Case study: Seattle Fremont bridge

Javier Monedero

Introduction

The data set comes from an automated bicycle counter installed in this bridge in 2012. It has inductive sensors to count the daily or hourly ciclists who use it. The data is provided by the Seattle government and the direct link is: https://data.seattle.gov/Transportation/Fremont-Bridge-Hourly-Bicycle-Counts-by-Month-Octo/65db-xm6k

Data overview

First, I write some functions to get an idea of the strucutre of this data set:

```
##                    Date Fremont.Bridge.West.Sidewalk
## 1 10/03/2012 12:00:00 AM                            4
## 2 10/03/2012 01:00:00 AM                            4
## 3 10/03/2012 02:00:00 AM                            1
## 4 10/03/2012 03:00:00 AM                            2
## 5 10/03/2012 04:00:00 AM                            6
## 6 10/03/2012 05:00:00 AM                           21
##   Fremont.Bridge.East.Sidewalk
## 1                            9
## 2                            6
## 3                            1
## 4                            3
## 5                            1
## 6                           10
```

```
##                      Date       Fremont.Bridge.West.Sidewalk
##  03/08/2015 03:00:00 AM:    2   Min.   :  0.00
##  03/09/2014 03:00:00 AM:    2   1st Qu.:  7.00
##  03/10/2013 03:00:00 AM:    2   Median : 32.00
##  01/01/2013 01:00:00 AM:    1   Mean   : 56.38
##  01/01/2013 01:00:00 PM:    1   3rd Qu.: 75.00
##  01/01/2013 02:00:00 AM:    1   Max.   :698.00
##  (Other)               :25503   NA's   :7
##  Fremont.Bridge.East.Sidewalk
##  Min.   :  0.00
##  1st Qu.:  7.00
##  Median : 28.00
##  Mean   : 53.36
##  3rd Qu.: 66.00
##  Max.   :667.00
##  NA's   :7
```

Interesting facts can be determined with the last function. First, there are two sensors on the bridge, one on the west sidewalk and the other on the east sidewalk. Second, there seem to be low differences between the

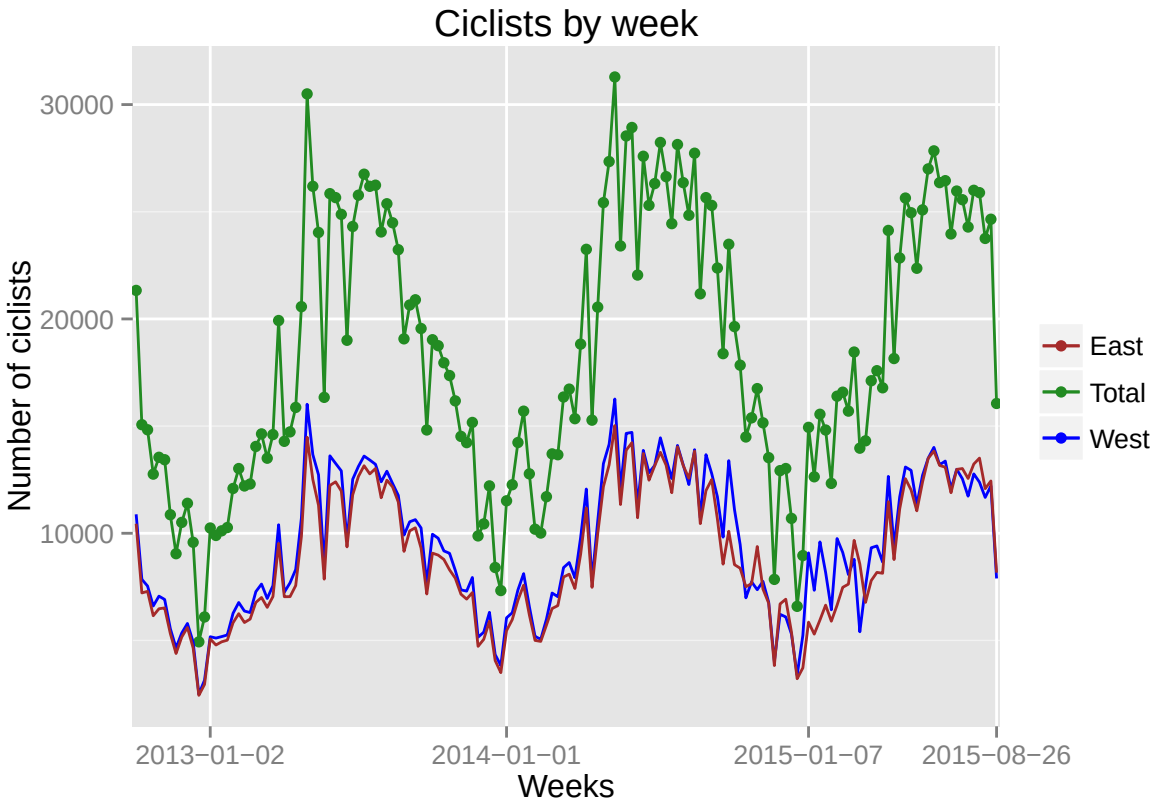
datum in both sides. As you can imagine, further analysis should be made to confirm it. Third, the data set contains NAs (undefined) values which could bring problems if I do not pay attention to it.

I am going to do a small data cleaning in order to make the programming assessment easy. I will rename the columns as West and East accodring to the sensors position, and I will add a new column called Total as the sum of every ciclist in every hour and day on both sides. Moreover, I will fill the NA values with 0.

```
##                      Date West East Total
## 1 10/03/2012 12:00:00 AM    4    9    13
## 2 10/03/2012 01:00:00 AM    4    6    10
## 3 10/03/2012 02:00:00 AM    1    1     2
## 4 10/03/2012 03:00:00 AM    2    3     5
## 5 10/03/2012 04:00:00 AM    6    1     7
## 6 10/03/2012 05:00:00 AM   21   10    31
```

Then, I make a general view of the data based on the ciclists weekly evolution along the years. This way, effects like the weekends, seasons, or whether problems like storms can be seem, and how they affected the number of ciclists in this part of Seattle.

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



The plot shows a strong seasonal variation as well as others factors influencing the number of ciclists who rided the bridge.

Exploratory data analysis

I will consider each day in the dataset as its own separate entity. It should be taken into account that for each day, there are 48 observations: two observations (east and west sidewalks sensors) by 24 hour-long periods. Thus, let's data tell statements.

Now, I want to determine whether the number of ciclists is statistically different on both sides of the bridge. For this purpose, I will conduct a t-test in which the null hypothesis will be there is no difference between the two groups means.

```
##
##  Welch Two Sample t-test
##
## data:  data$West and data$East
## t = 4.6262, df = 50827, p-value = 3.733e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.743420 4.306674
## sample estimates:
## mean of x mean of y
##  56.36787  53.34282
```

As can be seen, the p-value close to zero (less than 0.05) indicates there is a statistical difference. Thus, I reject the null, meaning there is a statistical difference between the two populations means. In other words,

the number of ciclists who used the west sidewalk statistically differed from the number of ciclists who rided the east one.

Going one step further, comparison in mean values points out west sidewalk is more used by west sidewalk ciclists.

```
## Mean West ciclists:   9460.901
```

```
## Mean East ciclists:   8953.171
```

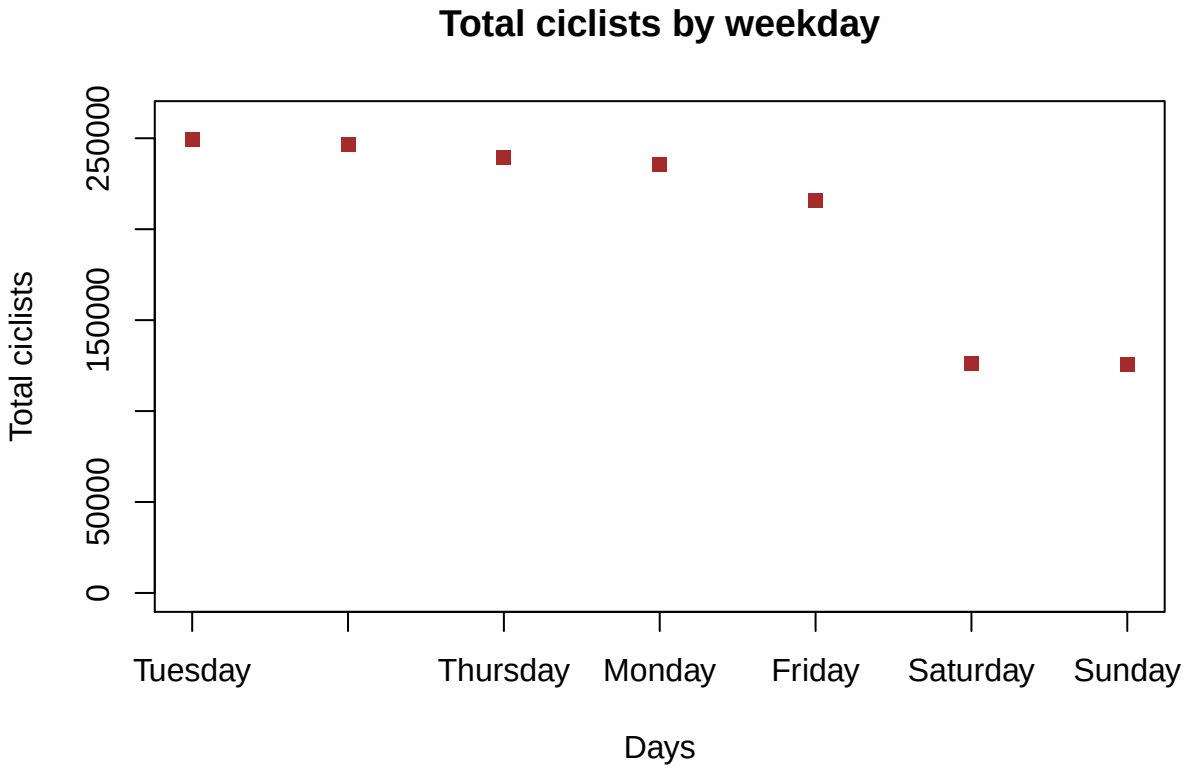
Continuing with the data analyses, I am going to determine if there are statistical differences between the days of the weeks under study.

```
##   Tuesday Wednesday  Thursday    Monday    Friday  Saturday    Sunday
##    249066    246782    239514    235628    215618    125940    125572
```

So, the number of ciclists on the weekends (Saturday + Sunday) highly decreases compared to the workdays. This indicates that most rides are due to trips to work.

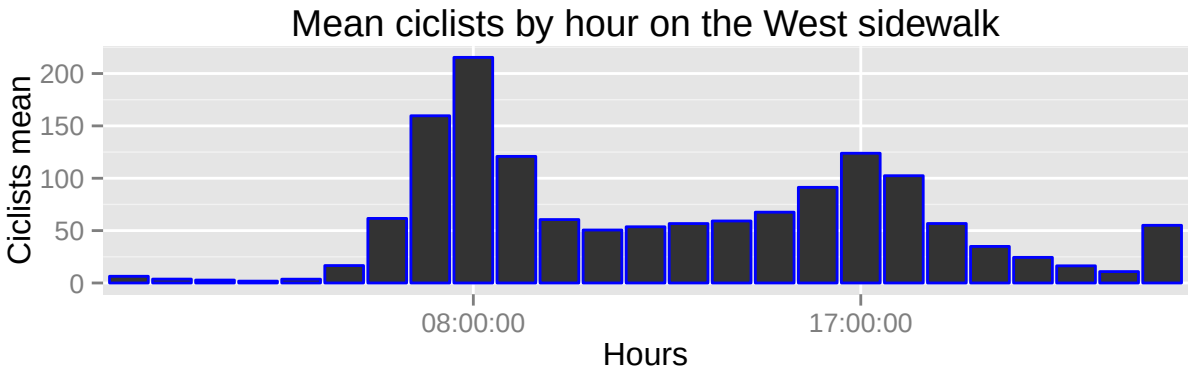
Visualizing data

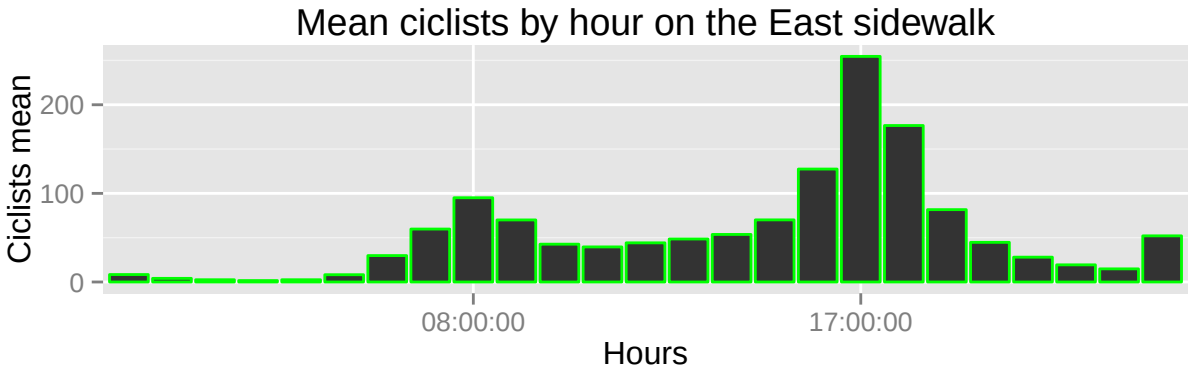
The previous table can se be seen in the following plot:



Here, rides on Friday decrease meaning there could be less ciclists on the road due to the fact that weekend is coming and they could have some kind of social advantage job. Central workdays in weeks show slightly higher trips in bicycle compared to Fridays.

```
## Loading required package: lattice
```





Plots show two time periods in which the ciclists attendance were higher. On the one hand, hours from 07:00:00 to 09:00:00 indicate the daily starting working hours and thus, the first peak corresponds to the job trips to go to work. On the other hand, the second peak around 17:00:00 refers to the time period in which workers exit their jobs. It also seems more ciclists used the west sidewalk to go to work and the east sidewalk to return home. Interestingly, there is a non-expected peak around 00:00 on both sides of the bridge.