CS2

Omar Shatrat, Ali Kamel, Abhishek Ramesh

October 20, 2018

```
#a
hour <- read.csv("https://raw.githubusercontent.com/abbiepopa/BSDS100/master/Data/hour.csv", sep = ",",
                 header = T)

day <- read.csv("https://raw.githubusercontent.com/abbiepopa/BSDS100/master/Data/day.csv", sep = ",",
                header = T)

#b
#str(hour)
#str(day)

#c

hour$holiday <- as.factor(hour$holiday)
day$holiday <- as.factor(day$holiday)

hour$season <- as.factor(hour$season)
day$season <- as.factor(day$season)

hour$yr <- as.factor(hour$yr)
day$yr <- as.factor(day$yr)

hour$mnth <- as.factor(hour$mnth)
day$mnth <- as.factor(day$mnth)

hour$hr <- as.factor(hour$hr) #

hour$weekday <- as.factor(hour$weekday)
day$weekday <- as.factor(day$weekday)

hour$workingday <- as.factor(hour$workingday)
day$workingday <- as.factor(day$workingday)

hour$weathersit <- as.factor(hour$weathersit)
day$weathersit <- as.factor(day$weathersit)

#d
hour$dteday <- as.Date(hour$dteday)
day$dteday <- as.Date(day$dteday)

str(hour)
```

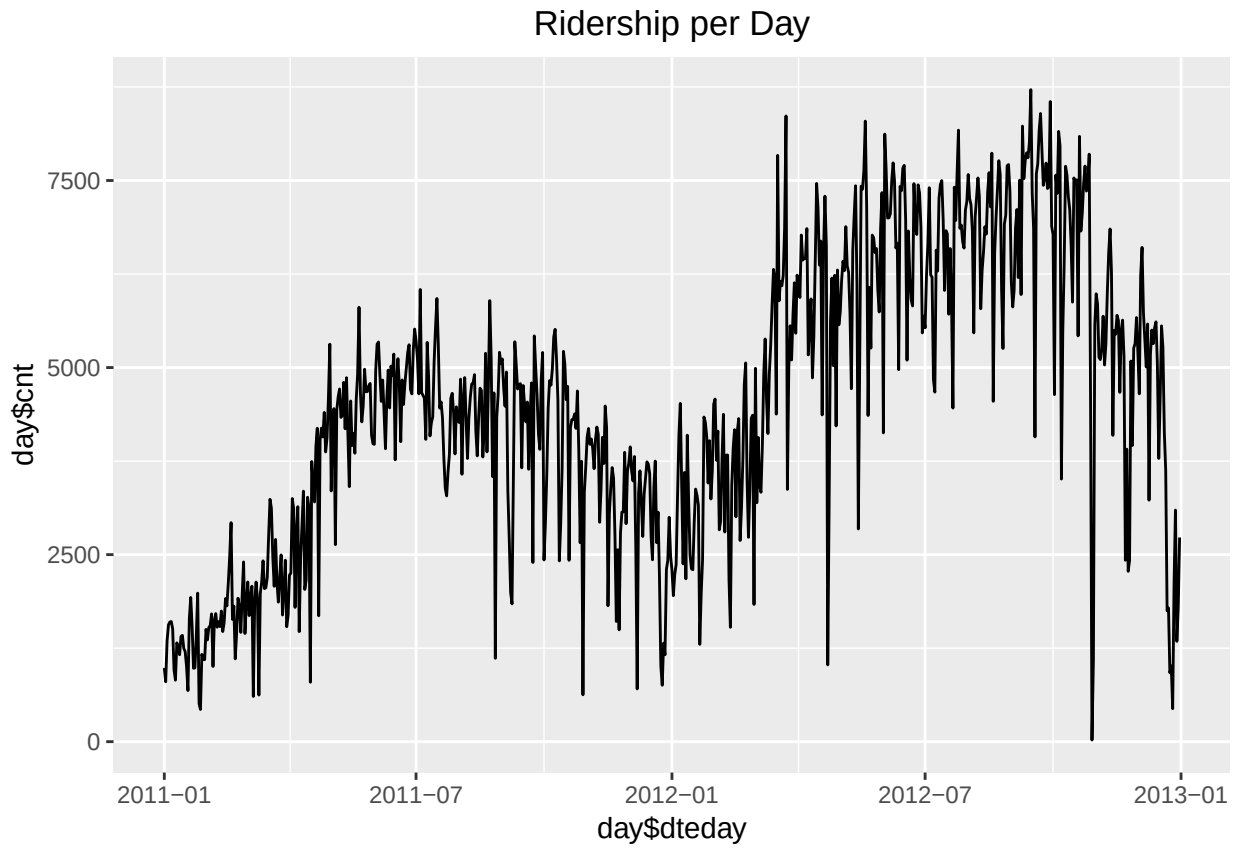
```
## 'data.frame':    17379 obs. of  17 variables:
##  $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : Date, format: "2011-01-01" "2011-01-01" ...
##  $ season    : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ yr        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ mnth      : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ hr        : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ holiday   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ weekday   : Factor w/ 7 levels "0","1","2","3",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ workingday: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ weathersit: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 2 1 1 1 1 ...
##  $ temp      : num  0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
##  $ atemp     : num  0.288 0.273 0.273 0.288 0.288 ...
##  $ hum       : num  0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
##  $ windspeed : num  0 0 0 0 0 0.0896 0 0 0 0 ...
##  $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
##  $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
##  $ cnt       : int  16 40 32 13 1 1 2 3 8 14 ...
```

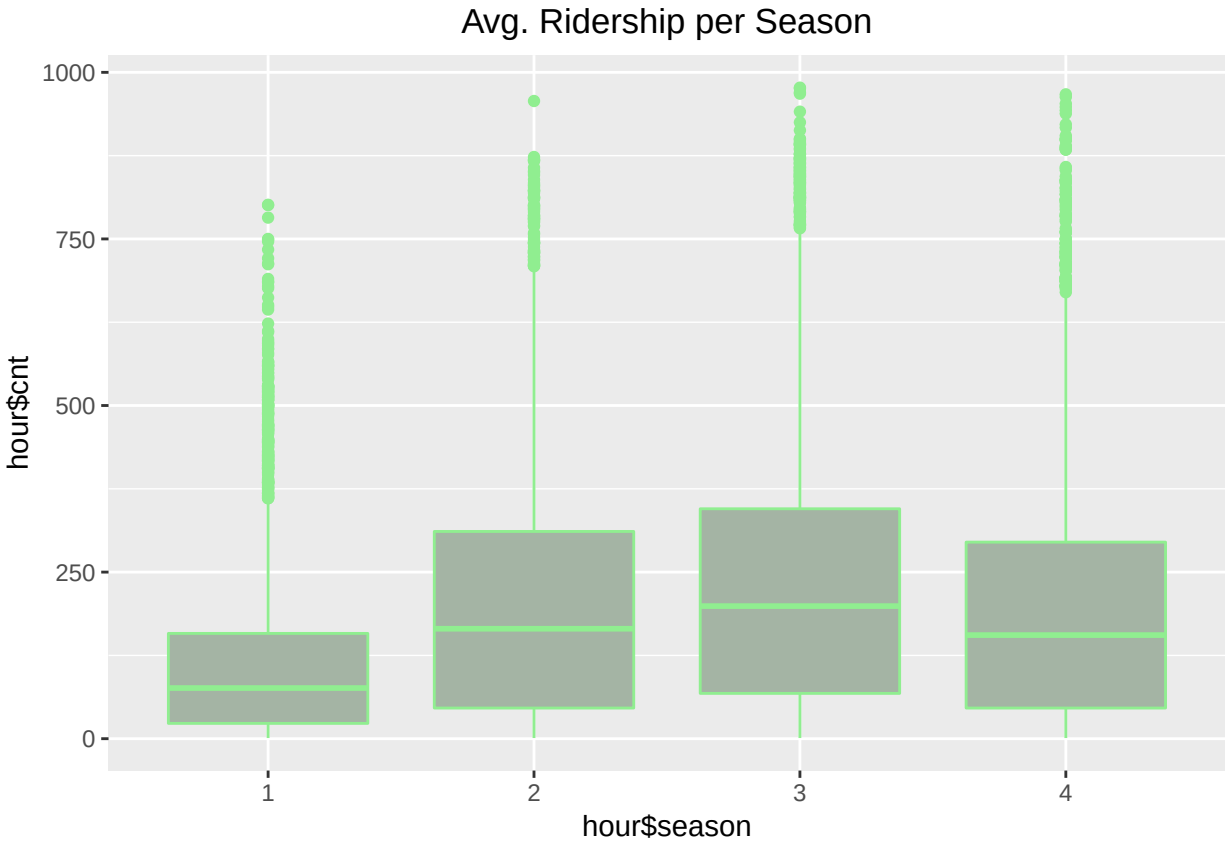
```
str(day)
```

```
## 'data.frame':    731 obs. of  16 variables:
##  $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : Date, format: "2011-01-01" "2011-01-02" ...
##  $ season    : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ yr        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ mnth      : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ weekday   : Factor w/ 7 levels "0","1","2","3",..: 7 1 2 3 4 5 6 7 1 2 ...
##  $ workingday: Factor w/ 2 levels "0","1": 1 1 2 2 2 2 2 1 1 2 ...
##  $ weathersit: Factor w/ 3 levels "1","2","3": 2 2 1 1 1 1 2 2 1 1 ...
##  $ temp      : num  0.344 0.363 0.196 0.2 0.227 ...
##  $ atemp     : num  0.364 0.354 0.189 0.212 0.229 ...
##  $ hum       : num  0.806 0.696 0.437 0.59 0.437 ...
##  $ windspeed : num  0.16 0.249 0.248 0.16 0.187 ...
##  $ casual    : int  331 131 120 108 82 88 148 68 54 41 ...
##  $ registered: int  654 670 1229 1454 1518 1518 1362 891 768 1280 ...
##  $ cnt       : int  985 801 1349 1562 1600 1606 1510 959 822 1321 ...
```

```
#2a
library(ggplot2)

ggplot(day, aes(x = day$dteday, y = day$cnt)) +
  ggtitle("Ridership per Day") + geom_line() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
#b
#Below

#c
ggplot(hour, aes(x = hour$season, y = hour$cnt))+
  geom_boxplot(fill='#A4B4A4', color="lightgreen") +
  ggtitle("Avg. Ridership per Season") +
  theme(plot.title = element_text(hjust = 0.5))
```

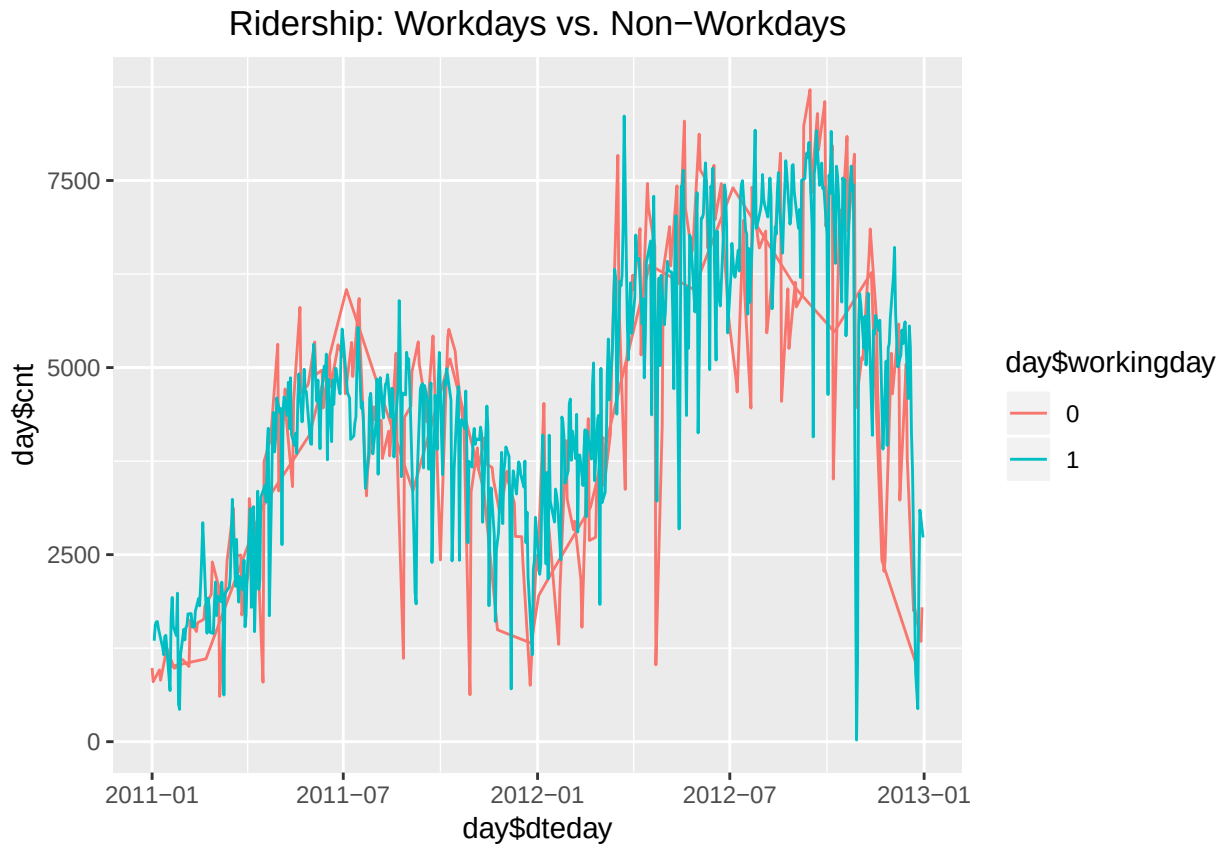



```
#d
#below
```

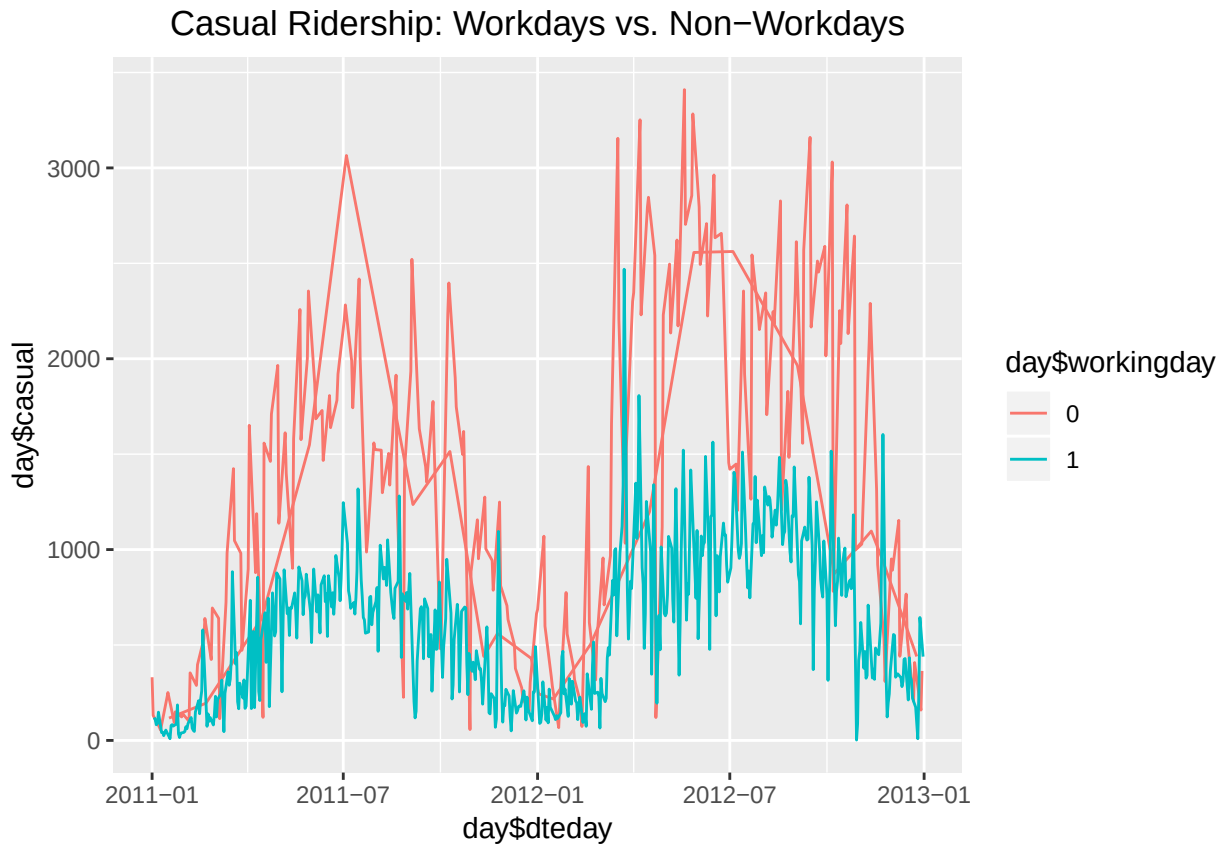
2b. The plot shows that ridership starts low and then gradually increases throughout the first half of 2011, decreases during the second half, rises immensely during the first half of 2012, and then plummets again during the second half. I hypothesize that this is because the weather is less pleasant during the beginning and end of the year; it is not "bike-riding weather".

2d. When looking at the seasonal breakdown of ridership per day, we find that Spring had the lowest average ridership. Summer and Fall see gradual increases and Winter sees a decrease. This data is consistent with out line chart.

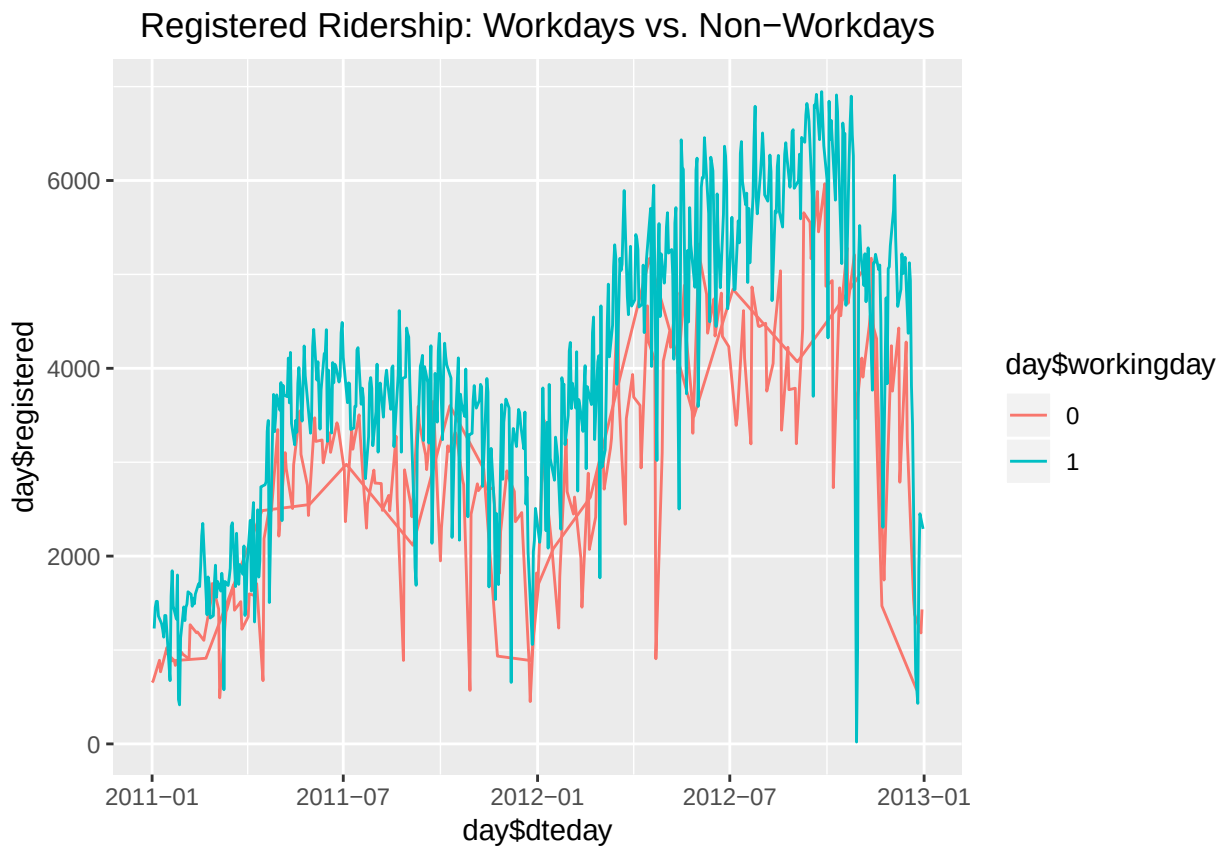
```
#3a
ggplot(day, aes(x = day$dteday, y = day$cnt, colour = day$workingday, shape = day$holiday)) +
  geom_line() +
  ggtitle("Ridership: Workdays vs. Non-Workdays") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
#b
#Below


#c

#Casual
ggplot(day, aes(x = day$dteday, y = day$casual, colour = day$workingday, shape = day$holiday)) +
  geom_line() +
  ggtitle("Casual Ridership: Workdays vs. Non-Workdays") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
#Registered
ggplot(day, aes(x = day$dteday, y = day$registered, colour = day$workingday, shape = day$holiday)) +
  geom_line() +
  ggtitle("Registered Ridership: Workdays vs. Non-Workdays") +
  theme(plot.title = element_text(hjust = 0.5))
```



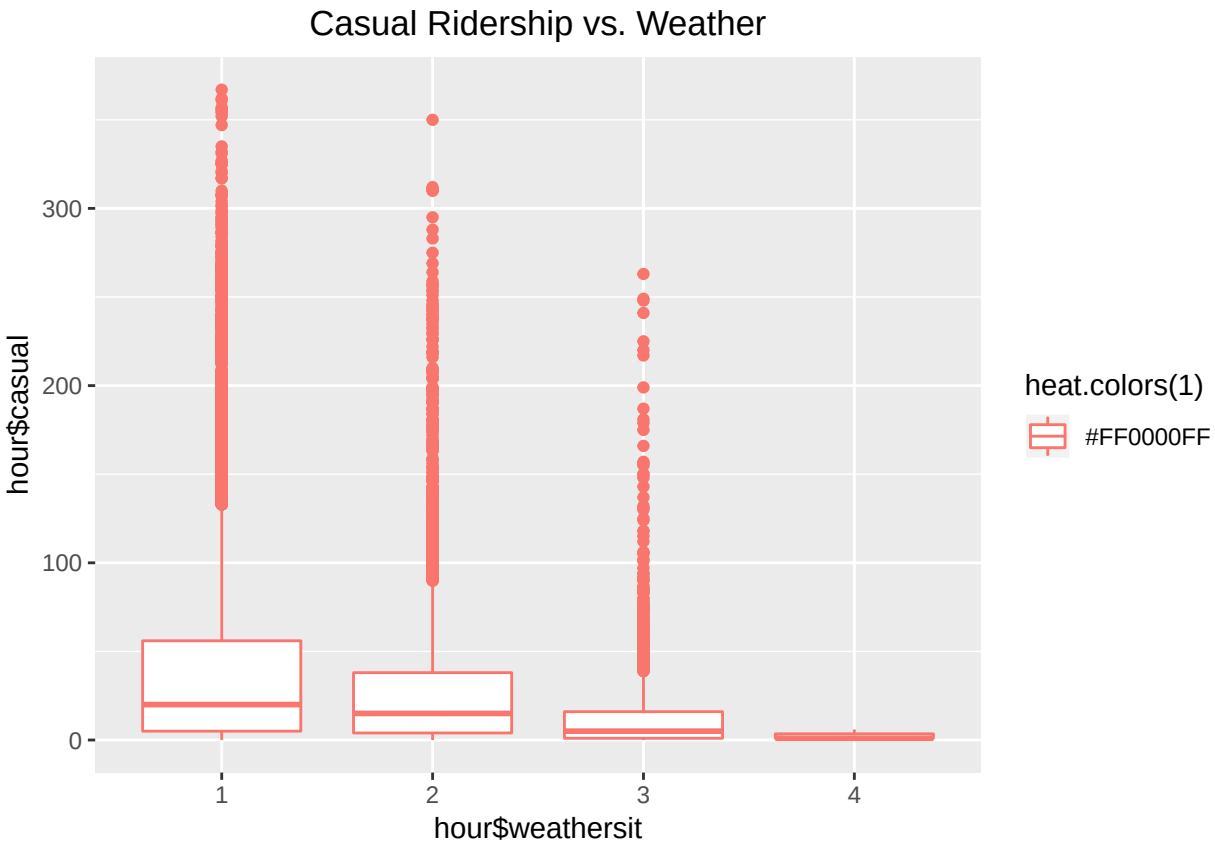
3b. There does not appear to be a difference in ridership trends depending on if the day is a workday or not. I feel that this is because people prefer to commute the same way regardless of if there is a holiday or not. When someone is used to driving everywhere, they are not suddenly just going to start biking unless there is a reason for them to do so.

3d. I immediately notice there is far less ridership among casual riders on workdays than casual riders on non-workdays. Despite the difference in proportion, both riderships tend to increase and decrease during similar parts of the year.

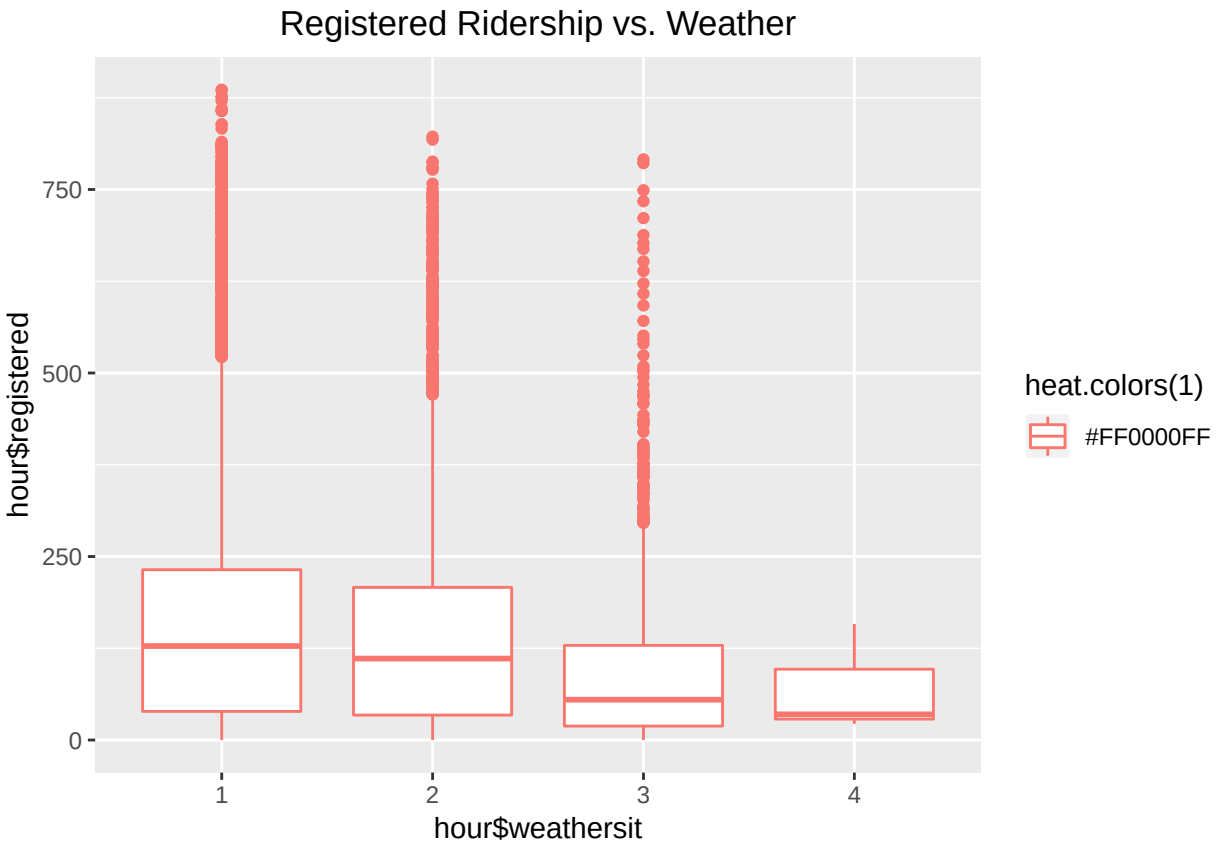
Among registered users, there does not seem to be significant variation in ridership depending on if the day is a workday or not. It appears that ridership is slightly higher on workdays. Both riderships tend to increase in the first half of the year and decrease in the second half.

From this, I conclude that registered users will ride bikes regardless of if they are going to work or somewhere else. Casual customers are far more likely to ride bikes on non-working days than on working days.

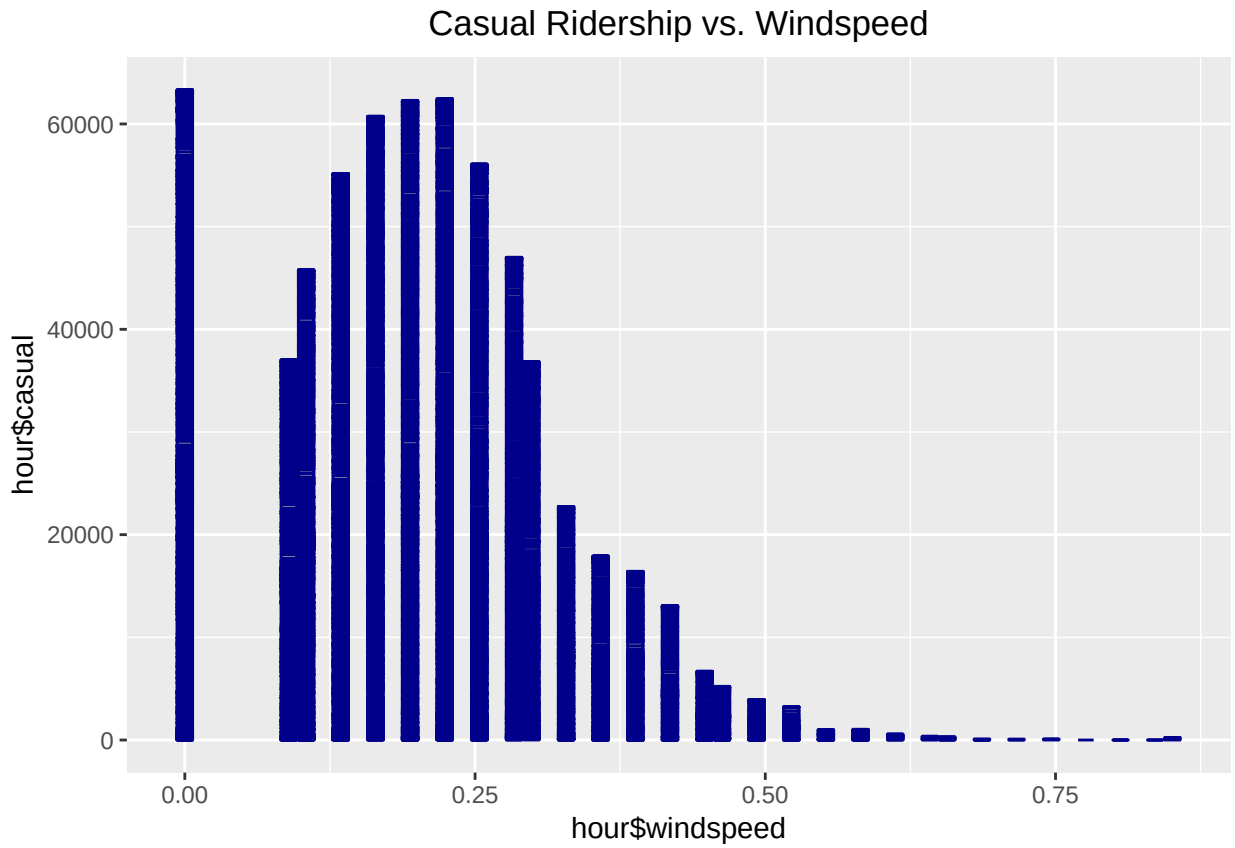
```
#4a:
ggplot(hour, aes(x = hour$weathersit, y = hour$casual, colour = heat.colors(1)))+
  geom_boxplot() +
  ggtitle("Casual Ridership vs. Weather") +
  theme(plot.title = element_text(hjust = 0.5))
```



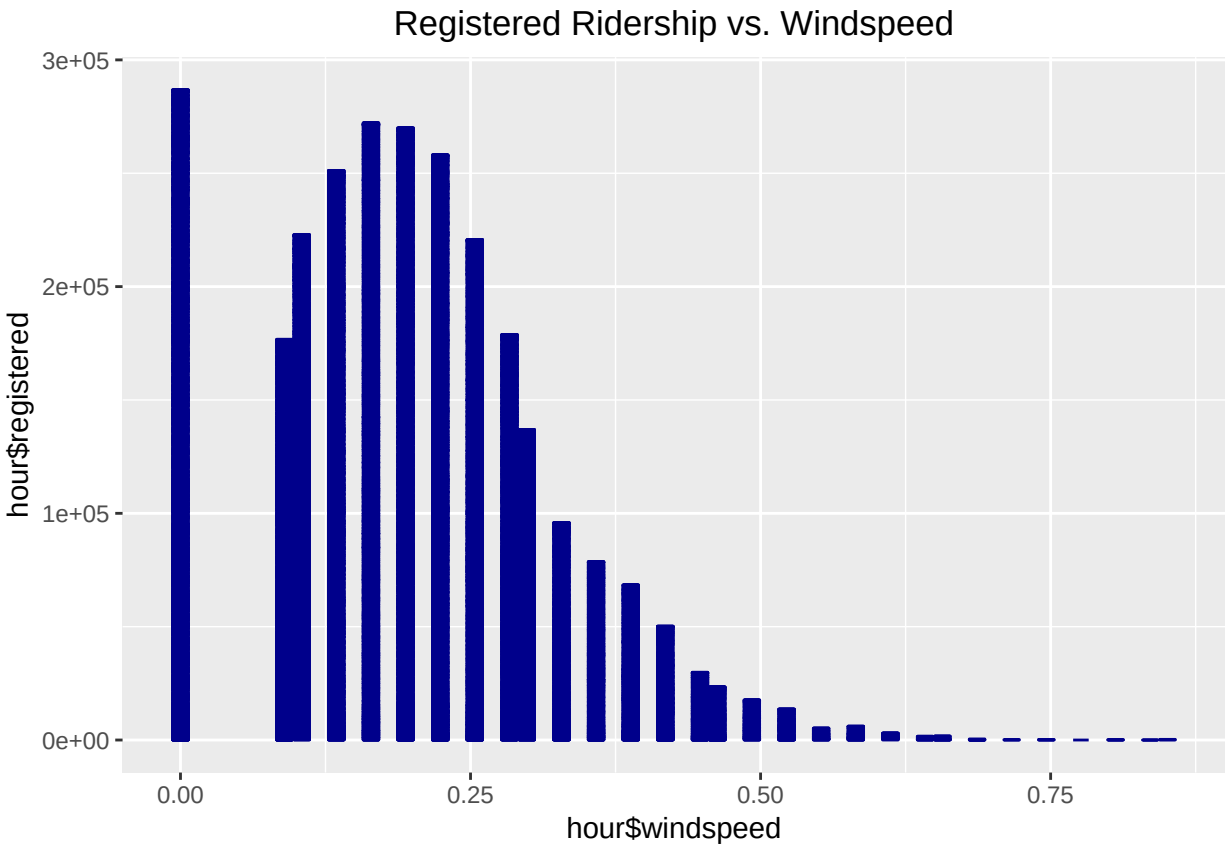

```
ggplot(hour, aes(x = hour$weathersit, y = hour$registered, colour = heat.colors(1)))+
  geom_boxplot() +
  ggtitle("Registered Ridership vs. Weather") +
  theme(plot.title = element_text(hjust = 0.5))
```



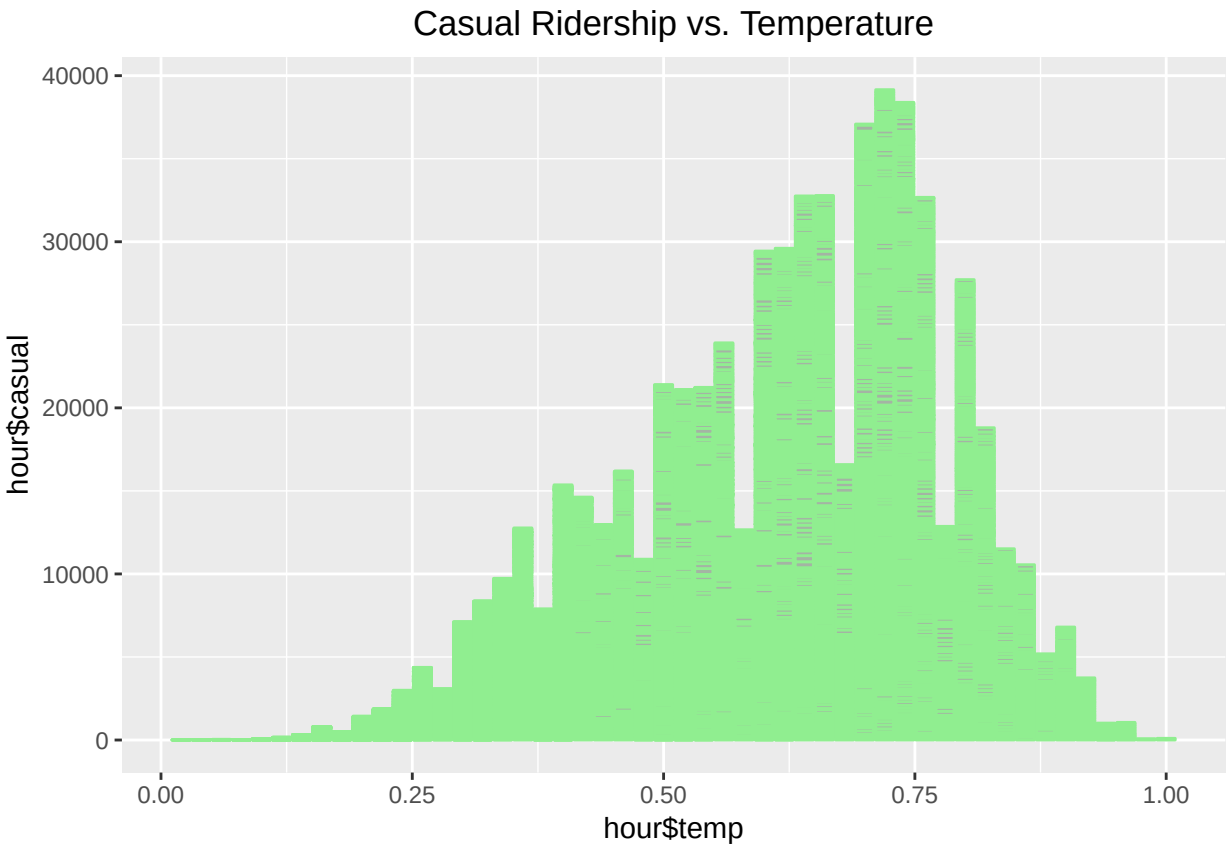

```
ggplot(hour, aes(x = hour$windspeed, y = hour$casual))+
  geom_col(fill='#A4B4A4', color="blue4") +
  ggtitle("Casual Ridership vs. Windspeed") +
  theme(plot.title = element_text(hjust = 0.5))
```



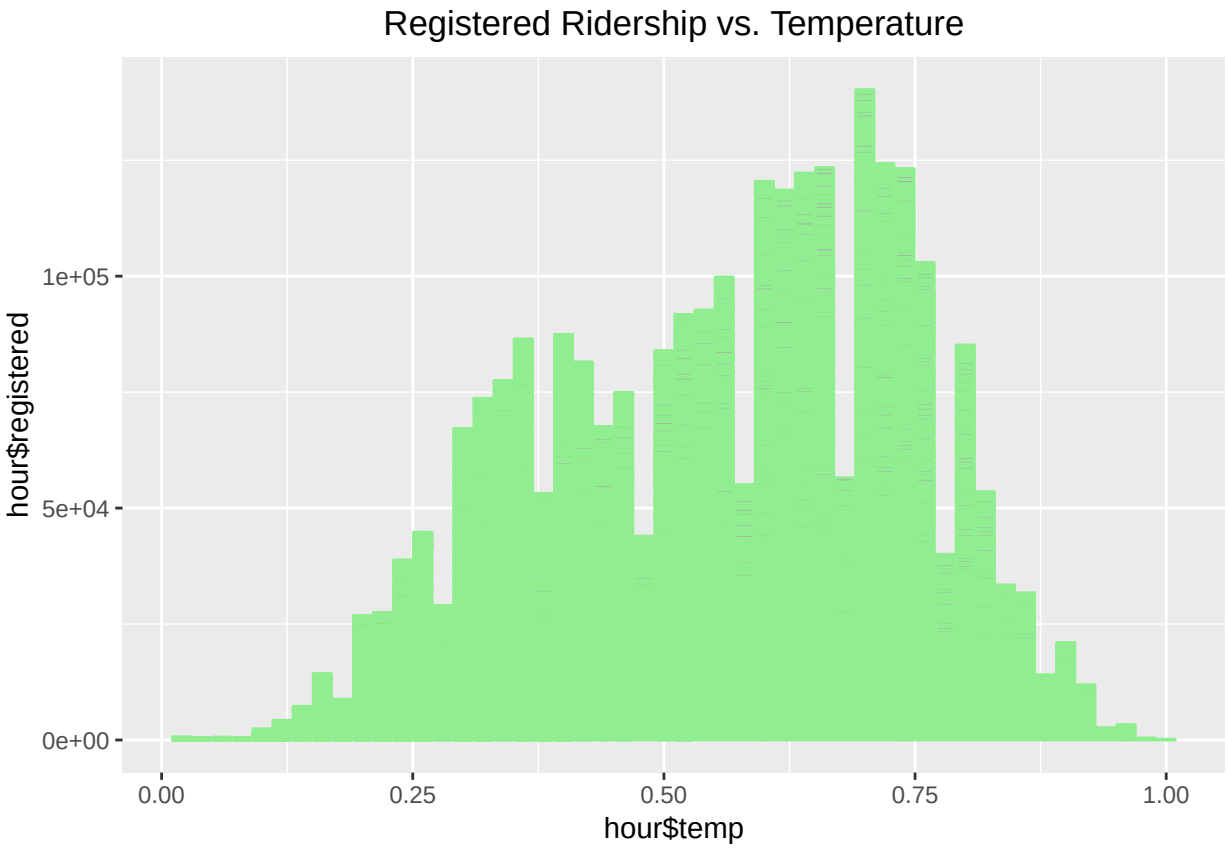

```
ggplot(hour, aes(x = hour$windspeed, y = hour$registered))+
  geom_col(fill='#A4B4A4', color="blue4") +
  ggtitle("Registered Ridership vs. Windspeed") +
  theme(plot.title = element_text(hjust = 0.5))
```



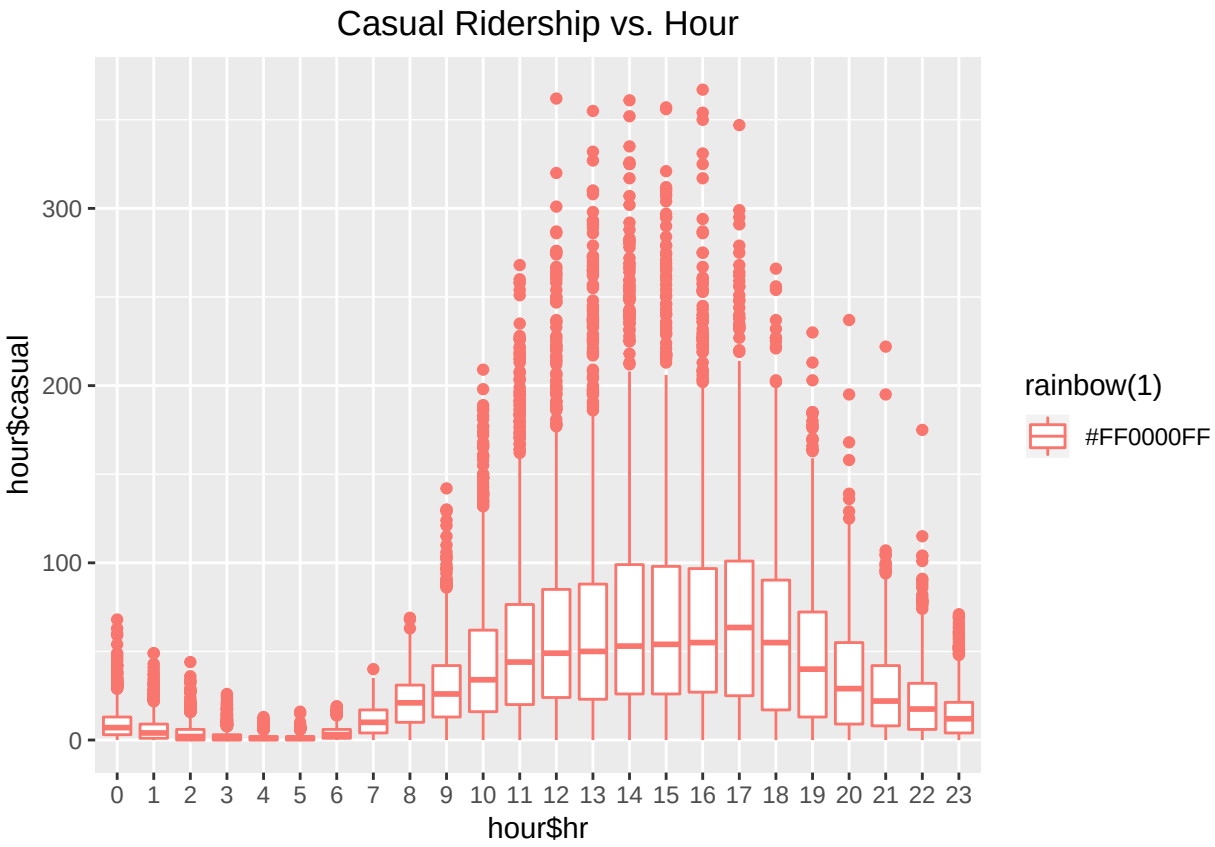

```
ggplot(hour, aes(x = hour$temp, y = hour$casual, group = hour$temp))+
  geom_col(fill='#A4B4A4', color="lightgreen") +
  ggtitle("Casual Ridership vs. Temperature") +
  theme(plot.title = element_text(hjust = 0.5))
```



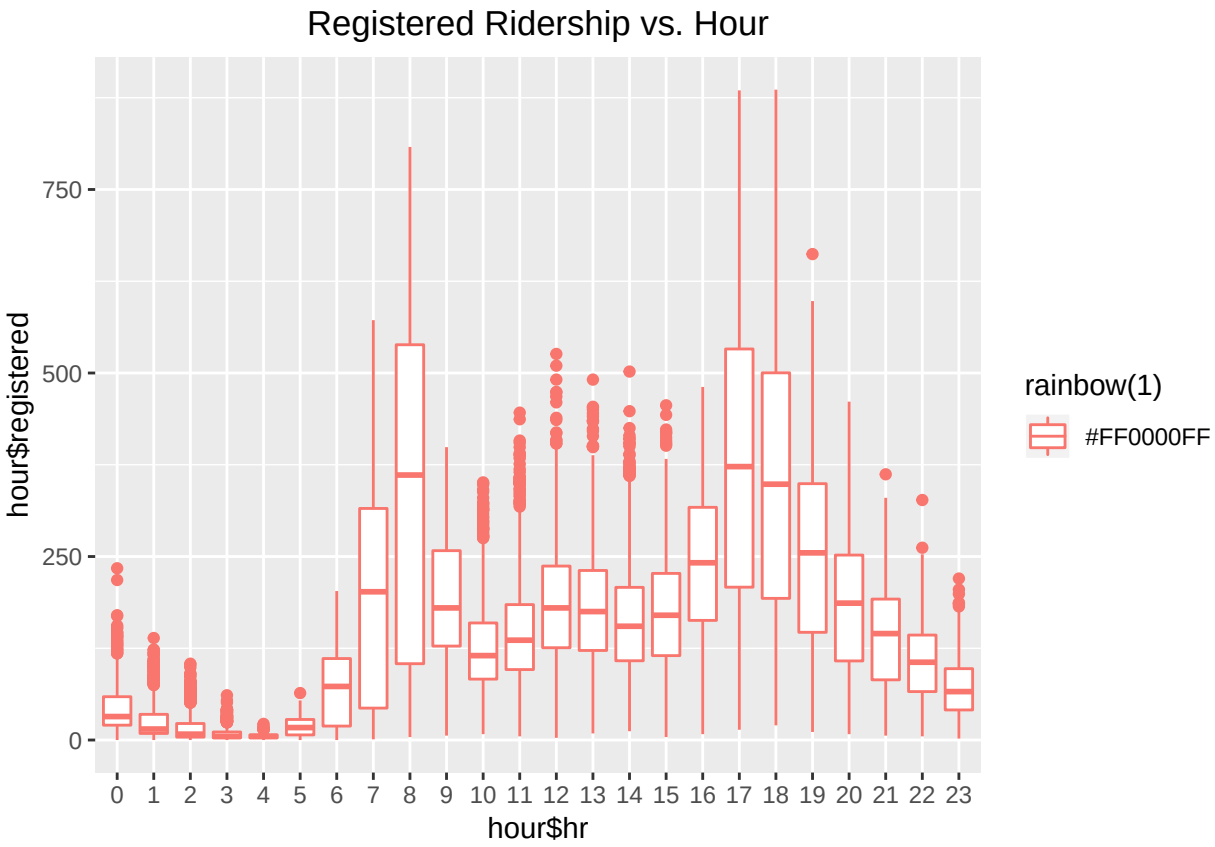

```
ggplot(hour, aes(x = hour$temp, y = hour$registered, group = hour$temp))+
  geom_col(fill='#A4B4A4', color="lightgreen") +
  ggtitle("Registered Ridership vs. Temperature") +
  theme(plot.title = element_text(hjust = 0.5))
```



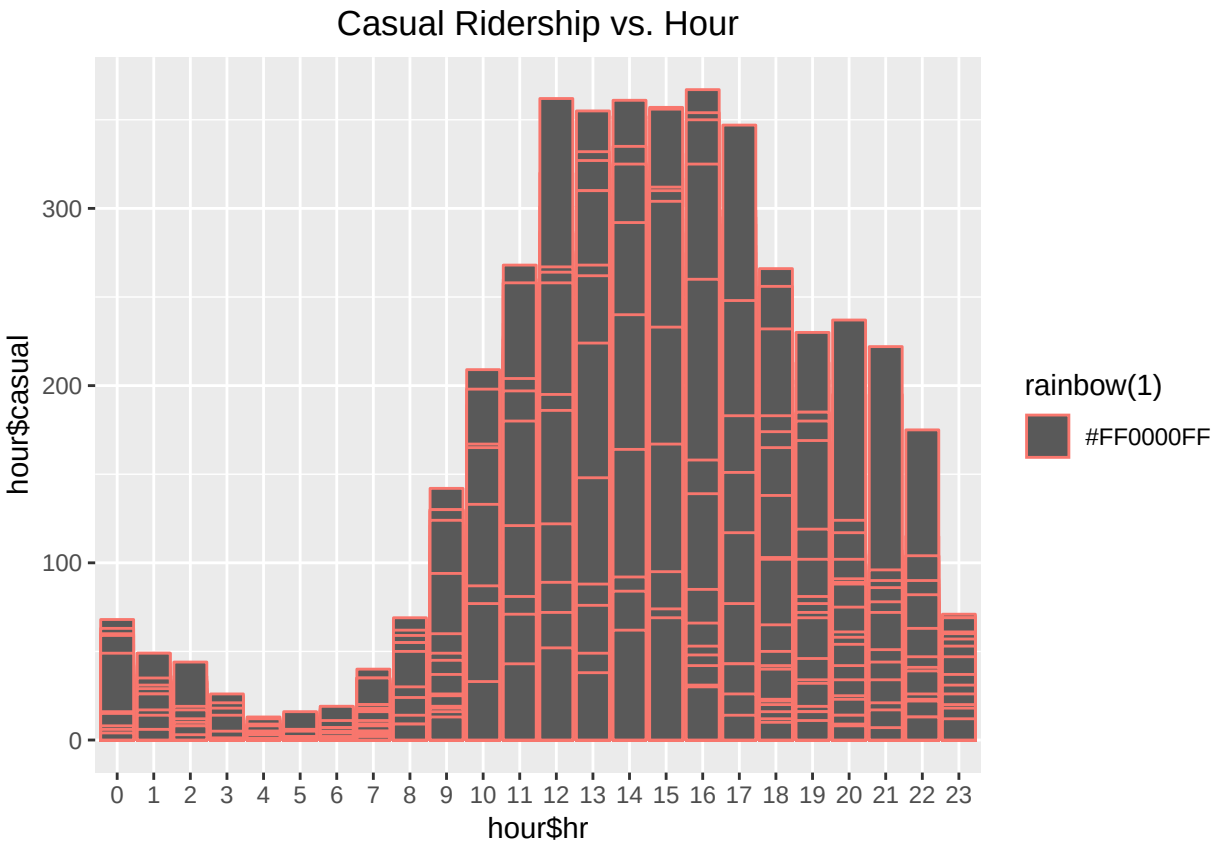

```
#b

ggplot(hour, aes(x = hour$hr, y = hour$casual, colour = rainbow(1)))+
  geom_boxplot() + ggtitle("Casual Ridership vs. Hour") +
  theme(plot.title = element_text(hjust = 0.5))
```



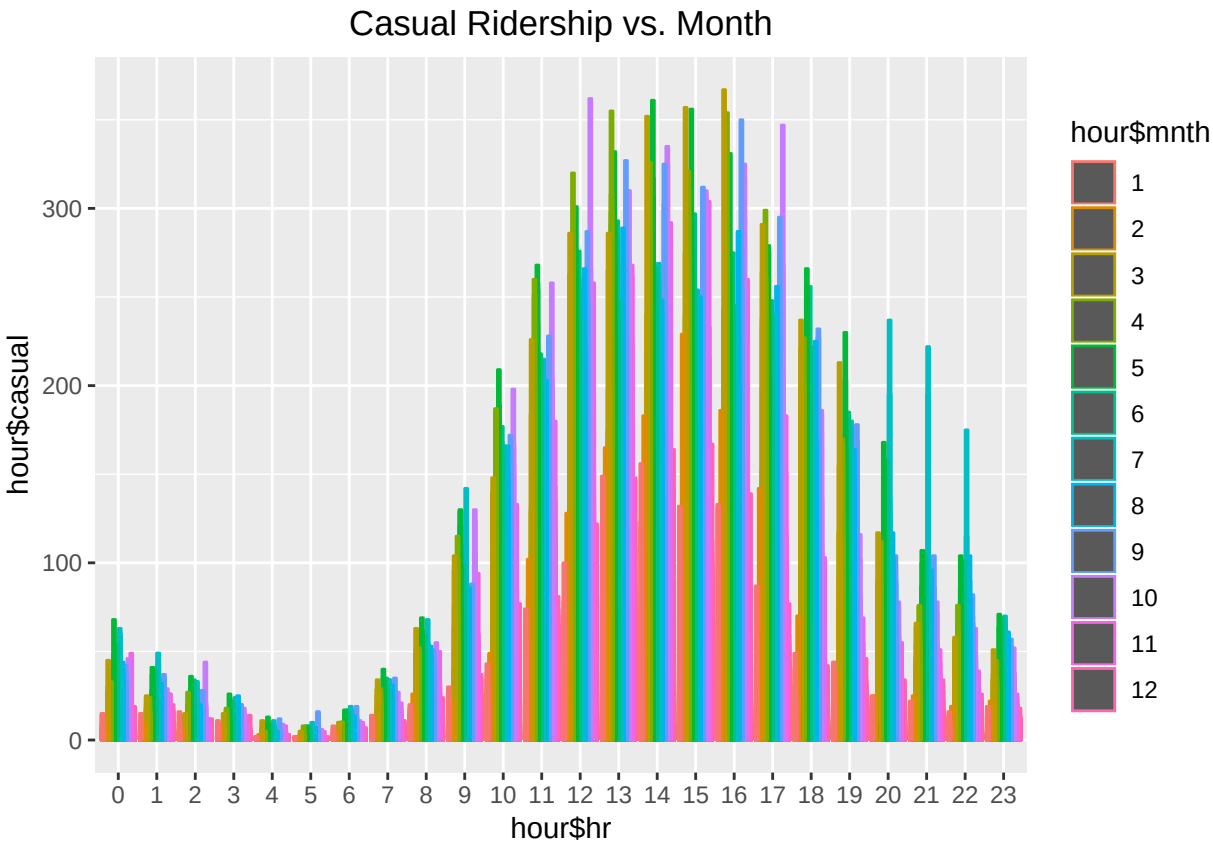

```
ggplot(hour, aes(x = hour$hr, y = hour$registered, colour = rainbow(1)))+
  geom_boxplot() +
  ggtitle("Registered Ridership vs. Hour") +
  theme(plot.title = element_text(hjust = 0.5))
```



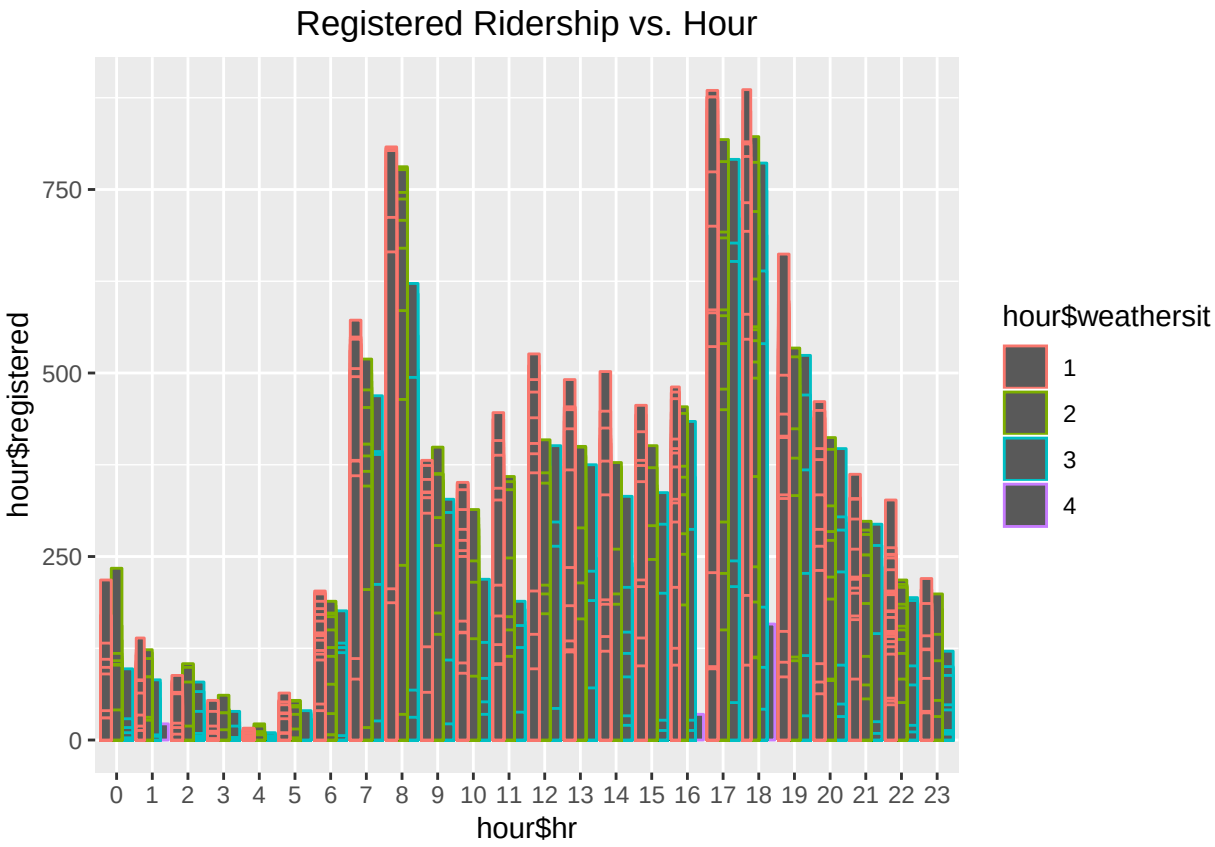

```
ggplot(hour, aes(x = hour$hr, y = hour$casual, colors(), colour = rainbow(1)))+
  geom_col(position = "dodge") +
  ggtitle("Casual Ridership vs. Hour") +
  theme(plot.title = element_text(hjust = 0.5))
```



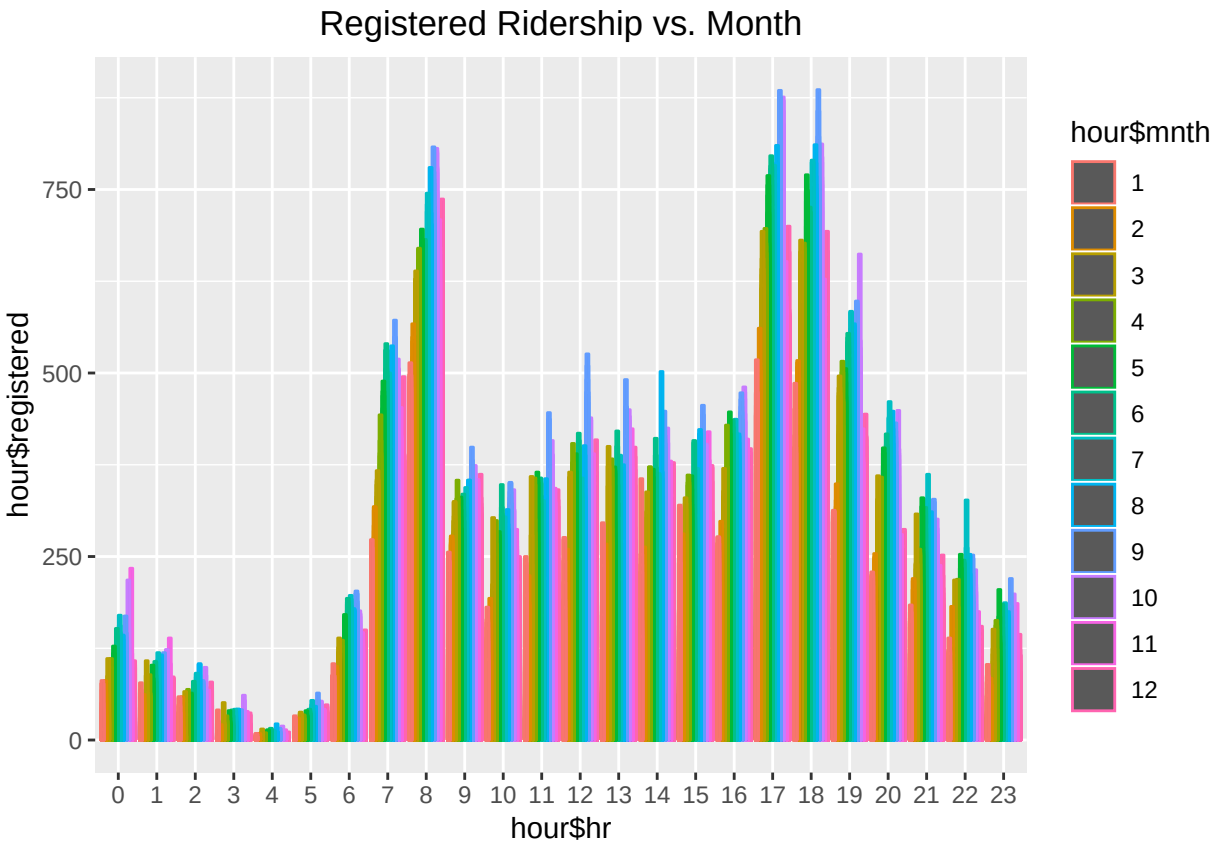

```
ggplot(hour, aes(x = hour$hr, y = hour$casual, color = hour$mnth))+
  geom_col(position = "dodge") +
  ggtitle("Casual Ridership vs. Month") +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_fill_hue(l = 10, c = 15)
```




```
ggplot(hour, aes(x = hour$hr, y = hour$registered, color = hour$weathersit))+
  geom_col(position = "dodge") +
  ggtitle("Registered Ridership vs. Hour") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(hour, aes(x = hour$hr, y = hour$registered, color = hour$mnth))+
  geom_col(position = "dodge") +
  ggtitle("Registered Ridership vs. Month") +
  theme(plot.title = element_text(hjust = 0.5))
```



4a. It appears as though when the weather is clearer and more stable, the ridership increases. Lower windspeeds correlate with higher ridership. Higher temperatures also correlate with higher ridership unless it gets too hot, in which case ridership decreases drastically. Interestingly, these trends do not vary significantly based on whether the rider is casual or registered.

4b. Casual riders tend to ride their bikes the most in the middle of the day, usually from 12-6. This is regardless of what weather or month it is. This being said, total casual ridership is most prominent during clear weather. Additionally, ridership ascends up into the middle months of the year and declines steadily afterwards.

Registered riders experience higher ridership in the early morning, likely because of commuting to work. The ridership also picks up a lot around 5:00, when most people tend to get off of work. This is regardless of what weather or month it is. Although total registered ridership is most prominent during times of clear weather, the drop-off between ridership in clear weather and other weather among registered riders is much smaller than that of the casual riders. This is likely because registered riders are more dependent on their bikes and need to ride them regardless of weather conditions. In terms of months, total registered ridership ascends throughout the year and drops off slightly at the very end of the year, during November and December. I predict once again that this drop off is much smaller compared to that of casual riders because registered riders are much more dependent on their bikes year-round and only experience a decrease in dependency during the holiday season, when people tend to work less and the weather is unfavorable.