Ford BikeShare Usership Analysis

```
trip <- read.csv('GoBike_data/trip.csv')

station <- read.csv('GoBike_data/station.csv')
str(trip)
```

```
## 'data.frame':    669959 obs. of  11 variables:
##  $ id                : int  4576 4607 4130 4251 4299 4927 4500 4563 4760 4258 ...
##  $ duration          : int  63 70 71 77 83 103 109 111 113 114 ...
##  $ start_date        : Factor w/ 361559 levels "1/1/2014 0:14",..: 319540 319554 319399 319421 31944
##  $ start_station_name: Factor w/ 74 levels "2nd at Folsom",..: 64 53 35 53 64 23 60 59 64 53 ...
##  $ start_station_id  : int  66 10 27 10 66 59 4 8 66 10 ...
##  $ end_date          : Factor w/ 357757 levels "1/1/2014 0:21",..: 316458 316474 316342 316358 31637
##  $ end_station_name  : Factor w/ 74 levels "2nd at Folsom",..: 64 53 35 53 28 23 5 59 64 33 ...
##  $ end_station_id    : int  66 10 27 10 67 59 5 8 66 11 ...
##  $ bike_id           : int  520 661 48 26 319 527 679 687 553 107 ...
##  $ subscription_type : Factor w/ 2 levels "Customer","Subscriber": 2 2 2 2 2 2 2 2 2 2 ...
##  $ zip_code          : Factor w/ 7440 levels "","0","1","100",..: 6452 6803 7145 6752 6419 6429 6776
```

```
trip$start_date <- as.character(trip$start_date)
trip$end_date <- as.character(trip$end_date)

to_time <- function(row){
  strsplit(row, ' ')[[1]][2]
}
trip$start_time <- sapply(trip$start_date, to_time)
trip$end_time <- sapply(trip$end_date, to_time)

trip$start_date <- as.Date(trip$start_date, '%m/%d/%Y %k')
trip$end_date <- as.Date(trip$end_date, '%m/%d/%Y %k')
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

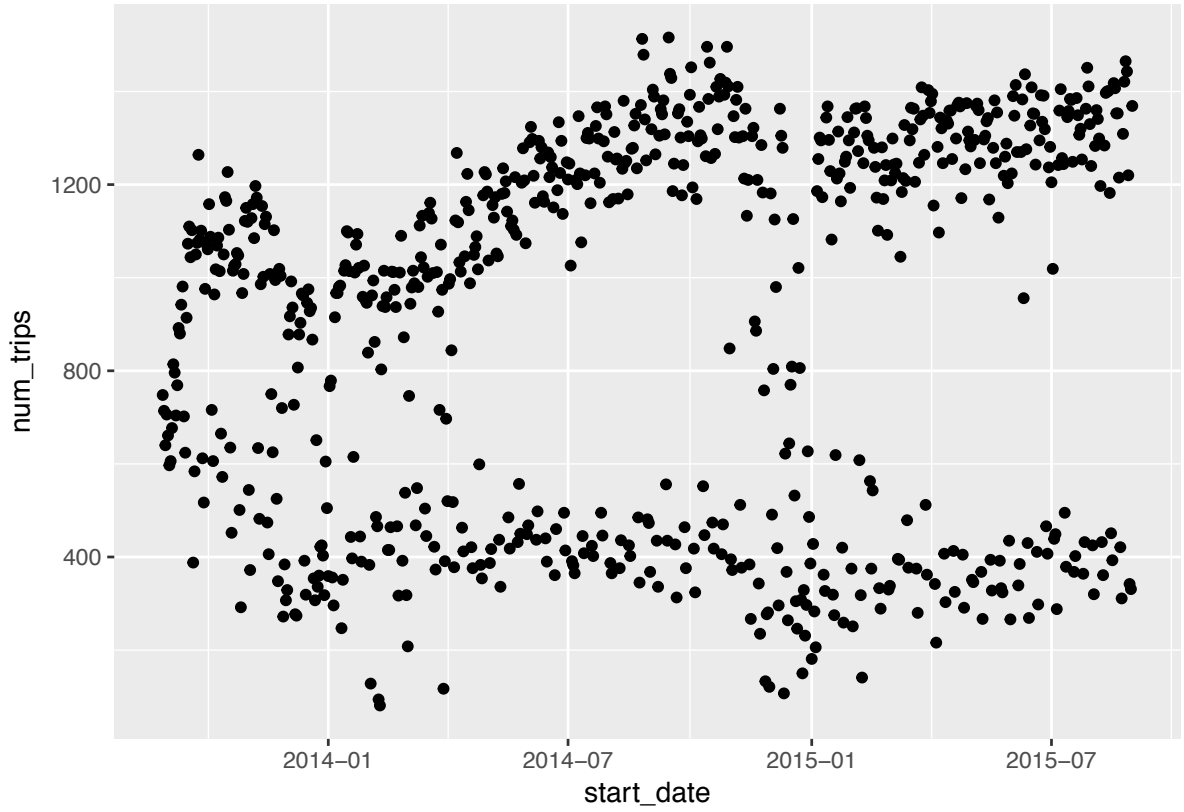
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

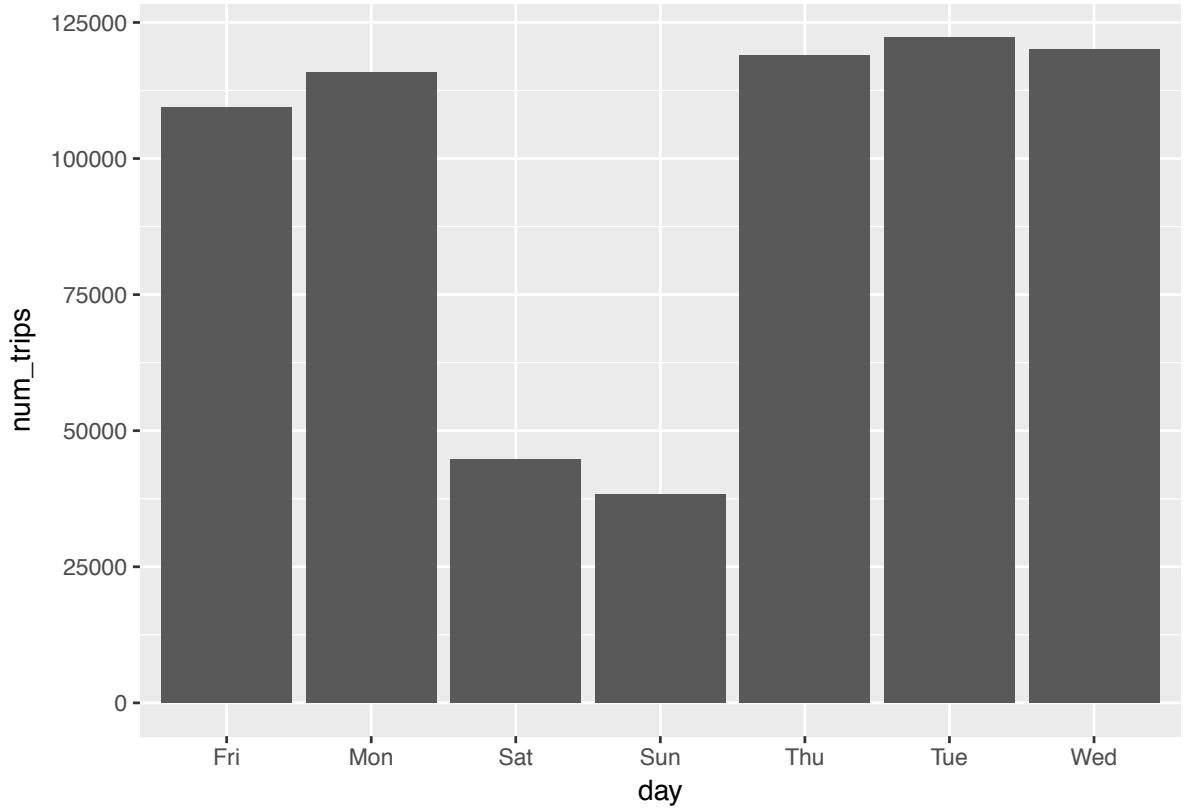
```
library(ggplot2)
sum(trip$start_date != trip$send_date)
```

```
## [1] 0
```

```
num_riders <- summarize(group_by(trip, start_date), num_trips=n())
ggplot(data = num_riders, aes(x= start_date, y= num_trips)) + geom_point()
```

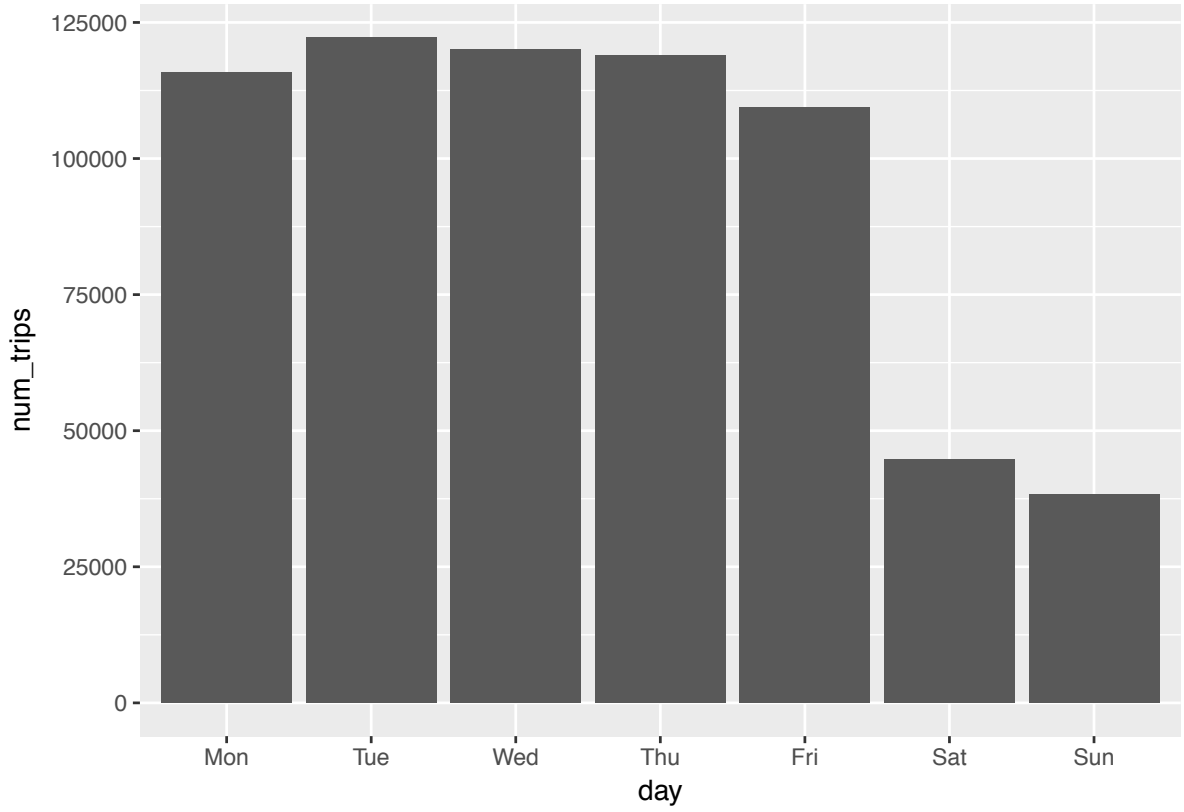



```
num_riders <- mutate(num_riders, day = weekdays(start_date, abbreviate = TRUE))
ggplot(num_riders, aes(x = day, y = num_trips)) +
  geom_bar(stat = 'identity')
```



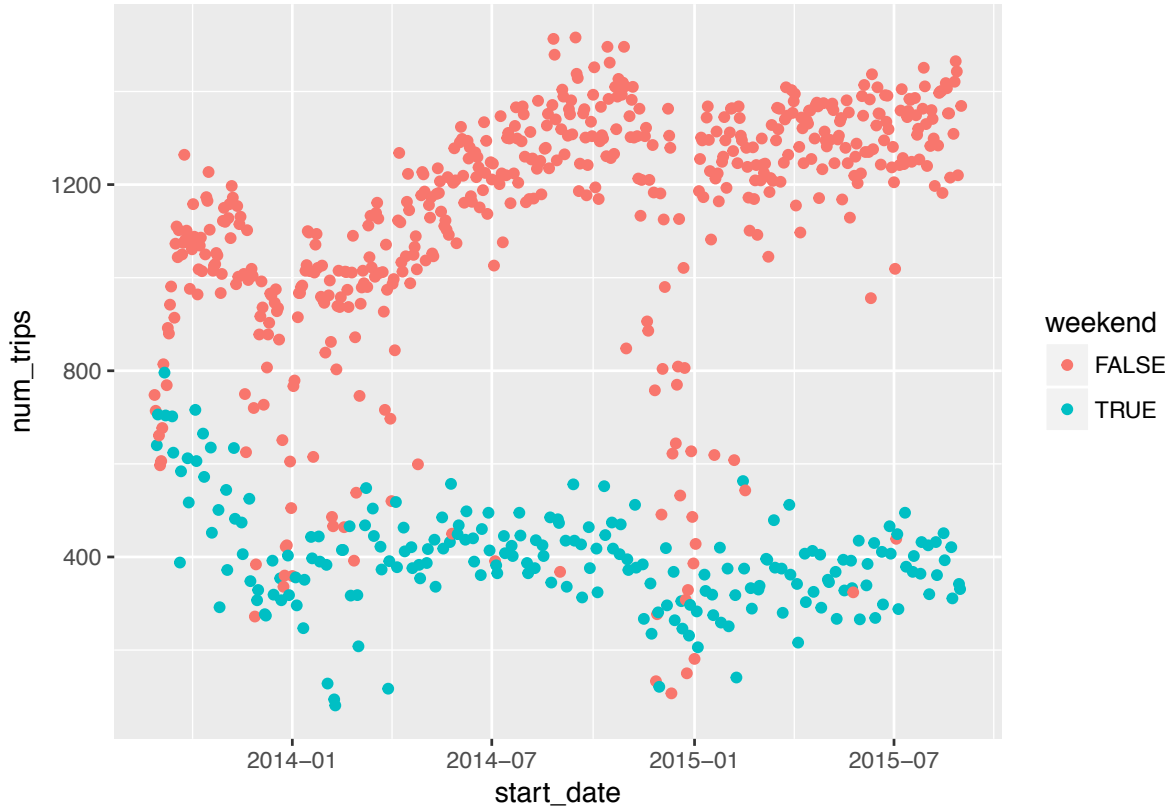
#Order the dates (x=axis) (Mon-Sun)

```
num_riders$day <- factor(num_riders$day, levels = c('Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun'))
ggplot(num_riders, aes(x = day, y = num_trips)) +
  geom_bar(stat = 'identity')
```



Specify color for weekend and weekday.

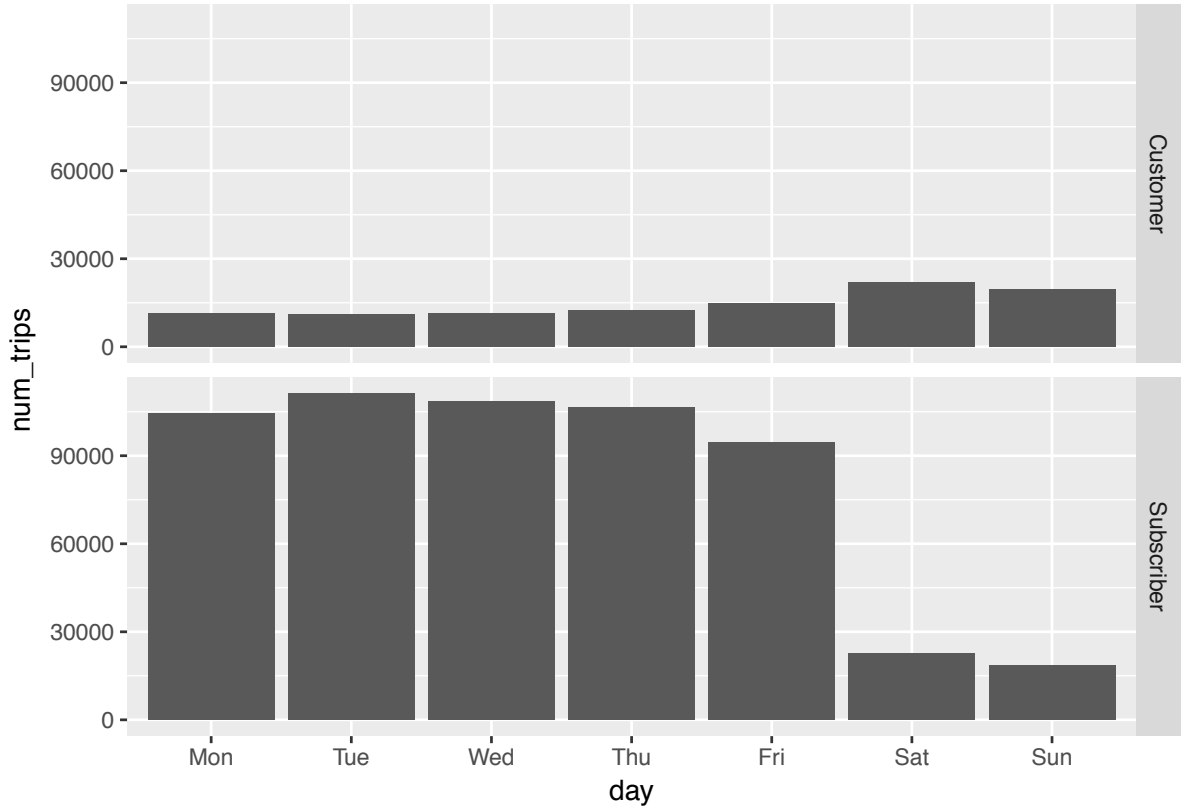
```
num_riders <- mutate(num_riders, 'weekend' = (day == 'Sat' | day == 'Sun' ))
ggplot(num_riders, aes(x = start_date, y = num_trips)) + geom_point(aes(color = weekend))
```




```
table(trip$subscription_type)
```

```
##
##   Customer Subscriber
##      103213     566746
```

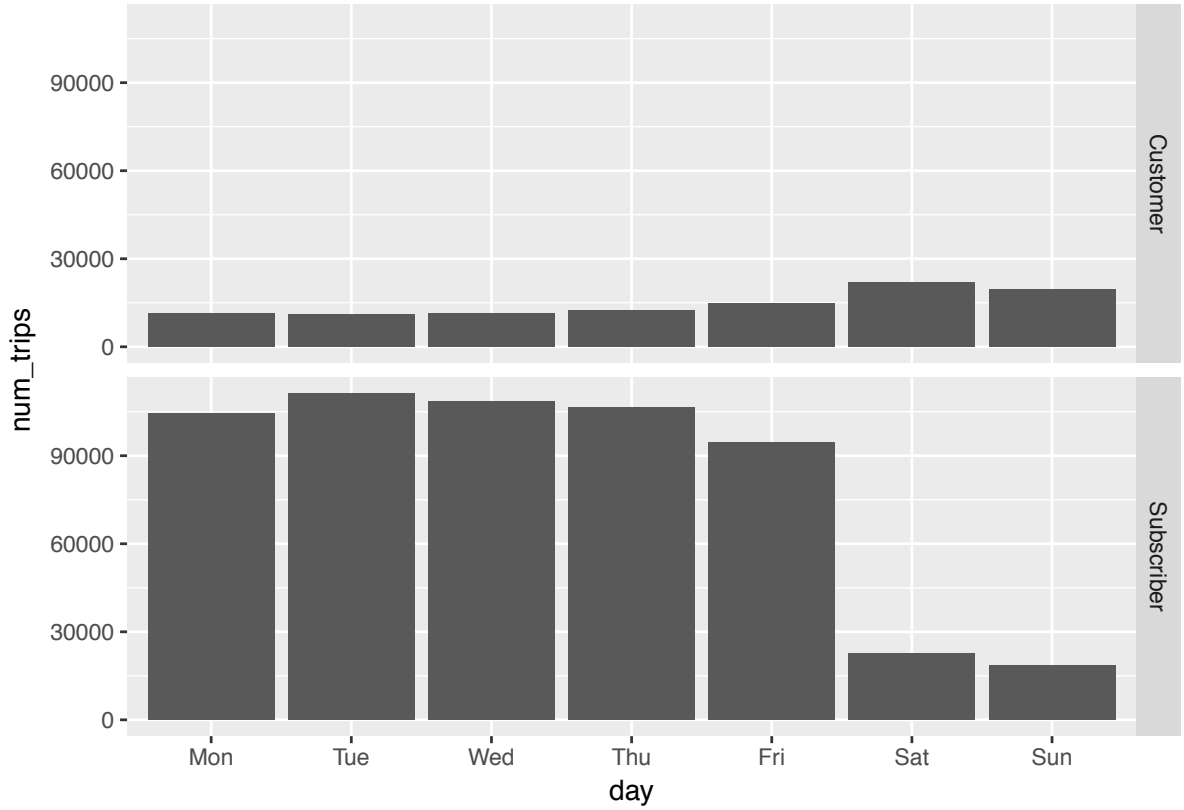
```
trip_subscriptions <- summarize(group_by(trip, start_date, subscription_type),
num_trips=n())
trip_subscriptions <- mutate(trip_subscriptions, 'day'=weekdays(start_date, abbreviate=TRUE))
trip_subscriptions$day <- factor(trip_subscriptions$day, levels=c('Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sa

trip_subscriptions <-mutate(trip_subscriptions, 'weekend'=(day== 'Sat' | day == 'Sun'))

ggplot(trip_subscriptions, aes(x=day, y=num_trips)) + geom_bar(stat='identity') + facet_grid(subscriptio
```



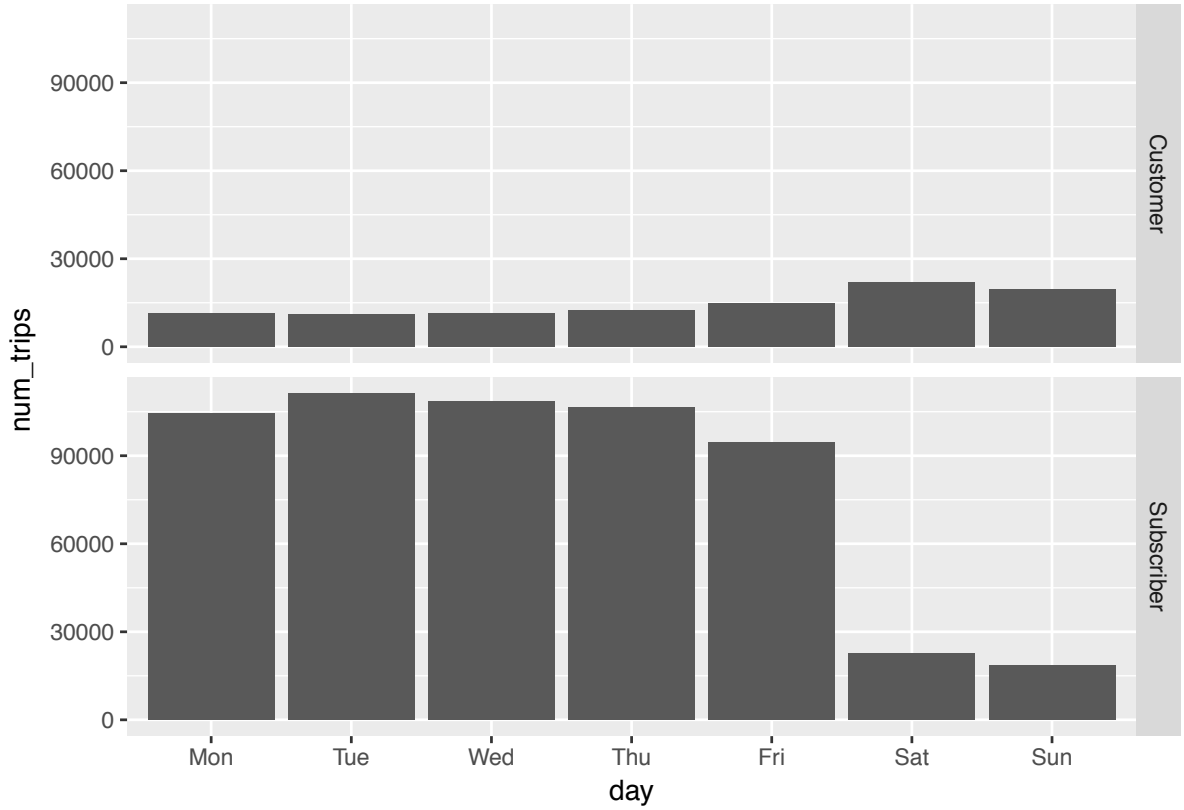

```
start_time <- strsplit(trip$start_time, ':')

# Explore Subscription Type
table(trip$subscription_type)
```

```
##
##   Customer Subscriber
##     103213     566746
```

```
trip_subscriptions <- trip %>% group_by(start_date, subscription_type) %>% summarize(num_trips = n())
trip_subscriptions <- trip_subscriptions %>% mutate('day'=weekdays(start_date, abbreviate=TRUE))
trip_subscriptions$day <- factor(trip_subscriptions$day, levels=c('Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sa

trip_subscriptions <- trip_subscriptions %>%
  mutate('weekend'=(day == 'Sat' | day == 'Sun'))

ggplot(trip_subscriptions, aes(x=day, y=num_trips)) +
  geom_bar(stat='identity') +
  facet_grid(subscription_type ~ .)
```

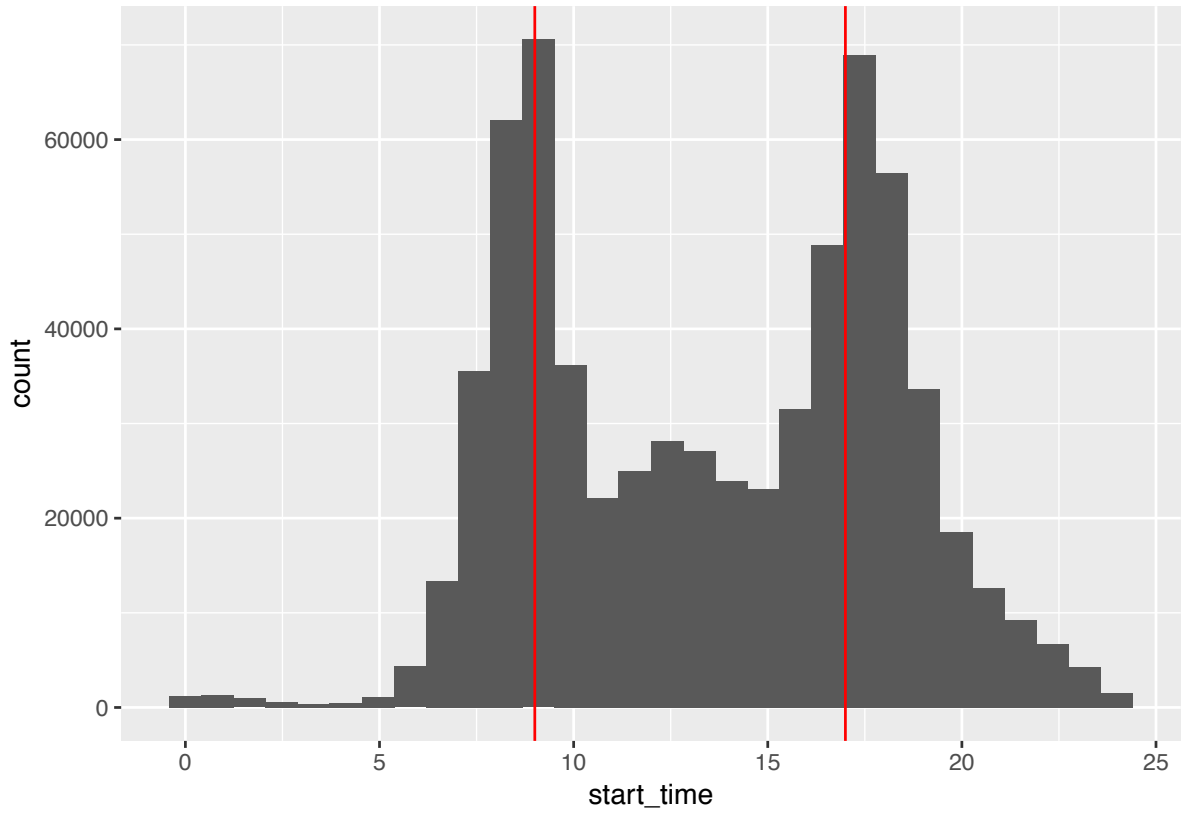



```
# trip_subscriptions <- trip_subscriptions %>% ungroup()

ggplot(trip_subscriptions, aes(x=day, y=num_trips)) +
  geom_bar(stat='identity') +
  facet_grid(subscription_type ~ .)
```




```
# Peak Times
start_time <- strsplit(trip$start_time, ':')
convert_time <- function(obs){
  split_time <- strsplit(obs, ':')[[1]]
  hour <- as.integer(split_time[1])
  min <- as.integer(split_time[2])

  return(hour + min/60)
}

trip$start_time <- sapply(trip$start_time, convert_time)

ggplot(trip, aes(start_time)) +
  geom_histogram() +
  geom_vline(xintercept=9, color='red') +
  geom_vline(xintercept=17, color='red')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
weather <- read.csv('GoBike_data/weather.csv')
```

Filter out non-SF city in dataframe

```
trip <- left_join(trip, station, by = c('start_station_id'='id' ))
trip <- filter(trip, city == 'San Francisco')
trip
```

```
##            id duration start_date
## 1        4576       63 2013-08-29
## 2        4299       83 2013-08-29
## 3        4927      103 2013-08-29
## 4        4760      113 2013-08-29
## 5        4549      125 2013-08-29
## 6        4557      130 2013-08-29
## 7        4386      134 2013-08-29
## 8        4749      138 2013-08-29
## 9        4329      142 2013-08-29
## 10       5097      142 2013-08-29
## 11       5084      144 2013-08-29
## 12       4982      146 2013-08-29
## 13       4265      151 2013-08-29
## 14       5093      160 2013-08-29
## 15       4168      161 2013-08-29
## 16       4533      165 2013-08-29
```