EXPLORE BIKESHARE DATA

PROGRAMMING LANGUAGE: R

Overview

Dataset Information

Prepare Data

Create New Datasets

Questions of exploration

Question 1

Question 2

Question 3

Overview

Over the past decade, bicycle-sharing systems have been growing in number and popularity in cities across the world. Bicycle-sharing systems allow users to rent bicycles on a very short-term basis for a price. This allows people to borrow a bike from point A and return it at point B, though they can also return it to the same location if they'd like to just go for a ride. Regardless, each bike can serve several users per day.

Thanks to the rise in information technologies, it is easy for a user of the system to access a dock within the system to unlock or return bicycles. These technologies also provide a wealth of data that can be used to explore how these bike-sharing systems are used.

In this project, I will use data provided by **Motivate**, a bike share system provider for many major cities in the United States, to uncover bike share usage patterns.

Dataset Information

Randomly selected data for the first six months of 2017 for three cities:

- Chicago
- Washington D.C
- New York City

The dataset includes the following columns:

- Start Time
- End Time
- Trip Duration (in seconds)
- Start Station
- End Station
- User Type (Subscriber or Customer)

The Chicago and New York City files also have the following two columns:

- Gender
- Birth Year

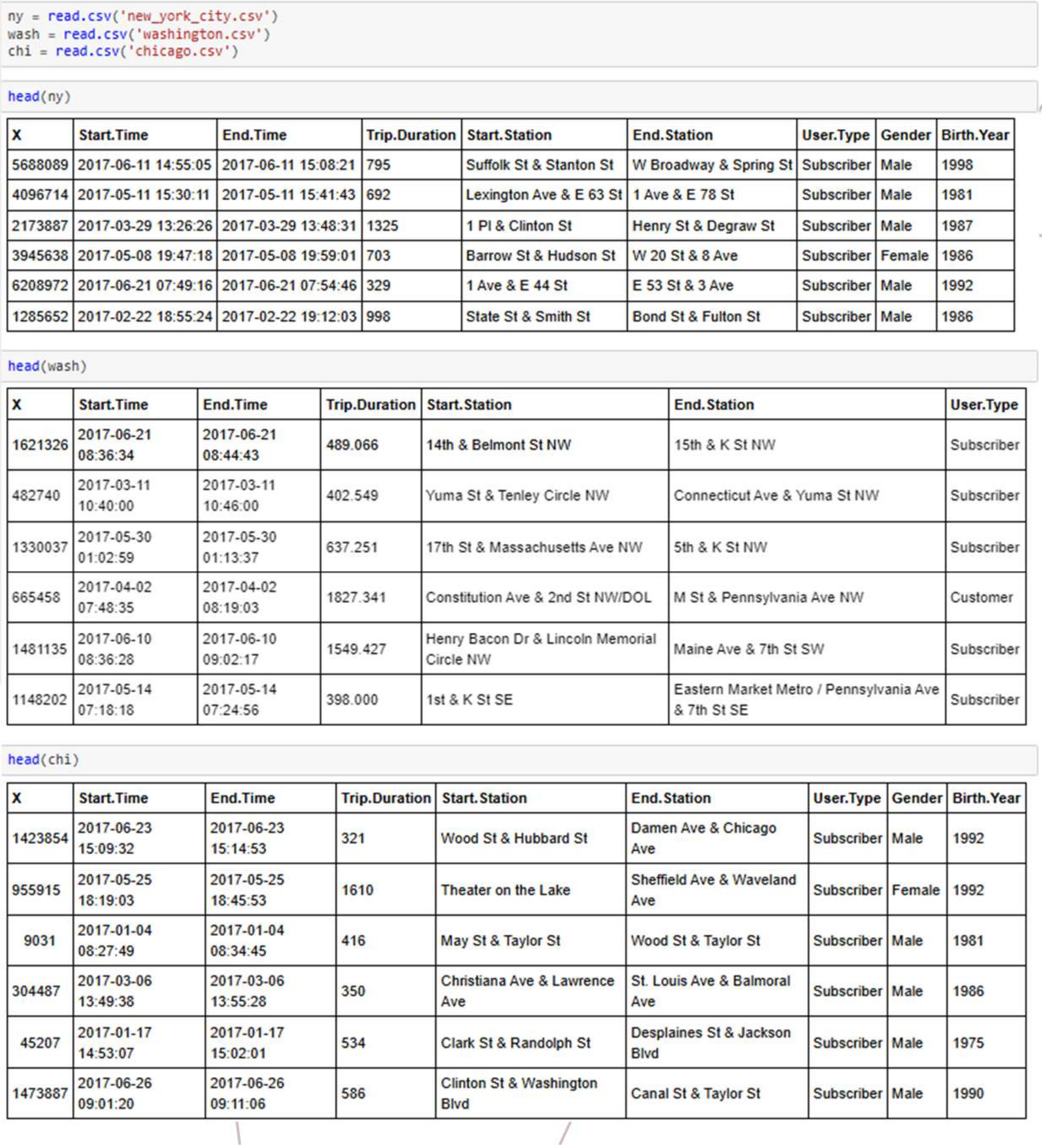
Questions of exploration

1. How do the number of subscribers and customers vary for each location

2. What is the distribution of trip durations in Washington?

3. What time of day is most common for users in Chicago?

PREPARE DATA



```
ny = read.csv('new_york_city.csv')
wash = read.csv('washington.csv')
chi = read.csv('chicago.csv')
```

```
head(ny)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station	User.Type	Gender	Birth.Year
5688089	2017-06-11 14:55:05	2017-06-11 15:08:21	795	Suffolk St & Stanton St	W Broadway & Spring St	Subscriber	Male	1998
4096714	2017-05-11 15:30:11	2017-05-11 15:41:43	692	Lexington Ave & E 63 St	1 Ave & E 78 St	Subscriber	Male	1981
2173887	2017-03-29 13:26:26	2017-03-29 13:48:31	1325	1 Pl & Clinton St	Henry St & Degraw St	Subscriber	Male	1987
3945638	2017-05-08 19:47:18	2017-05-08 19:59:01	703	Barrow St & Hudson St	W 20 St & 8 Ave	Subscriber	Female	1986
6208972	2017-06-21 07:49:16	2017-06-21 07:54:46	329	1 Ave & E 44 St	E 53 St & 3 Ave	Subscriber	Male	1992
1285652	2017-02-22 18:55:24	2017-02-22 19:12:03	998	State St & Smith St	Bond St & Fulton St	Subscriber	Male	1986

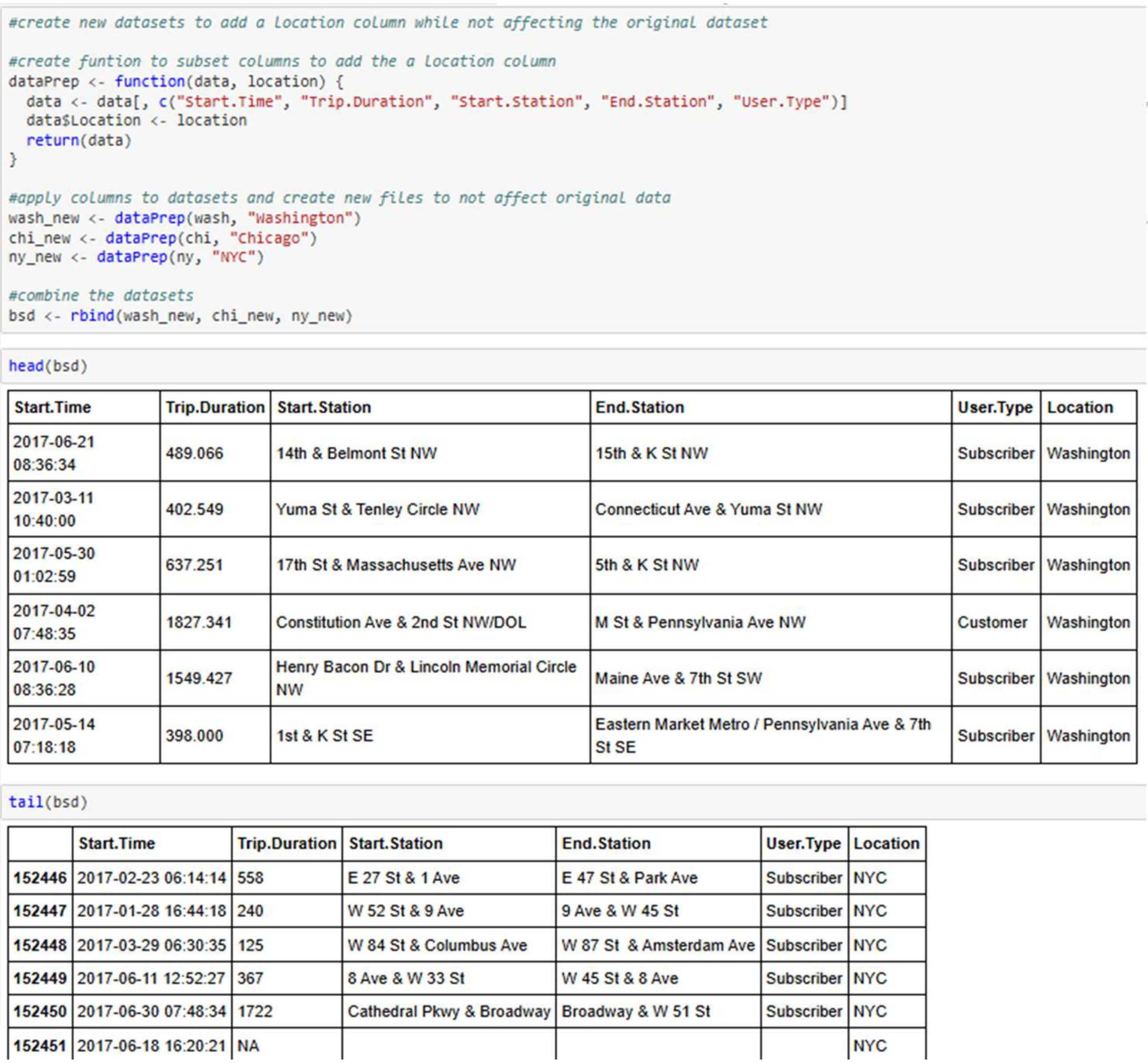
```
head(wash)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station	User.Type
1621326	2017-06-21 08:36:34	2017-06-21 08:44:43	489.066	14th & Belmont St NW	15th & K St NW	Subscriber
482740	2017-03-11 10:40:00	2017-03-11 10:46:00	402.549	Yuma St & Tenley Circle NW	Connecticut Ave & Yuma St NW	Subscriber
1330037	2017-05-30 01:02:59	2017-05-30 01:13:37	637.251	17th St & Massachusetts Ave NW	5th & K St NW	Subscriber
665458	2017-04-02 07:48:35	2017-04-02 08:19:03	1827.341	Constitution Ave & 2nd St NW/DOL	M St & Pennsylvania Ave NW	Customer
1481135	2017-06-10 08:36:28	2017-06-10 09:02:17	1549.427	Henry Bacon Dr & Lincoln Memorial Circle NW	Maine Ave & 7th St SW	Subscriber
1148202	2017-05-14 07:18:18	2017-05-14 07:24:56	398.000	1st & K St SE	Eastern Market Metro / Pennsylvania Ave & 7th St SE	Subscriber

```
head(chi)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station	User.Type	Gender	Birth.Year
1423854	2017-06-23 15:09:32	2017-06-23 15:14:53	321	Wood St & Hubbard St	Damen Ave & Chicago Ave	Subscriber	Male	1992
955915	2017-05-25 18:19:03	2017-05-25 18:45:53	1610	Theater on the Lake	Sheffield Ave & Waveland Ave	Subscriber	Female	1992
9031	2017-01-04 08:27:49	2017-01-04 08:34:45	416	May St & Taylor St	Wood St & Taylor St	Subscriber	Male	1981
304487	2017-03-06 13:49:38	2017-03-06 13:55:28	350	Christiana Ave & Lawrence Ave	St. Louis Ave & Balmoral Ave	Subscriber	Male	1986
45207	2017-01-17 14:53:07	2017-01-17 15:02:01	534	Clark St & Randolph St	Desplaines St & Jackson Blvd	Subscriber	Male	1975
1473887	2017-06-26 09:01:20	2017-06-26 09:11:06	586	Clinton St & Washington Blvd	Canal St & Taylor St	Subscriber	Male	1990

CREATE NEW DATASETS



```
#create new datasets to add a location column while not affecting the original dataset

#create funtion to subset columns to add the a location column
dataPrep <- function(data, location) {
  data <- data[, c("Start.Time", "Trip.Duration", "Start.Station", "End.Station", "User.Type")]
  data$Location <- location
  return(data)
}

#apply columns to datasets and create new files to not affect original data
wash_new <- dataPrep(wash, "Washington")
chi_new <- dataPrep(chi, "Chicago")
ny_new <- dataPrep(ny, "NYC")

#combine the datasets
bsd <- rbind(wash_new, chi_new, ny_new)
```

```
head(bsd)
```

Start.Time	Trip.Duration	Start.Station	End.Station	User.Type	Location
2017-06-21 08:36:34	489.066	14th & Belmont St NW	15th & K St NW	Subscriber	Washington
2017-03-11 10:40:00	402.549	Yuma St & Tenley Circle NW	Connecticut Ave & Yuma St NW	Subscriber	Washington
2017-05-30 01:02:59	637.251	17th St & Massachusetts Ave NW	5th & K St NW	Subscriber	Washington
2017-04-02 07:48:35	1827.341	Constitution Ave & 2nd St NW/DOL	M St & Pennsylvania Ave NW	Customer	Washington
2017-06-10 08:36:28	1549.427	Henry Bacon Dr & Lincoln Memorial Circle NW	Maine Ave & 7th St SW	Subscriber	Washington
2017-05-14 07:18:18	398.000	1st & K St SE	Eastern Market Metro / Pennsylvania Ave & 7th St SE	Subscriber	Washington

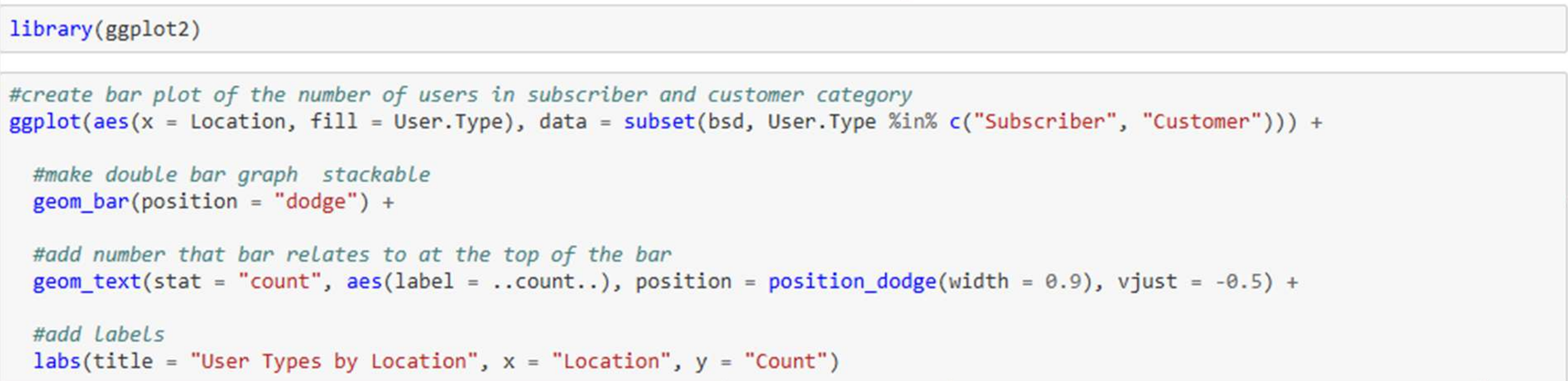
```
tail(bsd)
```

	Start.Time	Trip.Duration	Start.Station	End.Station	User.Type	Location
152446	2017-02-23 06:14:14	558	E 27 St & 1 Ave	E 47 St & Park Ave	Subscriber	NYC
152447	2017-01-28 16:44:18	240	W 52 St & 9 Ave	9 Ave & W 45 St	Subscriber	NYC
152448	2017-03-29 06:30:35	125	W 84 St & Columbus Ave	W 87 St & Amsterdam Ave	Subscriber	NYC
152449	2017-06-11 12:52:27	367	8 Ave & W 33 St	W 45 St & 8 Ave	Subscriber	NYC
152450	2017-06-30 07:48:34	1722	Cathedral Pkwy & Broadway	Broadway & W 51 St	Subscriber	NYC
152451	2017-06-18 16:20:21	NA				NYC

QUESTION 1:

HOW DO THE NUMBER OF SUBSCRIBERS AND CUSTOMERS VARY FOR EACH LOCATION?

BUILD DOUBLE BAR GRAPH



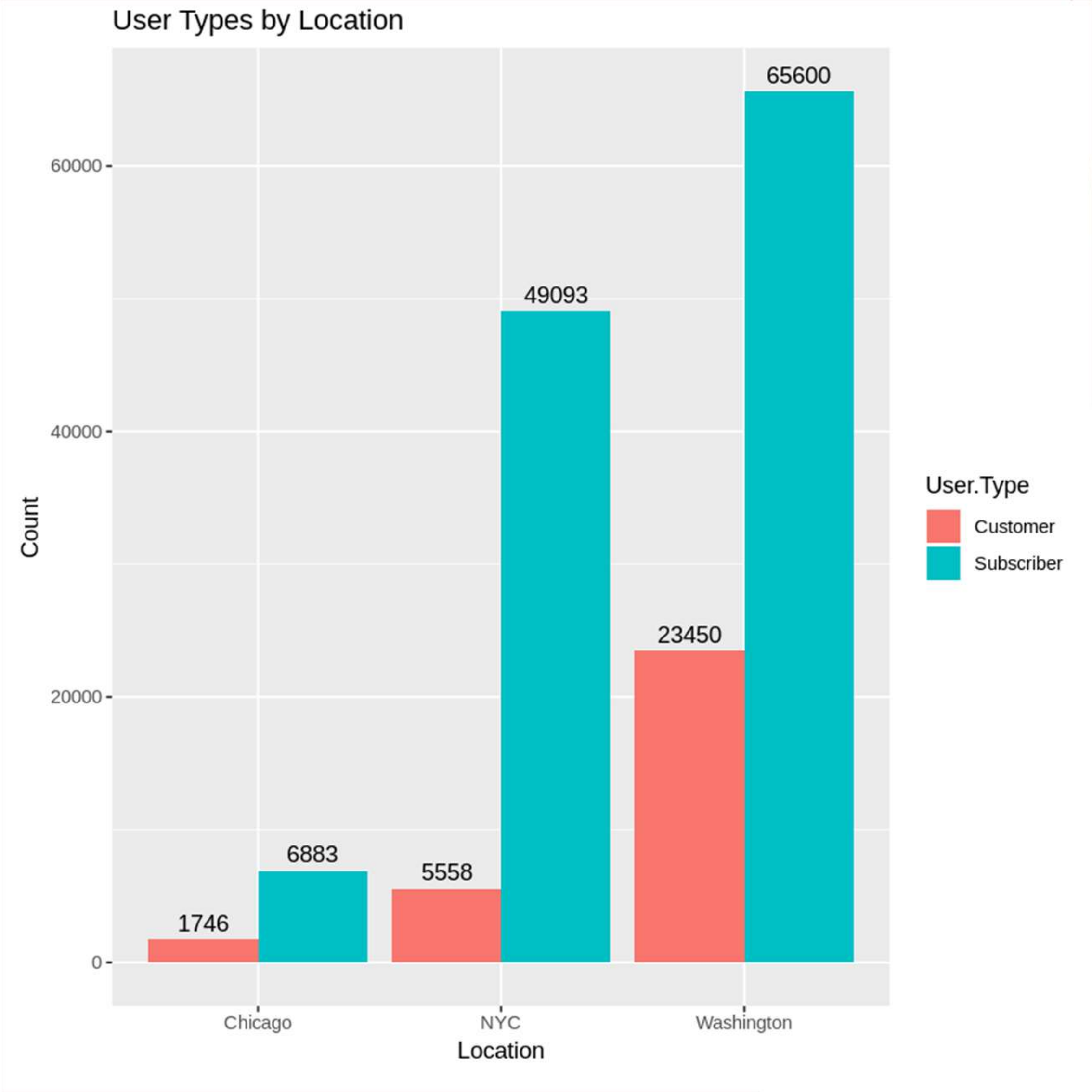
```
library(ggplot2)
```

```
#create bar plot of the number of users in subscriber and customer category
ggplot(aes(x = Location, fill = User.Type), data = subset(bsd, User.Type %in% c("Subscriber", "Customer"))) +

  #make double bar graph  stackable
  geom_bar(position = "dodge") +

  #add number that bar relates to at the top of the bar
  geom_text(stat = "count", aes(label = ..count..), position = position_dodge(width = 0.9), vjust = -0.5) +

  #add labels
  labs(title = "User Types by Location", x = "Location", y = "Count")
```

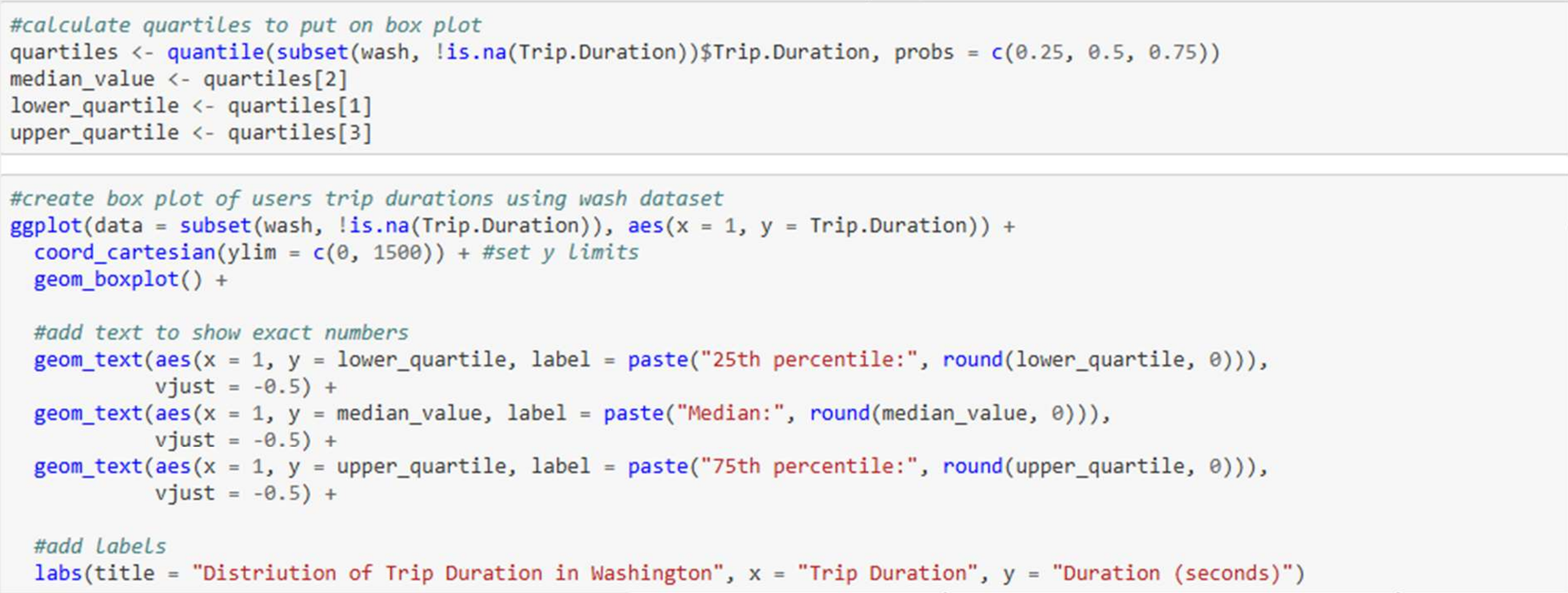


Bicycle sharing patterns vary significantly with New York City and Washington showing a strong dominance of subscribers over customers. In contrast, Chicago has a smaller margin of variety between subscribers and customers.

QUESTION 2:

WHAT IS THE DISTRIBUTION OF TRIP DURATION IN WASHINGTON?

BUILD DISTRIBUTION PLOT

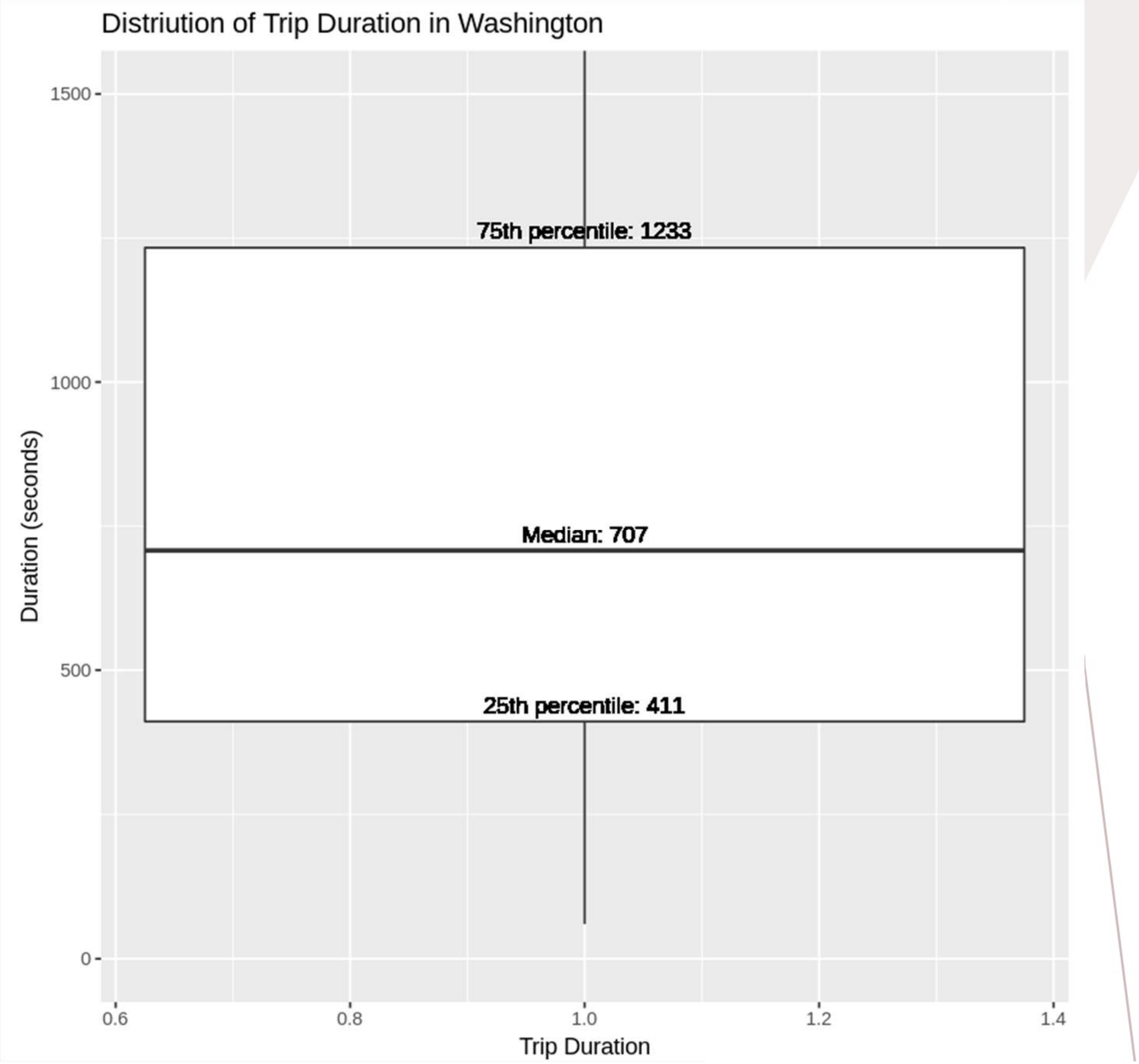


```
#calculate quartiles to put on box plot
quartiles <- quantile(subset(wash, !is.na(Trip.Duration))$Trip.Duration, probs = c(0.25, 0.5, 0.75))
median_value <- quartiles[2]
lower_quartile <- quartiles[1]
upper_quartile <- quartiles[3]
```

```
#create box plot of users trip durations using wash dataset
ggplot(data = subset(wash, !is.na(Trip.Duration)), aes(x = 1, y = Trip.Duration)) +
  coord_cartesian(ylim = c(0, 1500)) + #set y limits
  geom_boxplot() +

  #add text to show exact numbers
  geom_text(aes(x = 1, y = lower_quartile, label = paste("25th percentile:", round(lower_quartile, 0))),
            vjust = -0.5) +
  geom_text(aes(x = 1, y = median_value, label = paste("Median:", round(median_value, 0))),
            vjust = -0.5) +
  geom_text(aes(x = 1, y = upper_quartile, label = paste("75th percentile:", round(upper_quartile, 0))),
            vjust = -0.5) +

  #add labels
  labs(title = "Distriution of Trip Duration in Washington", x = "Trip Duration", y = "Duration (seconds)")
```

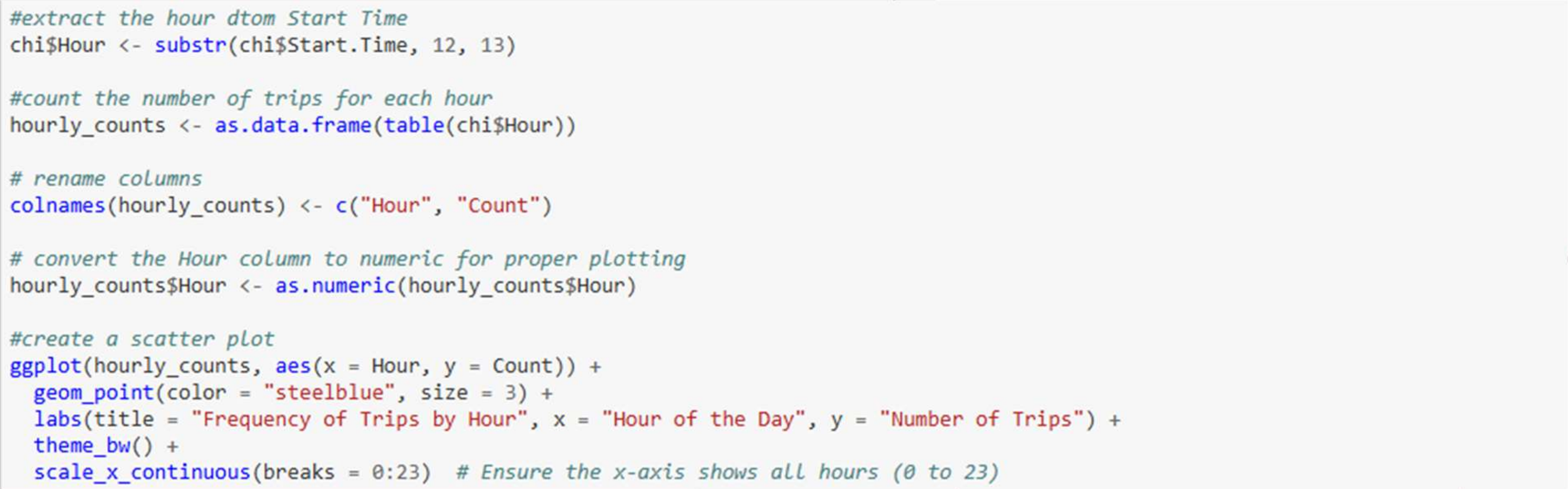


The distribution of trip durations in Washington show trips tend to fall between 7 minutes (411 seconds) and 20 Minutes (1233 seconds). Most trips fall below the 12-minute (707 seconds) mark.

QUESTION 3:

WHAT TIME OF DAY IS MOST COMMON FOR USERS IN CHICAGO?

BUILD SCATTER PLOT



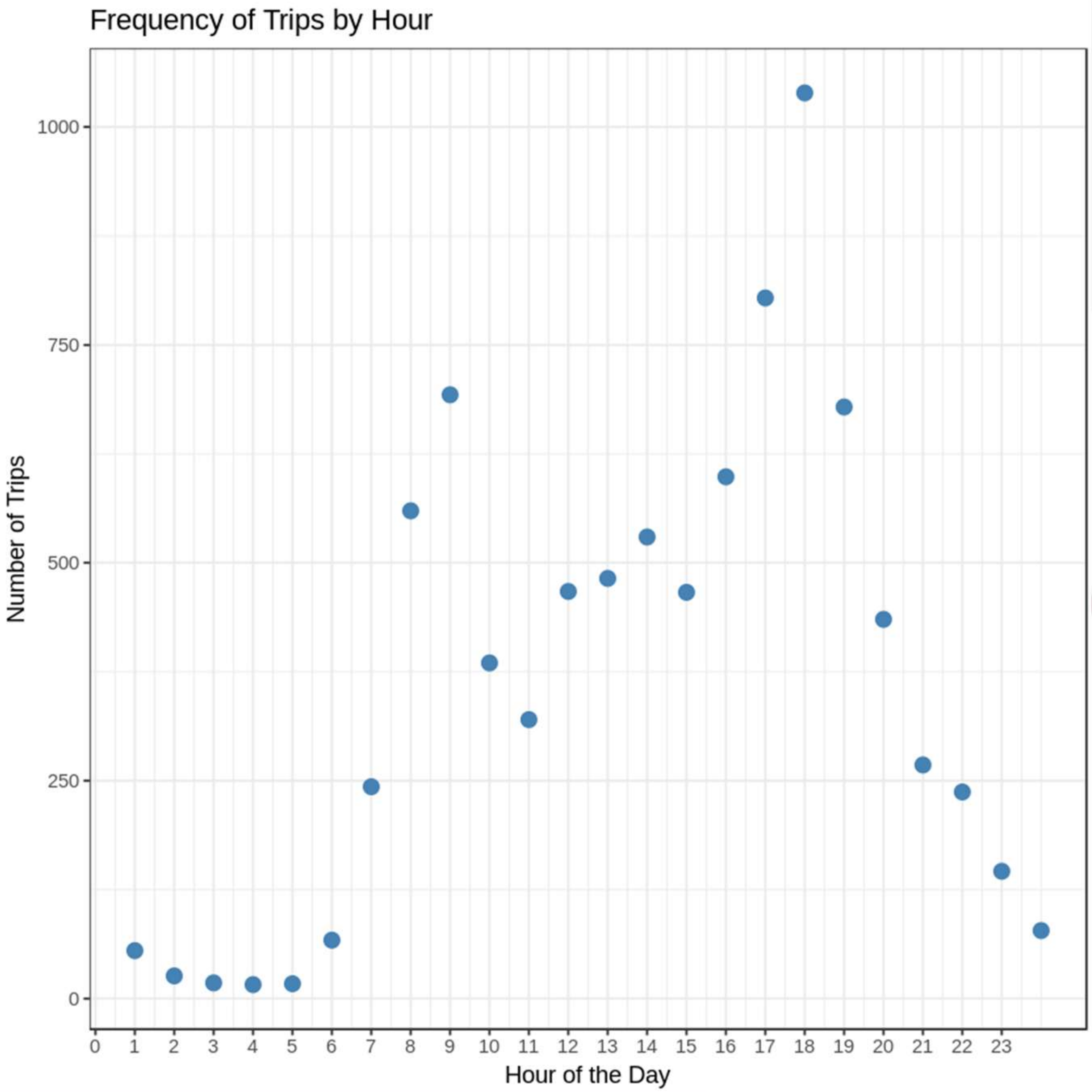
```r
#extract the hour dtom Start Time
chi$Hour <- substr(chi$Start.Time, 12, 13)

#count the number of trips for each hour
hourly_counts <- as.data.frame(table(chi$Hour))

# rename columns
colnames(hourly_counts) <- c("Hour", "Count")

# convert the Hour column to numeric for proper plotting
hourly_counts$Hour <- as.numeric(hourly_counts$Hour)

#create a scatter plot
ggplot(hourly_counts, aes(x = Hour, y = Count)) +
  geom_point(color = "steelblue", size = 3) +
  labs(title = "Frequency of Trips by Hour", x = "Hour of the Day", y = "Number of Trips") +
  theme_bw() +
  scale_x_continuous(breaks = 0:23)  # Ensure the x-axis shows all hours (0 to 23)
```



The most common time for users in Chicago is 1800 or 6:00 PM.

THANK YOU

Erica Greene