Bikes rentals prediction

Created by Marcos Ikino

Introduction

This study refers to a rental bikes system which contains hourly and daily counts information between the years 2011 and 2012. The datasets were obtained from the UCI repository (https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset) and the subject of this study is to predict the hourly number of bikes rentals regardless of the climatic conditions or period of the year.

For that, initially, we are going to do data analysis by comparing graphically the behaviors between the casual and the registered cyclist for better understanding the differences between them. To predict the output variable we will make the use of the machine learning model that better fits the predictions against the actual values.

Attribute information

- instant: record index
- dteday: date
- season: season (1:springer, 2:summer, 3:fall, 4:winter)
- yr: year (0: 2011, 1:2012)
- mnth: month (1 to 12)
- hr: hour (0 to 23)
- holiday: weather day is holiday or not (extracted from [Web Link])
- weekday: day of the week
- workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit:
  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
- temp: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users

- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Part I – Exploratory data analysis

1. Importing the data

```
data <- read.csv('hour.csv')
str(data)

## 'data.frame':    17379 obs. of  17 variables:
##  $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : Factor w/ 731 levels "2011-01-01","2011-01-02",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ yr        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mnth      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hr        : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday   : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weathersit: int  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 ...
```

2. Variables transformations

```
data$dteday <- NULL

# Defining the categorical variables
categ_vars <- c('season', 'yr', 'mnth', 'hr', 'holiday', 'weekday', 'workingday',
'weathersit')

for (i in categ_vars) {
    data[, i] <- factor(data[, i])
}
```

```r
# Renaming the categorical variables and their parameters

library(data.table)
library(plyr)

setnames(data, old = c('yr', 'mnth', 'hr', 'temp', 'atemp', 'hum', 'cnt'), new =
c('year', 'month', 'hour', 'temperature', 'atemperature', 'humidity', 'count'),
skip_absent = TRUE)
data$season <- mapvalues(data$season, from = c(1, 2, 3, 4), to = c('Spring',
'Summer', 'Fall', 'Winter'))
data$year <- mapvalues(data$year, from = c(0, 1), to = c(2011,2012))
data$holiday <- mapvalues(data$holiday, c(0, 1), to = c('No_holiday',
'Yes_holiday'))
data$weekday <- mapvalues(data$weekday, from = c(0,1,2,3,4,5,6), to = c("Sun",
"Mon", "Tue", "Wed", "Thu", "Fri", "Sat"))
data$workingday <- mapvalues(data$workingday, from = c(0 ,1), to = c('No_workday',
'Yes_workday'))
data$weathersit <- mapvalues(data$weathersit, from = c(1,2,3,4), to = c("Clear",
"Cloudy", "Light Rain/Snow", "Heavy Rain/Snow"))

str(data)
## 'data.frame':    17379 obs. of  16 variables:
##  $ instant     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ season      : Factor w/ 4 levels "Spring","Summer",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ year        : Factor w/ 2 levels "2011","2012": 1 1 1 1 1 1 1 1 1 1 ...
##  $ month       : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ hour        : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9 10
...
##  $ holiday     : Factor w/ 2 levels "No_holiday","Yes_holiday": 1 1 1 1 1 1 1 1
1 1 ...
##  $ weekday     : Factor w/ 7 levels "Sun","Mon","Tue",..: 7 7 7 7 7 7 7 7 7 7
...
##  $ workingday  : Factor w/ 2 levels "No_workday","Yes_workday": 1 1 1 1 1 1 1 1
1 1 ...
##  $ weathersit  : Factor w/ 4 levels "Clear","Cloudy",..: 1 1 1 1 1 2 1 1 1 1 ...
##  $ temperature : num  0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
##  $ atemperature: num  0.288 0.273 0.273 0.288 0.288 ...
##  $ humidity    : 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 ...
##  $ count       : int  16 40 32 13 1 1 2 3 8 14 ...
```

```r
summary(data)
```

```
##     instant         season       year          month           hour      
##  Min.   :    1   Spring:4242   2011:8645   5      :1488   16     :  730  
##  1st Qu.: 4346   Summer:4409   2012:8734   7      :1488   17     :  730  
##  Median : 8690   Fall  :4496               12     :1483   13     :  729  
##  Mean   : 8690   Winter:4232               8      :1475   14     :  729  
##  3rd Qu.:13034                             3      :1473   15     :  729  
##  Max.   :17379                             10     :1451   12     :  728  
##                                            (Other):8521   (Other):13004  
##         holiday      weekday          workingday   
##  No_holiday :16879   Sun:2502   No_workday : 5514  
##  Yes_holiday:  500   Mon:2479   Yes_workday:11865  
##                      Tue:2453                      
##                      Wed:2475                      
##                      Thu:2471                      
##                      Fri:2487                      
##                      Sat:2512                      
##            weathersit     temperature     atemperature       humidity     
##  Clear          :11413   Min.   :0.020   Min.   :0.0000   Min.   :0.0000  
##  Cloudy         : 4544   1st Qu.:0.340   1st Qu.:0.3333   1st Qu.:0.4800  
##  Light Rain/Snow: 1419   Median :0.500   Median :0.4848   Median :0.6300  
##  Heavy Rain/Snow:    3   Mean   :0.497   Mean   :0.4758   Mean   :0.6272  
##                          3rd Qu.:0.660   3rd Qu.:0.6212   3rd Qu.:0.7800  
##                          Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
##                                                                           
##    windspeed          casual         registered        count      
##  Min.   :0.0000   Min.   :  0.00   Min.   :  0.0   Min.   :  1.0  
##  1st Qu.:0.1045   1st Qu.:  4.00   1st Qu.: 34.0   1st Qu.: 40.0  
##  Median :0.1940   Median : 17.00   Median :115.0   Median :142.0  
##  Mean   :0.1901   Mean   : 35.68   Mean   :153.8   Mean   :189.5  
##  3rd Qu.:0.2537   3rd Qu.: 48.00   3rd Qu.:220.0   3rd Qu.:281.0  
##  Max.   :0.8507   Max.   :367.00   Max.   :886.0   Max.   :977.0  
## 
```

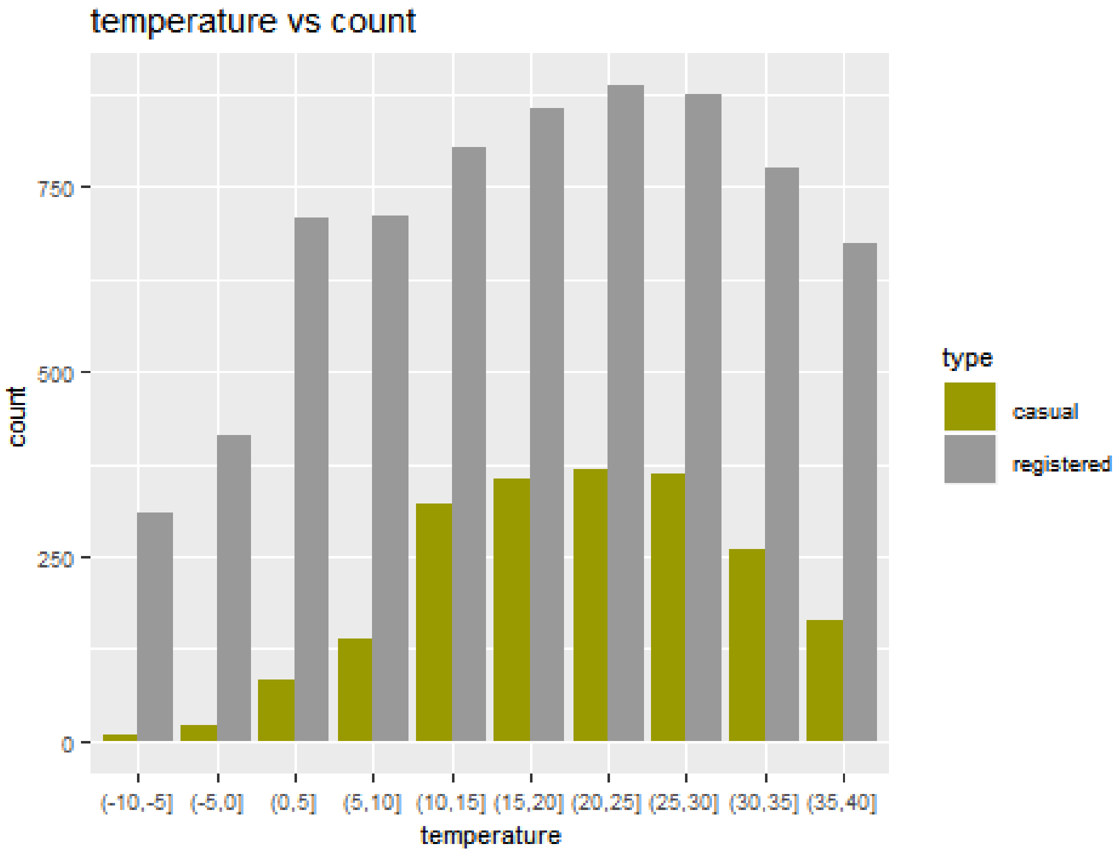
3. Data analysis

In this section, it will be displayed some graphs comparing the behaviors of the casual and registered users in the way that they rent the bikes according to climatic conditions, weekday, month or hour of the day. It is observable that the casual users are basically composed by the tourists, eventual users or the people that don´t utilize regularly the bikes for the working destination. Otherwise, the use proposal of the registered cyclists is essentially for the work locomotion.

a. temperature vs count

As observed, the general behavior between casual and registered cyclists is the same. When the temperature conditions are nice the bikes utilization increases, and on the other hand, in extreme conditions the usage drops accentually.

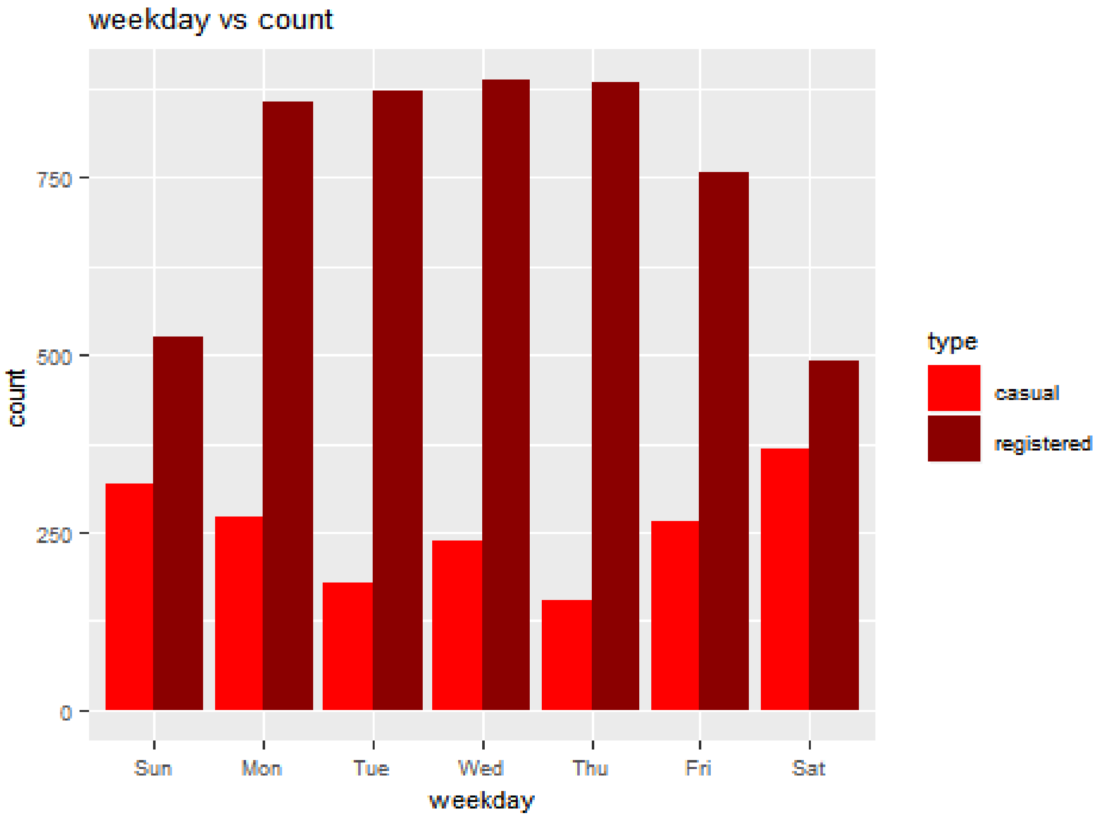
```r
library(dplyr)
library(ggplot2)
library(tidyr)

# Creating a new dataframe
casual <- data[, 'casual']
registered <- data[, 'registered']
temperature <- data[, 'temperature']
weekday <- data[, 'weekday']
weathersit <- data[, 'weathersit']
season <- data[, 'season']
hour <- data[, 'hour']
month <- data[, 'month']

data1 <- data.frame(casual, registered, temperature, weekday, weathersit, season, 
hour, month)

# Converting the temperaute data to Celsius
data1$temperature <- data1$temperature*(max(data1$temperature)-
min(data1$temperature))+min(data1$temperature)
data1$temperature <- round(data1$temperature*(39-(-8))+(-8),1)

# Plotting the graph
data2 <- data1 %>%
  gather(type, count, -c(temperature, weekday, weathersit, season, hour, month))

data3 <- data2 %>%
  mutate(temperature = cut(temperature, breaks= c(-10, -5, 0, 5, 10, 15, 20, 25, 
30, 35, 40)))

ggplot(data = data3, aes(x = temperature, y = count, fill = type)) +
  geom_bar(stat = 'identity', position = position_dodge())+
  scale_fill_manual(values = c("#999900", "#999999")) +
  ggtitle("temperature vs count") +
  theme(text = element_text(size=8.5),
        axis.text.x = element_text(angle=0))
```



b. weekday vs count

As predicted, the main utilization of the bikes for the registered users is for working, seen during the rentals of the week. In the opposite, on the weekend the rental proportion of casual users increases compared to a weekday, denoting the recreational use adopted by most of the casual users.

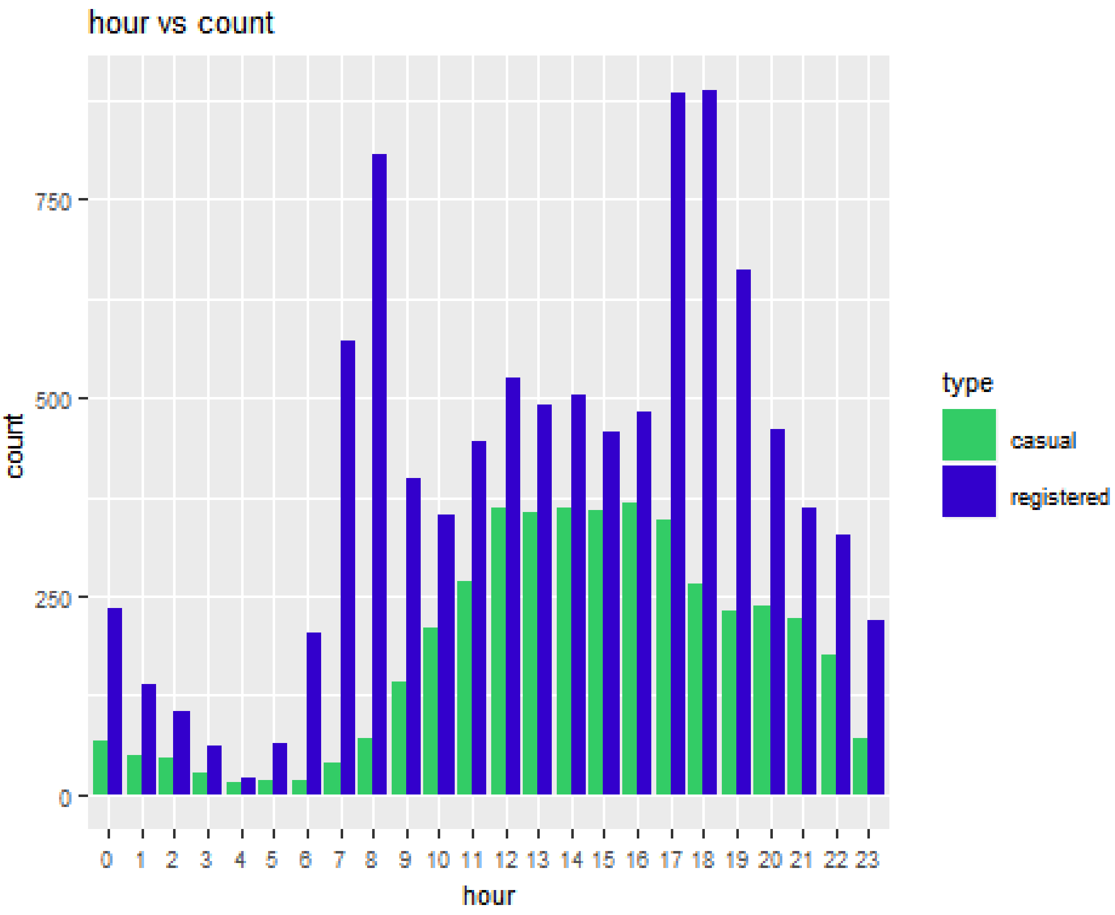
```
ggplot(data = data2, aes(x = weekday, y = count, fill = type)) +
  geom_bar(stat = 'identity', position = position_dodge()) +
  scale_fill_manual(values = c("red", "darkred")) +
  ggtitle("weekday vs count") +
  theme(text = element_text(size=8),
        axis.text.x = element_text(angle=0))
```



c. hour vs count

To corroborate the working utilization of the bikes by the registered users, the demand peaks can be clearly observed on the entrance and the exit regular working times, not noted for the casual cyclists that use the bikes in more affordable times.

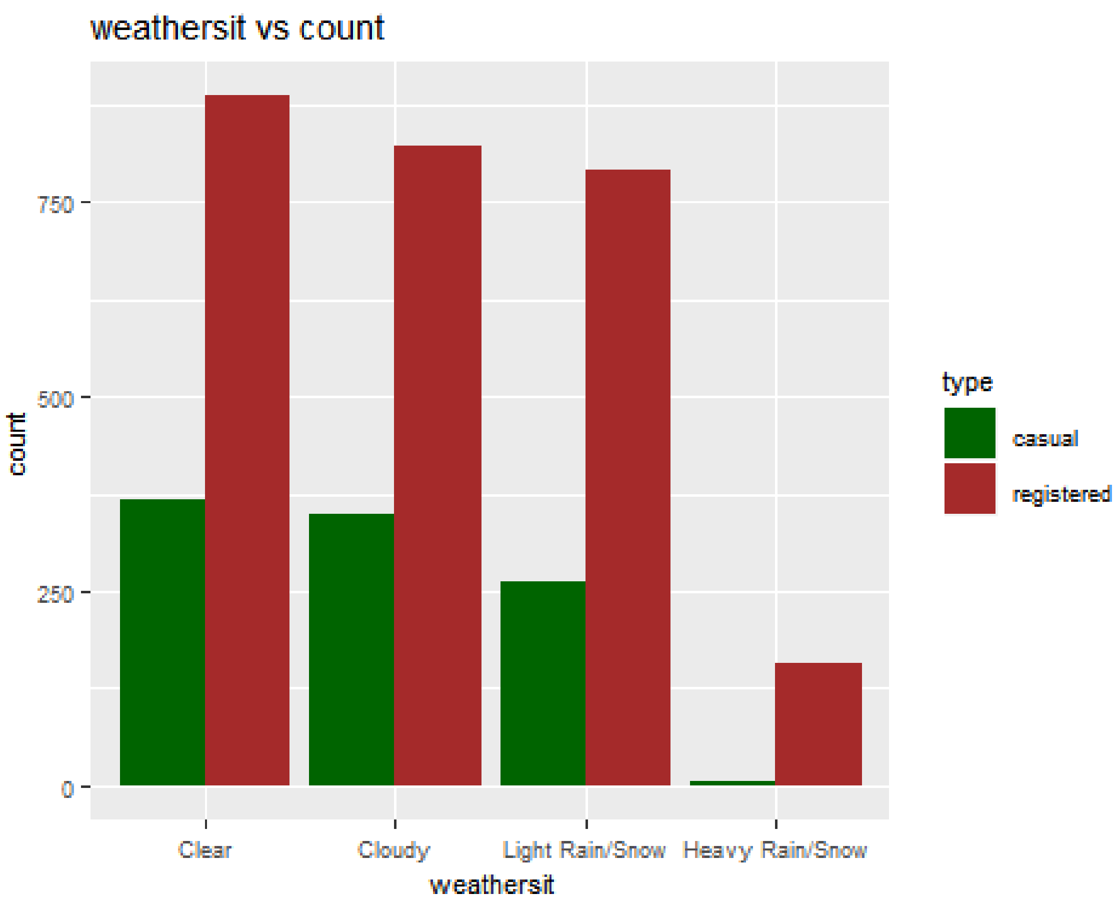
```r
ggplot(data = data2, aes(x = hour, y = count, fill = type)) +
  geom_bar(stat = 'identity', position = position_dodge())+
  scale_fill_manual(values = c("#33cc66", "#3300cc")) +
  ggtitle("hour vs count") +
  theme(text = element_text(size=8),
        axis.text.x = element_text(angle=0))
```



d. weathersit vs count

The weather conditions like the temperature have correlations for even casual and registered users to utilize or not the bikes due to climatic conditions. How worse it is, less they will rent it. Proportionally in the heavy rain/snow, the use by the casual cyclists drops more radically.

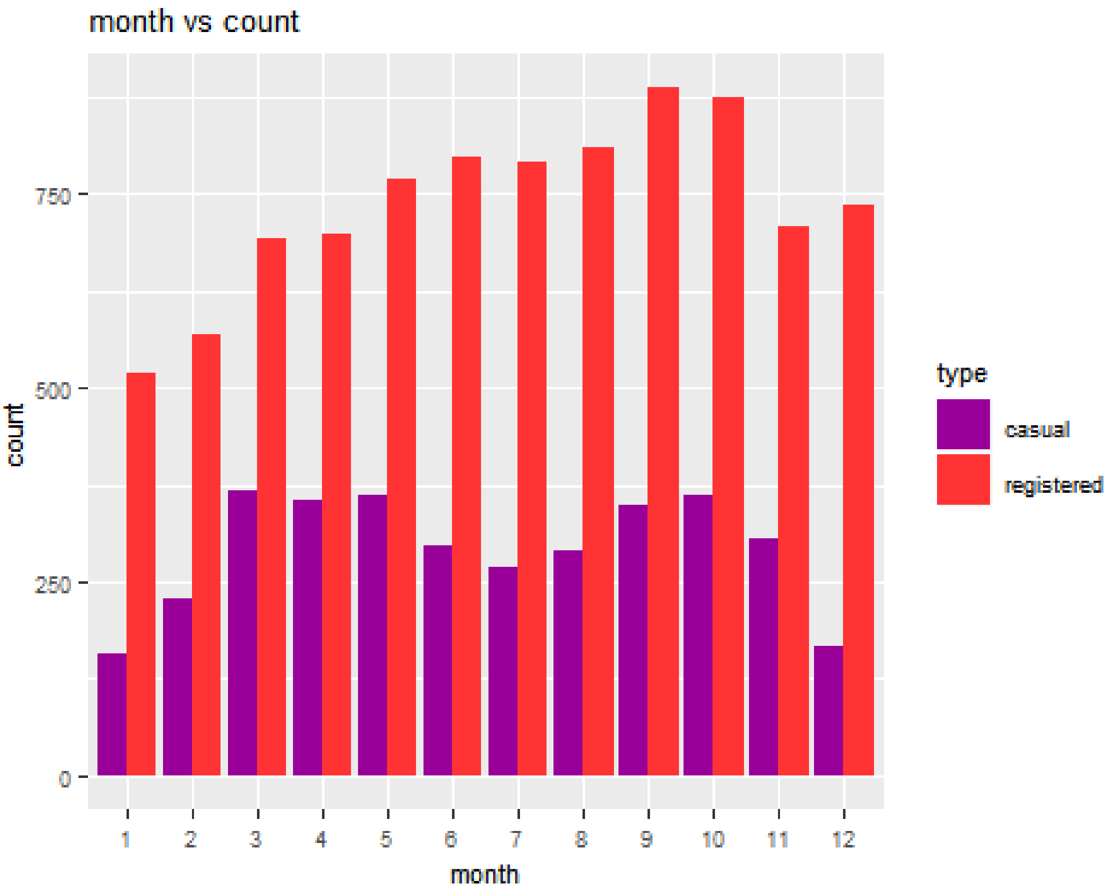
```r
ggplot(data = data2, aes(x = weathersit, y = count, fill = type)) +
  geom_bar(stat = 'identity', position = position_dodge())+
  scale_fill_manual(values = c("darkgreen", "brown")) +
  ggtitle("weathersit vs count") +
  theme(text = element_text(size=8.5),
        axis.text.x = element_text(angle=0))
```



e. month vs count

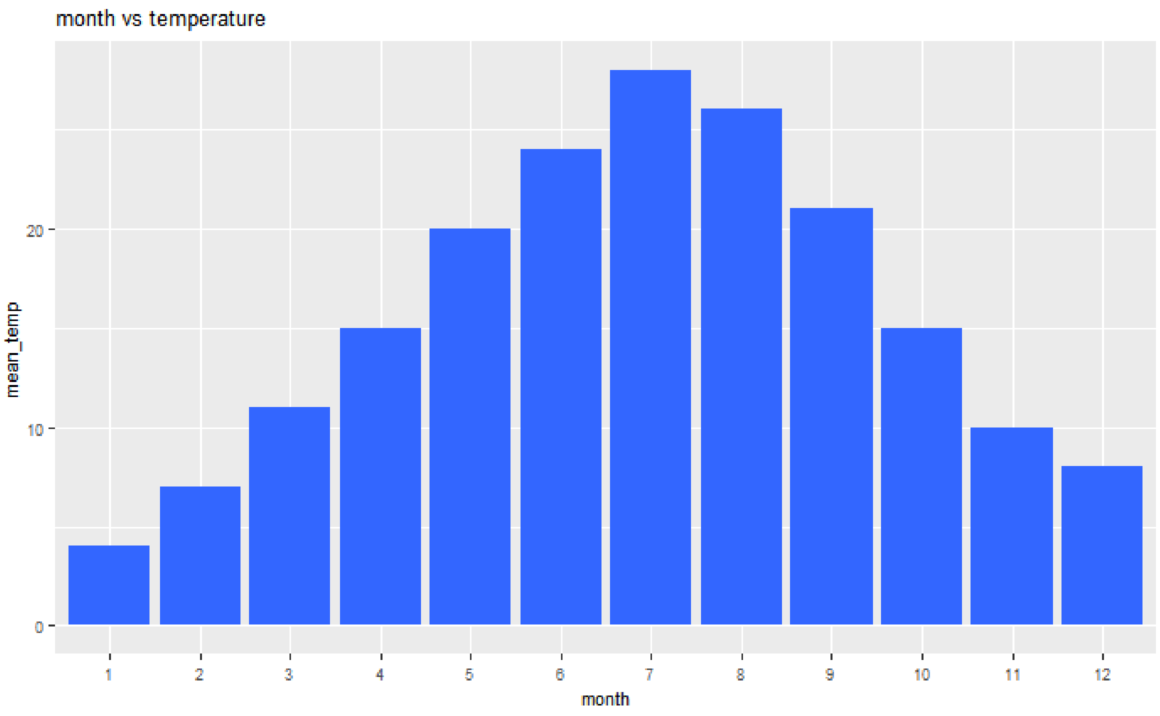
As mentioned previously, the bikes utilization is highly related to the temperature values (graph a). And the temperature is highly correlated with each respective month (graph e). On the extreme temperatures measured in the winter and summer, it can be noticed a natural decrease of the usage of the bikes, mainly detected on the casual cyclists during the months June through August, probably due to high summer temperatures in these months (see below month vs mean temperature graph) that can justify the casual cyclist to avoid the usage of the bikes. But there is no similar descent behavior on these months to the registered cyclists, once the employment is concentrated on the entrance and the exit regular working times, where the temperatures are normally lower (graph c).

```
ggplot(data = data2, aes(x = month, y = count, fill = type)) +
  geom_bar(stat = 'identity', position = position_dodge())+
  scale_fill_manual(values = c("#990099", "#FF3333")) +
  ggtitle("month vs count") +
  theme(text = element_text(size=8),
        axis.text.x = element_text(angle=0))
```



Month vs temperature

```
data4 <- data2 %>%
  select(month, temperature) %>%
  group_by(month) %>%
  summarise(mean_temp = round(mean(temperature)))

ggplot(data = data4, aes(x = month, y = mean_temp)) +
  geom_bar(stat = "identity", fill = "#FF6699") +
  ggtitle(“month vs temperature") +
  theme(text = element_text(size=8),
        axis.text.x = element_text(angle=0))
```



Part II - Prediction model

According to the analysis in the previous section, it was demonstrated based on the graphics that there is a reasonable difference among the **casual** and the **registered** cyclists, which for the proposal of this part II on the development of the predictive model, to distinct both of them for a better analytical solution where the peculiarities of each group does not influence the another one.

Targeting the simplification of this study, despite for the selecting the best model that better adjusts to the dataset is necessary to test many models, as were did, it will be only mentioned the best model that presented the best results with the lower root mean squared error (RMSE). The model that was chosen is the Extreme Gradient Boosting (XGBoost), which properly performed the model construction taking into account the tradeoff between willing results and computational needs.

1. Casual cyclists

1.1 Feature Selection

Through the testing of many machine models for the feature selection, all of them basically resumed similar results, where it was defined, which are the most and least important variables as shown below on the list of the importance of the variables. The result demonstrates that the holiday variable is not significative to predictive modeling. And besides, due to the high correlation between the temperature and atemperature variables, next calculated, the temperature variable was chosen to be removed to prevent the collinearity and for being a little less significative compared to the atemperature. As a reference to the importance of the variables, follows the demonstration by using the linear regression model and this methodology is valid for both, to casual and registered cyclists, by giving similar results.

1.2 Variables correlation

Correlation between the temperature and the atemperature variables.

```
cor(data$temperature, data$atemperature)
## [1] 0.9876721
```

1.3 Feature selection model

Constructing the model based on the linear regression algorithm for the determination of the most important variables.

```
library(caret)

trainIndex <- createDataPartition(data$casual,
                                  p = 0.7,
                                  list = FALSE,
                                  times = 1)
train_data <- data[trainIndex, ]
test_data <- data[-trainIndex, ]

control <- trainControl(savePredictions = TRUE)

model_lm <- train(log10(casual + 1) ~ season + year + workingday + month + hour +
holiday + weekday +
                  weathersit + atemperature +  temperature + humidity + windspeed,
                  data = train_data,
                  method = 'lm',
                  trControl = control
                  )

var_imp <- as.matrix(varImp(model_lm)$importance)
apply(var_imp, 2, sort)

##                                  Overall
## holidayYes_holiday             0.0000000
## `weathersitHeavy Rain/Snow`   0.5856824
## weekdaySat                     6.3749277
## month2                         8.3440858
## weathersitCloudy               9.3390970
## month7                        10.7360121
## month12                      11.0189204
## temperature                  14.3338940
## month8                       14.5333773
## seasonFall                   14.7344802
## month6                       15.5555542
## atemperature                 16.5017123
## hour7                        17.8057545
## weekdayMon                   17.8870968
## windspeed                    18.1255701
## month11                      18.9248888
## seasonWinter                 19.4652048
## month4                       22.1161728
## hour23                       22.4189548
## month9                       22.5077173
## seasonSummer                 24.2105851
## hour1                        25.2154517
## month5                       26.0657955
## weekdayThu                   26.8243072
## month10                      28.5656517
## weekdayTue                   30.1671879
## hour6                        30.8737872
## humidity                     32.0901420
## weekdayWed                   34.5713867
## month3                       40.4584366
## hour22                       40.9549333
## hour2                        42.4563236
## workingdayYes_workday        50.3697675
## hour21                       52.4758202
## hour8                        56.4770643
## year2012                     61.3792994
## hour20                       62.2761512
## `weathersitLight Rain/Snow`  67.2097087
## hour3                        67.8580033
## hour9                        68.9819392
## hour5                        74.8768907
## hour19                       76.7345243
## hour10                       81.8841268
## hour4                        82.4202365
## hour11                       89.6797475
## hour18                       89.8933836
## hour15                       91.1183775
## hour12                       92.5529811
## hour14                       92.7542905
## hour13                       92.9429975
## hour16                       94.8637583
## hour17                      100.0000000
```

1.4 Predicting model

Elaborating the predictive model to the casual cyclists using the Xgboost algorithm.

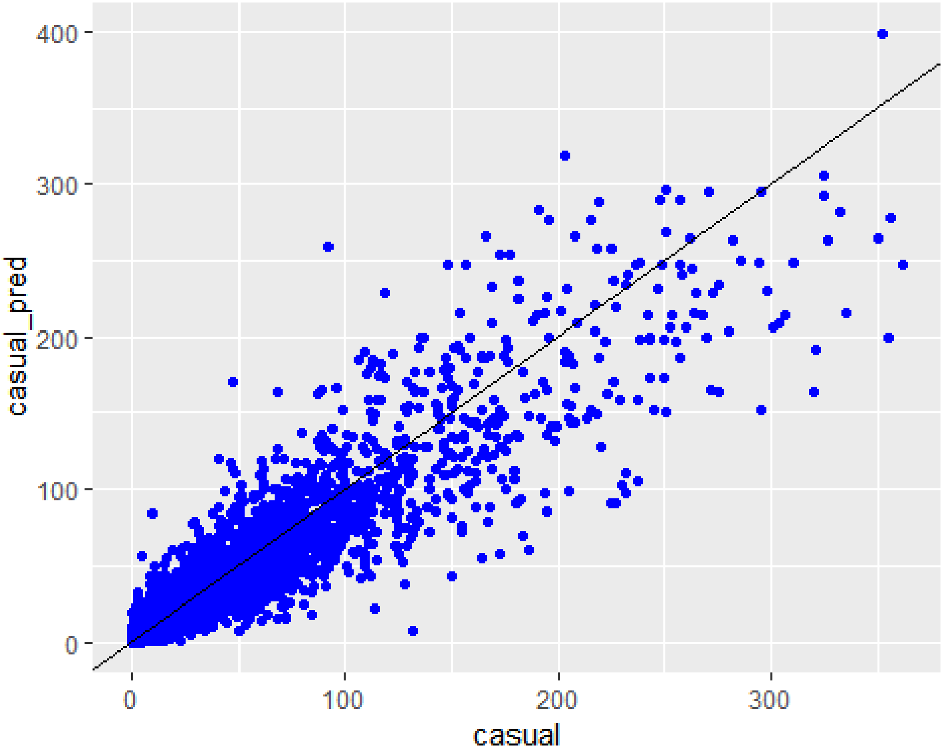
```
library(parallel)
library(iterators)
library(caret)
library(foreach)
library(doParallel)

cluster <- makeCluster(detectCores())
registerDoParallel(cluster)

control <- trainControl(savePredictions = TRUE, allowParallel = TRUE)
model_casual <- train(log10(casual + 1) ~ season + year + workingday + month + hour
+ weekday +
                       weathersit + atemperature + humidity + windspeed,
                       data = train_data,
                       method = 'xgbLinear',
                       trControl = control
                       )
stopCluster(cluster)
registerDoSEQ()

model_casual$results

##    lambda alpha nrounds eta      RMSE  Rsquared       MAE      RMSESD
## 27  1e-01 1e-01     150 0.3 0.2459743 0.8566191 0.1846733 0.003143472

##     RsquaredSD       MAESD
## 27 0.004209250 0.002135967

RMSE <- min(model_casual$results$RMSE)
RMSE
## [1] 0.2449309
```

1.5 Relative error calculation

Relative error calculation between the actual and the predictive values.

```
library(dplyr)
pred <- predict(model_casual, test_data)
test_data$casual_pred <- 10^(pred) - 1

test_data %>%
  select(casual, casual_pred) %>%
  summarise(error = round(mean((abs(sum(casual) -
sum(casual_pred))/sum(casual))*100),2))
```

```
##    error
## 1   5.71

library(ggplot2)
ggplot(test_data, aes(x = casual, y = casual_pred)) +
  geom_point(color = 'blue') +
  geom_abline()
```



2. Registered cyclists

2.1 Predicting model

Elaborating the predictive model to the registered cyclists using the Xgboost algorithm.

```
library(caret)

trainIndex <- createDataPartition(data$registered,
                                  p = 0.7,
                                  list = FALSE,
                                  times = 1)
```

```
train_data <- data[trainIndex, ]
test_data <- data[-trainIndex, ]

library(parallel)
library(iterators)
library(caret)
library(foreach)
library(doParallel)

cluster <- makeCluster(detectCores())
registerDoParallel(cluster)

control <- trainControl(savePredictions = TRUE, allowParallel = TRUE)
model_registered <- train(log10(registered + 1) ~ season + year + workingday +
month + hour + weekday +
                       weathersit + atemperature + humidity + windspeed,
                       data = train_data,
                       method = 'xgbLinear',
                       trControl = control
                       )
stopCluster(cluster)
registerDoSEQ()

model_registered$results

##    lambda alpha nrounds eta      RMSE  Rsquared       MAE      RMSESD
## 27  1e-01 1e-01     150 0.3 0.1675581 0.9239825 0.1146977 0.002583555

##     RsquaredSD       MAESD
## 27 0.002525471 0.001512846

RMSE <- min(model_registered$results$RMSE)
RMSE

## [1] 0.1675581
```

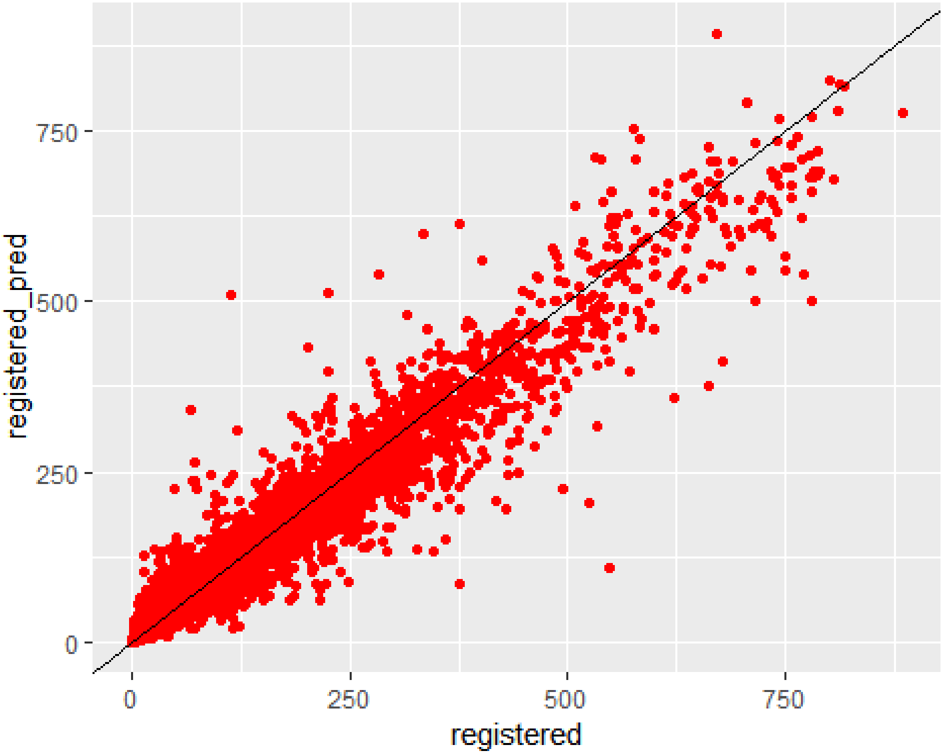
2.2 Relative error calculation

Relative error calculation between the actual and the predictive values.

```
library(dplyr)
pred <- predict(model_registered, test_data)
test_data$registered_pred <- 10^(pred) - 1

test_data %>%
  select(registered, registered_pred) %>%
  summarise(error = round(mean((abs(sum(registered) -
sum(registered_pred))/sum(registered))*100), 2))
```

```
##    error
## 1  3.84
```

```r
library(ggplot2)
ggplot(test_data, aes(x = registered, y = registered_pred)) +
  geom_point(color = 'red') +
  geom_abline()
```



Conclusion

By the data analysis, it was observed the main differences of the bikes employment between the casual and the registered users, the first one using occasionally and on the better climatic conditions, and the second group focusing as a vehicle to the working destination.

These characteristics can be confirmed when examining the RMSE and the relative error results and comparing the dispersion of the points on the casual and registered graphs. It can be seen respectively the lower error values and the bigger precision on graphs for the registered cyclists.

Thinking one step further for the discrepancies found among the predictive and actual values would be to analyze them, why they occurred, try to group them and discover if there is a noticeable pattern for the manifestation that is more prominent on the high values of casual and registered counts observable on the graphs.