ml-project

lohit adari

2022-12-03

#Bike rental prediction

```
rm(list=ls()) ; gc()
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 474397 25.4    1033320 55.2   644200 34.5
## Vcells 859439  6.6    8388608 64.0  1635008 12.5
```

```
library(car)
```

```
## Loading required package: carData
```

```
library(explore)
```

```
## Warning: package 'explore' was built under R version 4.2.2
```

```
library(ggplot2)
library(DataExplorer)
```

```
## Warning: package 'DataExplorer' was built under R version 4.2.2
```

```
library(MASS)
library(xgboost)
```

```
## Warning: package 'xgboost' was built under R version 4.2.2
```

```
library(caret)
```

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

```
library(lightgbm)
```

```
## Warning: package 'lightgbm' was built under R version 4.2.2
```

```
## Loading required package: R6
```

```
##
## Attaching package: 'lightgbm'
```

```
## The following objects are masked from 'package:xgboost':
##
##     getinfo, setinfo, slice
```

```
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 4.2.2
```

```
## Loaded gbm 2.1.8.1
```

```
library(gridExtra)
```

```
bikes_df<-read.csv("C:/Users/adari/Downloads/train (1).csv",sep=";",header = TRUE,skip = 1)
bikes_df_test<-read.table("C:/Users/adari/Downloads/test.csv",sep = ";",skip = 1)
```

```
## Warning in readLines(file, skip): line 1 appears to contain an embedded nul
```

```
colnames(bikes_df)<-c("id","year","hour","season","holiday","workingday","weather","temp","atemp","humi
colnames(bikes_df_test) <-c("id","year","hour","season","holiday","workingday","weather","temp","atemp"
```

```
head(bikes_df)
```

```
##   id year hour season holiday workingday weather  temp  atemp humidity
## 1  4 2011    8      3       0          0       1 27.88 31.820       57
## 2  5 2012    2      1       0          1       1 20.50 24.240       59
## 3  7 2011   20      3       0          1       3 25.42 28.790       83
## 4  8 2011   17      3       0          1       3 26.24 28.790       89
## 5  9 2011   19      2       0          1       2 34.44 37.120       39
## 6 10 2012   23      2       0          1       2 23.78 27.275       78
##   windspeed target
## 1    0.0000    132
## 2    0.0000     19
## 3   19.9995     58
## 4    0.0000    285
## 5   22.0028    326
## 6    7.0015     75
```

```
paste("Dimension of dataset: ", dim(bikes_df))
```

```
## [1] "Dimension of dataset:  7688" "Dimension of dataset:  12"
```

```
head(bikes_df_test)
```

```
##   id year hour season holiday workingday weather  temp  atemp humidity
## 1  1 2012   21      3       0          0       1 29.52 34.850       79
## 2  2 2012    3      2       0          0       1 23.78 27.275       83
```

```
## 3  6 2011   10      1       0          1       3 16.40 20.455        0
## 4 14 2012   19      1       0          1       1 13.94 15.150       46
## 5 17 2011   23      3       0          1       2 26.24 30.305       73
## 6 20 2012    6      2       0          0       1 21.32 25.000       72
##   windspeed
## 1    6.0032
## 2    0.0000
## 3   11.0014
## 4   19.9995
## 5   11.0014
## 6    7.0015
```

```
paste("Dimension of dataset: ", dim(bikes_df_test))
```

```
## [1] "Dimension of dataset:  3196" "Dimension of dataset:  11"
```

EDA: Exploratory Data Analysis

```
summary(bikes_df)
```

```
##        id              year           hour           season
##  Min.   :    4   Min.   :2011   Min.   : 0.00   Min.   :1.000
##  1st Qu.: 2772   1st Qu.:2011   1st Qu.: 6.00   1st Qu.:2.000
##  Median : 5478   Median :2011   Median :12.00   Median :3.000
##  Mean   : 5464   Mean   :2011   Mean   :11.56   Mean   :2.506
##  3rd Qu.: 8187   3rd Qu.:2012   3rd Qu.:18.00   3rd Qu.:4.000
##  Max.   :10886   Max.   :2012   Max.   :23.00   Max.   :4.000
##     holiday          workingday        weather          temp
##  Min.   :0.00000   Min.   :0.0000   Min.   :1.00   Min.   : 0.82
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.00   1st Qu.:13.94
##  Median :0.00000   Median :1.0000   Median :1.00   Median :20.50
##  Mean   :0.02901   Mean   :0.6774   Mean   :1.41   Mean   :20.27
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:2.00   3rd Qu.:26.24
##  Max.   :1.00000   Max.   :1.0000   Max.   :3.00   Max.   :41.00
##      atemp          humidity        windspeed         target
##  Min.   : 0.76   Min.   :  0.00   Min.   : 0.000   Min.   :  1.0
##  1st Qu.:16.66   1st Qu.: 46.00   1st Qu.: 7.002   1st Qu.: 41.0
##  Median :24.24   Median : 62.00   Median :12.998   Median :145.0
##  Mean   :23.70   Mean   : 61.77   Mean   :12.802   Mean   :191.4
##  3rd Qu.:31.06   3rd Qu.: 77.00   3rd Qu.:16.998   3rd Qu.:283.0
##  Max.   :45.45   Max.   :100.00   Max.   :56.997   Max.   :977.0
```

```
str(bikes_df)
```

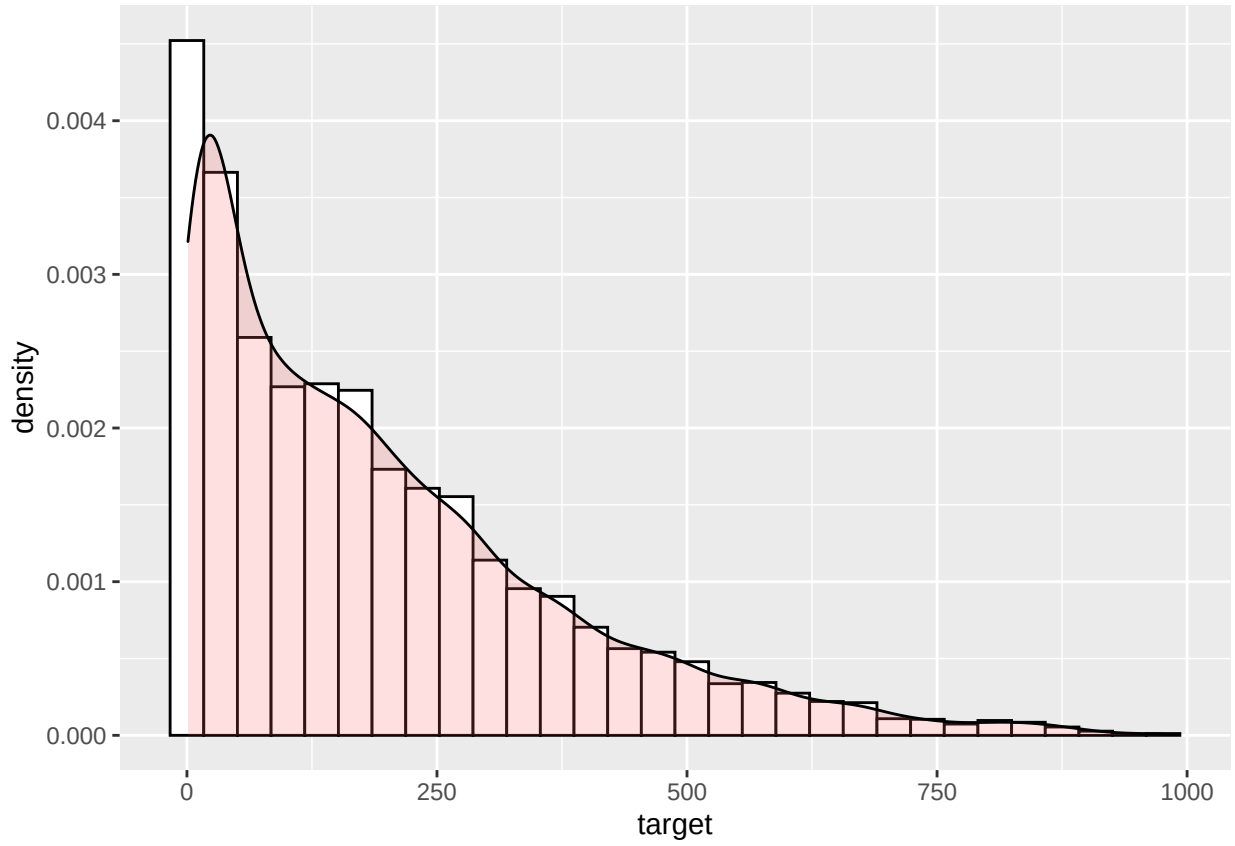
```
## 'data.frame':    7688 obs. of  12 variables:
##  $ id        : int  4 5 7 8 9 10 11 12 13 15 ...
##  $ year      : int  2011 2012 2011 2011 2011 2012 2011 2011 2011 2012 ...
##  $ hour      : int  8 2 20 17 19 23 22 14 13 15 ...
##  $ season    : int  3 1 3 3 2 2 3 3 1 2 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ workingday: int  0 1 1 1 1 1 1 1 0 0 ...
##  $ weather   : int  1 1 3 3 2 2 1 1 2 1 ...
```

```
##  $ temp      : num  27.9 20.5 25.4 26.2 34.4 ...
##  $ atemp     : num  31.8 24.2 28.8 28.8 37.1 ...
##  $ humidity  : int  57 59 83 89 39 78 94 53 72 50 ...
##  $ windspeed : num  0 0 20 0 22 ...
##  $ target    : int  132 19 58 285 326 75 160 134 94 463 ...
```

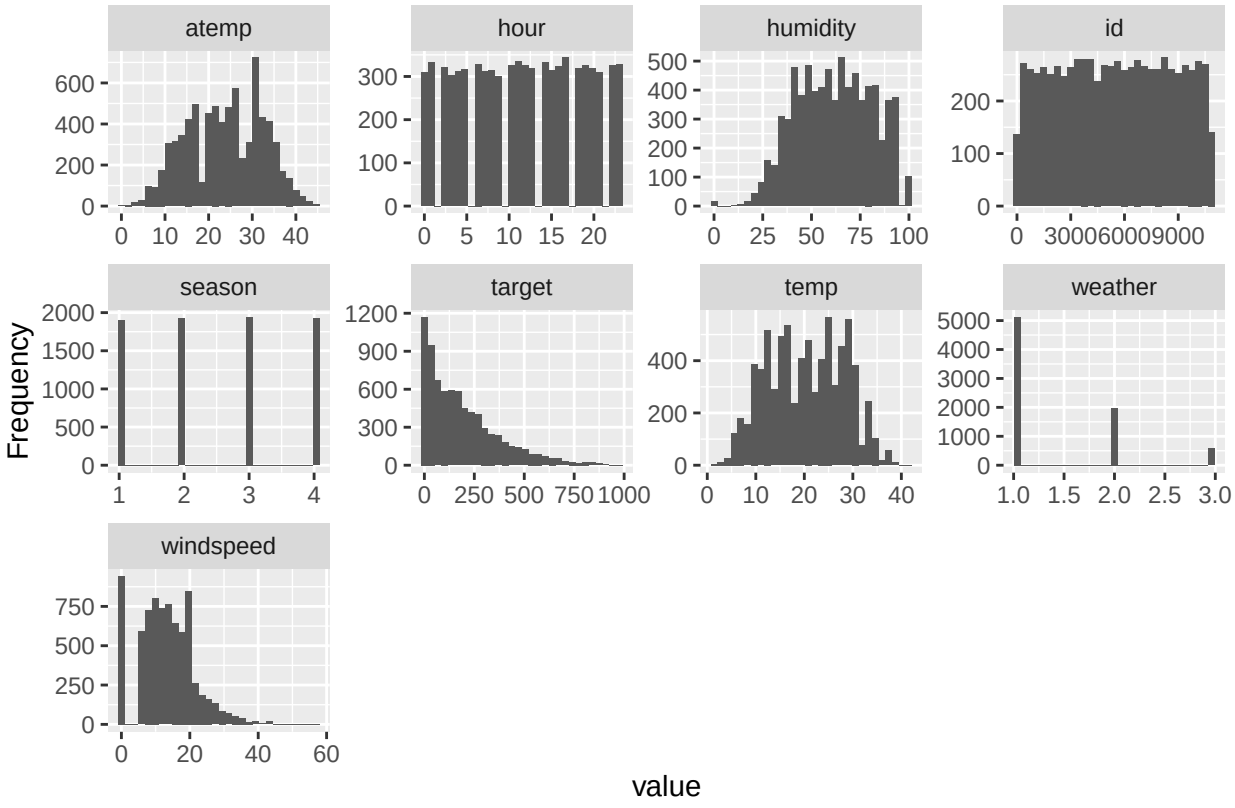
Target variabale explain its poission

```
ggplot(bikes_df, aes(x=target)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

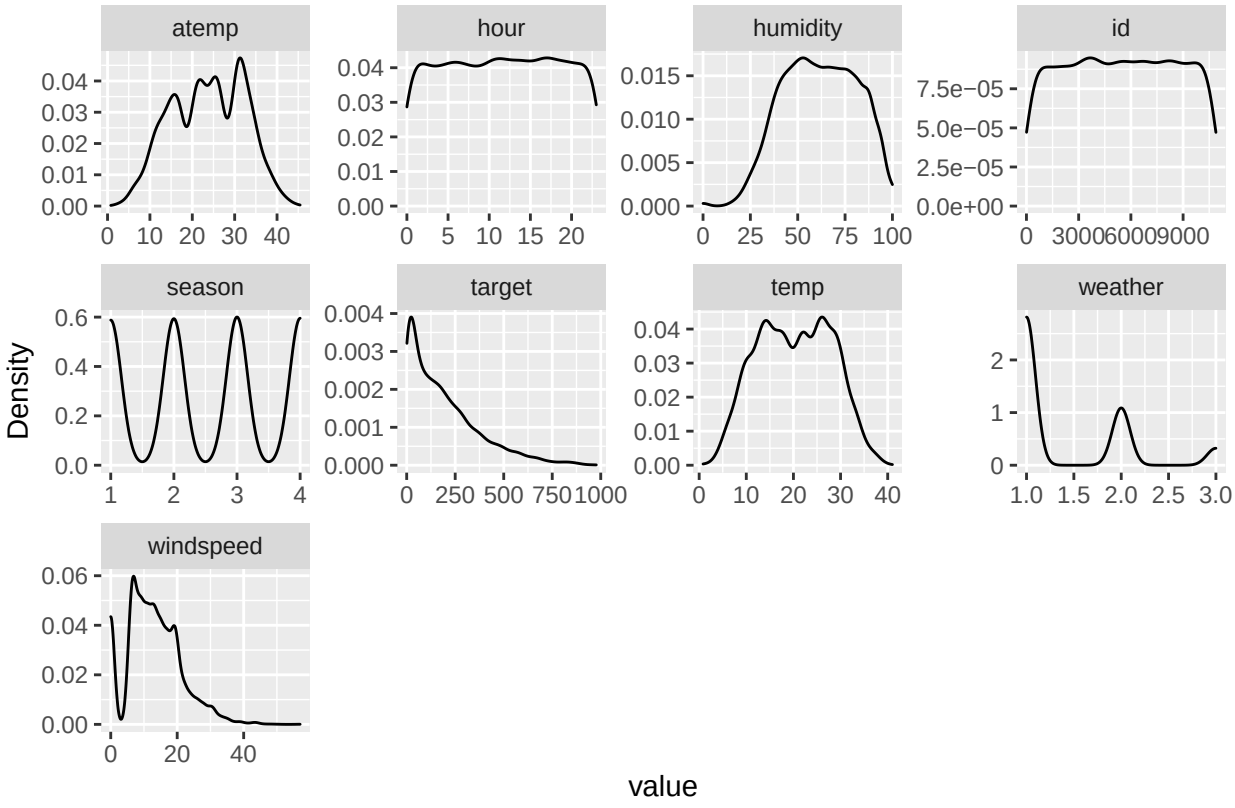
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



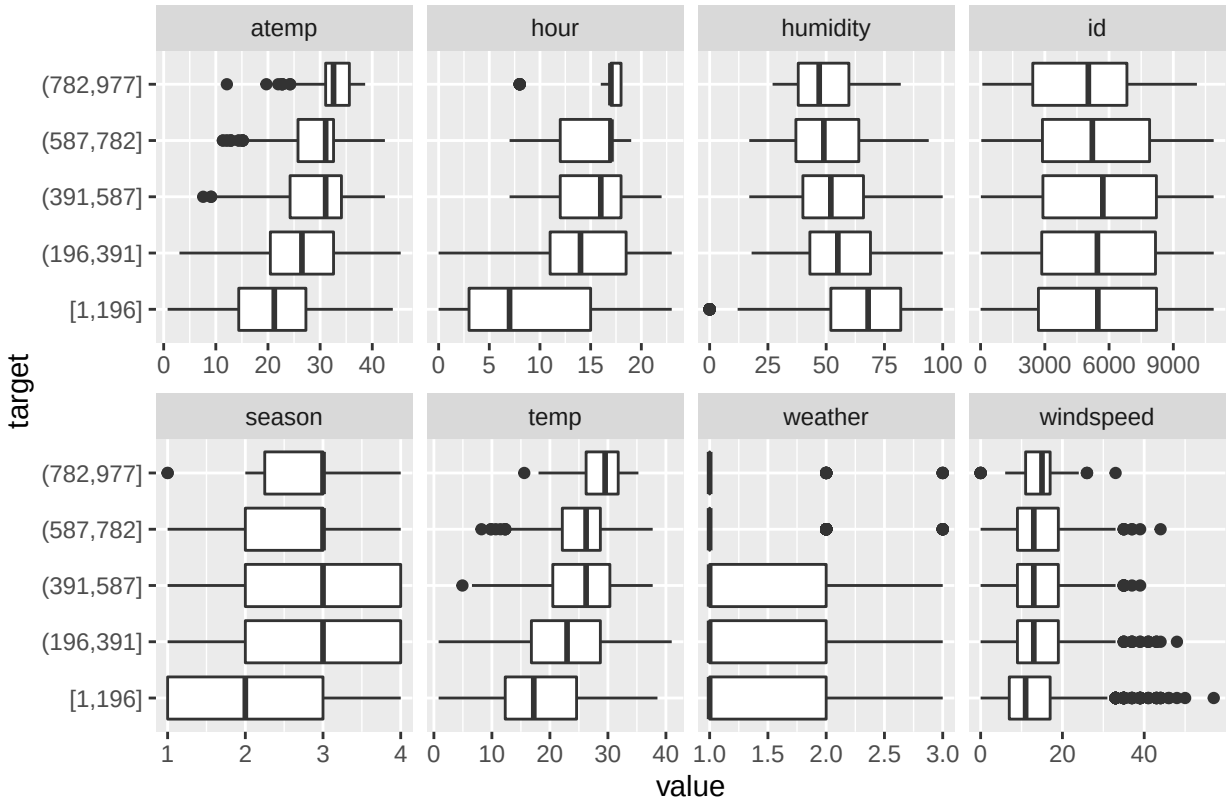

```
options(repr.plot.width=18, repr.plot.height=10)
plot_histogram(bikes_df)
```




```
plot_density(bikes_df)
```

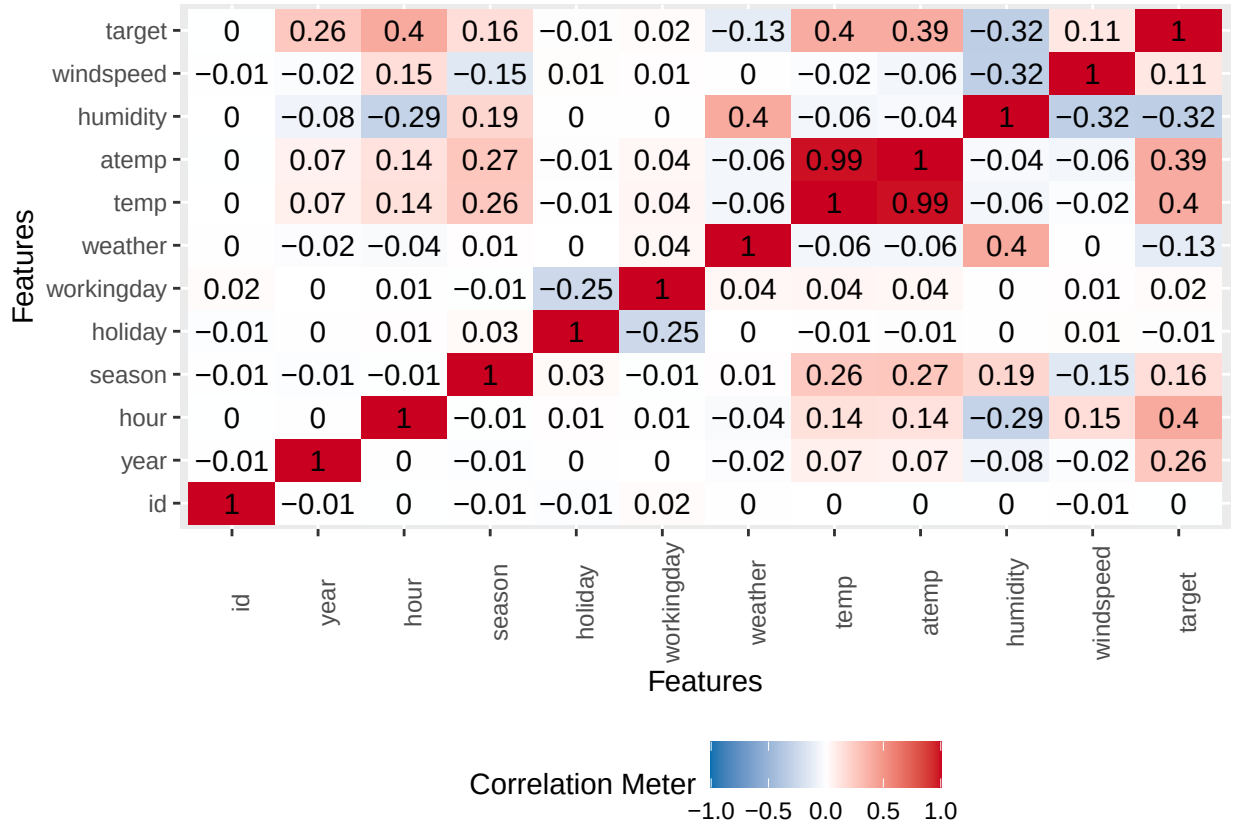



```
plot_boxplot(bikes_df,by="target")
```



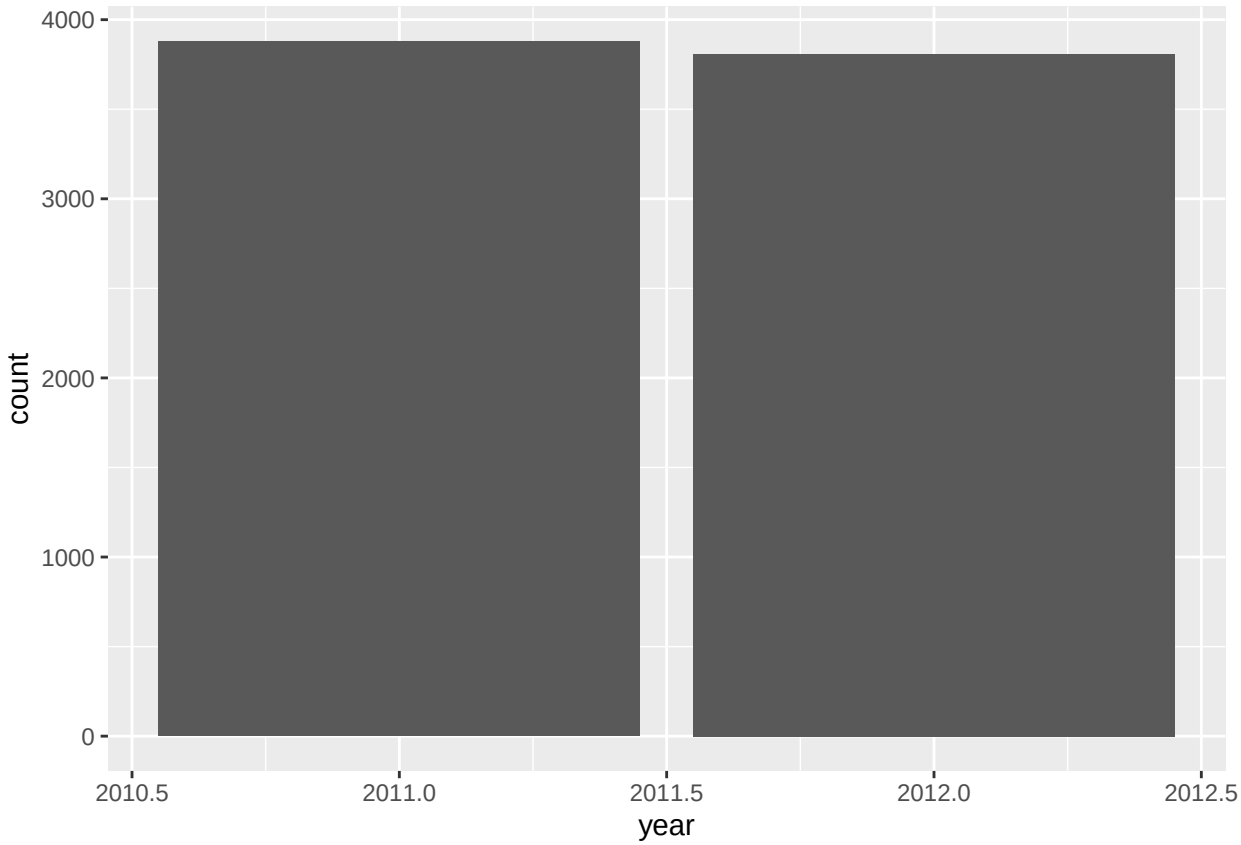
Drop Id column

```
plot_correlation(bikes_df, type = 'continuous')
```



Year Variable

```
options(repr.plot.width=8, repr.plot.height=6)
ggplot(bikes_df, aes(x=year)) +
  geom_bar()
```



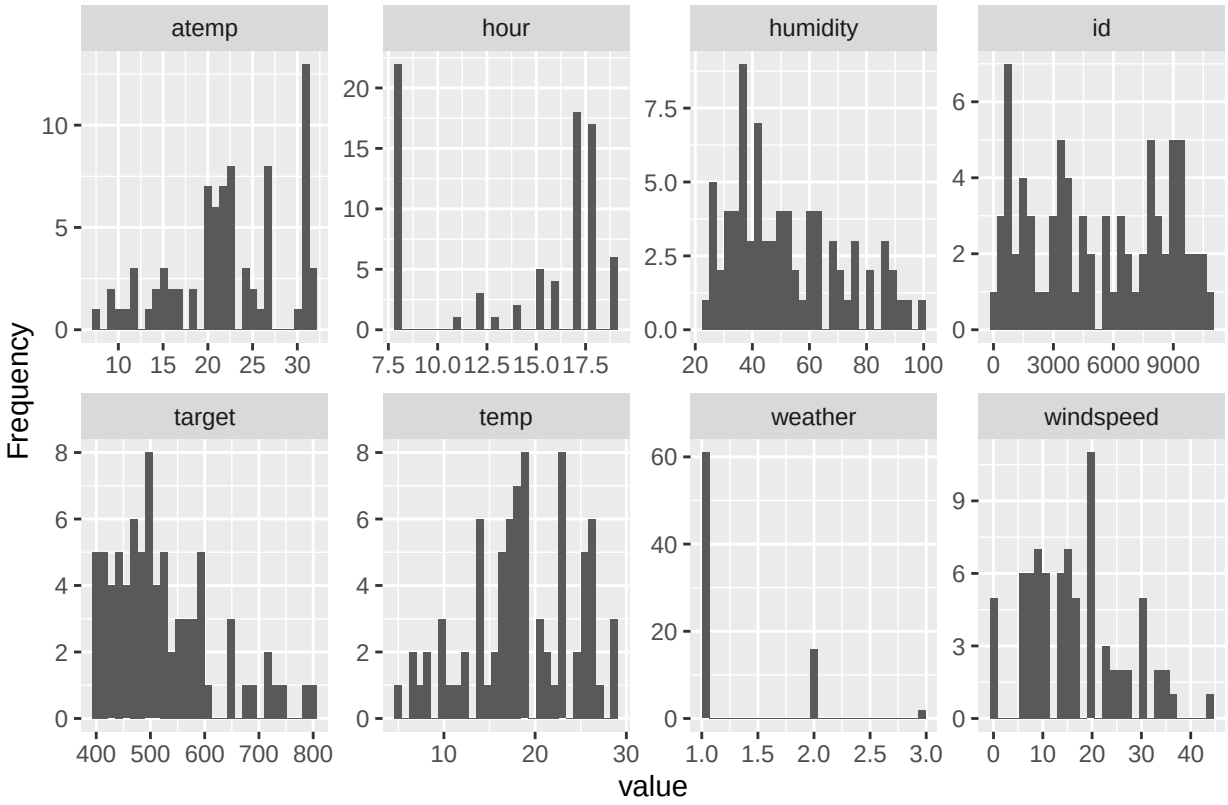

```
winter_high_demand_df <- subset(bikes_df, season == 1 & target > 400)
head(winter_high_demand_df)
```

```
##      id year hour season holiday workingday weather  temp  atemp humidity
## 18   25 2012   17      1       0          1       1 17.22 21.210       32
## 149 205 2012    8      1       0          1       3 16.40 20.455       87
## 263 364 2012    8      1       0          1       1  6.56  9.090       59
## 275 377 2012    8      1       0          1       1  8.20 11.365       82
## 409 563 2012    8      1       0          1       1 18.86 22.725       88
## 411 565 2012   18      1       0          1       1 16.40 20.455       43
##     windspeed target
## 18    22.0028    465
## 149    0.0000    445
## 263    7.0015    501
## 275    7.0015    499
## 409    7.0015    579
## 411   31.0009    410
```

```
paste("Number of rows: ", dim(winter_high_demand_df))
```

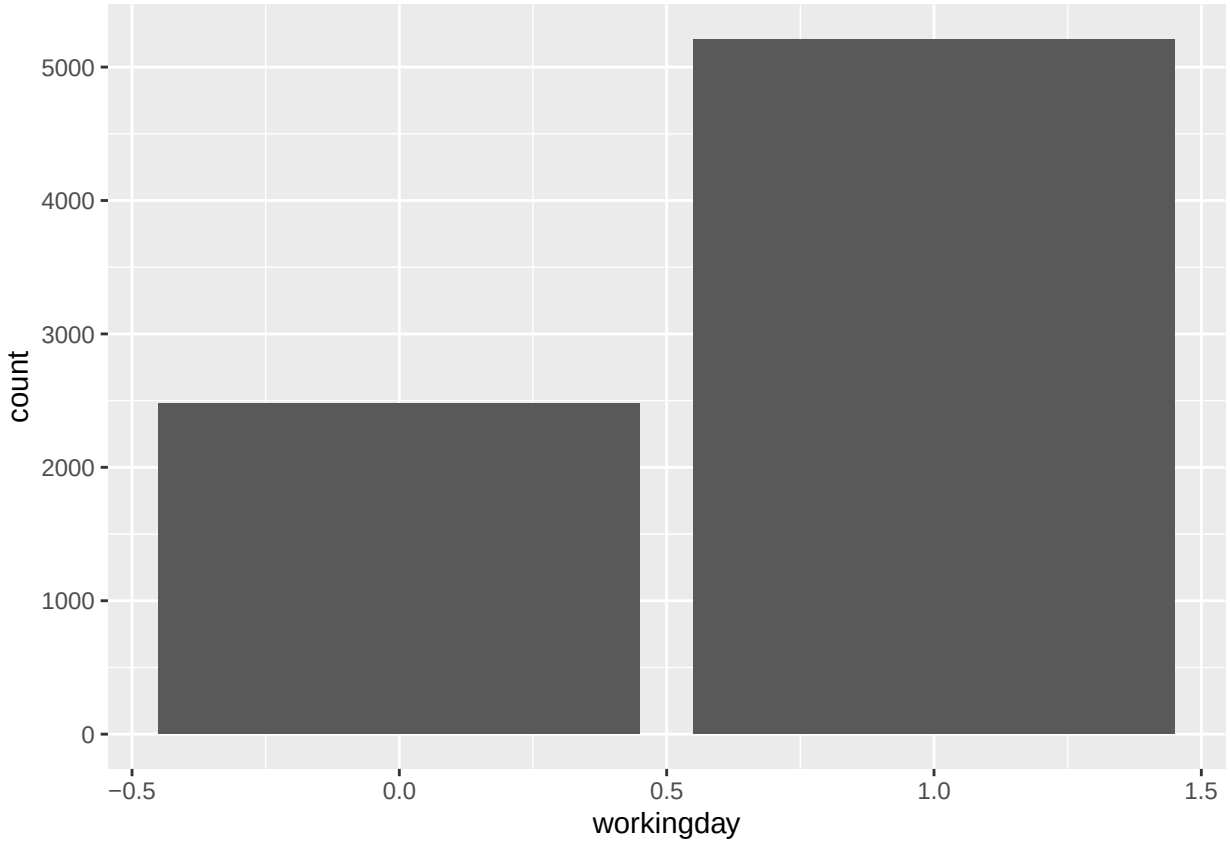
```
## [1] "Number of rows:  79" "Number of rows:  12"
```

```
options(repr.plot.width=14, repr.plot.height=10)
plot_histogram(winter_high_demand_df)
```



working day

```
ggplot(bikes_df, aes(x=workingday)) +
  geom_bar()
```

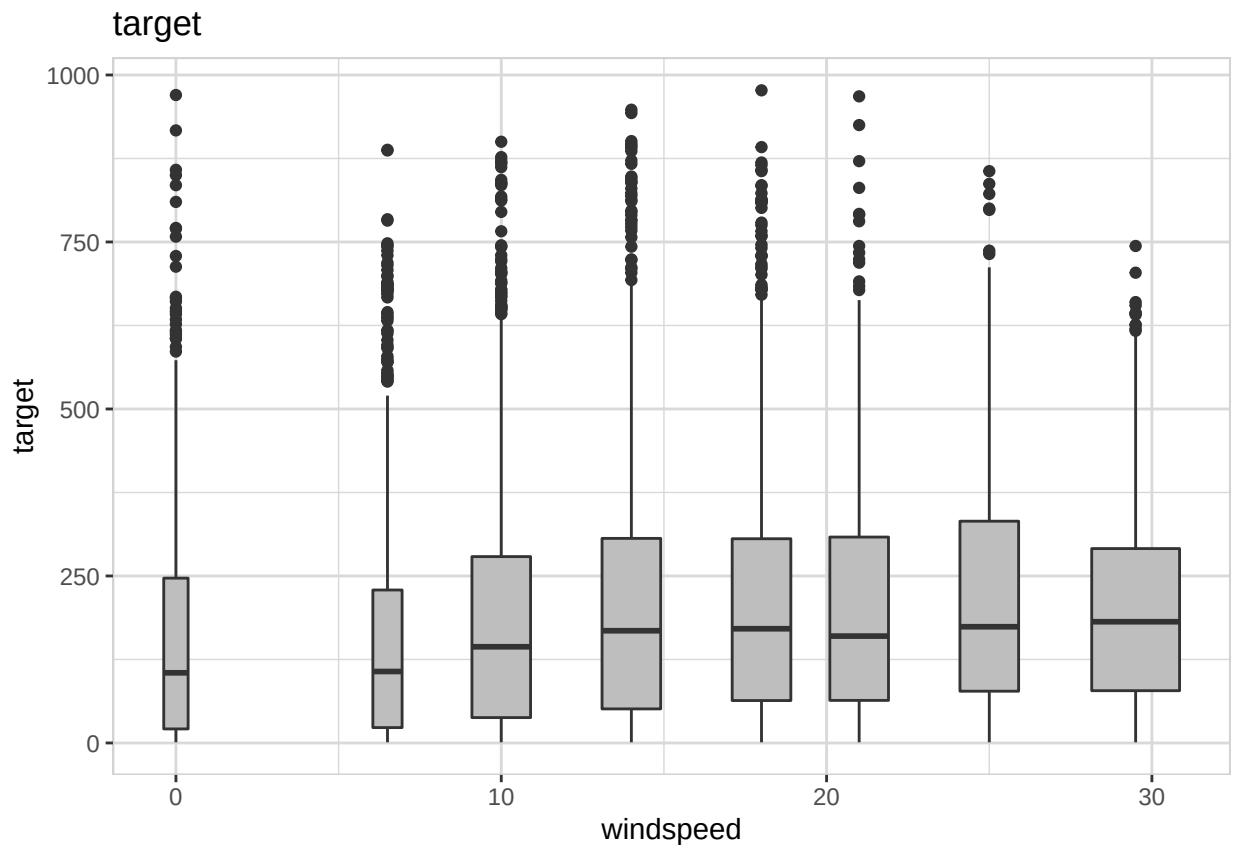



```
outiler_rows <- length(boxplot(subset(bikes_df,humidity>=0  & humidity <= 20)$target ,plot=FALSE)$out)+
length(boxplot(subset(bikes_df,humidity>=20 & humidity <= 40)$target ,plot=FALSE)$out)+
length(boxplot(subset(bikes_df,humidity>=40 & humidity <= 60)$target ,plot=FALSE)$out)+
length(boxplot(subset(bikes_df,humidity>=60 & humidity <= 80)$target ,plot=FALSE)$out)+
length(boxplot(subset(bikes_df,humidity>=80 & humidity <= 100)$target ,plot=FALSE)$out)

paste("We will drop about", outiler_rows, "outliers")
```

```
## [1] "We will drop about 277 outliers"
```

windspeed

```
explore(bikes_df,`windspeed`, target = target)
```




```
paste("Rows with >85 humdity and weather ==1 ",nrow(subset(bikes_df,humidity>85 & weather==1)))
```

```
## [1] "Rows with >85 humdity and weather ==1  289"
```

```
paste("Rows atemp >40 atemp and weather ==1 ",nrow(subset(bikes_df,atemp>40 & weather==1)))
```

```
## [1] "Rows atemp >40 atemp and weather ==1  96"
```

```
paste("# Duplicate ids: ",sum(duplicated(bikes_df$id)))
```

```
## [1] "# Duplicate ids:  0"
```

```
paste("# Duplicate rows: ",sum(duplicated(bikes_df)))
```

```
## [1] "# Duplicate rows:  0"
```

Data analysis

```
library(factoextra)
```

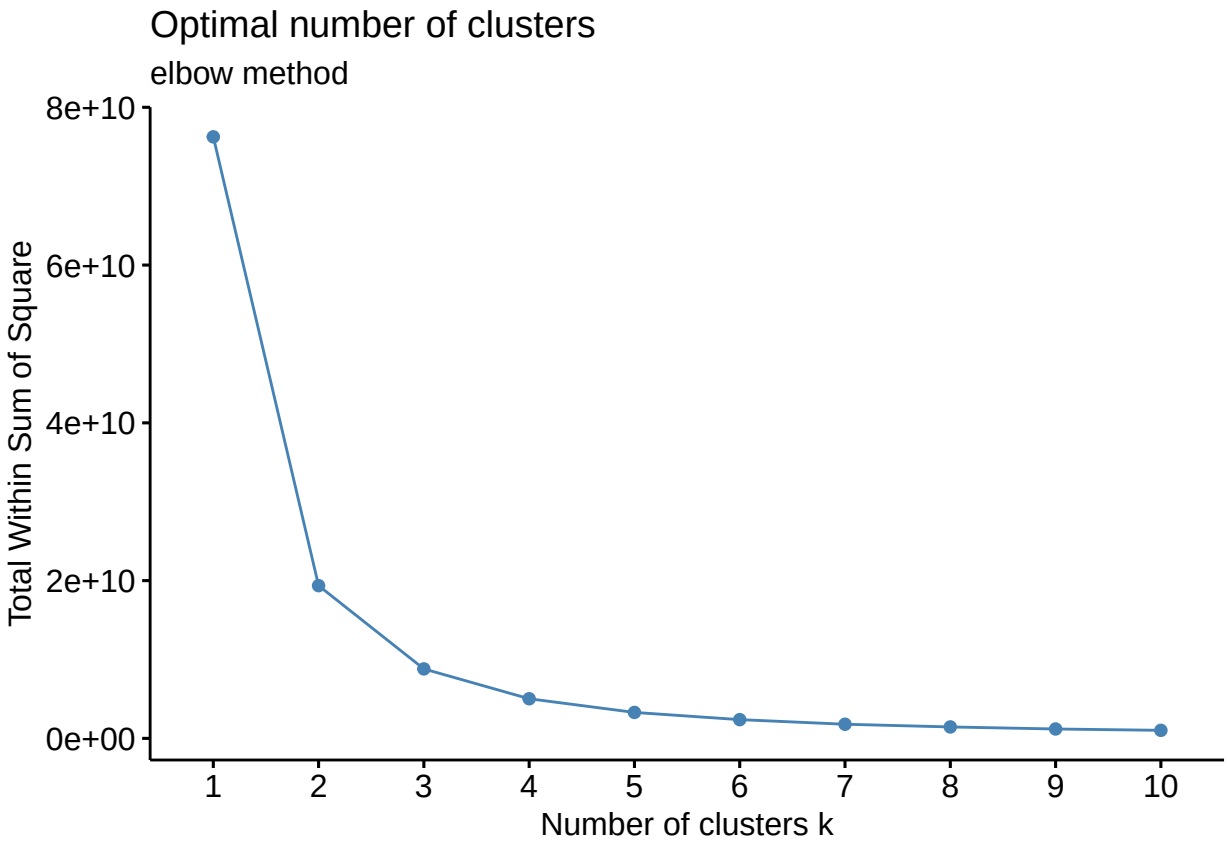
```
## Warning: package 'factoextra' was built under R version 4.2.2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

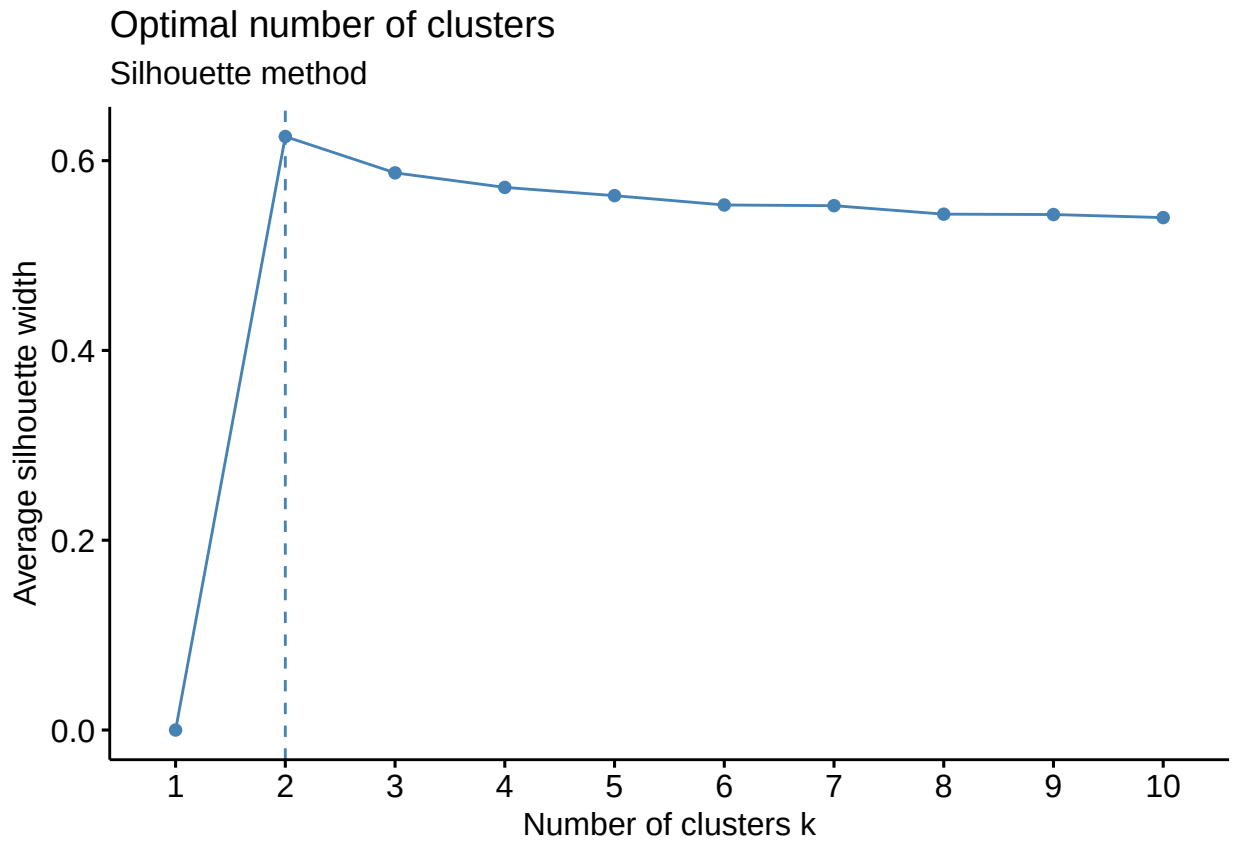
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:lightgbm':
##
##     slice

## The following object is masked from 'package:xgboost':
##
##     slice

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
fviz_nbclust(bikes_df,kmeans,method="wss")+labs(subtitle = "elbow method")
```




```
library(NbClust)
library(factoextra)
fviz_nbclust(bikes_df[,1:3], kmeans, method = "silhouette")+
labs(subtitle = "Silhouette method")
```




```
pcluster<-prcomp(bikes_df[,4:7],scale. = FALSE)
summary(pcluster)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4
## Standard deviation     1.1160 0.6285 0.4689 0.16168
## Proportion of Variance 0.6602 0.2094 0.1166 0.01386
## Cumulative Proportion  0.6602 0.8696 0.9861 1.00000
```

```
pcluster$rotation[,1:2]
```

```
##                     PC1          PC2
## season     -0.999952570  0.007415121
## holiday    -0.004145646  0.002794302
## workingday  0.004285773 -0.060675251
## weather    -0.007700824 -0.998126105
```