```
##
## 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
```

Research question

In New York City (NYC), there is since 2013 a paying bike sharing system "Citibike". Riders can rent bike at various docking stations throughout the city and returned them to another docking station. There are 2 main forms of payment; "pay as you go" meaning per ride or "Annual Subscription" meaning pay a flat fee for year with unlimited rides There is a time limit on how long the bike can be in use per ride; 30 minutes for non-subscribers and 45 minutes for subscribers. Financial penalties are applied in the cases the ride exceed these limits.

We are interested to explore whether there is any relationship between the age of rider and ride duration. Furthermore whether the gender of rider, or whether the ride is on weekday vs weekend day impact the ride duration. Since the data only contains additional information such as birth year and gender for rider that are annual subscribers, we will limit our analysis to this subset. In addition we will only limit the data set to the last 3 months (September, October, and December) of 2014.

Cases

The raw data is a record of every ride in the system (for the months of September, October, and December 2014) with the following characteristics;

- with a duration of > 1 minute
- that begin at publicly available stations (thereby excluding trips that originate at citibike depots for rebalancing or maintenance purposes).

The data includes the following fields:
Trip Duration (seconds) * Start Time and Date * Stop Time and Date * Start Station Name * End Station Name * Station ID * Station Lat/Long * Bike ID * User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member) * Gender (Zero=unknown; 1=male; 2=female) * Year of Birth

For the purpose of our analysis, we will subset the data to User Type = "Subscriber". In addition, we will calculate the age of rider as follows:
Age = Year of ride - Year of Birth

We will derive whether the ride occurred on a weekday or a weekend/holiday as well the new field: rideday will have values 1 = Weekday, 0 = Weekend or Holiday
Note: dayofweek is as follows: 1 - Sunday, 2 - Monday, 3 - Tuesday, 4 - Wednesday, 5 - Thursday, 6 - Friday, 7 - Saturday

Finally, the ride duration will be converted from seconds to minutes and round to nearest whole minute number (up or down). The new file durationminute will be added to data set.

Data collection

The data was collected by the operator of the system. The data is captured by the docking stations and centralized.

Type of study

The data is collected by "observations", it represents the actual experience of the riders that use bike sharing system.

Data Source

The raw data can be found at:
[http://www.citibikenyc.com/system-data]

Response

Ride duration in minutes, numerical This variable represents the ride duration in minutes.

Explanatory

Age (numerical), Gender of rider (categorical), weekday category (categorica) weekday category (rideday in data set) is a derive variable see R section above that indicate whether the ride took place on a weekday or a weekend day (or Holiday, in this case 10/13/2014 Columbus day)

Relevant summary statistics

Provide summary statistics relevant to your research question. For example, if you are comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Summary statistics on our selected data set:

```
##       age         durationminute        rideday          gender
##  Min.   : 16.00   Min.   :    1.00   Min.   :0.000   Min.   :0.000
##  1st Qu.: 29.00   1st Qu.:    6.00   1st Qu.:0.000   1st Qu.:1.000
##  Median : 35.00   Median :   10.00   Median :1.000   Median :1.000
##  Mean   : 37.59   Mean   :   13.28   Mean   :0.688   Mean   :1.226
##  3rd Qu.: 45.00   3rd Qu.:   15.00   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :115.00   Max.   :90277.00   Max.   :1.000   Max.   :2.000
##  NA's   :5
##    dayofweek
##  Min.   :1.00
##  1st Qu.:3.00
##  Median :4.00
##  Mean   :4.09
##  3rd Qu.:6.00
##  Max.   :7.00
##
```

From this summary, we can see that with have 5 rows with NA for age. The reason being is that the Year of Birth is not available for these entries. We will only consider complete row for our analysis.
Looking at duration in minute, we have a number that clearly represent an outlier with a value = 90277.00
Note: Subscriber my ride without monetary penalty for the first 45 minutes of every ride.

Also, we have some observations for which gender is not specified (gender = 0). The values for gender are Male = 1, Female = 2. We will remove the rows where gender = 0.

Pricing (Current rate as of 2016) $155 / year +$2.50 +$9 +$9 $14.95 / month with annual commitment +$2.50 +$9 +$9 between 45-75 minutes between 75-105 minutes every additional 30 minutes after

Checking the data set against "comple.case" function give us the 5 observation with missing age. We will remove these entries from data set.

```
##       age         durationminute        rideday          gender
##  Min.   : 16.00   Min.   :    1.00   Min.   :0.000   Min.   :1.000
##  1st Qu.: 29.00   1st Qu.:    6.00   1st Qu.:0.000   1st Qu.:1.000
##  Median : 35.00   Median :   10.00   Median :1.000   Median :1.000
##  Mean   : 37.59   Mean   :   13.28   Mean   :0.688   Mean   :1.227
##  3rd Qu.: 45.00   3rd Qu.:   15.00   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :115.00   Max.   :90277.00   Max.   :1.000   Max.   :2.000
##    dayofweek
##  Min.   :1.00
##  1st Qu.:3.00
##  Median :4.00
##  Mean   :4.09
##  3rd Qu.:6.00
##  Max.   :7.00
```

Some summary statistics, we will run some additional summary statistics for Age and Ride Duration in minutes. Also we will these statistics separating these by gender and/or ride is on a week day or weekend day.

```
##   max(age) min(age) mean(age)    n()
## 1      115       16  37.59325 759343
```

```
##   max(durationminute) min(durationminute) mean(durationminute)    n()
## 1               90277                   1             13.27842 759343
```

```
## Source: local data frame [2 x 5]
##
##   gender max(age) min(age) mean(age)    n()
##    (int)    (dbl)    (dbl)     (dbl)  (int)
## 1      1      115       16  37.93170 587232
## 2      2      115       16  36.43848 172111
```

```
## Source: local data frame [2 x 5]
##
##   gender max(durationminute) min(durationminute) mean(durationminute)
##    (int)               (dbl)               (dbl)                (dbl)
## 1      1               90277                   1             12.86131
## 2      2               13040                   1             14.70156
## Variables not shown: n() (int)
```

```
## Source: local data frame [2 x 5]
##
##   rideday max(age) min(age) mean(age)    n()
##     (dbl)    (dbl)    (dbl)     (dbl)  (int)
## 1       0      115       16  37.30408 236900
## 2       1      115       16  37.72437 522443
```

```
## Source: local data frame [2 x 5]
##
##   rideday max(durationminute) min(durationminute) mean(durationminute)
##     (dbl)               (dbl)               (dbl)                (dbl)
## 1       0                7659                   1             13.01223
## 2       1               90277                   1             13.39912
## Variables not shown: n() (int)
```
