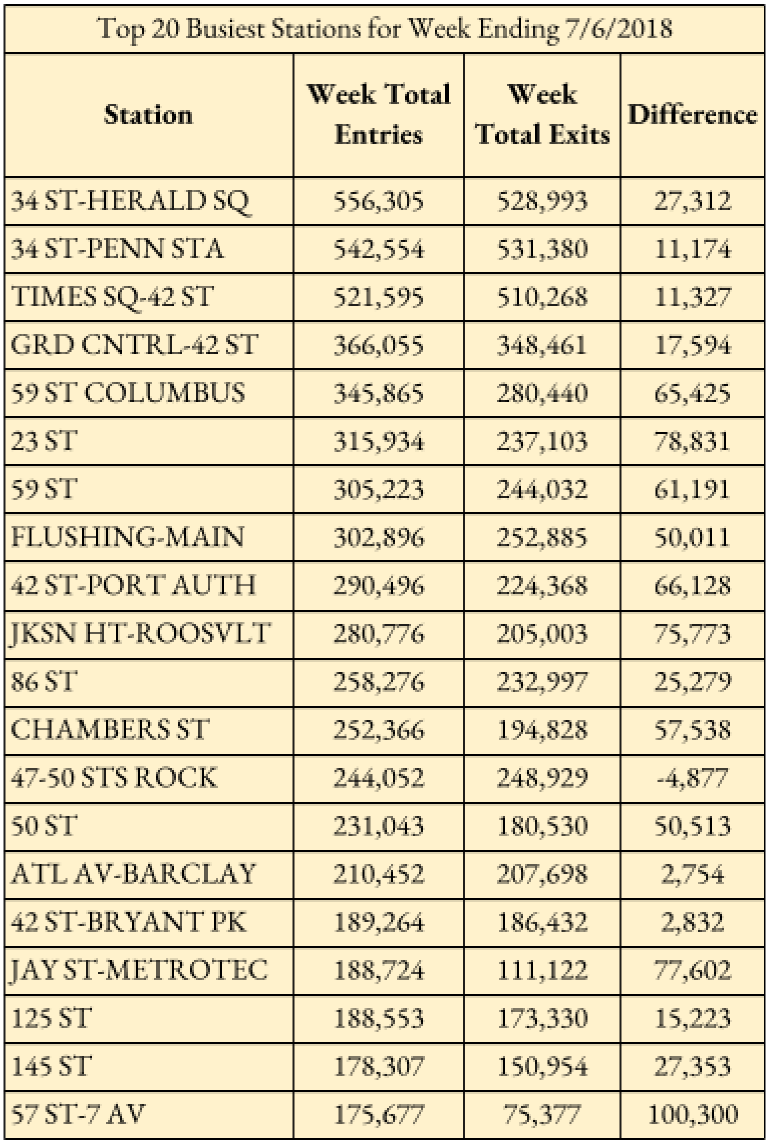
An Analysis of the MTA's Weekly Subway Turnstile Data
By Michael D'Acampora

The MTA's weekly turnstile report* is a data set describing customer entry and exit information for every turnstile in each of the 472 subway stations in New York City. The turnstiles are bidirectional and count each time a customer passes whether entering or exiting. The weekly data set gives you approximately 197,000 observations across 11 variables including station name, subway lines served, turnstile device number, date, time, audit event status, entry count, and exit count. Count audits are performed every four hours, starting at midnight on the first day the audit week (00:00:00 Saturday) and end ending 8pm of the last day (20:00:00 Friday). For the audit week June 30 - July 6, 2018 the twenty busiest stations as measured by entry and exit are shown in the table below.

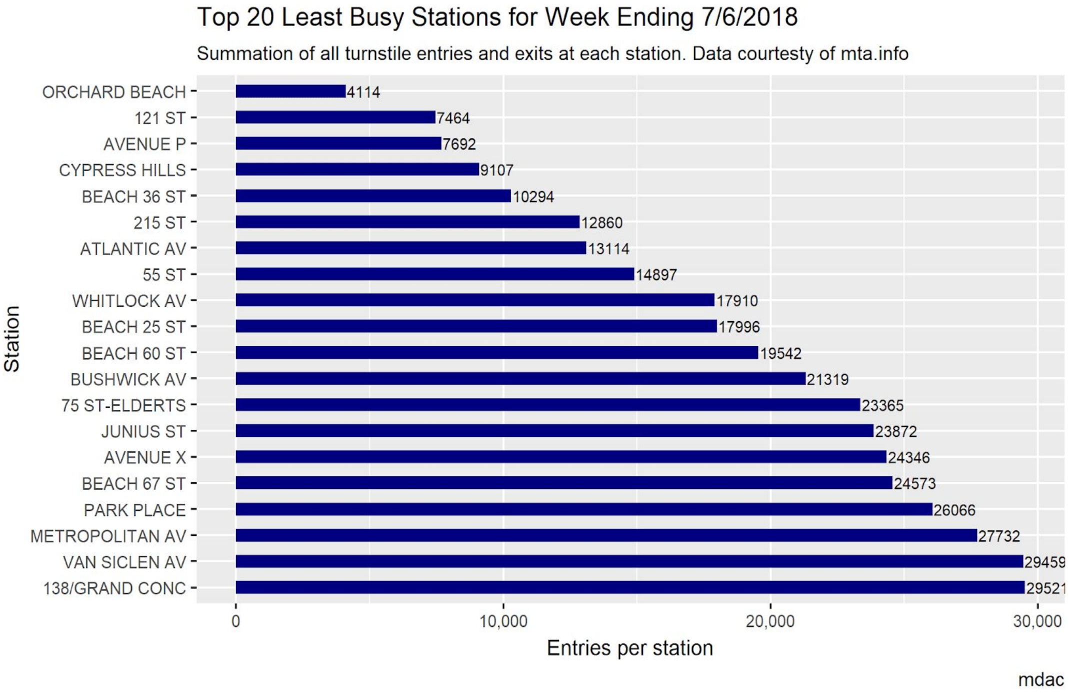


Top 20 Busiest Stations for Week Ending 7/6/2018

Station	Week Total Entries	Week Total Exits	Difference
34 ST-HERALD SQ	556,305	528,993	27,312
34 ST-PENN STA	542,554	531,380	11,174
TIMES SQ-42 ST	521,595	510,268	11,327
GRD CNTRL-42 ST	366,055	348,461	17,594
59 ST COLUMBUS	345,865	280,440	65,425
23 ST	315,934	237,103	78,831
59 ST	305,223	244,032	61,191
FLUSHING-MAIN	302,896	252,885	50,011
42 ST-PORT AUTH	290,496	224,368	66,128
JKSN HT-ROOSVLT	280,776	205,003	75,773
86 ST	258,276	232,997	25,279
CHAMBERS ST	252,366	194,828	57,538
47-50 STS ROCK	244,052	248,929	-4,877
50 ST	231,043	180,530	50,513
ATL AV-BARCLAY	210,452	207,698	2,754
42 ST-BRYANT PK	189,264	186,432	2,832
JAY ST-METROTEC	188,724	111,122	77,602
125 ST	188,553	173,330	15,223
145 ST	178,307	150,954	27,353
57 ST-7 AV	175,677	75,377	100,300

An interesting station to note is 47-50 Sts Rockefeller Center, which tallied 4,877 more exits than entries. Several hypotheses regarding the results include (1) an error in data collection and analysis (such as incomplete data), (2) an unauthorized access point allowing patrons to bypass turnstiles (such as a broken emergency exit door being used to avoid payment), or (3) the data is accurate and this station tends to see more one-way fares (being a high volume tourist area, visitors may exit at this station and visit various attractions that lead them out of the immediate area).

A second noteworthy station is 57 St - 7 Av in Manhattan, where there were over 100,000, or 57%, less exits than entrances. A potential reason could be the existence of exit-only high gate turnstiles that do no appear to record exit data, like the type that exist in the 77th St station in Brooklyn. Intermittent construction or excessive delays could have also caused trains to skip the station in one direction this week, resulting in less passengers exiting on their way home from work, for example.



Turning to the bottom half of the system, Orchard Beach station has the lowest recorded combined entry and exit counts at 4,114 for the week. But upon further inspection of the data set, this station corresponds to the 6 line, and there does not appear to be an Orchard Beach station when looking at the official NYCT subway map. Using the map, a search along the 6 line revealed that the Morrison Ave Soundview station was not present in this data set. Did this station undergo a name change some time ago and was never updated? It is a mystery that requires further investigation. No wonder it is the least busy!

Another station that personally catches my attention is the Park Place station. Since it is so close to the World Financial Center, the PATH train, and part of an express route from Penn Station, I would expect to see a higher amount of activity somewhere in the middle of the distribution. Especially since Chambers Street, the station one stop north, ranked as the 12th busiest station for the week. The two stations are only a few blocks apart and while the southbound E train ends at Park Place, the 2 and 3 lines continue through to Brooklyn.

There are plenty more things that can be done with this data. Solutions can be tailored for matching station GPS coordinates with turnstile data overlaid on a map, the data can be bucketed by size, by borough, by division, construction events, as well as situationally in the instance of peak/off-peak time.

References
*http://web.mta.info/developers/data/nyct/turnstile/turnstile_180707.txt
www.stackoverflow.com

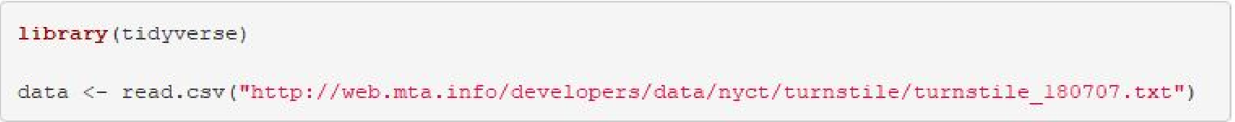
[END OF DOCUMENT]

APPENDIX

---
title: "MTA_Turnstile"
author: "Michael D'Acampora"
date: "7/8/2018"
output: html_document

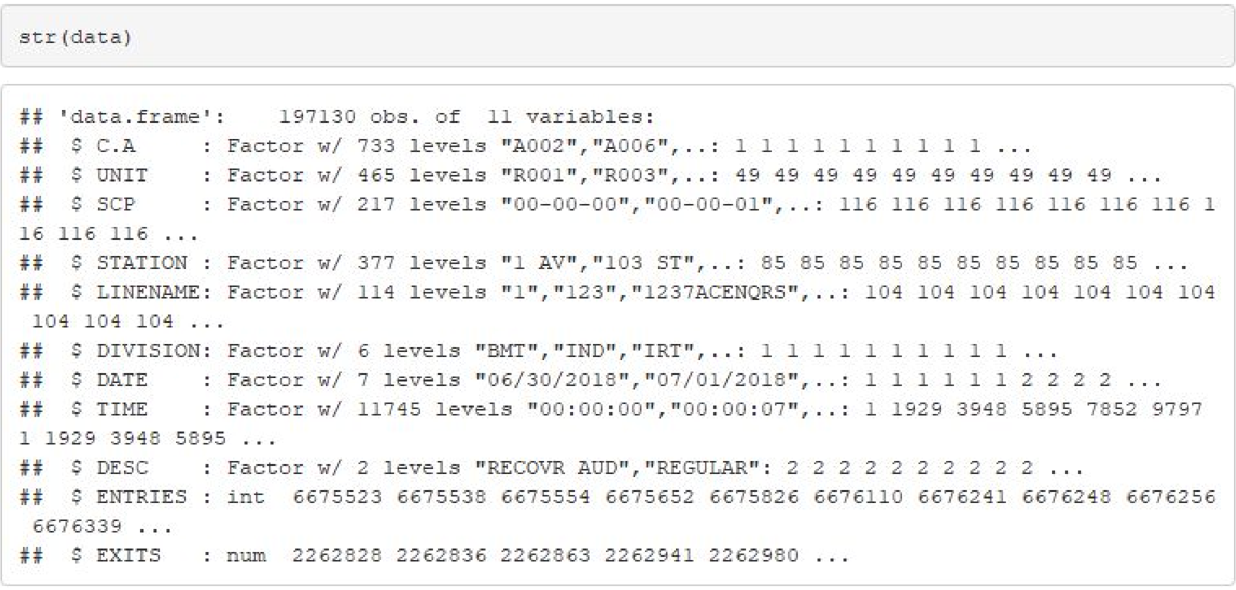
The following code was written in R in RStudio.

Load pertinent libraries, pull in the data from mta.info



```
library(tidyverse)

data <- read.csv("http://web.mta.info/developers/data/nyct/turnstile/turnstile_180707.txt")
```

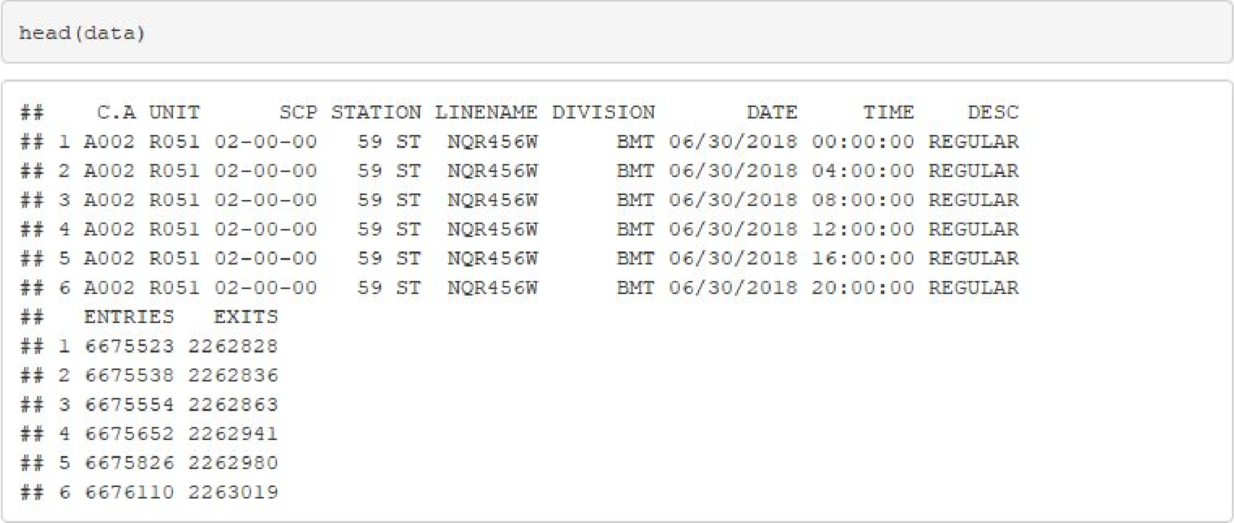
After the data are pulled, the structure of the data set is inspected, revealing 197,130 rows and 11 columns, as well as the data type of each column variable, which is important moving forward.



```
str(data)
```

```
## 'data.frame':    197130 obs. of  11 variables:
##  $ C.A     : Factor w/ 733 levels "A002","A006",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ UNIT    : Factor w/ 465 levels "R001","R003",..: 49 49 49 49 49 49 49 49 49 49 ...
##  $ SCP     : Factor w/ 217 levels "00-00-00","00-00-01",..: 116 116 116 116 116 116 116 1
16 116 116 ...
##  $ STATION : Factor w/ 377 levels "1 AV","103 ST",..: 85 85 85 85 85 85 85 85 85 85 ...
##  $ LINENAME: Factor w/ 114 levels "1","123","1237ACENQRS",..: 104 104 104 104 104 104 104
 104 104 104 ...
##  $ DIVISION: Factor w/ 6 levels "BMT","IND","IRT",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ DATE    : Factor w/ 7 levels "06/30/2018","07/01/2018",..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ TIME    : Factor w/ 11745 levels "00:00:00","00:00:07",..: 1 1929 3948 5895 7852 9797
1 1929 3948 5895 ...
##  $ DESC    : Factor w/ 2 levels "RECOVR AUD","REGULAR": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ENTRIES : int  6675523 6675538 6675554 6675652 6675826 6676110 6676241 6676248 6676256
 6676339 ...
##  $ EXITS   : num  2262828 2262836 2262863 2262941 2262980 ...
```

Now that structure is understood, get the head of the data set. The first six rows are displayed to give you a snapshot.



```
head(data)
```

```
##    C.A UNIT      SCP STATION LINENAME DIVISION       DATE     TIME    DESC
## 1 A002 R051 02-00-00   59 ST  NQR456W      BMT 06/30/2018 00:00:00 REGULAR
## 2 A002 R051 02-00-00   59 ST  NQR456W      BMT 06/30/2018 04:00:00 REGULAR
## 3 A002 R051 02-00-00   59 ST  NQR456W      BMT 06/30/2018 08:00:00 REGULAR
## 4 A002 R051 02-00-00   59 ST  NQR456W      BMT 06/30/2018 12:00:00 REGULAR
## 5 A002 R051 02-00-00   59 ST  NQR456W      BMT 06/30/2018 16:00:00 REGULAR
## 6 A002 R051 02-00-00   59 ST  NQR456W      BMT 06/30/2018 20:00:00 REGULAR
##   ENTRIES   EXITS
## 1 6675523 2262828
## 2 6675538 2262836
## 3 6675554 2262863
## 4 6675652 2262941
## 5 6675826 2262980
## 6 6676110 2263019
```

From here you can slowly tinker with and manipulate the data. The process was started after figuring out that audits were performed every four hours, so the most important times were midnight on the first day the audit week (00:00:00 saturday) and 8pm of the last day (20:00:00 friday), which we filtered for in the code below.

After filtering the important times, it was realized that turnstile values were continuous counts from the day they were installed, so you have to subtract the end of week turnstile count from the start of week. The mutate code below was used to create two new columns, "Total_Enter" and "Total_Exit". Since the values were in the same column, you have to subtract from the previous row, which is where you see the function "lag" being used.

To filter out erroneous values, which to this point are the odd rows, we again use the filter technique to pull only positive turnstile data on the date 07/06/2018, the end of week friday.

Thereafter you can group the data by station, and use the summarise function to sum up all turnstile values to get totals by station.

It should be noted that Lexington Av/63 and Beach 44 St stations had issues with normal auditing times and they were removed from the data set, as seen in the code.

Lastly the station entry values were arranged in descending order, and the first 20 rows were printed.

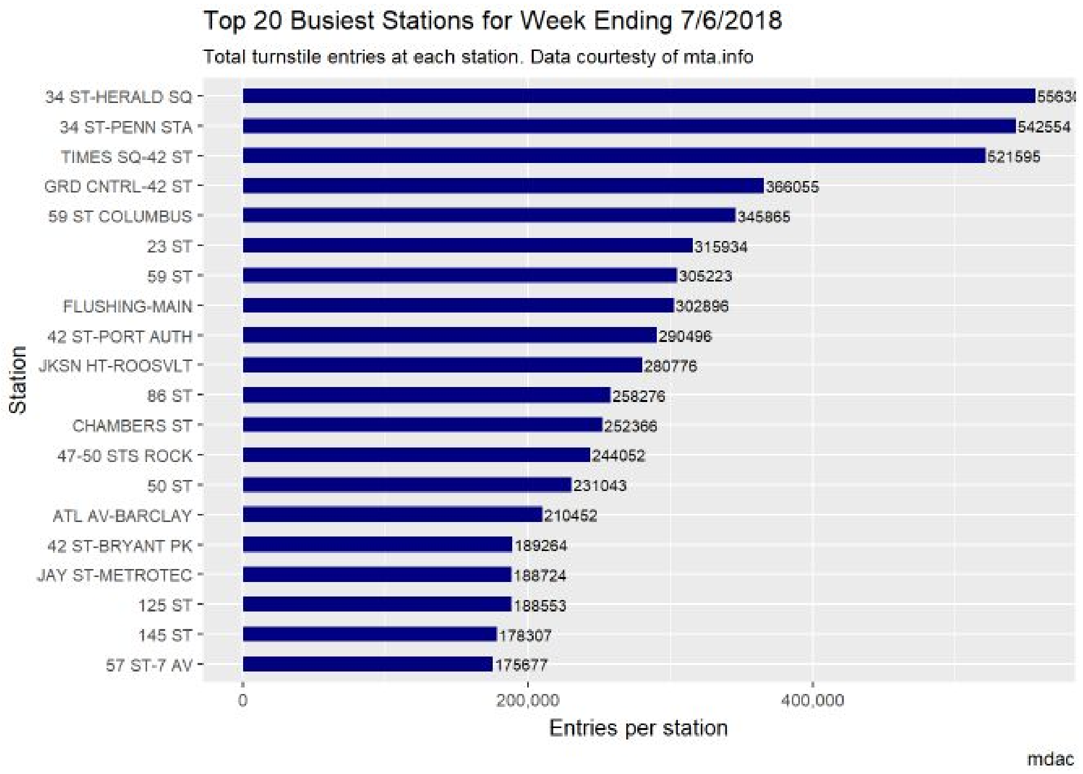
```
data2  <-  data %>%
  filter((DATE == "06/30/2018" & TIME == "00:00:00") | (DATE == "07/06/2018" & TIME == "20:0
0:00")) %>%
  mutate(Total_Enter = ENTRIES - lag(ENTRIES, 1),
         Total_Exit =  EXITS - lag(EXITS, 1)) %>%
  filter(Total_Enter > 0 & DATE == "07/06/2018") %>%
  group_by(STATION) %>%
  summarise(Week_Total_Entries = sum(Total_Enter),
            Week_Total_Exits = sum(Total_Exit)) %>%
  mutate(Difference = Week_Total_Entries - Week_Total_Exits) %>%
  filter(STATION != "LEXINGTON AV/63" & STATION != "BEACH 44 ST") %>%
  arrange(desc(Week_Total_Entries))

data2 %>%
  head(20)
```

```
## # A tibble: 20 x 4
##    STATION         Week_Total_Entries Week_Total_Exits Difference
##    <fct>                        <int>            <dbl>      <dbl>
##  1 34 ST-HERALD SQ             556305           528993      27312
##  2 34 ST-PENN STA              542554           531380      11174
##  3 TIMES SQ-42 ST              521595           510268      11327
##  4 GRD CNTRL-42 ST             366055           348461      17594
##  5 59 ST COLUMBUS              345865           280440      65425
##  6 23 ST                       315934           237103      78831
##  7 59 ST                       305223           244032      61191
##  8 FLUSHING-MAIN               302896           252885      50011
##  9 42 ST-PORT AUTH             290496           224368      66128
## 10 JKSN HT-ROOSVLT             280776           205003      75773
## 11 86 ST                       258276           232997      25279
## 12 CHAMBERS ST                 252366           194828      57538
## 13 47-50 STS ROCK              244052           248929      -4877
## 14 50 ST                       231043           180530      50513
## 15 ATL AV-BARCLAY              210452           207698       2754
## 16 42 ST-BRYANT PK             189264           186432       2832
## 17 JAY ST-METROTEC             188724           111122      77602
## 18 125 ST                      188553           173330      15223
## 19 145 ST                      178307           150954      27353
## 20 57 ST-7 AV                  175677            75377     100300
```

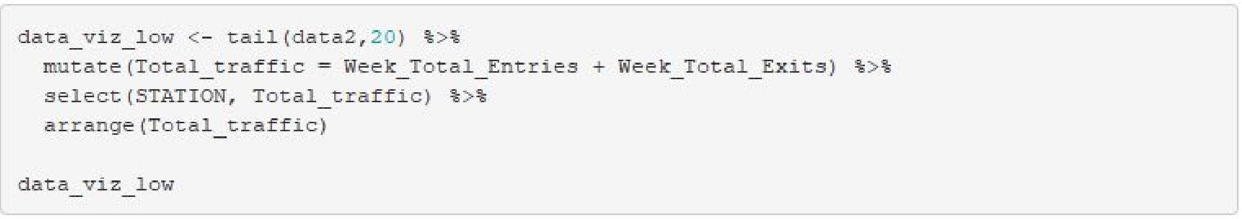
The next set of code subsets the data frame into the first 20 rows, and afterwards we display a horizontal bar chart. **This bar chart was omitted from the report in favor of the simple table.**

```r
data_viz <- data2[1:20,] %>% arrange(desc(Week_Total_Entries))
```

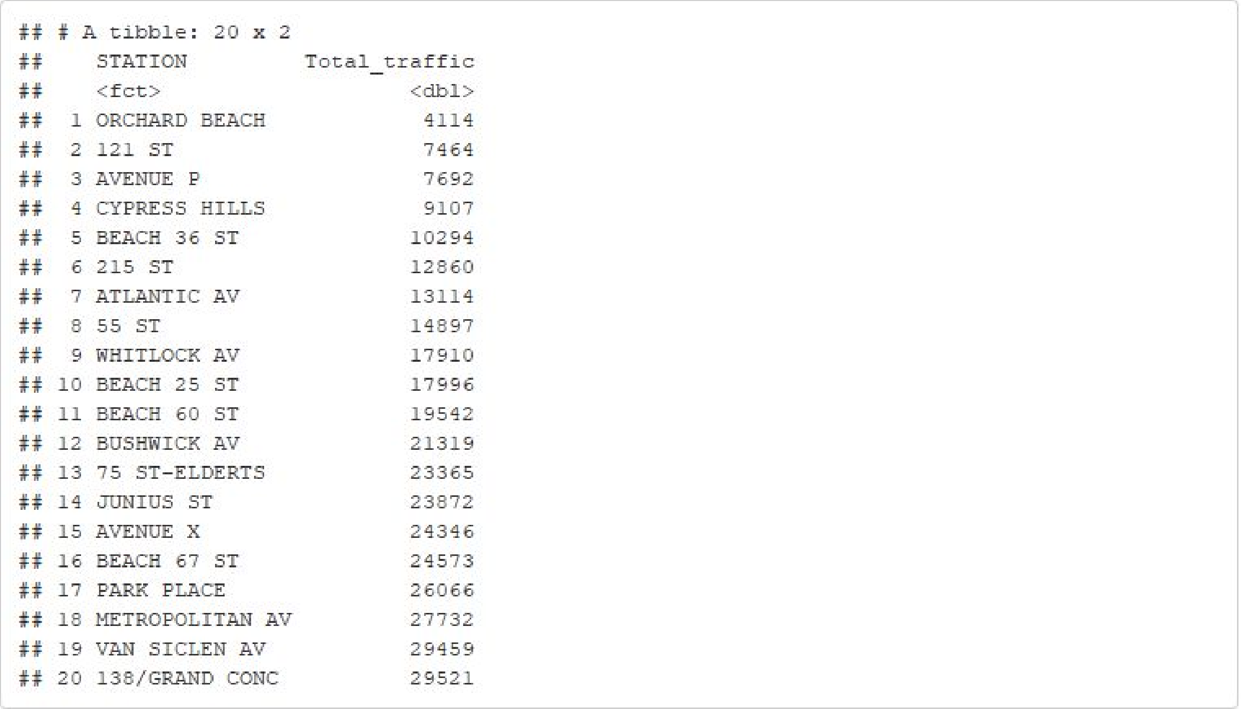
```r
ggplot(data_viz, aes(x = reorder(STATION, Week_Total_Entries), y = Week_Total_Entries)) +
  geom_bar(stat = "identity", fill = "navy", width = .5) +
  geom_text(aes(label=Week_Total_Entries),
            position=position_dodge(width = 0.1), hjust = -0.03, vjust = 0.5, size = 2.5) +
    coord_flip() +
  scale_y_continuous(labels = scales::comma) +
  theme(text = element_text(size = 10)) +
  labs(title = "Top 20 Busiest Stations for Week Ending 7/6/2018",
       subtitle = "Total turnstile entries at each station. Data courtesty of mta.info",
       caption = "mdac",
       x = "Station",
       y = "Entries per station")
```



The data frame was afterwards altered to take a look at the bottom bucket. This time total entries and exits are combined.



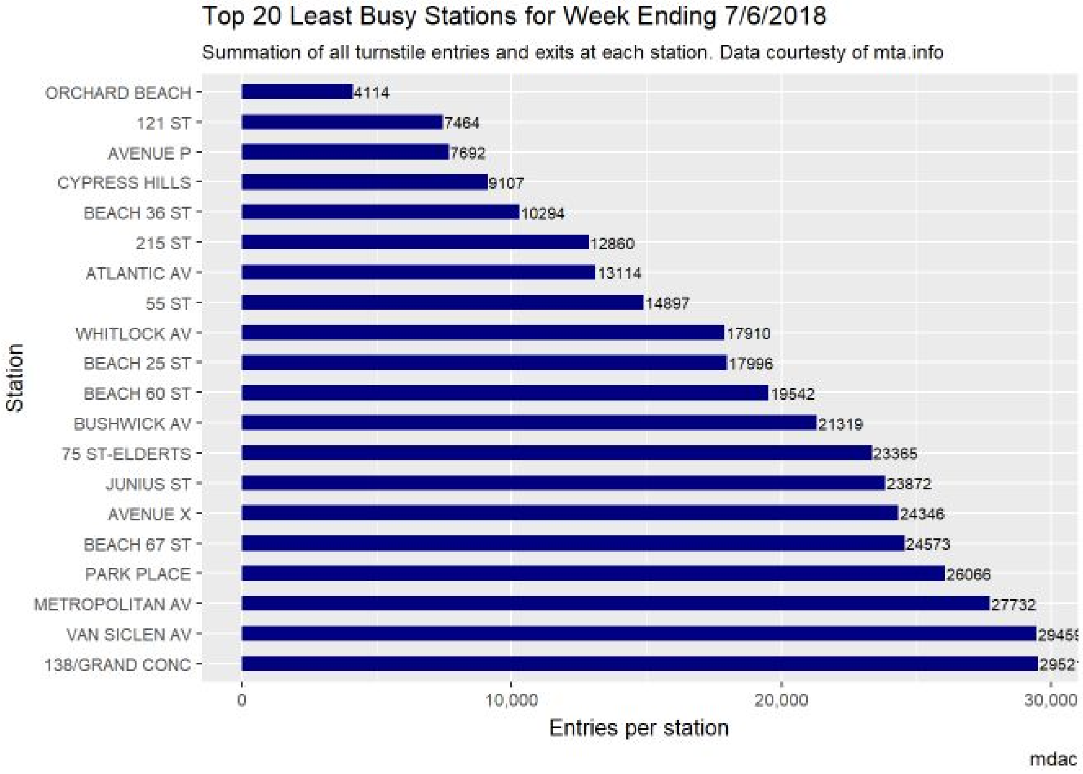
```
data_viz_low <- tail(data2,20) %>%
  mutate(Total_traffic = Week_Total_Entries + Week_Total_Exits) %>%
  select(STATION, Total_traffic) %>%
  arrange(Total_traffic)

data_viz_low
```



```
## # A tibble: 20 x 2
##    STATION         Total_traffic
##    <fct>                   <dbl>
##  1 ORCHARD BEACH            4114
##  2 121 ST                   7464
##  3 AVENUE P                 7692
##  4 CYPRESS HILLS            9107
##  5 BEACH 36 ST             10294
##  6 215 ST                  12860
##  7 ATLANTIC AV             13114
##  8 55 ST                   14897
##  9 WHITLOCK AV             17910
## 10 BEACH 25 ST             17996
## 11 BEACH 60 ST             19542
## 12 BUSHWICK AV             21319
## 13 75 ST-ELDERTS           23365
## 14 JUNIUS ST               23872
## 15 AVENUE X                24346
## 16 BEACH 67 ST             24573
## 17 PARK PLACE              26066
## 18 METROPOLITAN AV         27732
## 19 VAN SICLEN AV           29459
## 20 138/GRAND CONC          29521
```

And the chart is displayed.

```
ggplot(data_viz_low, aes(x = reorder(STATION, -Total_traffic), y = Total_traffic)) +
  geom_bar(stat = "identity", fill = "navy", width = .5) +
  geom_text(aes(label = Total_traffic),
            position = position_dodge(width = 0.1), hjust = -0.03, vjust = 0.5, size = 2.5)
+
  coord_flip() +
  scale_y_continuous(labels = scales::comma) +
  theme(text = element_text(size = 10)) +
  labs(title = "Top 20 Least Busy Stations for Week Ending 7/6/2018",
       subtitle = "Summation of all turnstile entries and exits at each station. Data courte
sty of mta.info",
       caption = "mdac",
       x = "Station",
       y = "Entries per station")
```



[END APPENDIX]