Overview

Background

With increasing urban population, traffic congestion and saturation and/or lack of public transportation bike sharing proved to be an ingenious environment friendly solution for daily commuters. There has been steady increase in the number of bike share programs worldwide reaching 1608 bike share programs with a fleet of 18.2 million bikes in 2018.

Despite the steady growth in bike sharing programs one of the key challenges faced by aggregators is to estimate the demand for bikes and allocate resources accordingly as the usage rates vary from around three to eight trips per bicycle per day globally2. The variation in usage could be due to multitude of factors one of which we believe are the prevalent weather conditions. We can expect that passengers are more likely to choose bike rides on days when the weather is pleasant without snowfall and/or heavy winds. Another important factor is time during the day. The demand is more during morning and evening peak traffic hours, and lesser during other times of the day.

Further, a study carried out by Bowman Cutter and Matthew Neidell's on the effect of voluntary information disclosure of information on air quality urging people to reduce ozone emissions found that there is an increase in people choosing alternate methods of transportation on days such warnings are issued, supporting the idea that weather parameters have an effect on individual's behavior and choices.

Data Description

The response variable is:
Y (Cnt): Total bikes rented by both casual & registered users together

The predicting variables are:
X_1 (Instant): Record index
X_2 (Dteday): Day on which the observation is made
X_3 (Season): Season which the observation is made (1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall)
X_4 (Yr): Year on which the observation is made
X_5 (Mnth): Month on which the observation is made
X_6 (Hr): Day on which the observation is made (0 through 23)
X_7 (Holiday): Indictor of a public holiday or not (1 = public holiday, 0 = not a public holiday)
X_8 (Weekday): Day of week (0 through 6)
X_9 (Working day): Indicator of a working day (1 = working day, 0 = not a working day)
X_{10} (Weathersit): Weather condition (1 = Clear, Few clouds, Partly cloudy, Partly cloudy, 2 = Mist & Cloudy, Mist & Broken clouds, Mist & Few clouds, Mist, 3 = Light Snow, Light Rain, Thunderstorm & Scattered clouds, Light Rain & Scattered clouds, 4 = Heavy Rain, Ice Pallets, Thunderstorm & Mist, Snow & Fog)
X_{11} (Temp): Normalized temperature in Celsius
X_{12} (Atemp): Normalized feeling temperature in Celsius
X_{13} (Hum): Normalized humidity
X_{14} (Windspeed): Normalized wind speed
X_{15} (Casual): Bikes rented by casual users in that hour
X_{16} (Registered): Bikes rented by registered users in that hour

---

Exploratory Data Analysis

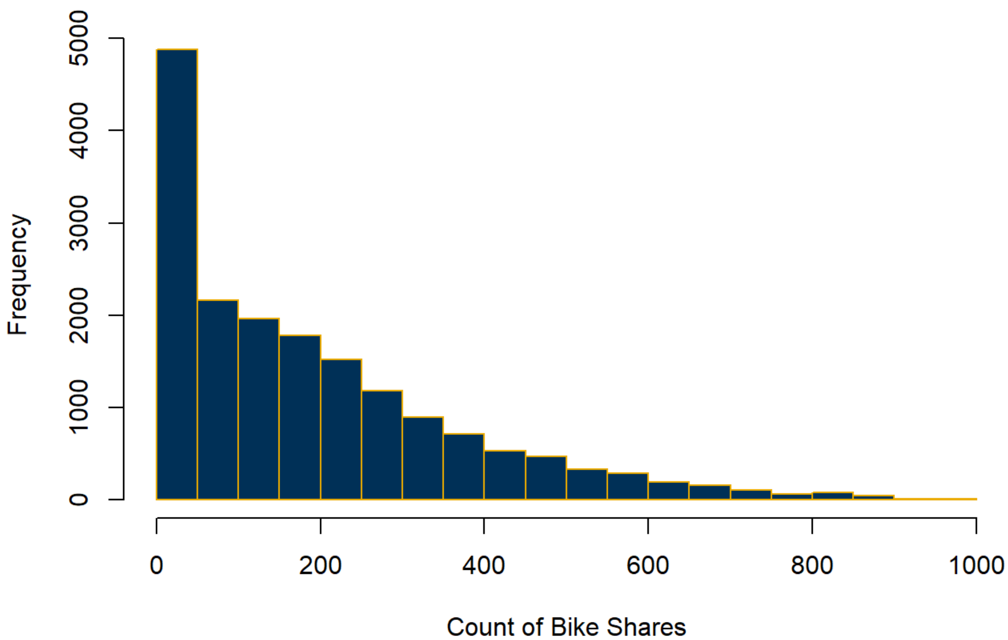
Reading data

```
# Set colors
gtblue = rgb(0, 48, 87, maxColorValue = 255)
techgold = rgb(179, 163, 105, maxColorValue = 255)
buzzgold = rgb(234, 170, 0, maxColorValue = 255)
bobbyjones = rgb(55, 113, 23, maxColorValue = 255)
# Read the data using read.csv
data = read.csv("Bikes.csv")
# Show the number of observations
obs = nrow(data)
cat("There are", obs, "observations in the data")
```

```
## There are 17379 observations in the data
```

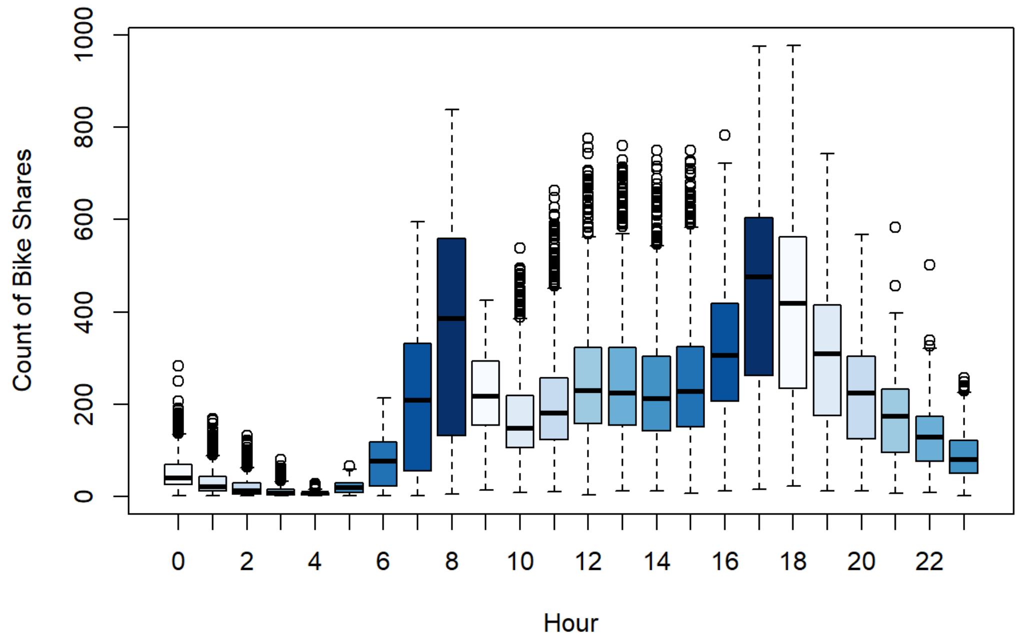
Response Data Distribution

```
# Check the distribution of the response, cnt
hist(data$cnt,
     main="",
     xlab="Count of Bike Shares",
     border=buzzgold,
     col=gtblue)
```



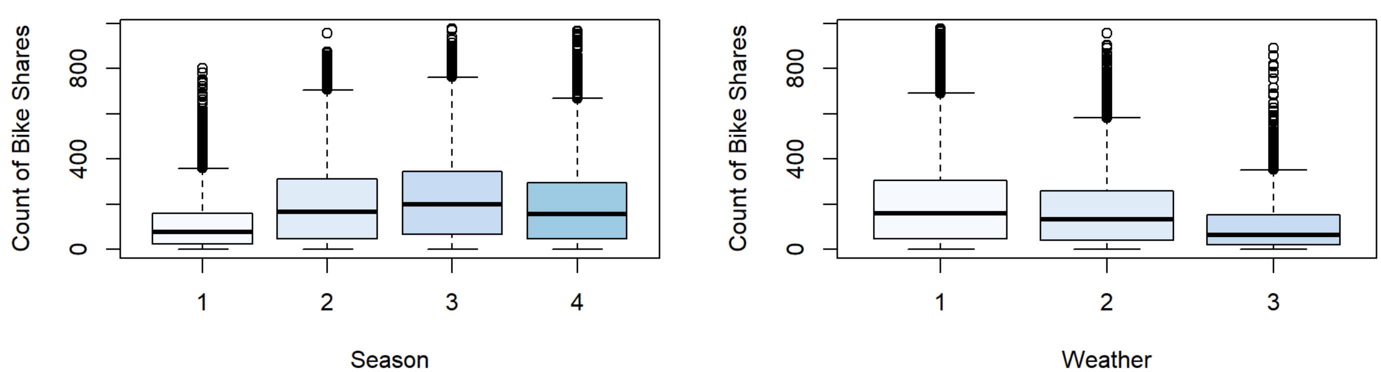
- The frequency of zero bike shares is high, which skews the demand data.

```
# Check the response, cnt, against time of day
boxplot(cnt~hr,
        main="",
        xlab="Hour",
        ylab="Count of Bike Shares",
        col=blues9,
        data=data)
```



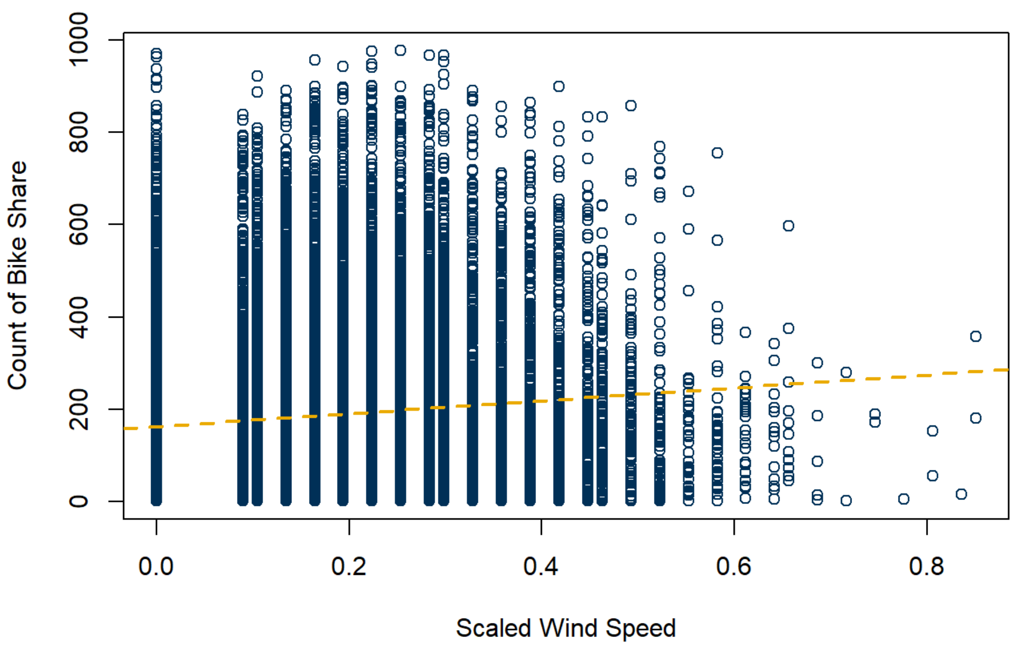
The number of bike shares between hour 0 and hour 6 is low. The majority activity as expected is focused between hour 7 and hour 23, peaking at hour 8 and hour 17.

```
par(mfrow=c(1, 2))

# Plot cnt against season
boxplot(cnt~season,
        main="",
        xlab="Season",
        ylab="Count of Bike Shares",
        col=blues9,
        data=data)

# Plot cnt against weather
boxplot(cnt~weathersit,
        main="",
        xlab="Weather",
        ylab="Count of Bike Shares",
        col=blues9,
        data=data)
```



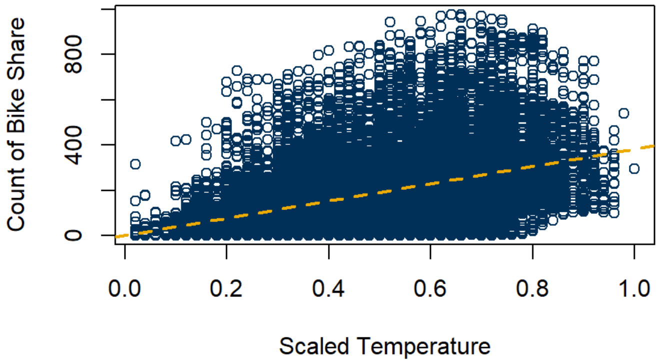
The number of bikes rented during winter are the lowest. The number of bikes decreases as the weather becomes unfavorable.

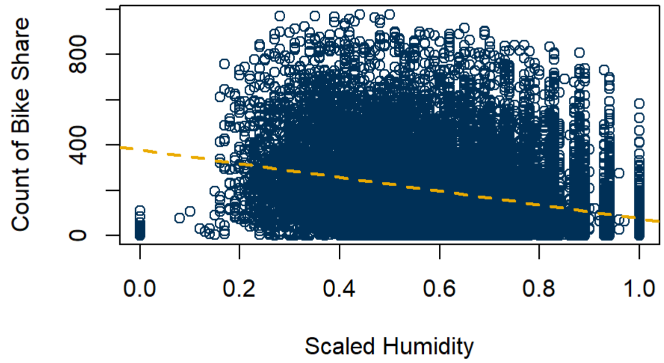
```
plot(data$windspeed,
     data$cnt,
     xlab="Scaled Wind Speed",
     ylab="Count of Bike Share",
     main="",
     col=gtblue)
abline(lm(cnt~windspeed, data=data),
       col=buzzgold,
       lty=2, lwd=2)
```



The count of rental bikes seems to decrease as windspeed increases. <- ***Need to discuss this as the OLS line contradicts this statement.***

```
par(mfrow=c(1, 2))

plot(data$temp,
     data$cnt,
     xlab="Scaled Temperature",
     ylab="Count of Bike Share",
     main="",
     col=gtblue)
abline(lm(cnt~temp, data=data), col=buzzgold, lty=2, lwd=2)

plot(data$hum,
     data$cnt,
     xlab="Scaled Humidity",
     ylab="Count of Bike Share",
     main="",
     col=gtblue)
abline(lm(cnt~hum, data=data), col=buzzgold, lty=2, lwd=2)
```





The count of rental bikes seems to decrease as humidity increases although the demand varies within similar ranges at varying humidity levels.
The count of rental bikes seems to increase as temperature increases however with much wider variability at larger temperature levels.