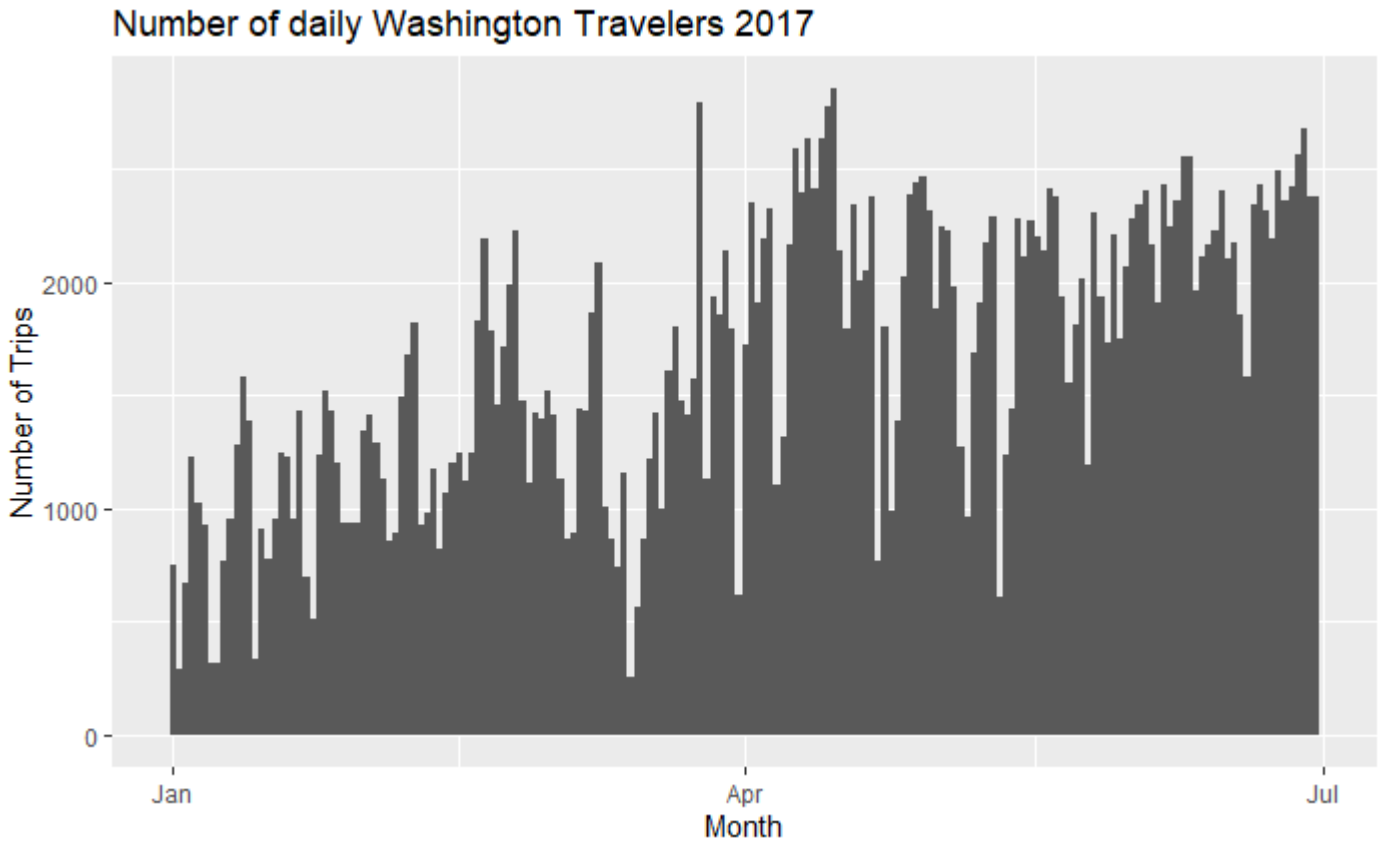
Explore Bike Share Data

Code

````{ Initialization and Importing} library(ggplot2) nyc <- read.csv('new-york-city.csv', sep = ',') wash <- read.csv('washington.csv', sep = ',') chi <- read.csv('chicago.csv', sep = ',')

names(nyc) names(wash) head(chi) ````

## Question 1




Hide

```
# Washington #
wash.Start.Date <- as.Date(c(wash$Start.Time)) # Get Year-Month-Day info from NYC
ggplot(data = wash, aes(x = wash.Start.Date)) + # Plot the number of riders per day/month of 201
7
  geom_histogram(binwidth = 1) +
  # Setting the graph title and axis labels
  ggtitle('Number of daily Washington Travelers 2017') +
  xlab('Month') +
  ylab('Number of Trips')

# Summary
wash$Start.Month <- format(as.Date(wash$Start.Time), "%m")
table(unlist(wash$Start.Month))
```
```

```
   01    02    03    04    05    06
30053 38932 41863 62620 58193 68339
```

Hide

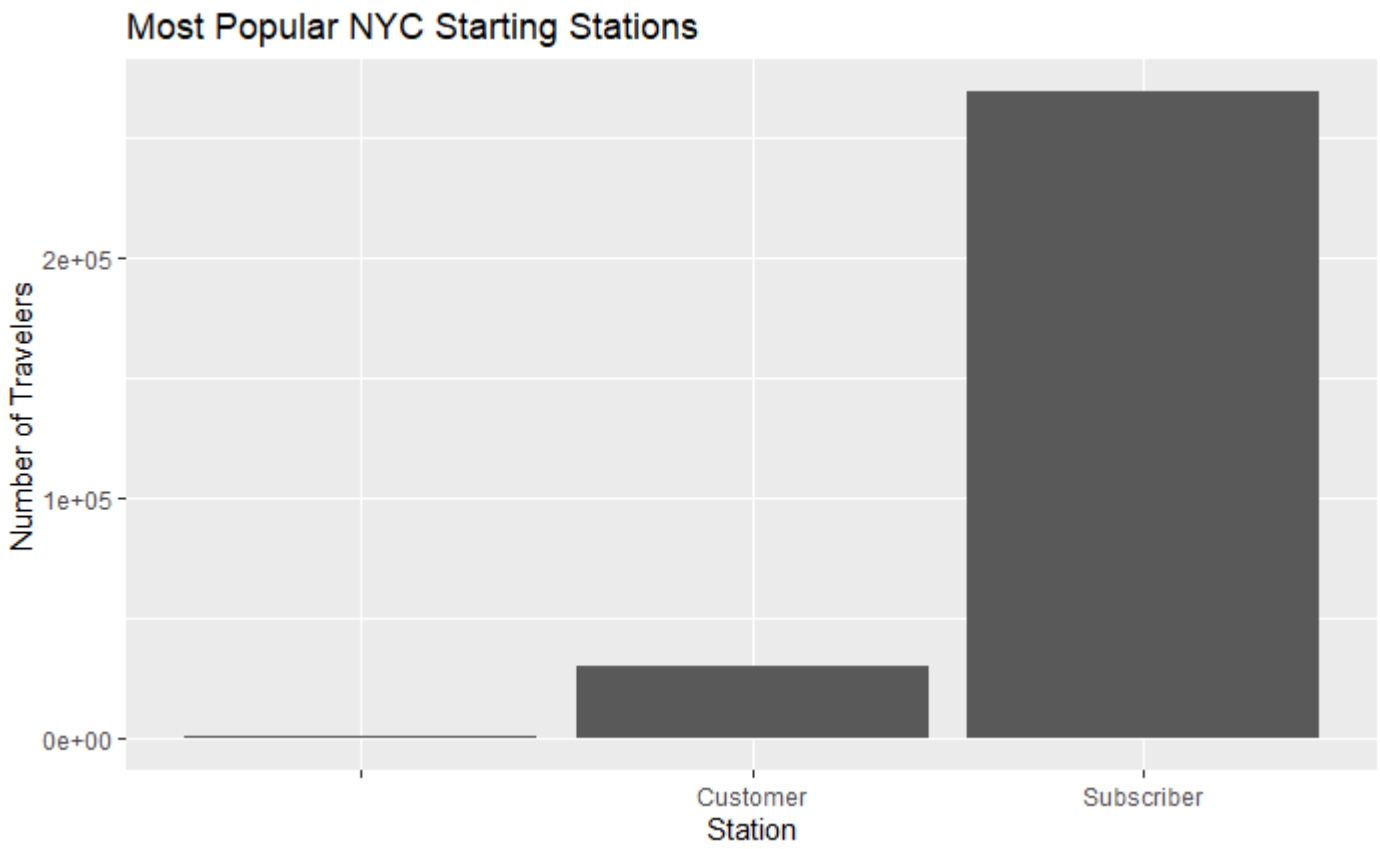
```
# Analysis:
  # Based on the chart the most popular time to travel in Washington is around April - May. This
is supported by the summary table that counts June as having 68339 travelers and January as the
 least popular with 30053
  # There is an increase in travel closer to the summer.
  # The chart and summary makes it apparent that only data from January 2017 to the end of June
 was collected for this dataset.
# Issues:
  # Unfortunately the summary seems to count the dates as integer values with maximum and minimu
m values being the numeric value of the start and end month/day without taking into consideratio
n of their occurrences.
```

## Question 2

Hide

```
ggplot(data = subset(nyc, !is.na(User.Type)), aes(x = nyc$User.Type)) + # Plot the number of cus
tomers and subscribers
  geom_histogram(binwidth = 1, stat='count') +
  # Setting the graph title and axis labels
  ggtitle('Most Popular NYC Starting Stations') +
  xlab('Station') +
  ylab('Number of Travelers')
```

```
Ignoring unknown parameters: binwidth, bins, pad
```




Hide

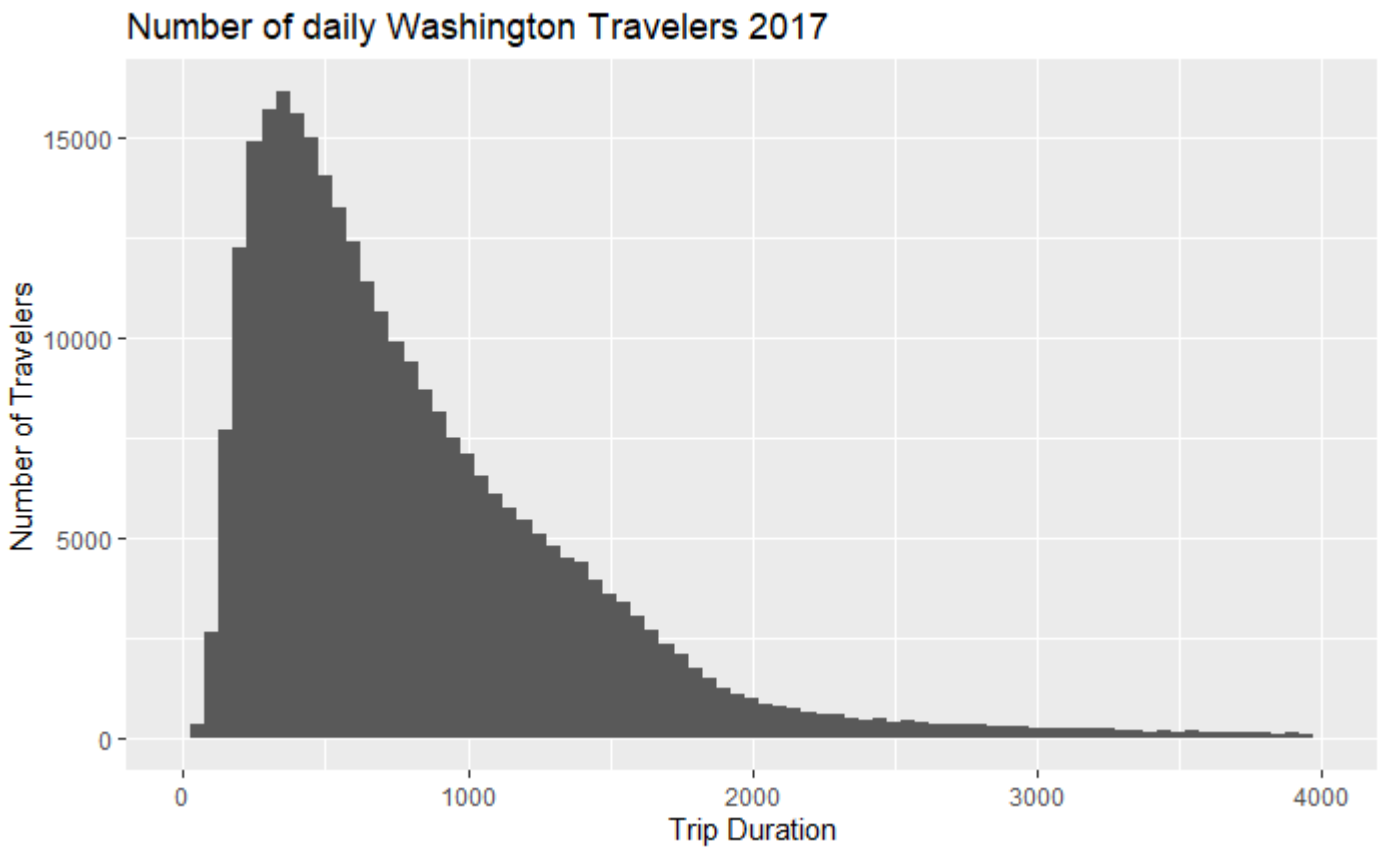
```
# Summary
table(unlist(nyc$User.Type))
```

```
            Customer Subscriber
      692      30159     269149
```

Hide

```
# Analysis
  # There are many more subscribers, 269149, than customers, 30159.
# Issues:
  # Even though a subset filter was used to factor out the null variables, there are still some
 blank values.
```

####Question 3




Hide

```
#Plotting the graph while removing null values where gender is NA
ggplot(aes(x = chi$Trip.Duration), data = chi ) +
  geom_histogram(binwidth = 50) +
  scale_x_continuous(limits = c(0, 4000)) +
  # Giving graph title and axis labels
  ggtitle('Number of daily Washington Travelers 2017') +
  xlab('Trip Duration') +
  ylab('Number of Travelers')

# Summary to find stats of female versus male travel duration
table(chi$Gender)
```

```
      Female   Male
 61052  57758 181190
```

Hide

```
by(chi$Trip.Duration,chi$Gender,summary)
```

```
chi$Gender:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     60     928    1380    1863    1961   86224
------------------------------------------------------------------------------------------------
----------
chi$Gender: Female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   400.0   648.0   781.5  1012.0 85742.0
------------------------------------------------------------------------------------------------
----------
chi$Gender: Male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   339.0   542.0   673.2   860.0 85572.0
```

Hide

```
# Analysis: Looking at the graph one can tell that most riders travel for less than 750 minutes.
  # Looking at the summary, it appears that women have a higher average for travel duration than
men even though there tend to be more male travelers.
  # The summary also shares that the longest travel duration by a female was 85,742.
# Issues:
  # Since the dataset was so large, a facet_wrap/facet_grid to show female versus male was takin
g too long.
  # I decided to not remove null genders because I noticed users who aren't subscribers still co
ntributed quite a bit.
```