*Mini Project – Bike Rentals*

*By Sakshi Jain*

*Table of Contents*

1. *Project Objective*
2. *Exploratory Data Analysis – Step by step approach*
3. *Correlation Analysis*
4. *Regression Analysis*
5. *Conclusion*

*Project Objective*

* *The objective of the report is to explore the data set “Bike Rentals and Weather” in R and generate insights about the data set. This exploration report will consists of the following:*
* *Importing the dataset in R*
* *Understanding the structure of dataset*
* *Graphical exploration*
* *Descriptive statistics*
* *Advanced statistics*
* *Insights from the dataset*

*Exploratory Data Analysis*

*Bike Rentals & Weather ==================================================================================*

* *The data came from a two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA. The Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto, aggregated the data on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information that were extracted from http://www.freemeteo.com*
* *The dataset contains of 731 observations (rows) of 16 original variables + 5 new variables (columns).*

*Original columns:*

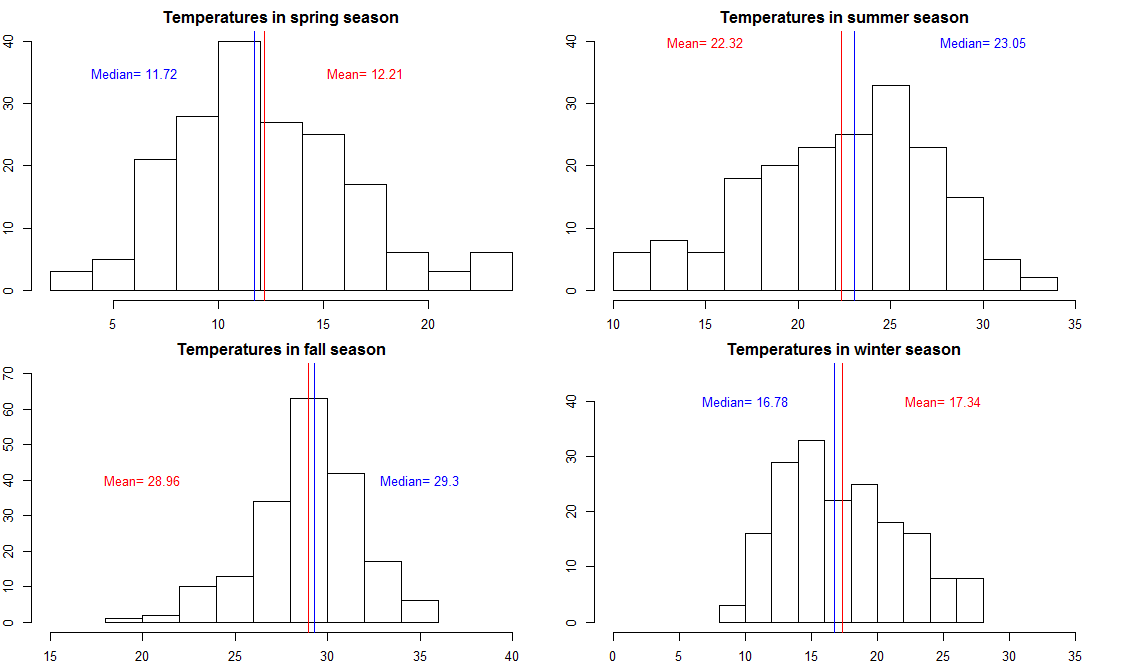
* *instant: record index*
* *dteday: date*
* *season : season (1:springer, 2:summer, 3:fall, 4:winter)*
* *yr : year (0: 2011, 1:2012)*
* *mnth : month ( 1 to 12)*
* *holiday : weather day is holiday or not (extracted from*[*http://dchr.dc.gov/page/holiday-schedule*](http://dchr.dc.gov/page/holiday-schedule)*)*
* *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 clouds  
  -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*

*New columns:*

* *weather: recoded weathersituations in nice, cloudy, wet and lousy*
* *raw.temp: raw values for temperature in Celsius*
* *raw.atemp: raw values for feeling temperature in Celsius*
* *raw.hum: raw values for humidity in %*
* *raw.windspeed: raw values for windspeed in km/h*
* *How do the temperatures change across the seasons? What are the mean and median temperatures?*

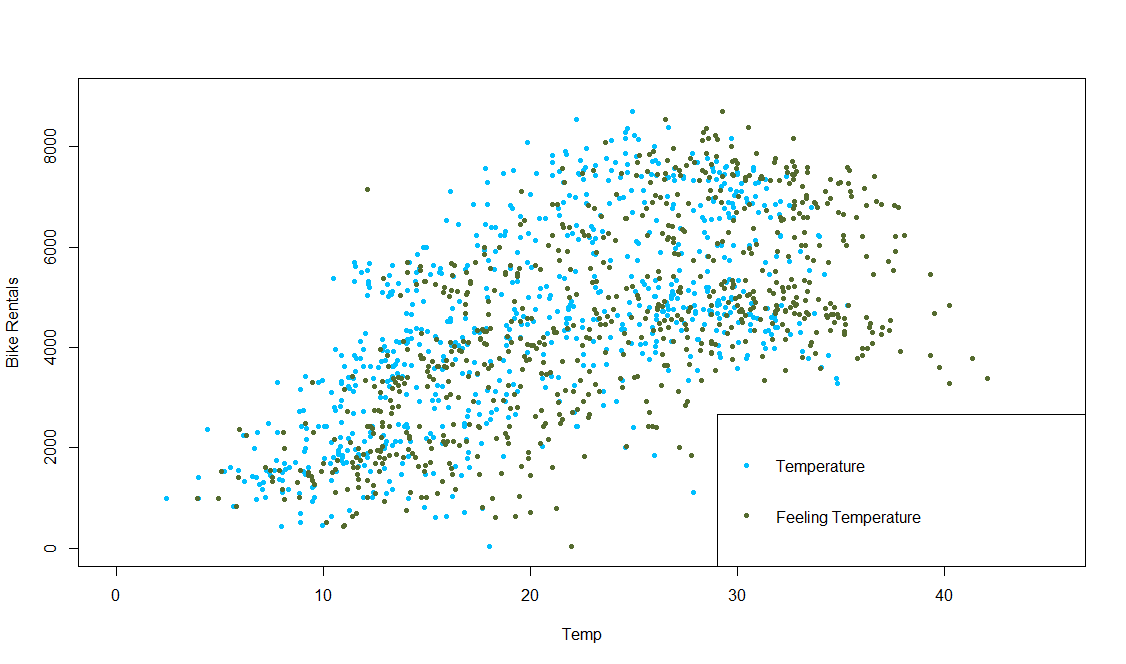
First we converted the temperature, because the data of temperature was divided to 41. Secondly we calculated the mean, the median and the standard deviation of all seasons

* *Spring:*
  + *The mean temperature of spring was 12.21.*
  + *The median temperature of spring was 11.72.*
  + *The standard deviation of the temperature in spring was 4.21.*
* *Summer:*
  + *The mean temperature of summer was 22.32.*
  + *The median temperature of summer was 23.05.*
  + *The standard deviation of the temperature in summer was 5.03.*
* *Fall:*
  + *The mean temperature of fall was 28.96.*
  + *The median temperature of fall was 29.3.*
  + *The standard deviation of the temperature in fall was 2.9.*
* *Winter:*
  + *The mean temperature of winter was 17.34.*
  + *The median temperature of winter was 16.78.*
  + *The standard deviation of the temperature in winter was 4.42*

**

*Correlation Analysis*

* + *Is there a correlation between the real temperature and the total count of bike rentals and feeling temperature and the total count of bike rentals? Do the correlations differ?*
  + *We run a correltaion to find out if there is a connection between the temperature and bike rentals. The correlation between bike rentals and real vs. feeling temperature reveal the same value of r= 0.63. This can be interpretet as a moderate to strong correlation between two parameters. That is, there is a high probability that bike rentals are dependent on temperature. Although a correlation does not tell us something about the causality of different parameters and which one influences the other, in this case it seems unrealistic that the number of bike rentals have an impact on the weather. In the scatterplot we can see that feeling temperature on an average is rated about 3 °C higher than the real temperature*.

**

* *We conducted a two-sample t-test with a p< 0,05 level to check for differences in the means of two groups. Results revealed a significant difference between the means of feeling temperature and temperature, in general (t= -8,32; df= 1450,2; p< 0,001) and for each season. This means, unregarding if the temperature is cold or hot, the negative t-values on tests for each seasons show that feeling temperature is always rated higher, with biggest difference in fall (t= -11,37; df= 357,39; p< 0,001; mean temp= 28,96; mean atemp= 32,79) and smallest in spring (t= -5,46; df= 350,98; p< 0,001; mean temp= 12,02; mean atemp= 14,85), even though all differences were highly significant (p-value below 0,001).*

*Regression Analysis*

* *Linear regression analysis was used to test if holiday and nice weather predicted the number of total bike rentals. Test results with a p< 0,05 level show that nice (b= 848,5; p< 0,001) and wet weather (b= -2255,2; p< 0,001) compared to cloudy weather significantly predicted the number of bike rentals. Also did holiday, even if less significant (b= -929,5; p= 0,02). The two predictors explained a significant proportion of variances in number of total bike rentals (R²= 0,11; F(3, 727)= 28,61; p< 0,001). To find out if the variance in weather explaines the variance of number of bike rentals, we conducted an ANOVA. Results revealed that there was a main effect of weather (F(2, 728)= 40,07; p< 0,001). At last, we conducted post-hoc tests to test pair-wise differences between the weather conditions with a Tukey test. The results indicated that there are significant differences between the means of all three groups (nice-cloudy, p< 0,001; wet-cloudy, p< 0,001; wet-nice, p< 0,001). (Comment: P-value for comparison of nice-cloudy and wet-nice was 0e+00. As I read, reporting a p= 0,000 is not correct, that is why I converted it into p< 0,001.)Mul*ticollinearity between the independent variables
* *To find out about mean values for temperature, humidity, windspeed and bike rentals we conducted a calculation with the dplyr function which allows us the execute an aggregate command for each dependent variable. Results show that Mean Temperature and Mean Bike Rentals are higher in summer months*

*Conclusion*

* *After conducting a variety of statistical analyses on the day dataset that gives us data on weather situations and number of bike rentals in Washington D.C. from years 2011 and 2012, I can conclude that temperatures change across the seasons and vary regarding their means. Results indicated that there is a moderate to strong correlation between feeling as well as real temperature and the total number of bike rentals, although feeling temperature and real temperature significantly* *differ in their means. We found out that holiday and especially weather situation are good predictors for the number of bike rentals. In execution of mean values for temperature, humidity, windspeed and bike rentals, similarities of a bell-shaped distribution could be found between temperature and bike rentals. In the end, we showed with a custom function that on strict limits for weather conditions with temperatures with a minimum of* *10 degrees and exclusion of rainy conditions, only about half of the days across the year were suitable for biking*