Washington Bike Sharing Service

Maurizio Murino

16 February 2016

# 1. Introduction

*This project is about the Capital Bikeshare (CaBi) scheme (*[*http://www.capitalbikeshare.com/*](http://www.capitalbikeshare.com/)*) in Washington DC, launched in 2008 by the government of District of Columbia. In this project, we focus on the operations of CaBi in 2011 and 2012. In the Excel spreadsheet data.xlsx, it contains hourly rental data spanning these two years. Two types of rental data were recorded: casual renters and registered renters. A registered renter is a pre-registered member of the bike sharing program, while a casual renter pays for the use of a bike on the spot using a credit card. In addition to the time/date information (including holidays), the spreadsheet also records the weather condition in each hour (e.g., temperature, wind speed and humidity).*

# 2. Loading the data

We are going to load in R the whole Excel workbook. Since it is composed by three separated sheets, the package XLConnect comes in use. It allow to read the entire workbook, and then load the single worksheets. Please note that XLConnect requires Java to be pre-installed.

library(XLConnect)

## Loading required package: XLConnectJars  
## XLConnect 0.2-11 by Mirai Solutions GmbH [aut],  
## Martin Studer [cre],  
## The Apache Software Foundation [ctb, cph] (Apache POI, Apache Commons  
## Codec),  
## Stephen Colebourne [ctb, cph] (Joda-Time Java library)  
## http://www.mirai-solutions.com ,  
## http://miraisolutions.wordpress.com

work\_book <- loadWorkbook("raw data.xlsx")

Since the work\_book is quite small (1728 bytes) we are going to load in memory every sheet like a separate object.

object.size(work\_book)

## 1728 bytes

note <- readWorksheet(work\_book, sheet = "Note")  
all\_data <- readWorksheet(work\_book, sheet = "All data")  
testing <- readWorksheet(work\_book, sheet = "Testing")

# 3. Data description

The Excel workbook is composed by three sheets: "Note" is the coodebook, "All data" is the main dataset, "Testing" is a second and smaller dataset.

#summary(work\_book) [check!]

### 3.1 "Note"

It is the codebook. It describes how the content of the dataset is coded. Specifically, it reports how seasons (season) and weekdays (Weekday) are numerically coded in the datasets.

str(note)

## 'data.frame': 22 obs. of 5 variables:  
## $ Season : chr "1" "2" "3" "4" ...  
## $ Col2 : chr "Spring" "Summer" "Autumn" "Winter" ...  
## $ Col3 : logi NA NA NA NA NA NA ...  
## $ Weekday: num 1 2 3 4 5 6 7 NA NA NA ...  
## $ Col5 : chr "Monay" "Tuesday" "Wednesday" "Thursday" ...

### 3.2 "all\_data"

The dataset all\_data is composed by 17379 observation with 13 variables. datetime is profitably flanked by variables providing further information on the day of the week, the season and if we are talking about a working or a resting day.

str(all\_data)

## 'data.frame': 17379 obs. of 13 variables:  
## $ datetime : POSIXct, format: "2011-01-01 00:00:00" "2011-01-01 01:00:00" ...  
## $ Weekday : num 6 6 6 6 6 6 6 6 6 6 ...  
## $ season : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Holiday : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Workingday : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ weather : num 1 1 1 1 1 2 1 1 1 1 ...  
## $ Temp : num 9.84 9.02 9.02 9.84 9.84 ...  
## $ Feel.like.temp : num 14.4 13.6 13.6 14.4 14.4 ...  
## $ Humidity : num 81 80 80 75 75 75 80 86 75 76 ...  
## $ Windspeed : num 0 0 0 0 0 ...  
## $ Casual.renter : num 3 8 5 3 0 0 2 1 1 8 ...  
## $ Registered.renter: num 13 32 27 10 1 1 0 2 7 6 ...  
## $ Total.renter : num 16 40 32 13 1 1 2 3 8 14 ...

### 3.3 "testing"

testing is a smaller 688 observations dataset. On this data, we will try do develop a model that could forecast the daily rental demand for both casual and registered renters.

str(testing)

## 'data.frame': 688 obs. of 15 variables:  
## $ datetime : POSIXct, format: "2013-01-01 00:00:00" "2013-01-01 01:00:00" ...  
## $ Weekday : num 2 2 2 2 2 2 2 2 2 2 ...  
## $ season : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Holiday : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Workingday : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ weather : num 1 1 1 1 1 2 1 1 1 1 ...  
## $ Temp : num 13.1 12.8 11.8 11.5 10.9 ...  
## $ Feel.like.temp : num 17.4 16.2 16.2 15.9 14.9 ...  
## $ Humidity : num 71 70.7 77.3 79 79 ...  
## $ Windspeed : num 0 6 0 0 4 ...  
## $ Col11 : logi NA NA NA NA NA NA ...  
## $ Date : POSIXct, format: "2013-01-01" "2013-01-02" ...  
## $ Casual.renter : chr NA NA NA NA ...  
## $ Registered.renter: logi NA NA NA NA NA NA ...  
## $ Total.renter : logi NA NA NA NA NA NA ...

# 4. Part 1 (50%) - Data exploration

*Use data visualisation to explore the data as much as possible. The top management also welcomed any interesting insights that you might observe from the data. The top management also indicated that they were interested in knowing more about their customers, including the behavioural differences between casual renters and registered renters. Although you were only given limited data (not including the data by stations), the top management wondered if the data could reveal any information about the purposes of the bike rentals by casual renters and registered renters.*

### 4.1 Casual renters and regular renters

In first place we load the xts package, that will help us in manage the time series. Then we create a new xts object containing the number of bike renters along the time horizon considered, that is, from January 1st 2011 and December 31st 2012.

library(xts)

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

casual\_xts<- xts(all\_data$Casual.renter, all\_data$datetime)  
regular\_xts<- xts(all\_data$Regular.renter, all\_data$datetime)

We think that the weather could be could be a relevat incentive to the use of the bike. Hence, we are going to plot the total number of bike renter per day. The plot could be more readable by plotting their number by week, but we choose this time format in order to make the plots in line with the requests of the second part of the assignment.

**NOTE:** The legend box is too compact to allocate the weather conditions as reported in note:

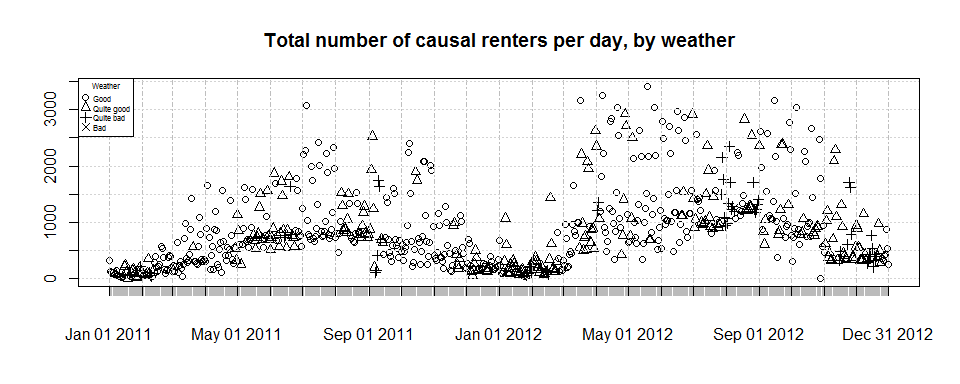
**Good:** Clear, Few clouds, Partly cloudy, Partly cloudy

**Quite good:** Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

**Quite bad:** Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

**Bad:** Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

plot(apply.daily(casual\_xts, sum), type = "p", pch = all\_data$weather, main = "Total number of causal renters per day, by weather")  
legend('topleft', legend=c("Good", "Quite good", "Quite bad","Bad"), pch=1:4, title = "Weather", cex = 0.5, pt.cex = 1)



## ADJUST TIME SERIES  
#plot(apply.daily(regular\_xts, sum), type = "p", pch = all\_data$weather, main = "Total number of regular renters per day, by weather")  
#legend('topleft', legend=c("Good", "Quite good", "Quite bad","Bad"), pch=1:4, title = "Weather", cex = 0.5, pt.cex = 1))

The bike renting follows, unsurprisingly, a seasonal trend. At the same time, the weather condition is likely to be a strong element concurring in determining the amount of rents.

Such an hypothesis is supported by a simple observation of the frequency of the rents according to the weather conditions. Even if the numbers between the registred users are greater, good weather conditions reveal a remakbly higher number of bike rents.

all\_data$weather2 <- all\_data$weather  
all\_data$weather2 <- factor(all\_data$weather2, levels = c(1,2,3,4), labels = c("good", "quite\_good", "quite\_bad", "bad"))  
  
# Total number of Casual renters by weather conditions  
  
aggregate(all\_data$Casual.renter,by=list(Category=all\_data$weather2), FUN=sum)

## Category x  
## 1 good 456624  
## 2 quite\_good 139107  
## 3 quite\_bad 35529  
## 4 bad 16

# Total number of Registered renters by weather conditions  
  
aggregate(all\_data$Registered.renter ,by=list(Category=all\_data$weather2), FUN=sum)

## Category x  
## 1 good 1862136  
## 2 quite\_good 670905  
## 3 quite\_bad 171894  
## 4 bad 243

At the same time, could be interesting an observation of the number of rents according to the day of the week.

all\_data$Weekday2 <- all\_data$Weekday  
all\_data$Weekday2 <- factor(all\_data$Weekday2, levels = c(1,2,3,4,5,6,7), labels = c("monday", "tuesday", "wednesday", "thursday", "friday", "saturday", "sunday"))  
  
# Total number of Casual renters by weekday  
  
aggregate(all\_data$Casual.renter,by=list(Category=all\_data$Weekday2), FUN=sum)

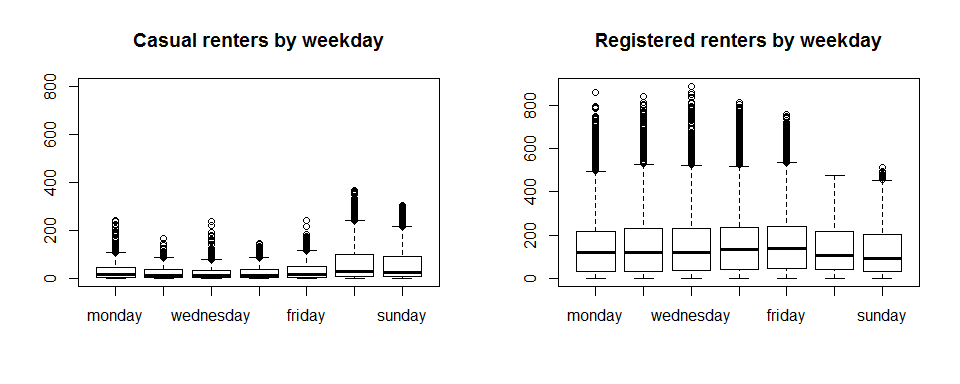
## Category x  
## 1 monday 71580  
## 2 tuesday 56901  
## 3 wednesday 56888  
## 4 thursday 60736  
## 5 friday 78610  
## 6 saturday 163371  
## 7 sunday 143190

# Total number of Registered renters by weekday  
  
aggregate(all\_data$Registered.renter,by=list(Category=all\_data$Weekday2), FUN=sum)

## Category x  
## 1 monday 392051  
## 2 tuesday 408325  
## 3 wednesday 412762  
## 4 thursday 428263  
## 5 friday 416272  
## 6 saturday 338152  
## 7 sunday 309353

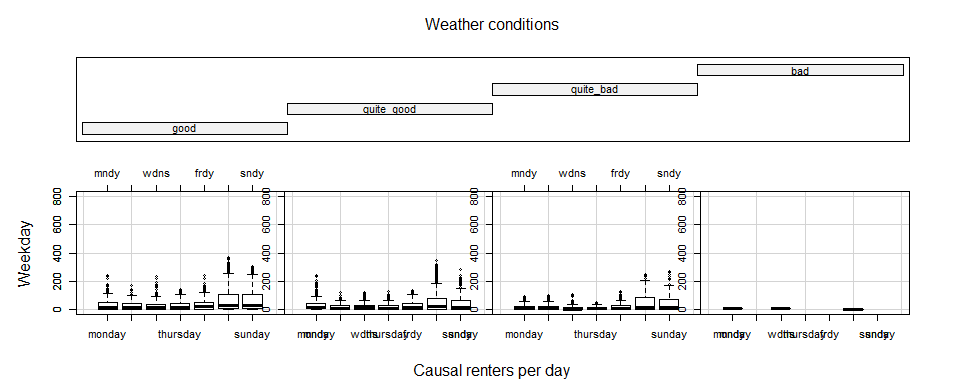
We can see that there is a relevant peak if rents during the week end. Even if there are many outliers, the result is indicative of the different behaviour between registered and casual users. Registerd users show a stronger activities during the workind days of the week. It is possible that the rent bike represents the daily vehicle for the home - office travel. Causal users show a sigtly superior use of the bike during the weekend. It suggest an use of the bike of touristic nature.

par(mfrow = c(1, 2))  
with(all\_data, plot(Weekday2, Casual.renter, main = "Casual renters by weekday", ylim=c(0, 800)))  
with(all\_data, plot(Weekday2, Registered.renter, main = "Registered renters by weekday"))



At the same time, the weather conditions still suggests us a major role in determining the rent of a bike by casual bikers.

coplot(Casual.renter ~ Weekday2 | weather2, data = all\_data, rows=1, xlab = c("Causal renters per day", "Weather conditions"), ylim=c(0, 800), ylab = "Weekday", xlim = c(0, 8), panel = function(x, y, ...){boxplot(y ~ x, add=TRUE)})



Similarly, the weather conditions affect the behaviour of the registred users.

coplot(Registered.renter ~ Weekday2 | weather2, data = all\_data, rows=1, xlab = c("Registered renters per day", "Weather conditions"), ylab = "Weekday", xlim = c(0, 8), panel = function(x, y, ...){boxplot(y ~ x, add=TRUE)})

