Semiparametric Regression

Project - Bike Rentals

Gogołowicz Diana, Makowski Michał

15 June 2017

Spis treści

Introduction

This report was made as a summary of the grading project on *Semiparametric Regression*, the course conducted by Prof. Jarosław Harężlak at University of Wrocław. On the following pages we will focus on the data we choose, try to find some obvius and less obvious dependencies beetwen them and, at the end, we will build some mathematical models to *squeeze* data even more and generalize our knowledge about them.

The course covered several types of regression models, from simple lienear model, through generalized linear models, ending at generalized additive model, smoothing splines.

Goal

The goal of the final project was not clearly specifed, we have choose data on our own and then try to analyze them as best as we could. That is interesting approach, beacuse every group came up with something different. Our efforts was summed up in the presentation and in this report.

Data overview

Discription

We choose the dataset which had been merged from three datasources. It contains information about public bike rentals (from company called Capital Bikes) in Washington, DC. Data was recorded over the period of two years (2011-2012). There are two datasets, first which countain hourly records, second contaning cumulated daily data.

The raw data contained only information from rental system, but the information about weather and holidays was added. All data are available at following sources:

- Original Source: (http://capitalbikeshare.com/system-data)
- Weather Information: (http://www.freemeteo.com)
- Holiday Schedule: (\textit{}http://dchr.dc.gov/page/holiday-schedule)

The raw merged data have following structure:

Data dictionary:

- instant: record index
- dteday : date
- season : season (1:spring, 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
- 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
  - 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 divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- 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

We have several information about date&time, rentals and weather. All of them do not need any more clarification.

Preparation

As we can see above, there were some data preparation required: all data which are discrete characteristics were converted to factor variables. The temperatures, wind speed and humidity were converted to their real vaules (we "unnormalize" them, it will be easier to interpret them). We decide not to factorize *month* variable, we suppose there might be trend visible over the whole year. We added *nextworking* variable, it descibes whether following day is working or not (we would like to invistigate rentiers behavior at the evenings/nights before day off). As there were only 4 "Very bad" days, we merge them with "Bad" category.

Below short summary of data is presented. It gives some overview how they look like, we will do this only for cumulated daily data.

```
      dateday       season      year         month       holiday       weekday    workingday
 2011-01-01:  1   Winter:181   2011:365   Min.   : 1.00   No :710   Sunday   :105   No :231
 2011-01-02:  1   Spring:184   2012:366   1st Qu.: 4.00   Yes: 21   Monday   :105   Yes:500
 2011-01-03:  1   Summer:188              Median : 7.00             Tuesday  :104
 2011-01-04:  1   Fall  :178              Mean   : 6.52             Wednesday:104
 2011-01-05:  1                           3rd Qu.:10.00             Thursday :104
 2011-01-06:  1                           Max.   :12.00             Friday   :104
 (Other)   :725                                                     Saturday :105
    weather          temp            atemp           humidity       windspeed          casual
 Good    :463   Min.   : 2.424   Min.   : 3.953   Min.   : 0.00   Min.   : 1.500   Min.   :   2.0
 Mediocre:247   1st Qu.:13.820   1st Qu.:16.892   1st Qu.:52.00   1st Qu.: 9.042   1st Qu.: 315.5
 Bad     : 21   Median :20.432   Median :24.337   Median :62.67   Median :12.125   Median : 713.0
                Mean   :20.311   Mean   :23.718   Mean   :62.79   Mean   :12.763   Mean   : 848.2
                3rd Qu.:26.872   3rd Qu.:30.430   3rd Qu.:73.02   3rd Qu.:15.625   3rd Qu.:1096.0
                Max.   :35.328   Max.   :42.045   Max.   :97.25   Max.   :34.000   Max.   :3410.0

  registered       count      nextworking
 Min.   :  20   Min.   :  22   No :231
 1st Qu.:2497   1st Qu.:3152   Yes:500
 Median :3662   Median :4548
 Mean   :3656   Mean   :4504
 3rd Qu.:4776   3rd Qu.:5956
 Max.   :6946   Max.   :8714
```

Data exploration

We divide available information in 3 categories for better paper orginization:

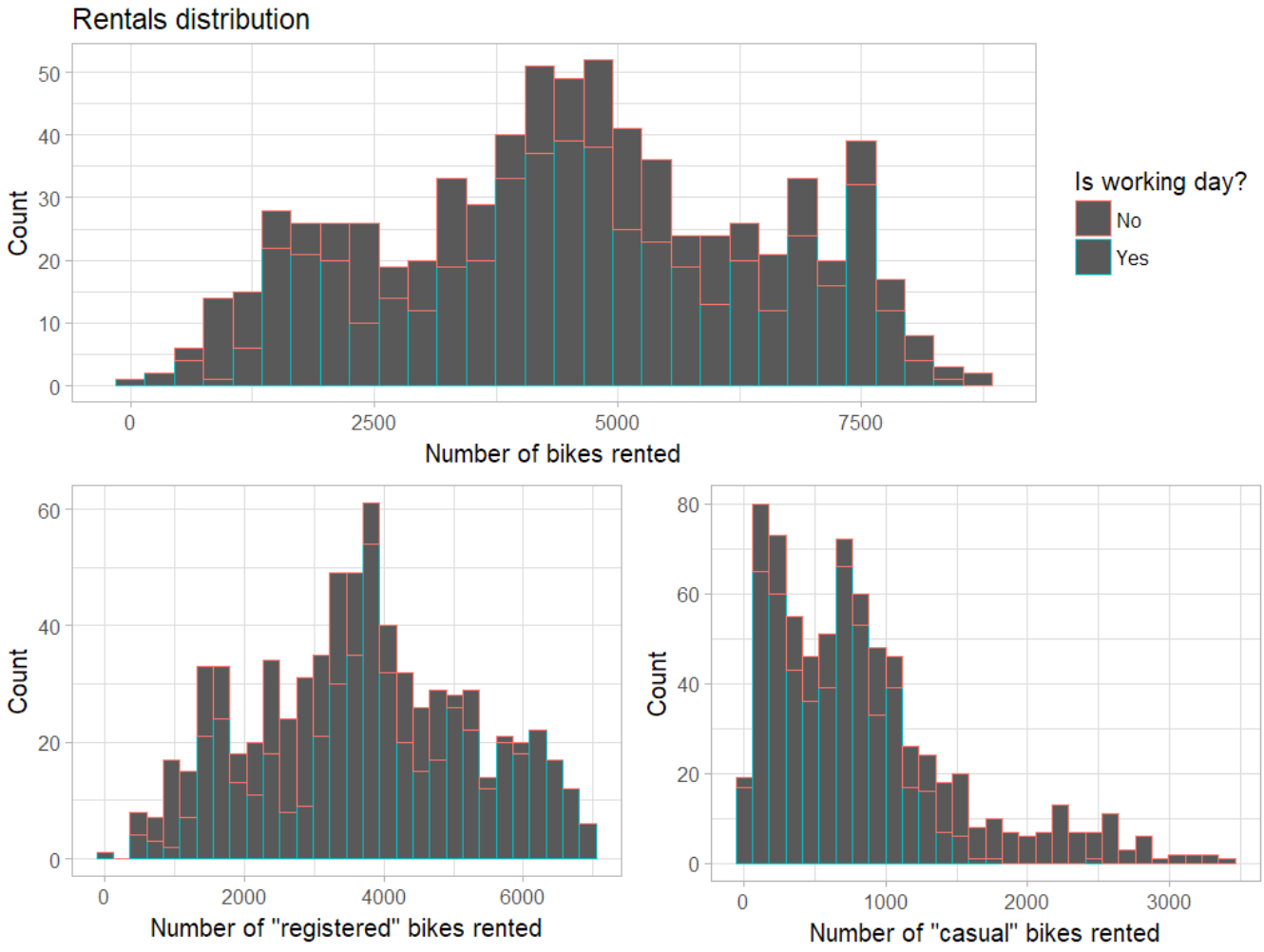
- Rental data
  count, casual, registered
- Date&Time
  date, year, month, day, hour, season, isworkingday, isweekend
- Weather
  weather, temp, humidity, windspeed

Of course some of the information will "jump" into different categories.

Rentals

Daily

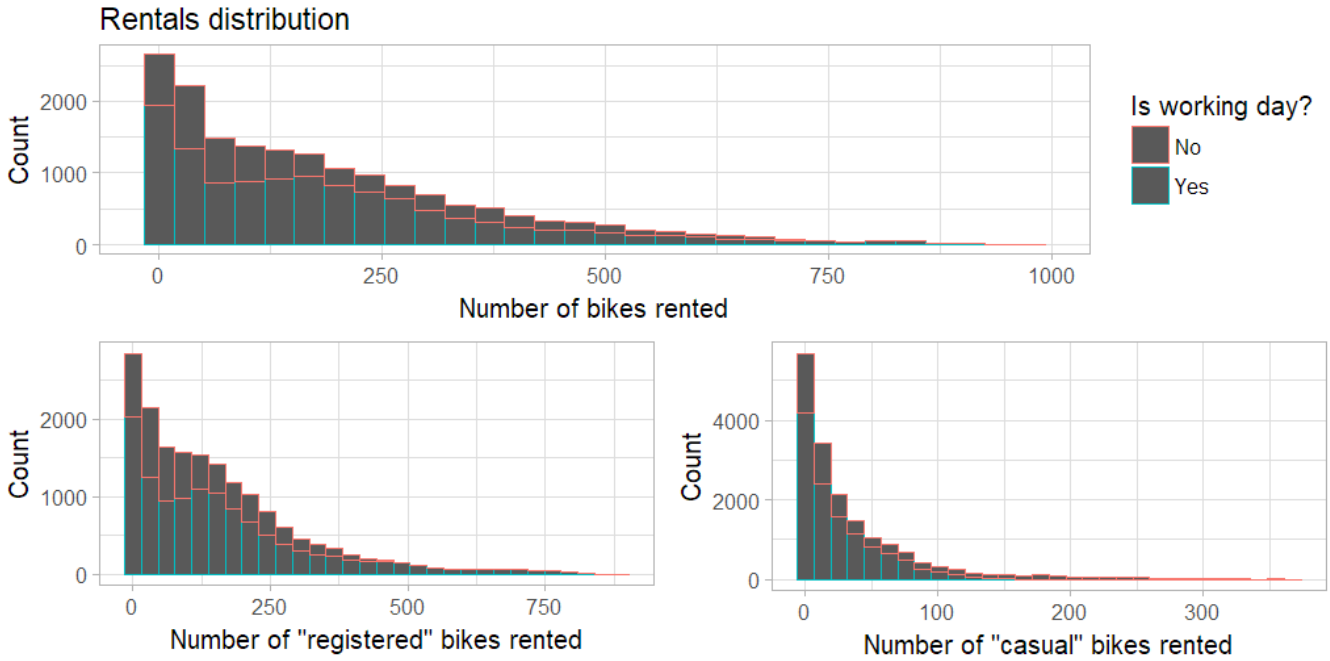
We begin with raw analysis of daily rentals number, let's plot some histograms.



We can see there is significant diffrence is distribution of each group. Registered user seems to be more "bounded" to city bikes, the situation where none of them rent the bike happens very rarely. The distribution of casuals is much more skewed toward zero. We can observe important dependency: highest number of bikes rented for each group are achieved on diffrent days: for casuals it is on holidays, for registered it is on working days. We will dive into this later.

Hourly

Now it is time for hourly rentals.

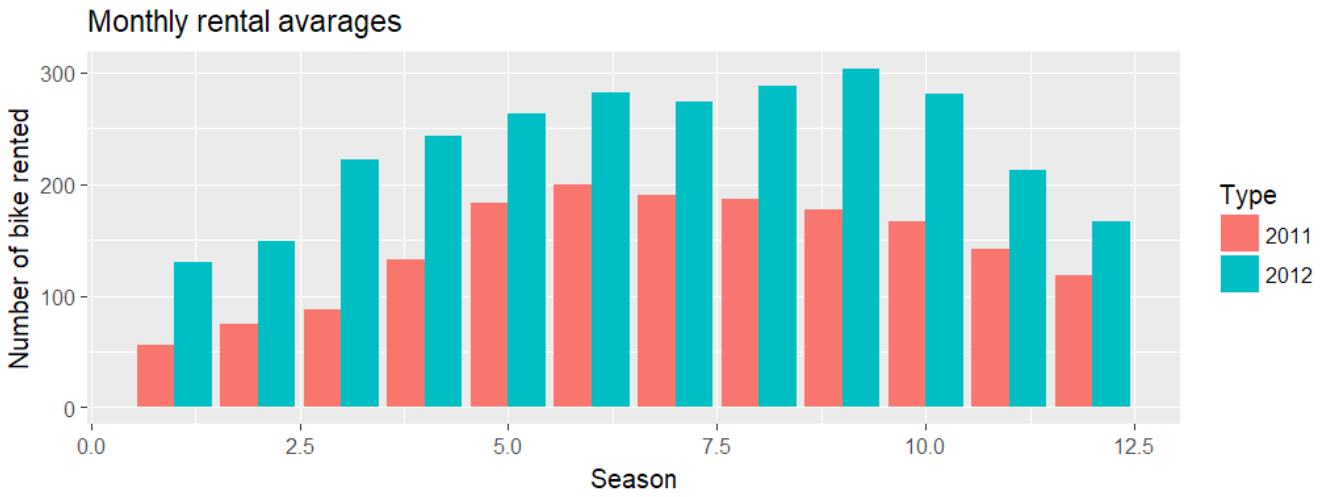


The histograms are completely diffrent comparing them to the last ones. but we could observe the same dependecy (for working days and holidays), as for hourly rentals. That shows it might be worth deeper analysis.

Bins for each histograms are the same.

Monthly

Let's plot monthly avareges.



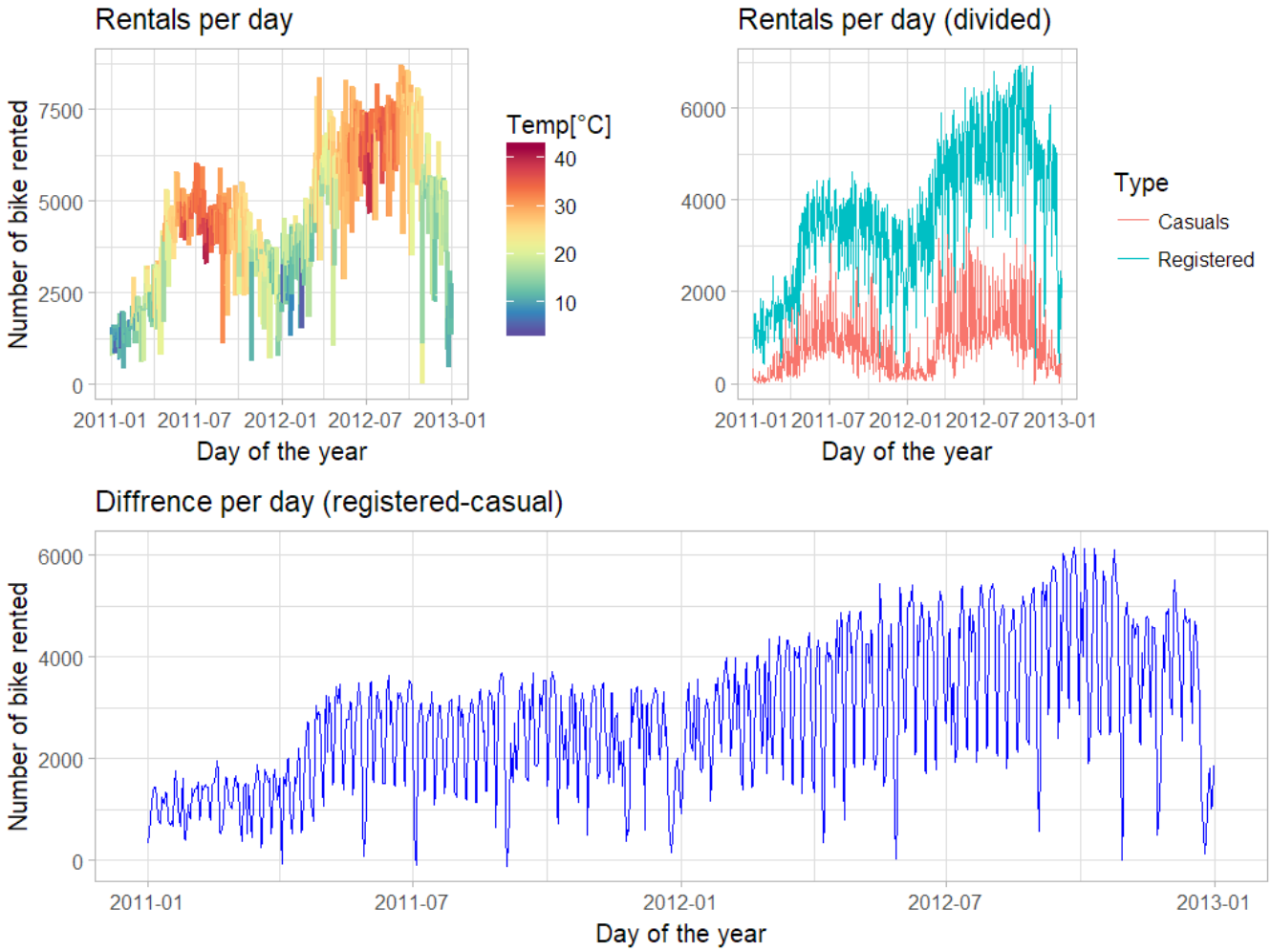
We can see visible change of rentals over a year. For each month number of bikes rented increased by tens of percents by year. If we had more data, we would invistigate this trend better. Nonetheless, year seems to be important factor in modeling bikes popularity.

Date&Time + Weather

On the following pages we will try to show dependecies beetween number of bikes rented per hour/day and weather.

Daily

Let's add information about **Date&Time** and **Weather** to raw rental data. On the following plots we will present how daily rentals distribute over two years.



First plot (upper left) shows how many people were using city bikes each day, of course seasons are well visible there, people prefer using publiic bikes during spring and summer. We also could see about 50% increase in overal popularity of bikes (from ~4500 to ~7000).

From second, upper right plot, which divide rentals into this done by registered and casual users, who we could deduce that over time bigger and bigger fraction of users are those registered in the system. Among registered users the increase of rentals over two years is much more visible.

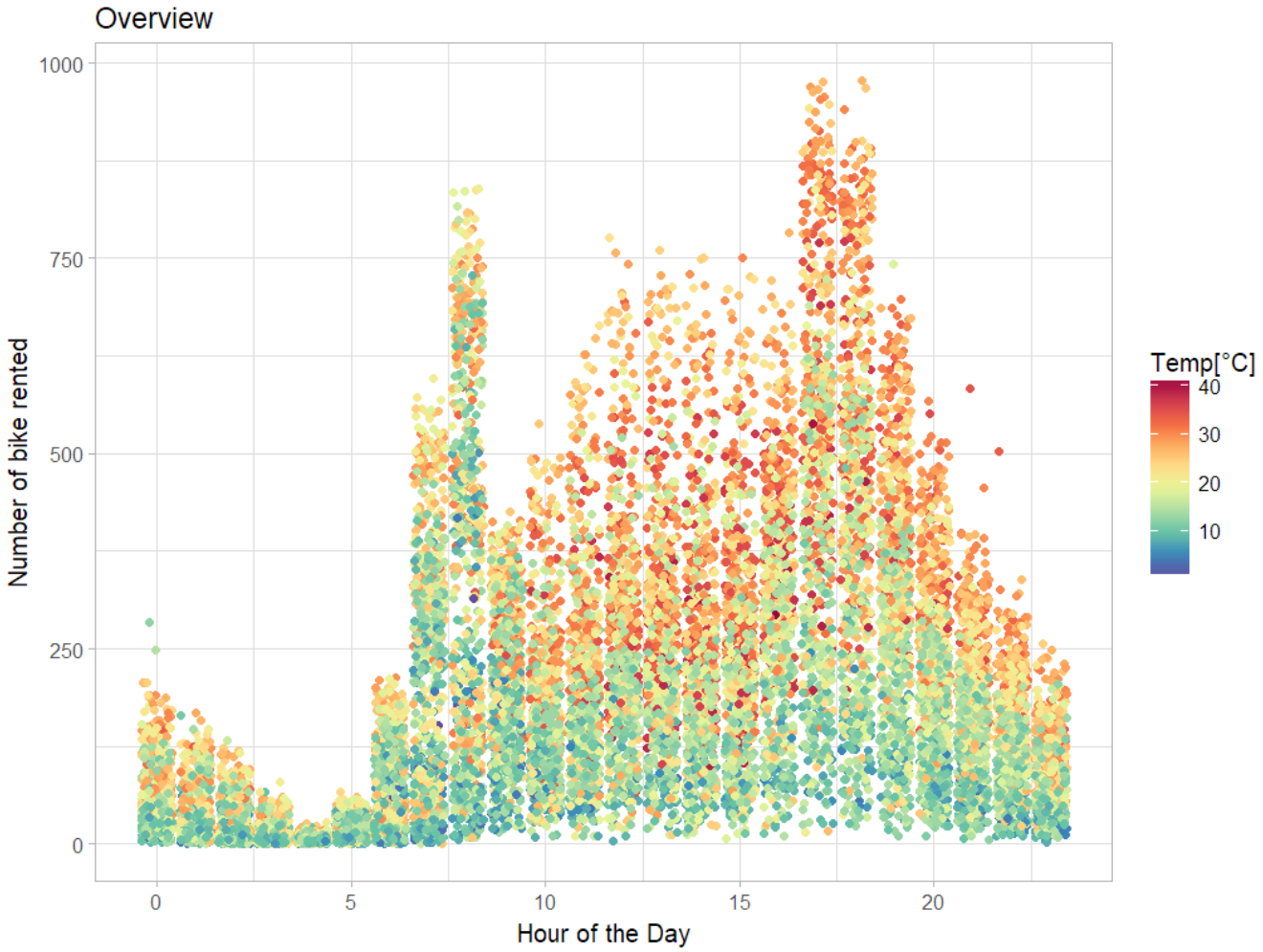
Last, lower, plot, proof this observation - over time the diffrence beetween registered and casuals increse. Unfortunetelly, we do not have data to invistigate how many casuals began registered.

There is strange jump at the and of 2012, we do not have any information what could cause it. We tried to google it, nothing informative came out. It might be preferable rental conditions for those who decide to sign up for the Capital Bikes program.

Hourly

Temperature

Now we plot how number of rented bikes distributes over a day, taking into account temperature, wind speed and humidity.

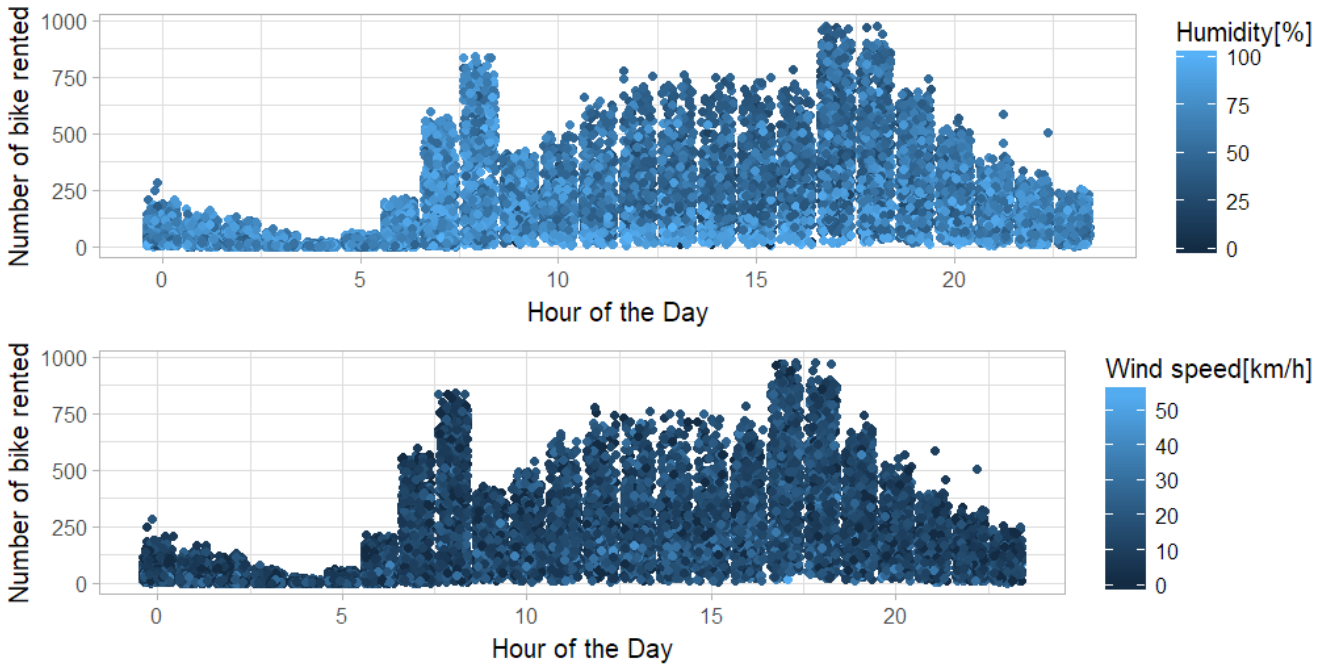


Firstly we could observe that number of bikes rented is very high just before 8AM and after 4PM. We will focus on this later, but it suggest that people tend to use public bikes to go to/from work. Moreover people tend to use bikes more often after work, that might suggest they do no want to use them early in the morning, when they are in a hurry. After work, they are more relaxed and use them more gladly (this statement might be too brave!).

Secondly, what is obvious, people prefer using Capital Bikes when the temperature is higher, nobody likes to cycle in freezing cold. :)

Humidity&Wind

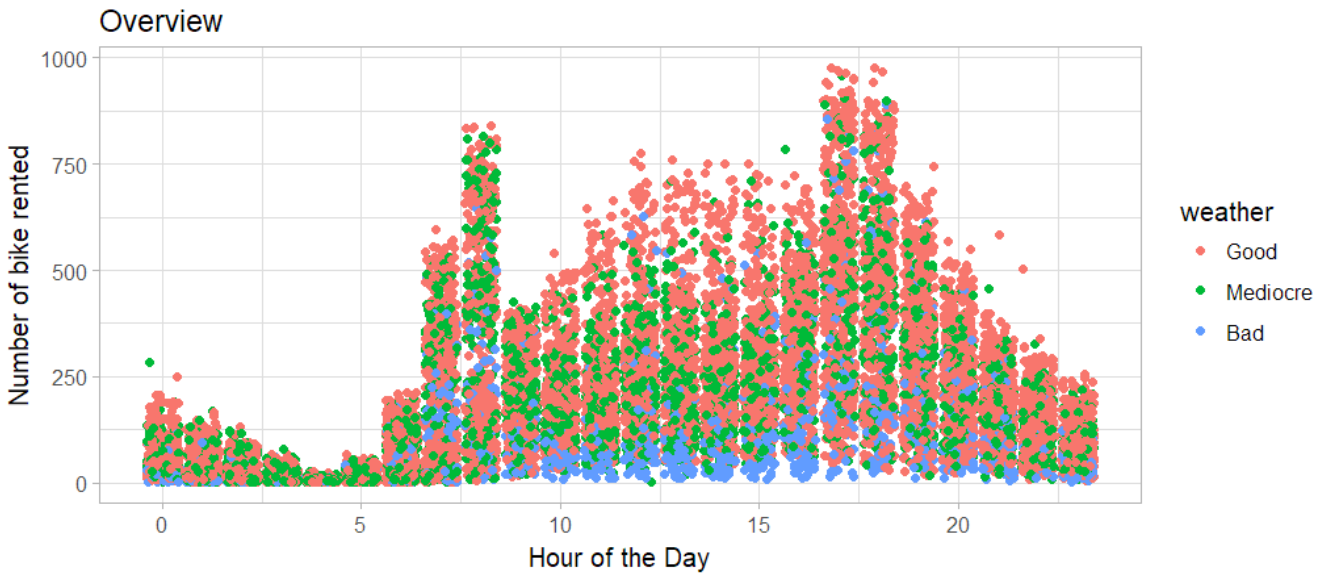
Let's is there any influence of wind and humidity on rentals.



There is no visible influence, humidity and wind speed do not have very strong impact on people willingess to cycle. Or we just could not find it. :) There are some little patterns, but we will not focus on them.

Weather

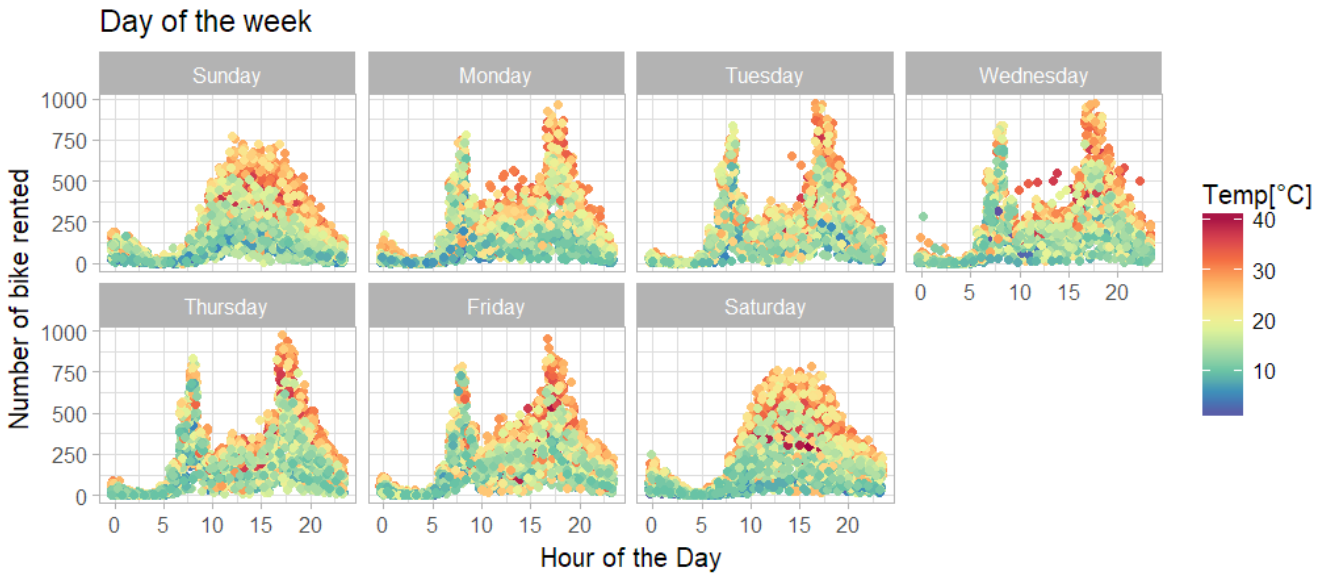
Last but not least, we plot rentals number taking *weather* variable into account.



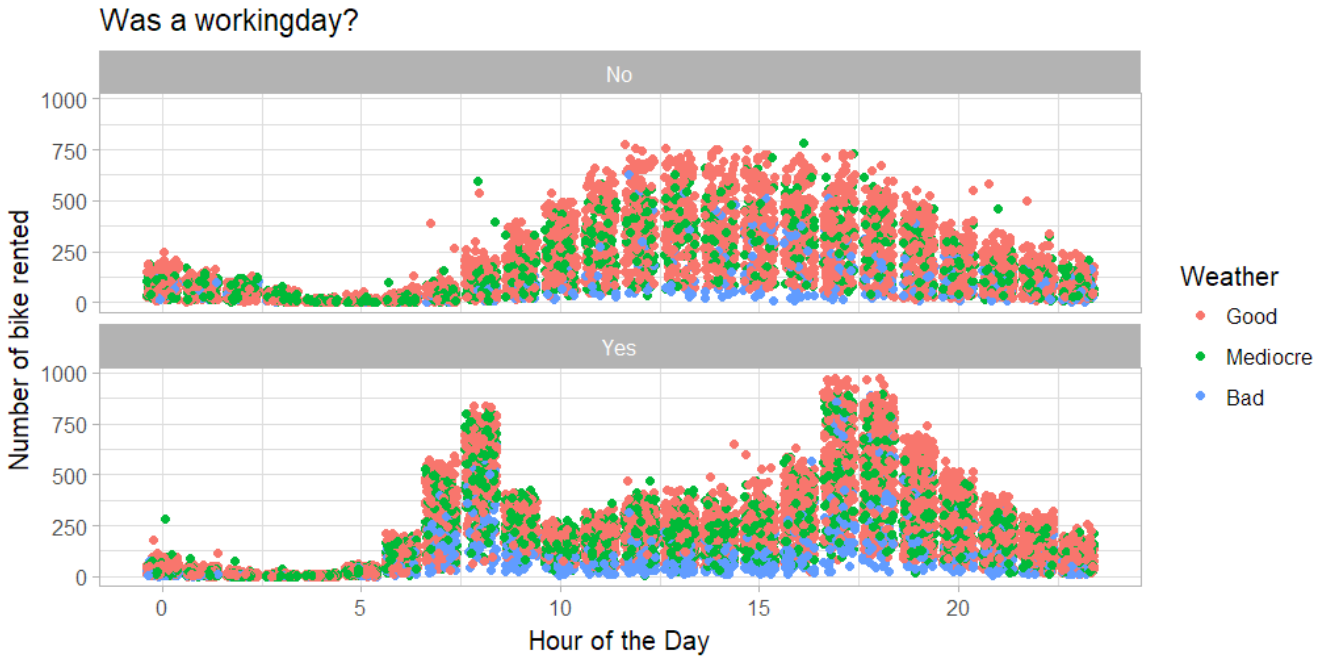
The dependency is visible, but it is not so strong as in temperature plot. People rather do not want to ride during bad weather, but mediocre and good seems to be at the same level of popularity.

Weekday

The plot which shows popularity of bikes during each day of the week is presented below.

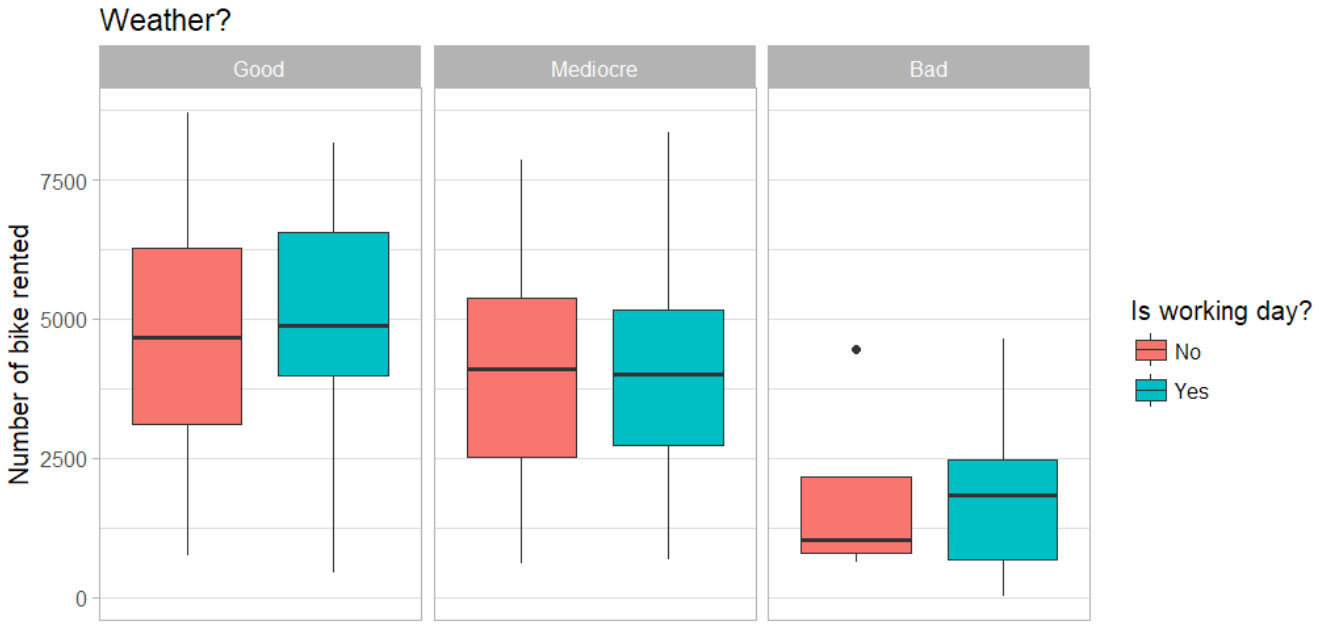


Thesis, that people seems to use city bikes to ride to work it even stronger now. The distribution of bikes popularity on weekdays is completely diffrent from this on weekends. Let's go further, crete only two plots, for working days and holidays.



Plots above somehow proof our concept, on holidays people use public days in completely diffrent way, there are no peaks before and after work (of course, people are not going there!). What is also visible, on working days weather has much bigger impact on people decision about using Capital Bikes. On working days the "sandwich" structure is much more visible, on holidays citizens rather do not use bikes in bad weather, but mediocre and good levels seems to be comparable popular.

We invistigate obserwations from last paragraph a little deeper.



The concept of lack of sandwich structure during holidays seems to be fault. There is visible shift beetween each weather level, no matter at what type of the day we are focusing. The diffrence might be smaller for holidays, but it is stille there. We could perform statistical tests to check it more deeply.

Next working

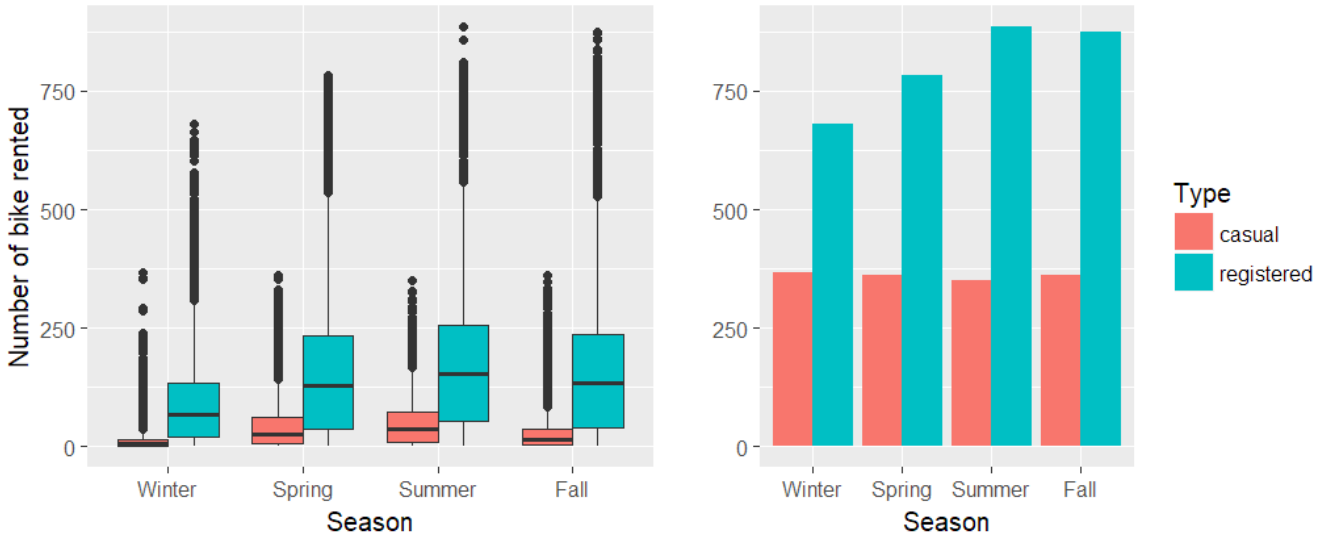
We invistigate whether the next working day have impact of citizens willingness of using public bikes.

There is little diffrence, but we think it is not important. It would rather do not have any impact on citizens decision.

Casual vs Registered

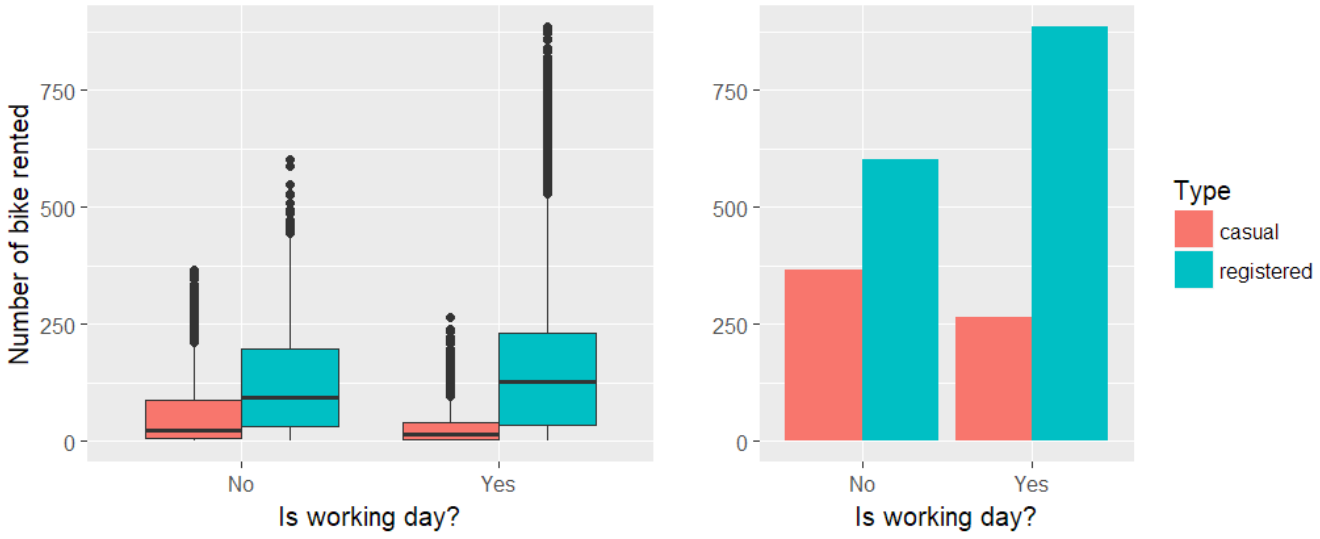
In this part we will focus on partition of the data which might be useful in bussinnes approach - on **casual** and **registered** user. We will plot diffrent characteristics for those two groups. On following pages several plots will be presented, each followed by brief summary. Plots on the right show average number of bike rented in each subgroup.

Seasons



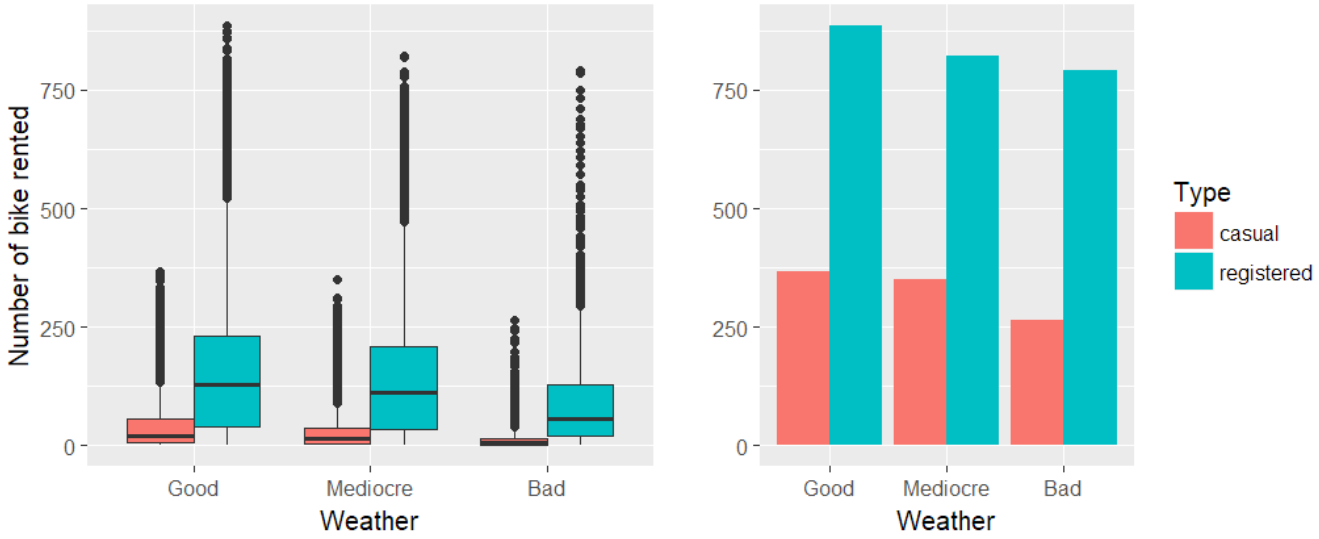
Casual riders seems to choose Capital Bikes similiary often each season, while registered ones has diffrent preferations.

Working days



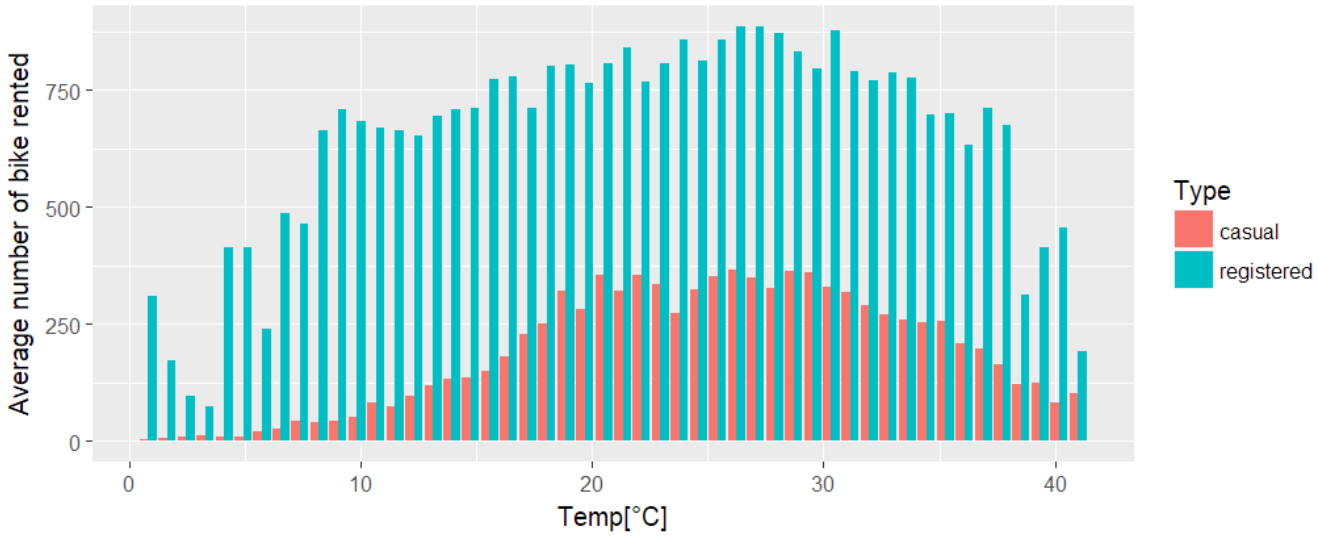
That is very interesting plot. We could observe that casual riders prefer to use bikes on holidays, register ones use the more on workingdays. That could be very useful information for marketing department. The diffrence beetween casuals and registered is much smaller on holidays.

Weather



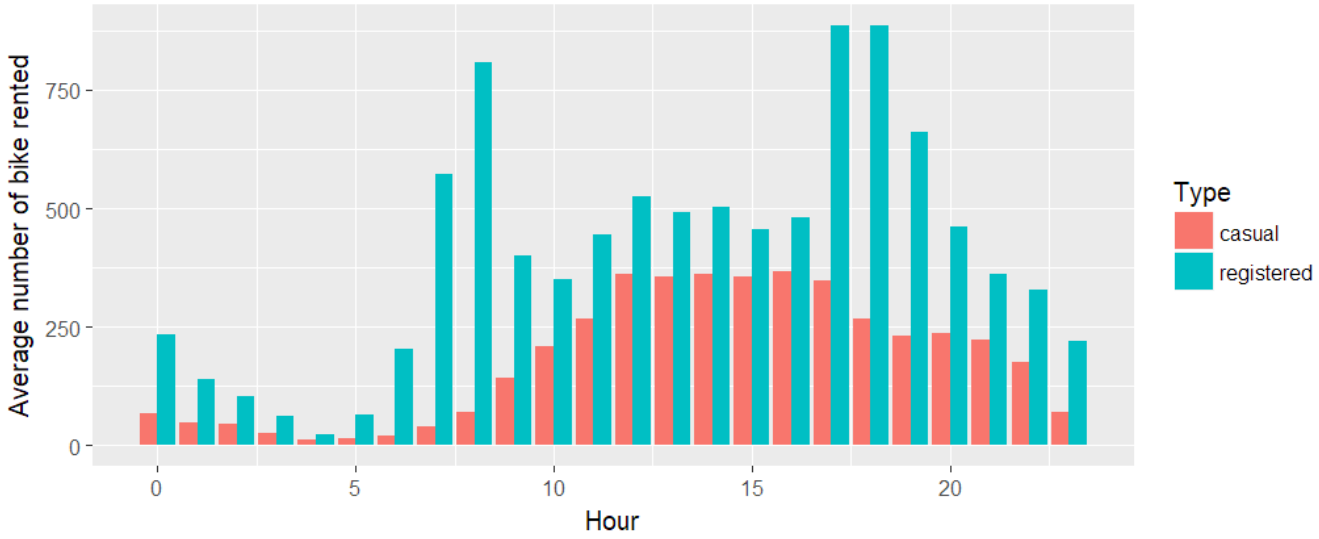
We tried to state that on holidays poeple do not care whether weather is good or mediocre. Then, boxplots rather bring this thesis down. If we compare casuals and registered then that might be true. There should be some statistical testing applied to verify this thesis, but that is not subject of this report.

Temperature



We could observe that casuals avoid riding in lower temperatures, for registered ones the distribution of mean is more uniform.

Hour



The pattern similiar to this obtained by dividing data into holidays and working days is visible.

Conclusion

Capital Bikes are very popular among casual riders during holidays. On the other hand, registered ones tend to use bikes on the working days. In following section we will build some models to prove this consept.

Models

Summary

There are many relationships between number of rentals and environment characteristics. Most important factors are: weather, year, and working day.

Models build above could be used for predicting future interest in public bikes, best time for servicing bikes and, if additional data will be provided, managing bike transportation. Also, different interest among different types of users was shown.