Bike sharing competition

## Introduction

This document is created for the purpose of Bike Sharing Demand completion on Kaggle. The point of the competition is to predict the demand on the rented bikes based on the historical data.

There were 10886 training cases and 6493 test cases in the data.

## Data Quality Assesment

The structure of the training set in R looks as follows:

## 'data.frame': 10886 obs. of 12 variables:  
## $ datetime : Factor w/ 10886 levels "2011-01-01 00:00:00",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ season : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ holiday : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ workingday: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ weather : int 1 1 1 1 1 2 1 1 1 1 ...  
## $ temp : num 9.84 9.02 9.02 9.84 9.84 ...  
## $ atemp : num 14.4 13.6 13.6 14.4 14.4 ...  
## $ humidity : int 81 80 80 75 75 75 80 86 75 76 ...  
## $ windspeed : num 0 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 ...

Comparing this structure to feature description provided by Kaggle:

Data Fields

datetime - hourly date + timestamp  
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter  
holiday - whether the day is considered a holiday  
workingday - whether the day is neither a weekend nor holiday  
weather - 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 - temperature in Celsius  
atemp - "feels like" temperature in Celsius  
humidity - relative humidity  
windspeed - wind speed  
casual - number of non-registered user rentals initiated  
registered - number of registered user rentals initiated  
count - number of total rentals

### Data representation fixes

We can observe some problems with data representation in followig variables:

* datetime
* season
* holiday
* workingday
* weather

Let's look at the data fields structure in more detail

## datetime season holiday   
## 2011-01-01 00:00:00: 1 Min. :1.000 Min. :0.00000   
## 2011-01-01 01:00:00: 1 1st Qu.:2.000 1st Qu.:0.00000   
## 2011-01-01 02:00:00: 1 Median :3.000 Median :0.00000   
## 2011-01-01 03:00:00: 1 Mean :2.507 Mean :0.02857   
## 2011-01-01 04:00:00: 1 3rd Qu.:4.000 3rd Qu.:0.00000   
## 2011-01-01 05:00:00: 1 Max. :4.000 Max. :1.00000   
## (Other) :10880   
## workingday weather temp atemp   
## Min. :0.0000 Min. :1.000 Min. : 0.82 Min. : 0.76   
## 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:13.94 1st Qu.:16.66   
## Median :1.0000 Median :1.000 Median :20.50 Median :24.24   
## Mean :0.6809 Mean :1.418 Mean :20.23 Mean :23.66   
## 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:26.24 3rd Qu.:31.06   
## Max. :1.0000 Max. :4.000 Max. :41.00 Max. :45.45   
##   
## humidity windspeed casual registered   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. : 0.0   
## 1st Qu.: 47.00 1st Qu.: 7.002 1st Qu.: 4.00 1st Qu.: 36.0   
## Median : 62.00 Median :12.998 Median : 17.00 Median :118.0   
## Mean : 61.89 Mean :12.799 Mean : 36.02 Mean :155.6   
## 3rd Qu.: 77.00 3rd Qu.:16.998 3rd Qu.: 49.00 3rd Qu.:222.0   
## Max. :100.00 Max. :56.997 Max. :367.00 Max. :886.0   
##   
## count   
## Min. : 1.0   
## 1st Qu.: 42.0   
## Median :145.0   
## Mean :191.6   
## 3rd Qu.:284.0   
## Max. :977.0   
##

Below the prposed approach for different inccorectly represented data fields:

* **datetime** - variable is stored as Factor variable with 10886, so one level for every observations. We can solve this problem by extracting the time using some kind of string operations or using data handling packages. I will use R package *lubridate* for this purpose.
* **season** - this is 4 level factor variable, represented as a integer. The type change will be required for analysis
* **holiday** - this is binnary variable, it might be interesting to connect this information with the day of the week extracted from datetime.
* **workingday** - another binnary variable, should contain same information as day of the week excluding the holiday
* **weather** - factor variable with four levels represented as integer. Simillar to the **season**. The commonsense would suggest that this variable is ordinal but results below does not support this assumption. While these results contradicting a little commonsense additional analysis should be performed.

## [1] "Average number of bikes rented by weather"

## 1 2 3 4   
## 205.2368 178.9555 118.8463 164.0000

The average number of bikes rented for **weather** category "4" which corresponds to really bad weather is higher then for category "3" which corresponds to slightly better weather. It could suggest the problems with the data or that snow alone is not real problem for cyclist. It would be interesting idea to confirm if this weather variable can describe the situtation when snow is just on the ground and not snowing.

### Correlation Analysis

The next step is to visualise correlation between variables to evaluate which of them could be removed before the model creation step is performed. The *corrplot* package is used for visualizastion, and the Peason Correlation Coefficient was used for this case.

## Warning: package 'corrplot' was built under R version 3.1.3