Economy of the Big Data World

HW 2 – Bike sharing systems analysis

Given Data Base:

Raw table (CSV Format) that contains an hourly data on Bike renting service in Washington D.C between 2011-2012. The variables are as following:

- Date & Time
- Season
- Holiday
- Working day
- Weather
- Temp
- Atemp (feels like)
- Humidity
- Windspeed
- Count (number of rentals per hour)

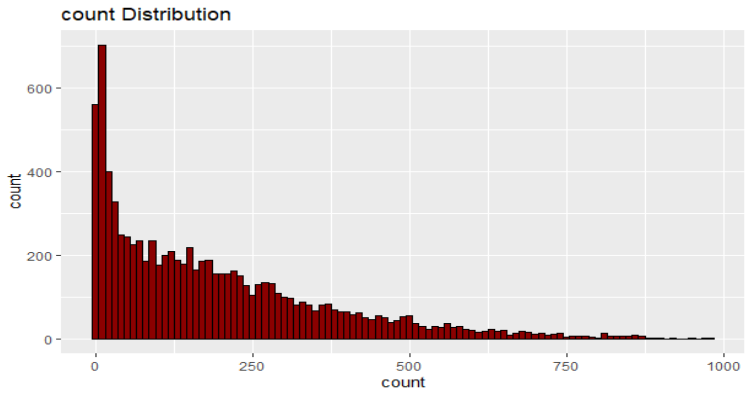
The task:

Predict number of rentals per hour.

Part I – Descriptive Statistics

Number of Rentals (Count) -

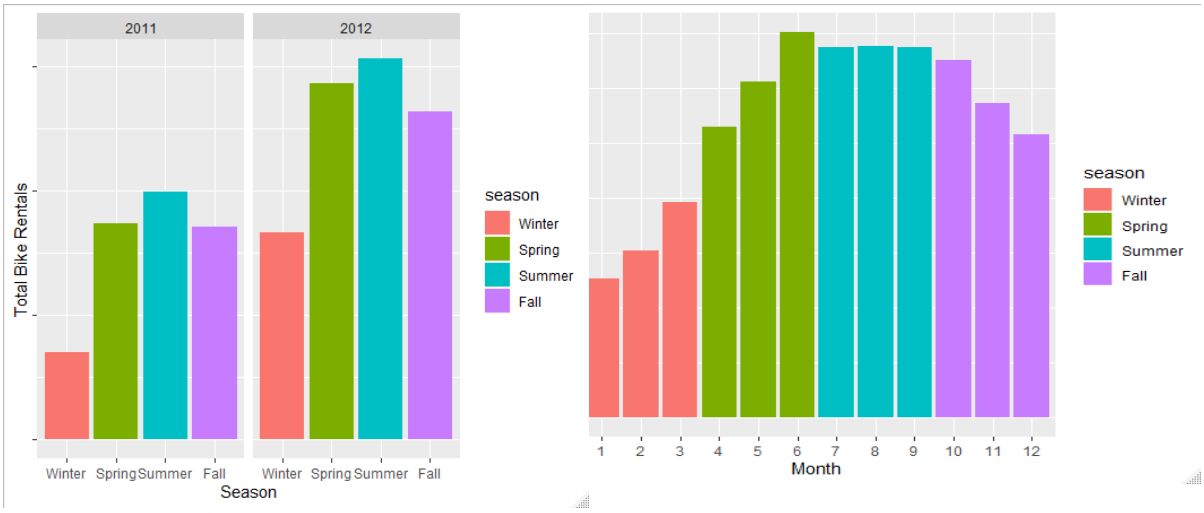
According to the distribution of Count, we can see that the number of rentals per hour is usually between 0 to 400 as we got a 'right-tail' distribution.




```
ggplot(mydata)+aes(count)+
  geom_histogram(color="black",fill="darkred",binwidth = 10)+
  labs(title = "count Distribution")
```

Seasonality –

The number of rentals is affected by seasons, as we can see clearly that during the winter, the rentals' number is relatively low. We also compared between 2011 and 2012 for consistency check:

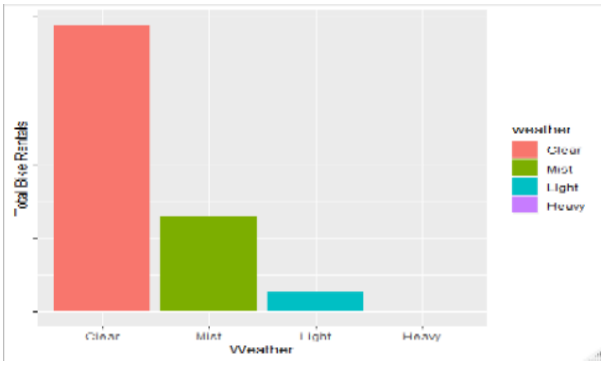


We could argue that the variable 'Season' contains within itself many aspects of the Weather. Meaning, during the winter, it's colder, darker, etc. Therefore, it's reasonable that there will be less rentals during the winter due to the typically weather, and not because it's officially wintertime.

```
mydata$season <- factor(mydata$season,
                        levels = c(1,2,3,4),
                        labels = c("winter", "spring", "summer","fall"))
ggplot(data = Dataset_train,aes(x = season, y = count)) +
  geom_bar(stat = 'identity', aes(fill = season),) +
  xlab('Season') +
  ylab('Total Bike Rentals') +
  theme(axis.text.y = element_blank()) +
  facet_wrap(Dataset_train$Year)
```

Weather –

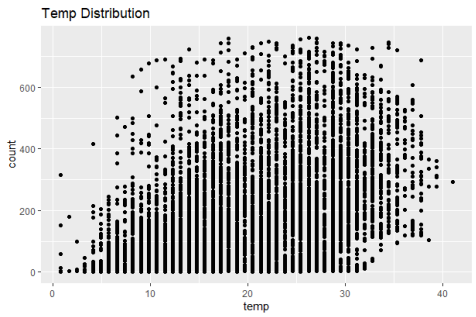
Following our conclusion above, we can see clearly that the number of rentals is affected by the Weather; when it's clear, there are much more rentals comparing to the less comfortable weather conditions rain wise.




```
mydata$weather<-as.factor(mydata$weather)
mydata$weather<-factor(mydata$weather, labels=c("clear","mist","light","heavy"))

ggplot(mydata,aes(x = weather, y = count)) +
  geom_bar(stat = 'identity', aes(fill = weather)) +
  xlab('Weather') +
  ylab('Total Bike Rentals') +
  theme(axis.text.y = element_blank())
```

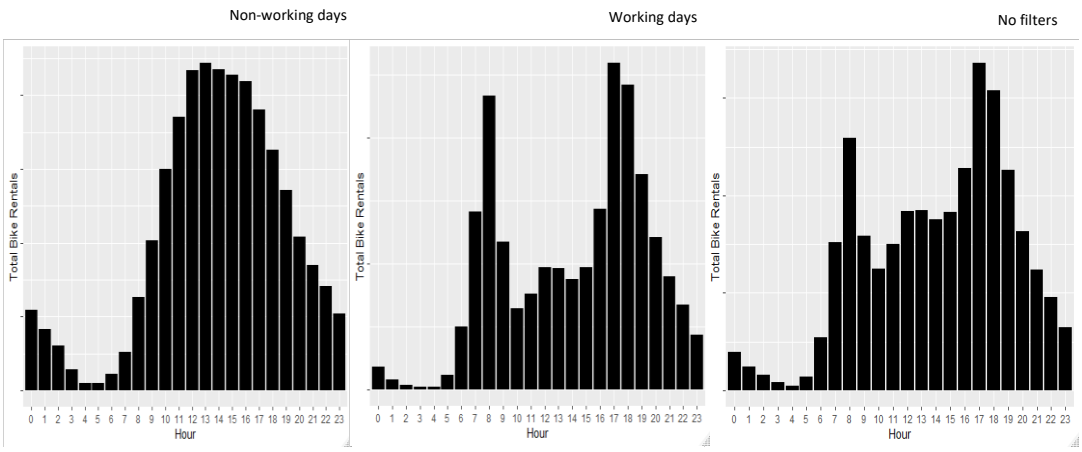
Also, we can see a positive correlation between number of rentals and temperature:




```
ggplot(mydata,aes(x = temperatue, y = count)) +
  geom_point(stat = 'identity')
```

Date & Time –

According to the data, the busiest hours are around 8 AM and 5-6 PM. That's reasonable as people are getting to work around 8 AM and leaving around 5-6 PM. This assumption is matching the fact that on the weekend, the rentals' distribution is different than on the workdays.



Final conclusions –

The bikes' rental service is affected by two main elements:

1. Weather – there is a positive correlation between better weather and rentals. When it's clearer, there are more rentals.
2. Workdays – there are much more rentals during the working days than during the weekends.

Part II – Linear Regression Model

Linear Regression -

$$Count_i = \beta_0 + \beta_1 * Weather + \beta_2 * Hour * Wday + \ u_i$$

We decided to create our model based on the two main variables; weather and Date & Time. We interact Hour and Wday as we saw above that the Hours on a workday are behaving different than on the weekend.

The results are:

R^2 - 0.578

P-value – 2.2e-16