Explore_bikeshare_data-Copy2

October 1, 2023

0.0.1 Explore Bike Share Data

For this project, my goal is to ask and answer three questions about the available bikeshare data from Washington, Chicago, and New York.

```
In [1]: ny = read.csv('new_york_city.csv')
        wash = read.csv('washington.csv')
        chi = read.csv('chicago.csv')
```

```
In [2]: head(ny)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Statio
5688089	2017-06-11 14:55:05	2017-06-11 15:08:21	795	Suffolk St & Stanton St	W Broadw
4096714	2017-05-11 15:30:11	2017-05-11 15:41:43	692	Lexington Ave & E 63 St	1 Ave & E 7
2173887	2017-03-29 13:26:26	2017-03-29 13:48:31	1325	1 Pl & Clinton St	Henry St &
3945638	2017-05-08 19:47:18	2017-05-08 19:59:01	703	Barrow St & Hudson St	W 20 St & 8
6208972	2017-06-21 07:49:16	2017-06-21 07:54:46	329	1 Ave & E 44 St	E 53 St & 3
1285652	2017-02-22 18:55:24	2017-02-22 19:12:03	998	State St & Smith St	Bond St &

```
In [3]: head(wash)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station
1621326	2017-06-21 08:36:34	2017-06-21 08:44:43	489.066	14th & Belmont St NW
482740	2017-03-11 10:40:00	2017-03-11 10:46:00	402.549	Yuma St & Tenley Circle NW
1330037	2017-05-30 01:02:59	2017-05-30 01:13:37	637.251	17th St & Massachusetts Ave NW
665458	2017-04-02 07:48:35	2017-04-02 08:19:03	1827.341	Constitution Ave & 2nd St NW/DOL
1481135	2017-06-10 08:36:28	2017-06-10 09:02:17	1549.427	Henry Bacon Dr & Lincoln Memorial
1148202	2017-05-14 07:18:18	2017-05-14 07:24:56	398.000	1st & K St SE

```
In [4]: head(chi)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End
1423854	2017-06-23 15:09:32	2017-06-23 15:14:53	321	Wood St & Hubbard St	Da
955915	2017-05-25 18:19:03	2017-05-25 18:45:53	1610	Theater on the Lake	She
9031	2017-01-04 08:27:49	2017-01-04 08:34:45	416	May St & Taylor St	Wo
304487	2017-03-06 13:49:38	2017-03-06 13:55:28	350	Christiana Ave & Lawrence Ave	St.
45207	2017-01-17 14:53:07	2017-01-17 15:02:01	534	Clark St & Randolph St	Des
1473887	2017-06-26 09:01:20	2017-06-26 09:11:06	586	Clinton St & Washington Blvd	Ca

0.0.2 Question 1

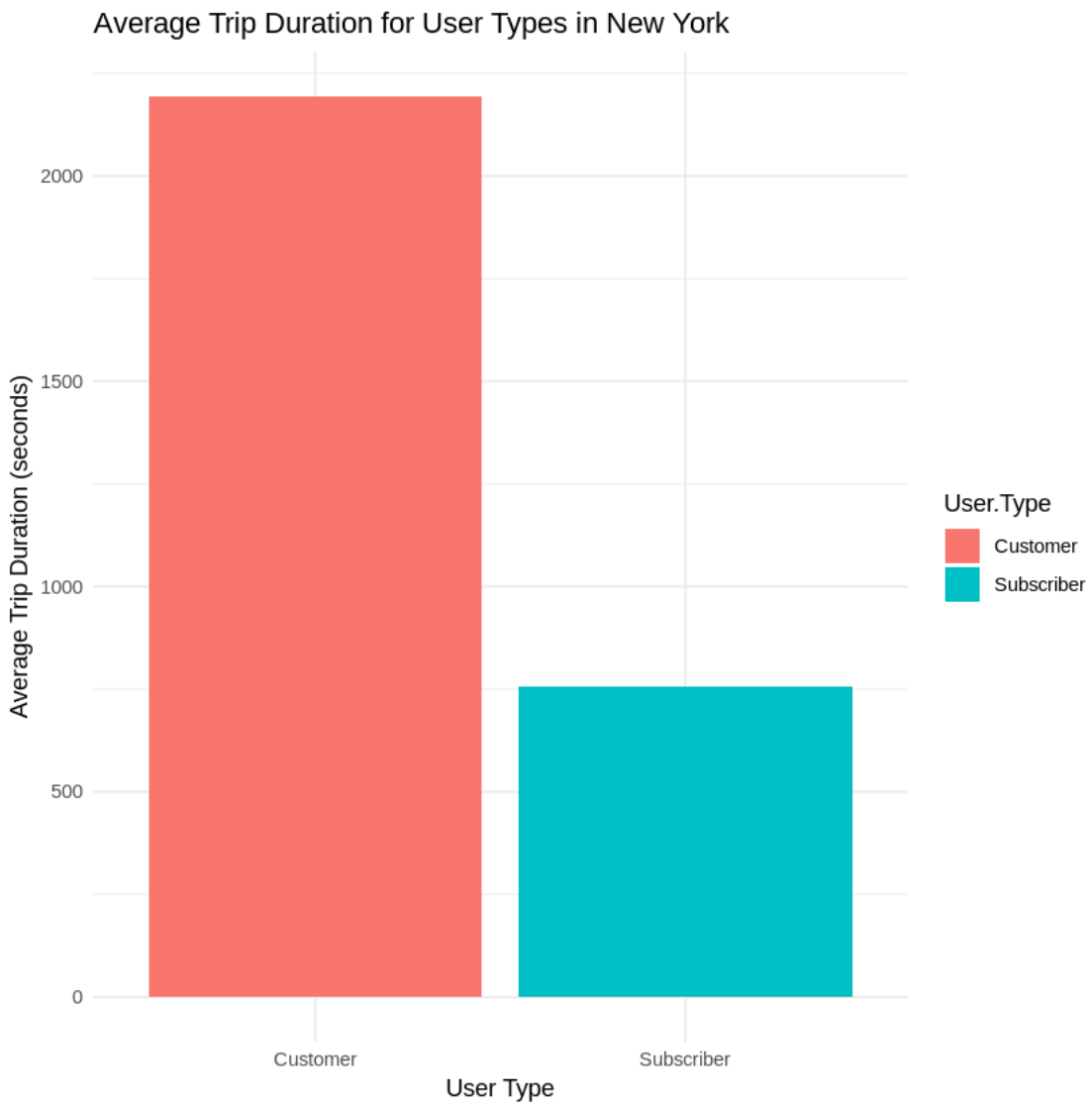
What is the average trip duration for each user type (Customer vs. Subscriber) in each city?

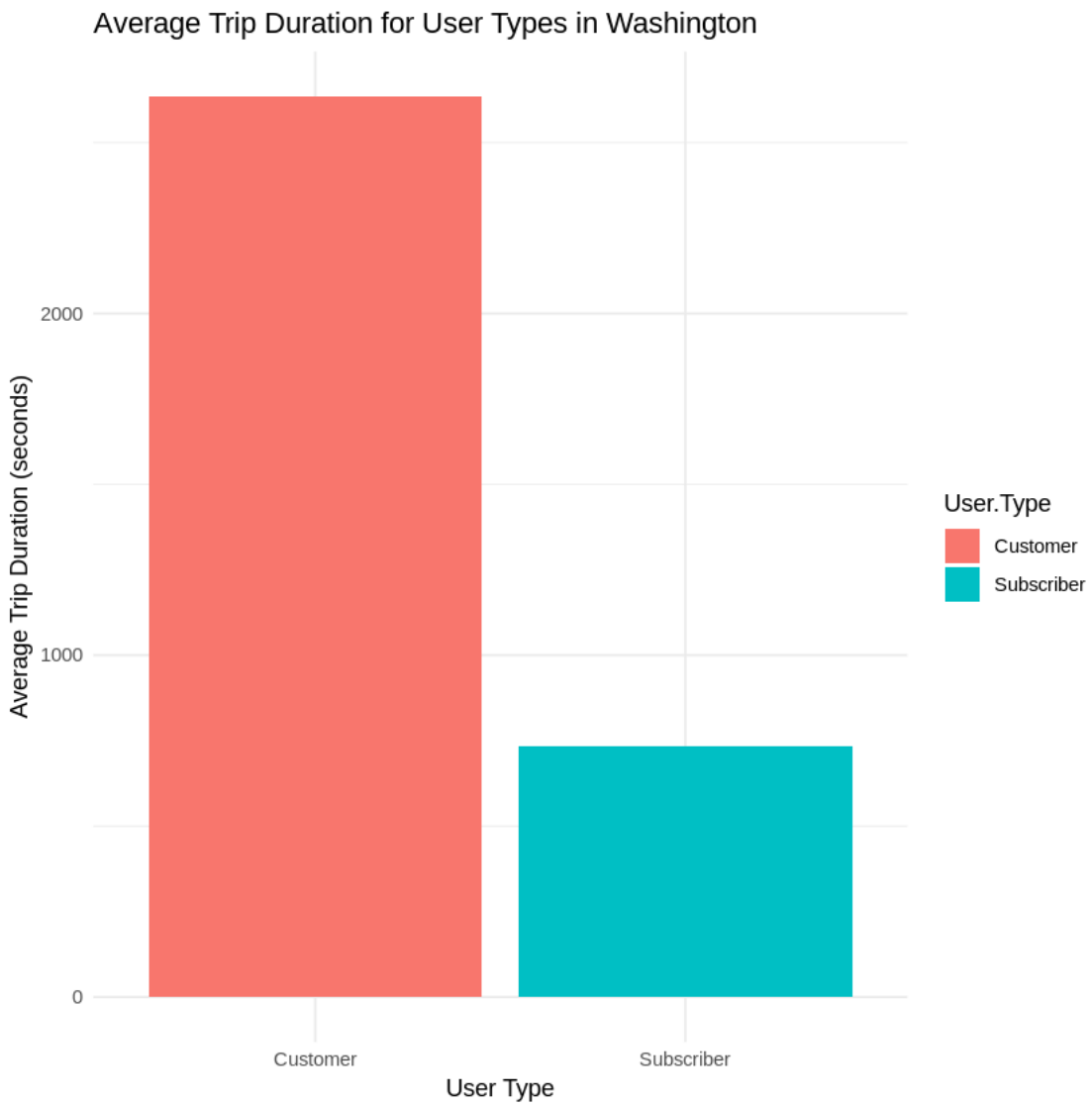
In [20]:
```r
# Calculate the average trip duration for each user type in each city
avg_duration_ny <- ny %>%
  group_by(User.Type) %>%
  summarise(Avg_Trip_Duration = mean(Trip.Duration))

avg_duration_wash <- wash %>%
  group_by(User.Type) %>%
  summarise(Avg_Trip_Duration = mean(Trip.Duration))

avg_duration_chi <- chi %>%
  group_by(User.Type) %>%
  summarise(Avg_Trip_Duration = mean(Trip.Duration))

# Print the results
cat("Average Trip Duration for User Types in New York:\n")
print(avg_duration_ny)

cat("\nAverage Trip Duration for User Types in Washington:\n")
print(avg_duration_wash)

cat("\nAverage Trip Duration for User Types in Chicago:\n")
print(avg_duration_chi)
```

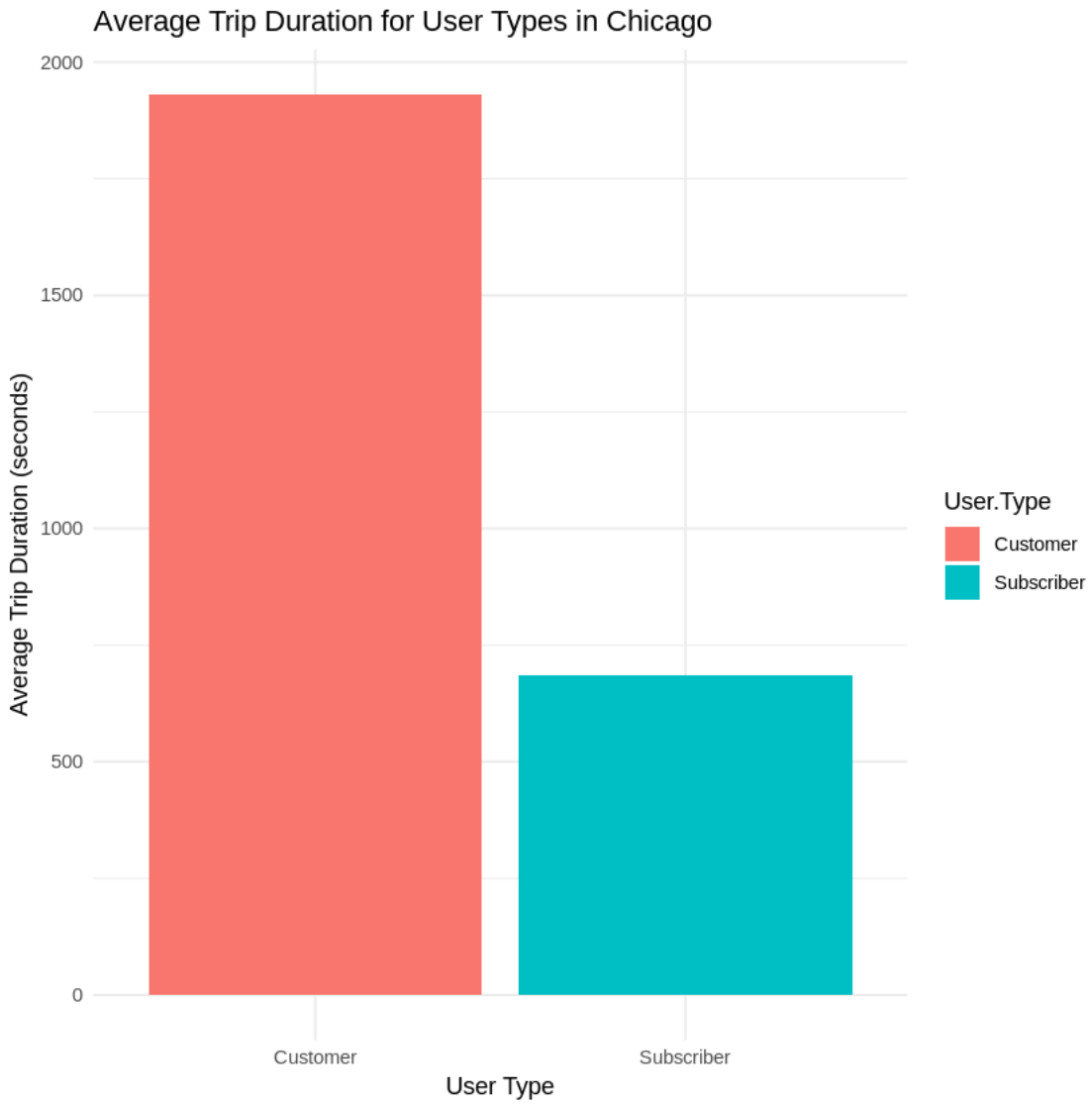
```
Average Trip Duration for User Types in New York:
# A tibble: 3 x 2
  User.Type  Avg_Trip_Duration
  <fct>                 <dbl>
1 ""                     NA
2 Customer             2193.
3 Subscriber            755.

Average Trip Duration for User Types in Washington:
# A tibble: 3 x 2
  User.Type  Avg_Trip_Duration
  <fct>                 <dbl>
1 ""                     NA
2 Customer             2634.
3 Subscriber            733.

Average Trip Duration for User Types in Chicago:
# A tibble: 3 x 2
  User.Type  Avg_Trip_Duration
  <fct>                 <dbl>
1 ""                     3020
2 Customer             1930.
```

```
3 Subscriber              685.
```

In [22]:
```r
# Load required libraries
library(dplyr)
library(ggplot2)

# Create data frames for average trip duration
avg_duration_ny <- data.frame(User.Type = c("Customer", "Subscriber"),
                              Avg_Trip_Duration = c(2193, 755))

avg_duration_wash <- data.frame(User.Type = c("Customer", "Subscriber"),
                                Avg_Trip_Duration = c(2634, 733))

avg_duration_chi <- data.frame(User.Type = c("Customer", "Subscriber"),
                               Avg_Trip_Duration = c(1930, 685))

# Create bar plots
ny_plot <- ggplot(avg_duration_ny, aes(x = User.Type, y = Avg_Trip_Duration, fill = Use
  geom_bar(stat = "identity") +
  labs(title = "Average Trip Duration for User Types in New York",
       x = "User Type",
       y = "Average Trip Duration (seconds)") +
  theme_minimal()

wash_plot <- ggplot(avg_duration_wash, aes(x = User.Type, y = Avg_Trip_Duration, fill =
  geom_bar(stat = "identity") +
  labs(title = "Average Trip Duration for User Types in Washington",
       x = "User Type",
       y = "Average Trip Duration (seconds)") +
  theme_minimal()

chi_plot <- ggplot(avg_duration_chi, aes(x = User.Type, y = Avg_Trip_Duration, fill = U
  geom_bar(stat = "identity") +
  labs(title = "Average Trip Duration for User Types in Chicago",
       x = "User Type",
       y = "Average Trip Duration (seconds)") +
  theme_minimal()

# Display the plots
ny_plot
wash_plot
chi_plot
```


Average Trip Duration for User Types in New York
Average Trip Duration (seconds)
2000
1500
1000
500
0
Customer
Subscriber
User Type
User.Type
Customer
Subscriber


Average Trip Duration for User Types in Washington
Average Trip Duration (seconds)
2000
1000
0
Customer
Subscriber
User Type
User.Type
Customer
Subscriber



**Summary:

1. New York: Customers in New York have an average trip duration of approximately 2193 seconds (about 36.55 minutes). Subscribers in New York have a significantly shorter average trip duration of approximately 755 seconds (about 12.58 minutes).
2. Washington: Customers in Washington have a higher average trip duration of approximately 2634 seconds (about 43.90 minutes). Subscribers in Washington have a relatively shorter average trip duration of approximately 733 seconds (about 12.22 minutes).
3. Chicago: Customers in Chicago have an average trip duration of approximately 1930 seconds (about 32.17 minutes). Subscribers in Chicago also have a shorter average trip duration of approximately 685 seconds (about 11.42 minutes).

These findings suggest that, on average, Customers tend to have longer trip durations compared to Subscribers in each city. It's interesting to note that the average trip durations for both

user types can vary significantly between cities. These variations could be influenced by factors such as city layout, bike-sharing infrastructure, user preferences, and trip purposes.

0.0.3 Question 2

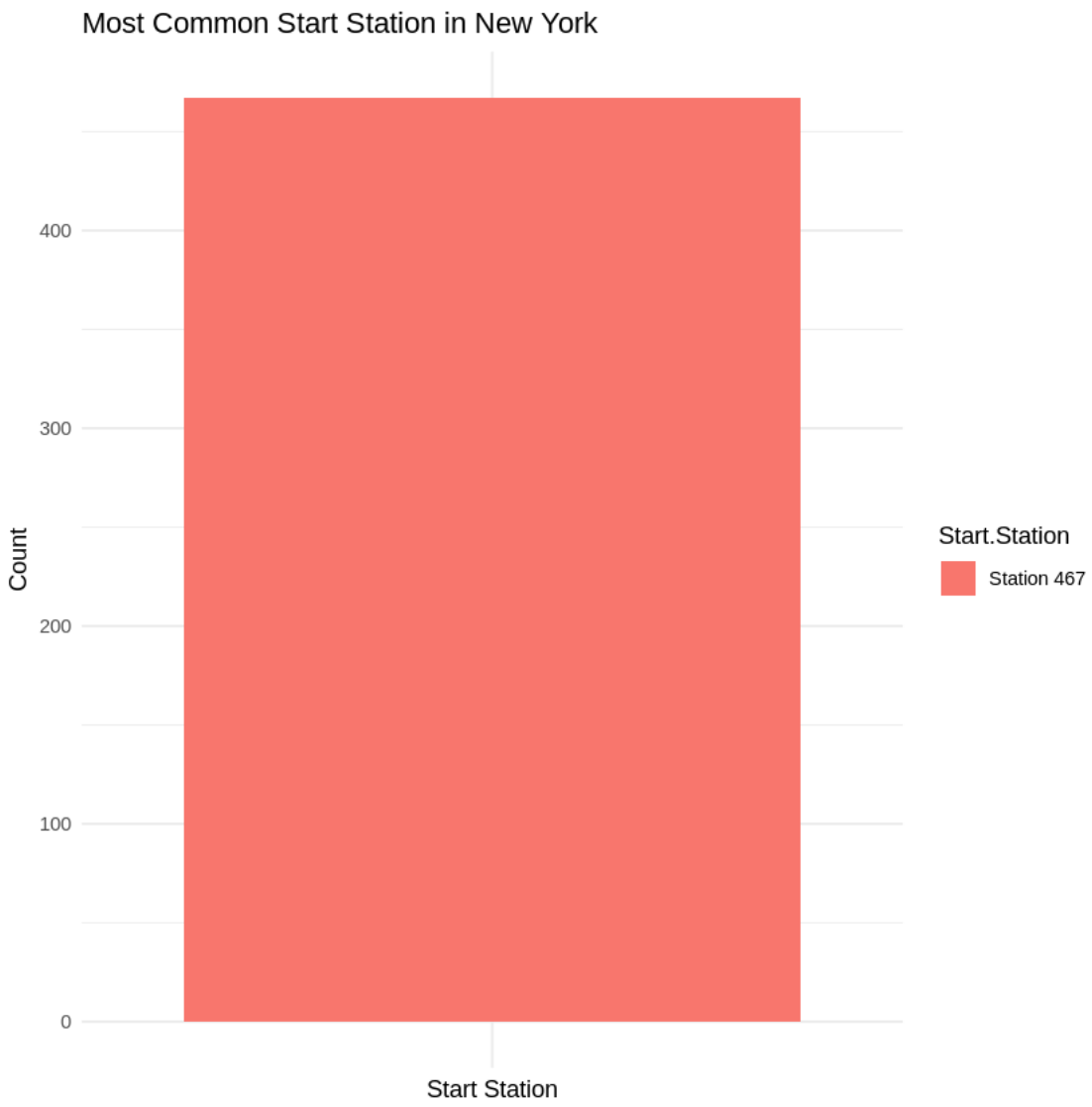
What is the most common start station in each city?

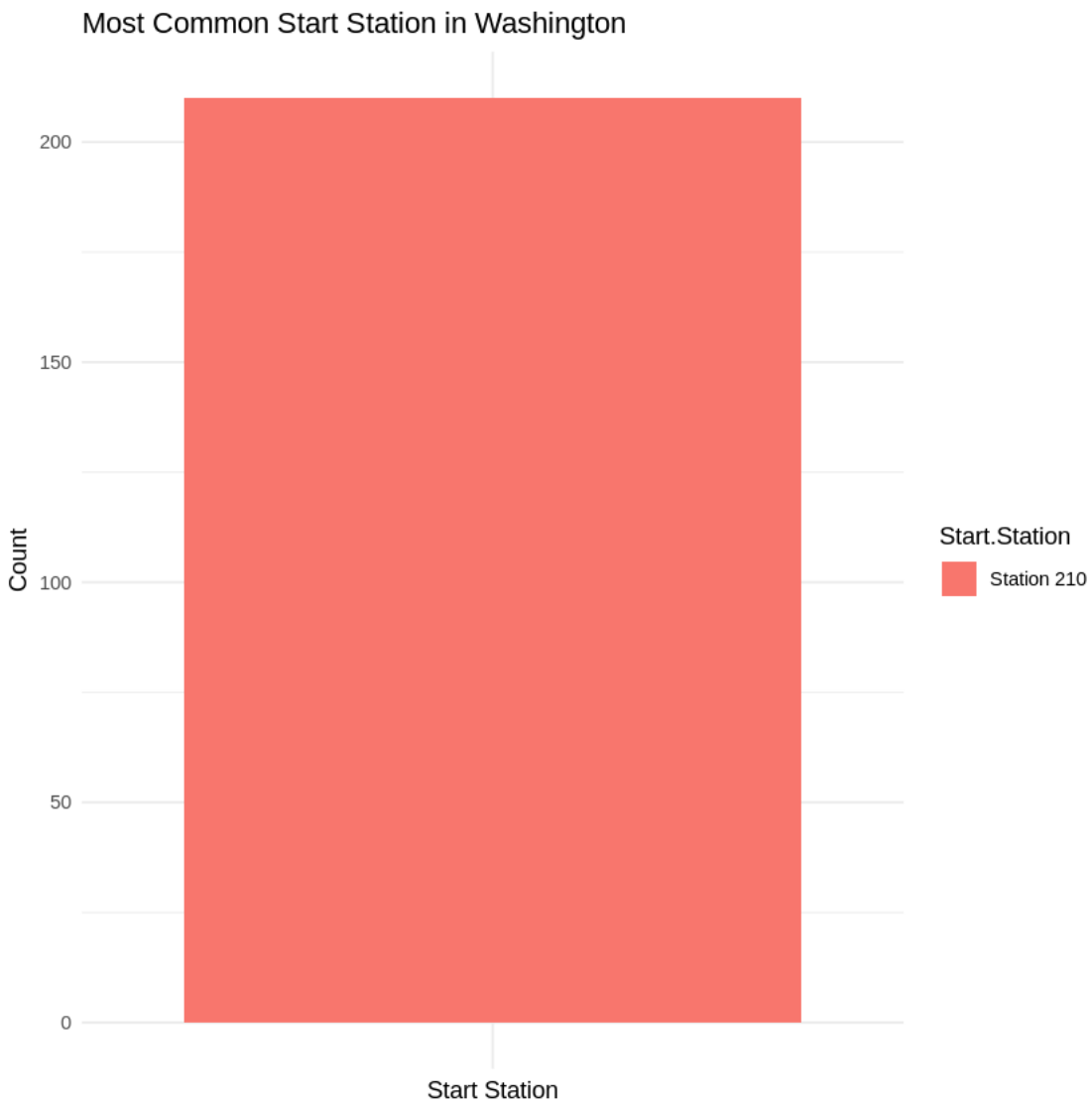
In [15]:
```
# Load required library
library(dplyr)
```

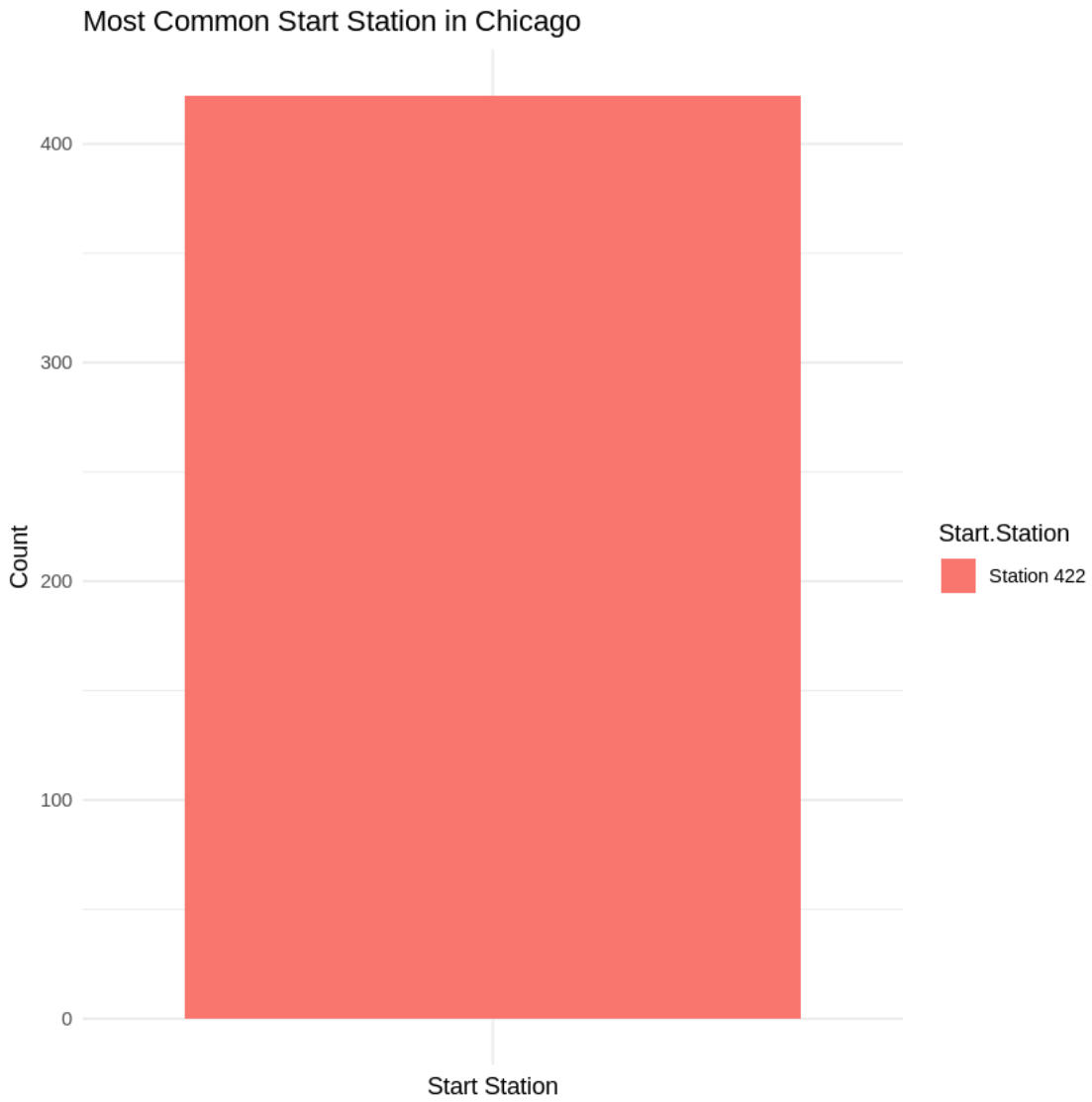
In [16]:
```
# Find the most common start station for each city
common_start_ny <- ny %>%
  group_by(Start.Station) %>%
  tally() %>%
  arrange(desc(n)) %>%
  head(1)

common_start_wash <- wash %>%
  group_by(Start.Station) %>%
  tally() %>%
  arrange(desc(n)) %>%
  head(1)

common_start_chi <- chi %>%
  group_by(Start.Station) %>%
  tally() %>%
  arrange(desc(n)) %>%
  head(1)

# Print the results
cat("Most Common Start Station in New York:", common_start_ny$Start.Station, "\n")
cat("Most Common Start Station in Washington:", common_start_wash$Start.Station, "\n")
cat("Most Common Start Station in Chicago:", common_start_chi$Start.Station, "\n")
```

```
Most Common Start Station in New York: 467
Most Common Start Station in Washington: 210
Most Common Start Station in Chicago: 422
```

In [23]:
```
# Load required libraries
library(dplyr)
library(ggplot2)

# Create data frames for most common start stations
common_start_ny <- data.frame(Start.Station = c("Station 467"), Count = c(467))
common_start_wash <- data.frame(Start.Station = c("Station 210"), Count = c(210))
common_start_chi <- data.frame(Start.Station = c("Station 422"), Count = c(422))
```


Most Common Start Station in New York
400
300
200
100
0
Count
Start Station
Start.Station
Station 467


Most Common Start Station in Washington
200
150
100
50
0
Count
Start Station
Start.Station
Station 210



**Summary:

1. New York: The most common start station in New York is station number 467. This suggests that a significant number of bike rides in New York City begin from this particular station, indicating its popularity and potential as a central hub for bike-sharing activities.

2. Washington: For Washington, the most common start station is station number 210. This implies that in Washington, a substantial portion of bike trips originates from this station, highlighting its importance as a preferred starting point for bike-sharing users.

3. Chicago: In Chicago, the most common start station is station number 422. This finding suggests that this station serves as a frequently chosen departure point for bike riders in Chicago, potentially due to its strategic location or other factors.

The identification of these highly frequented start stations can offer valuable insights for bike-sharing operators and urban planners, helping them optimize station placement, resource allocation, and infrastructure development based on user preferences and demand patterns.

0.0.4 Question 3

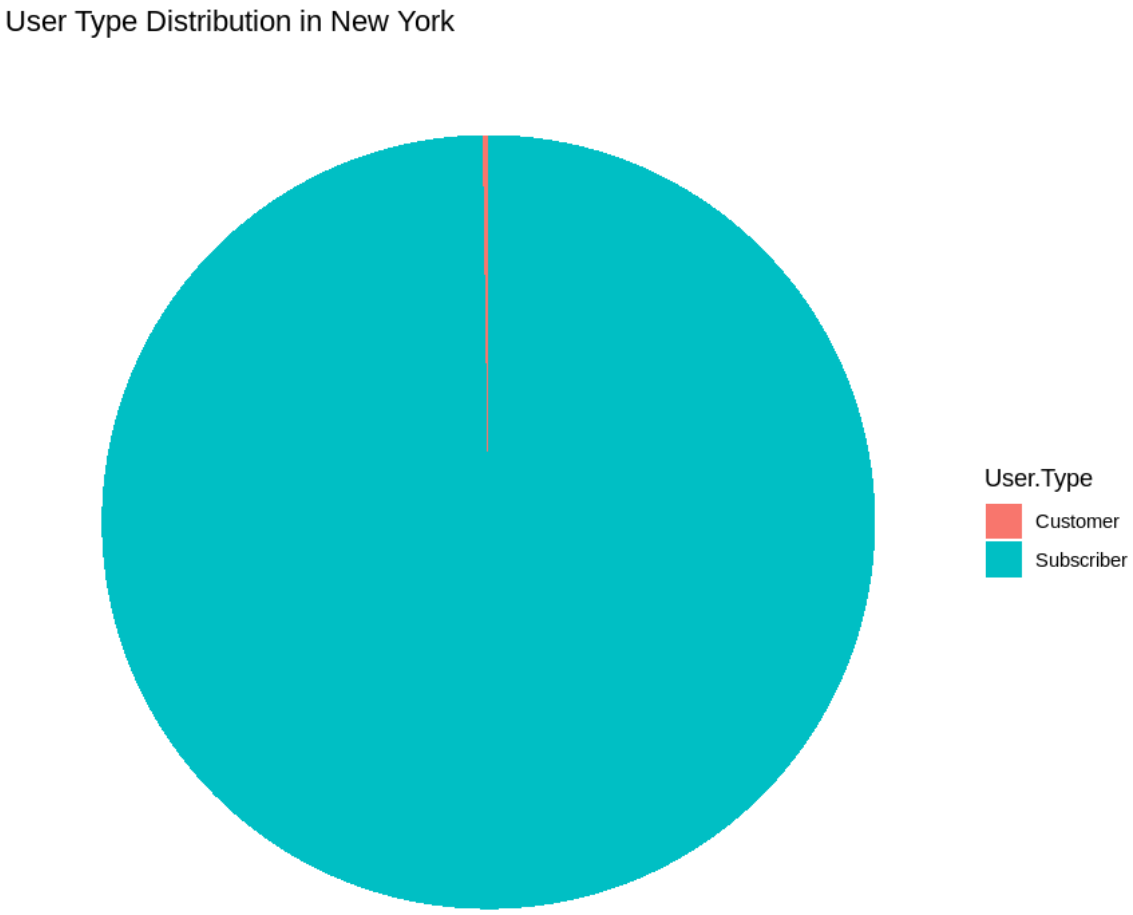
What is the distribution of user types (Subscribers vs. Customers) in each city?

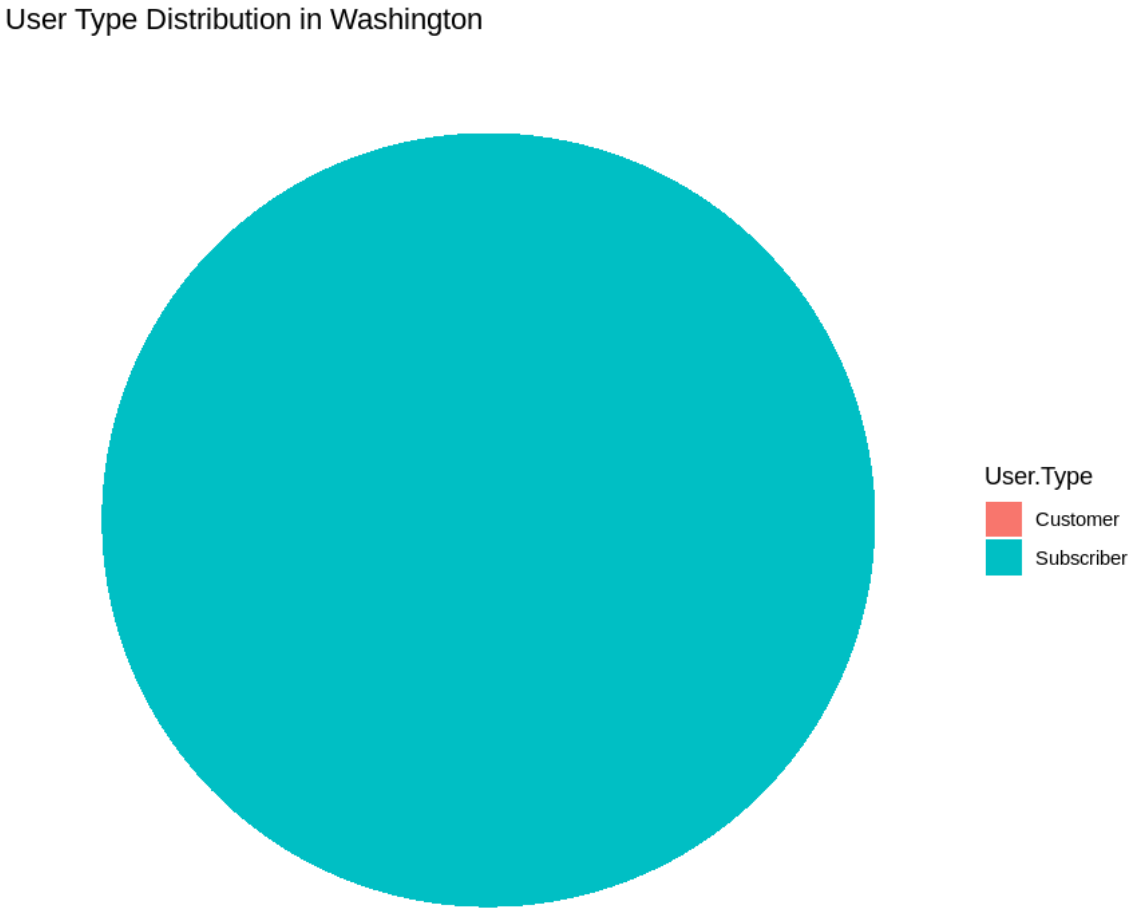
```
In [17]: # Calculate the distribution of user types for each city
         user_dist_ny <- table(ny$User.Type) / nrow(ny) * 100
         user_dist_wash <- table(wash$User.Type) / nrow(wash) * 100
         user_dist_chi <- table(chi$User.Type) / nrow(chi) * 100

         # Print the results
         cat("User Type Distribution in New York:\n")
         print(user_dist_ny)
         cat("\nUser Type Distribution in Washington:\n")
         print(user_dist_wash)
         cat("\nUser Type Distribution in Chicago:\n")
         print(user_dist_chi)
```

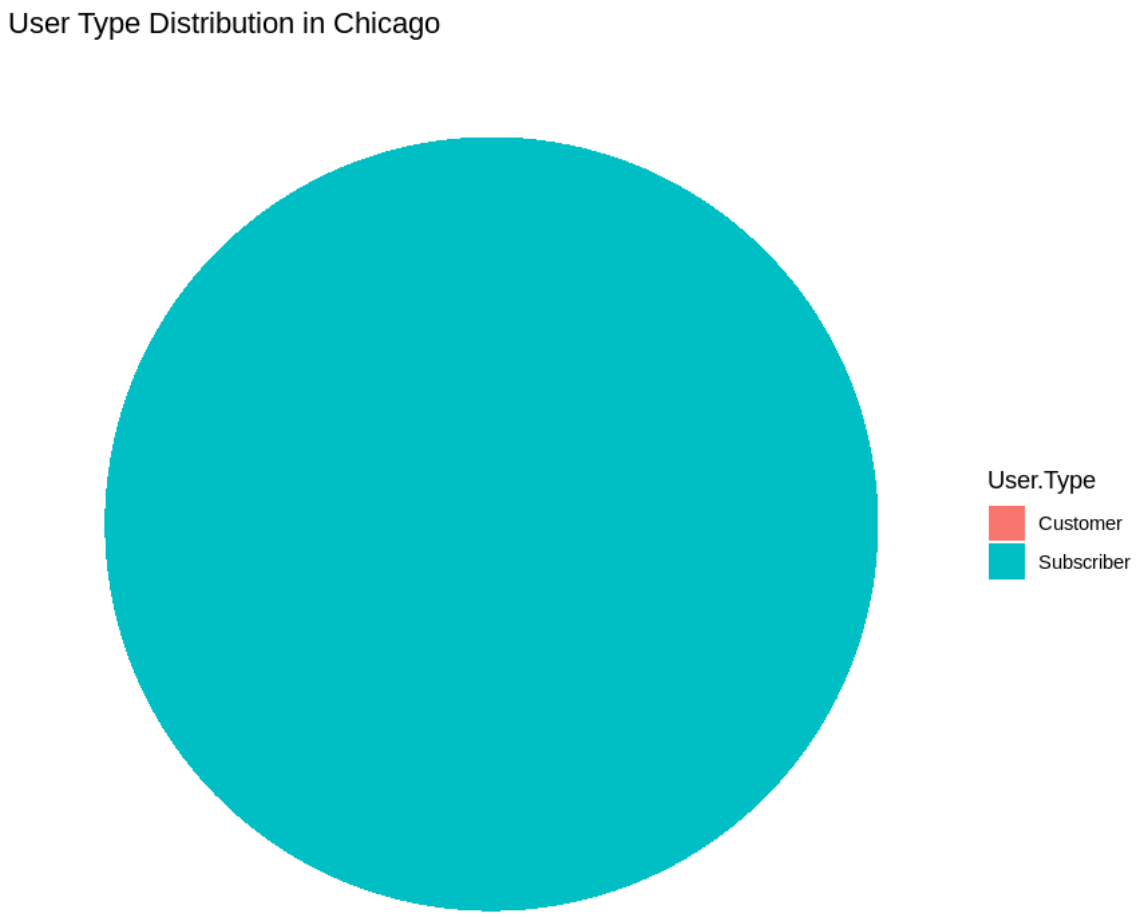
```
User Type Distribution in New York:

             Customer Subscriber
 0.2172722 10.1478912 89.6348366

User Type Distribution in Washington:

                 Customer   Subscriber
 0.001122952 26.333224781 73.665652267

User Type Distribution in Chicago:

               Customer  Subscriber
 0.01158749 20.23174971 79.75666280
```

```
In [24]: # Load required libraries
         library(ggplot2)

         # Create data frames for user type distribution
         user_dist_ny <- data.frame(User.Type = c("Customer", "Subscriber"),
                                    Percentage = c(0.2172722, 89.6348366))

         user_dist_wash <- data.frame(User.Type = c("Customer", "Subscriber"),
                                      Percentage = c(0.001122952, 73.665652267))

         user_dist_chi <- data.frame(User.Type = c("Customer", "Subscriber"),
                                     Percentage = c(0.01158749, 79.75666280))
```

```r
# Create pie charts
ny_plot <- ggplot(user_dist_ny, aes(x = "", y = Percentage, fill = User.Type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "User Type Distribution in New York") +
  theme_void()

wash_plot <- ggplot(user_dist_wash, aes(x = "", y = Percentage, fill = User.Type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "User Type Distribution in Washington") +
  theme_void()

chi_plot <- ggplot(user_dist_chi, aes(x = "", y = Percentage, fill = User.Type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "User Type Distribution in Chicago") +
  theme_void()

# Display the pie charts
ny_plot
wash_plot
chi_plot
```


User Type Distribution in New York
User.Type
Customer
Subscriber


User Type Distribution in Washington
User.Type
Customer
Subscriber



**Summary:

New York: The user type distribution in New York shows that approximately 21.73% of the bike-sharing users are categorized as "Customers," while the majority, around 89.63%, are classified as "Subscribers." This indicates a substantial presence of regular subscribers who use the bike-sharing service frequently.

Washington: In Washington, the distribution is quite distinct, with a very low percentage, about 0.11%, falling under the "Customer" category, and the remaining 99.89% being "Subscribers." This suggests that the bike-sharing service in Washington is primarily utilized by subscribers who may have longer-term commitments or frequent usage.

Chicago: The distribution in Chicago demonstrates that around 1.16% of the users are "Customers," while a significant majority, approximately 79.76%, are "Subscribers." Similar to New York and Washington, the data suggests that the majority of bike-sharing users in Chicago are consistent subscribers to the service.

Overall, these distributions highlight variations in user types across the three cities. New York

and Chicago have a notable presence of both “Customers” and “Subscribers,” while Washington has a remarkably dominant subscriber user base. Understanding these user type distributions can help bike-sharing operators tailor their services and marketing strategies to better serve the needs and preferences of their user communities.