Final Project - Group 14

2022-12-04

1. "Read data and some base plots"

Read data and simple processing.

```
suppressPackageStartupMessages(library(tidyverse)) # just in case
library(ISLR2)
library(tidyverse)
library(dplyr)
library(naniar)
library(lubridate)
```

```
## Loading required package: timechange
```

```
##
## Attaching package: 'lubridate'
```

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

```
suppressPackageStartupMessages(library(glmnet)) # penalized linear models
suppressPackageStartupMessages(library(glmnetUtils))  # for quality of life functions over glmnet
suppressPackageStartupMessages(library(corrplot)) # correlation plots
suppressPackageStartupMessages(library(pls)) # for pcr
setwd("~/Semester files/STA 545/STA545_Final_Project")
#call data
origin_data=read_csv('Bike-Sharing-Dataset/hour.csv',show_col_types = FALSE)
#Check how many predictors have NAs
origin_data%>%miss_var_summary()%>%filter(n_miss!=0)%>%nrow()%>%print()
```

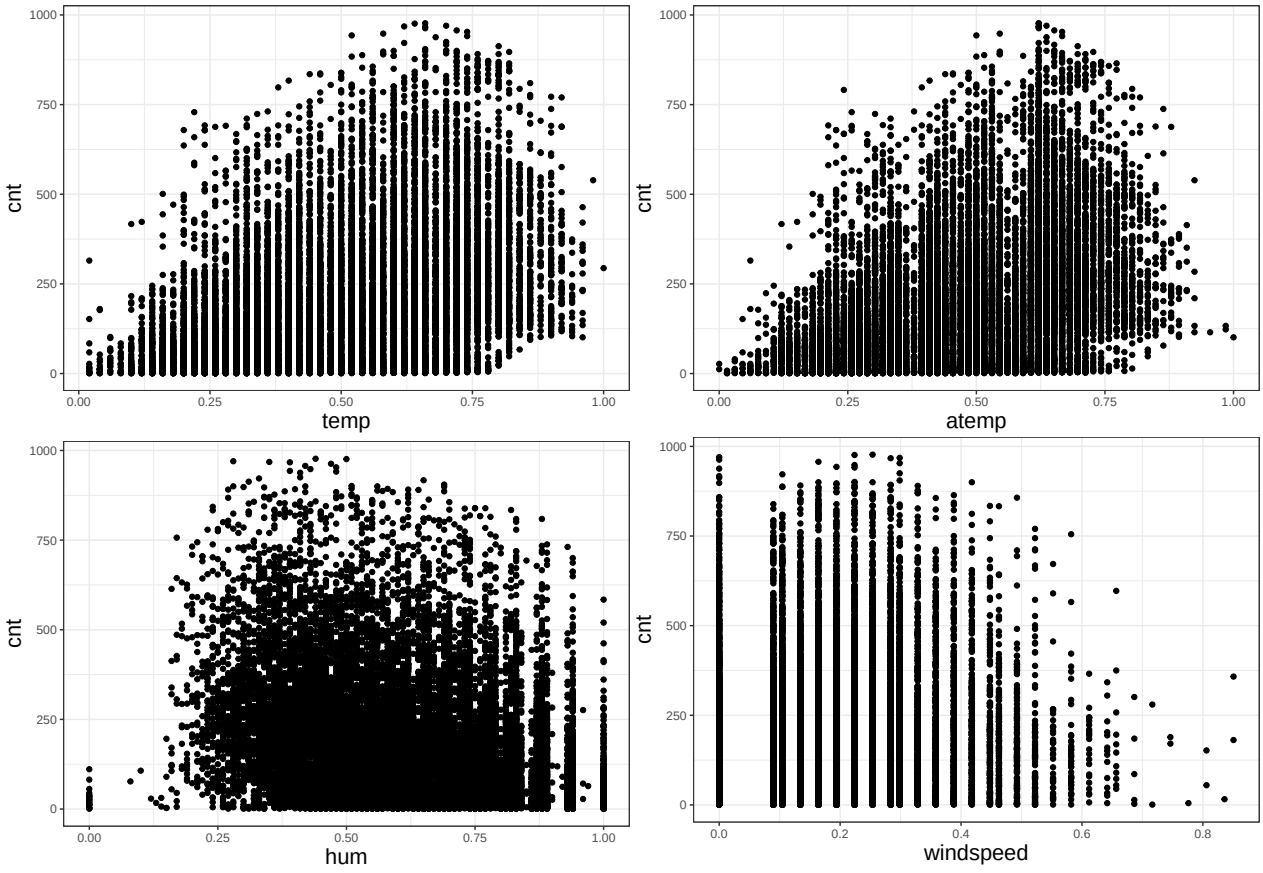
```
## [1] 0
```

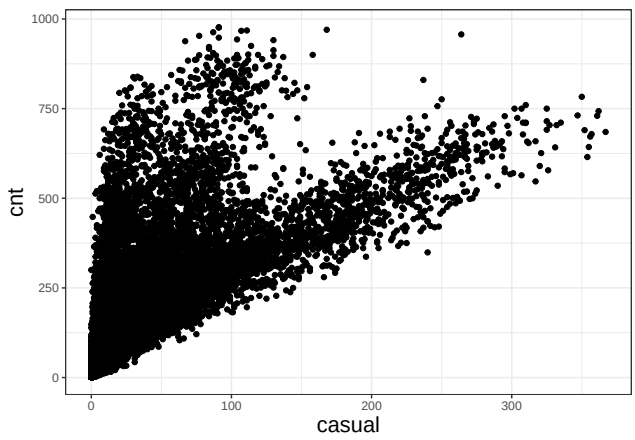
```
#Avoid changing original data
bs_hour=origin_data%>%mutate(dteday=as.Date(dteday))%>%select(-instant)
#Add one hourly identifiable column to identify every row
bs_hour=bs_hour%>%mutate(hourly_id=paste(as.character(dteday),as.character(hr)))%>%mutate(hourly_id=ymd
bs_hour=bs_hour[,c(1:15,17,16)]
bs_hour$windspeed=as.numeric(bs_hour$windspeed)
```

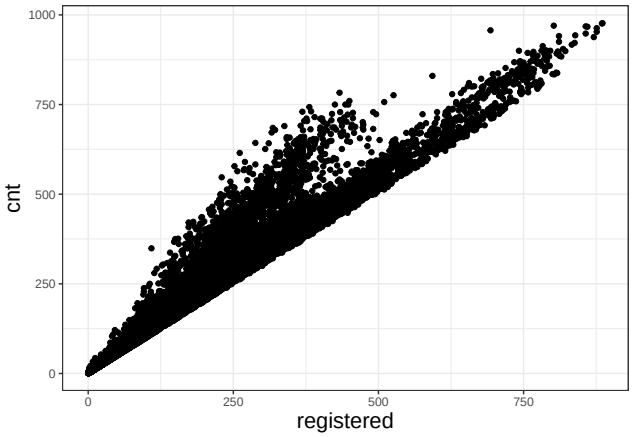
Scatter plots & Box plots for total counts.

```
col_vec_scatter=colnames(bs_hour)[10:15]
col_vec_box=colnames(bs_hour)[2:9]
for (value in col_vec_scatter) {
    print(ggplot(bs_hour)+geom_point(aes_string(value,'cnt'))+theme_bw()+
            theme(axis.title.y=element_text(size=16),
        axis.title.x=element_text(size=16)))
}
```

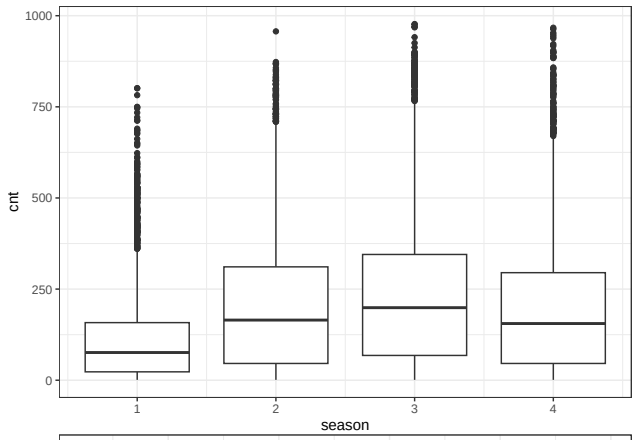
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation ideoms with `aes()`
```

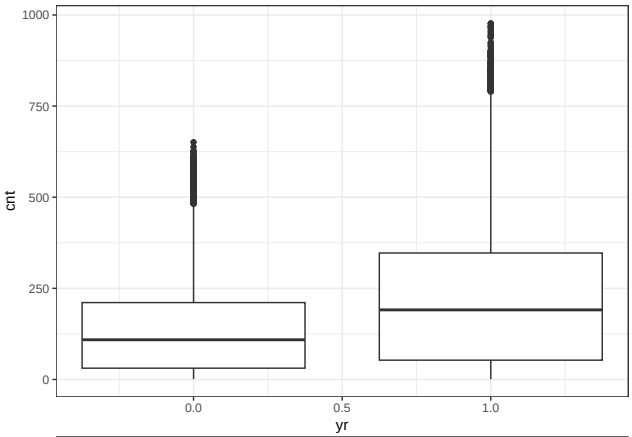


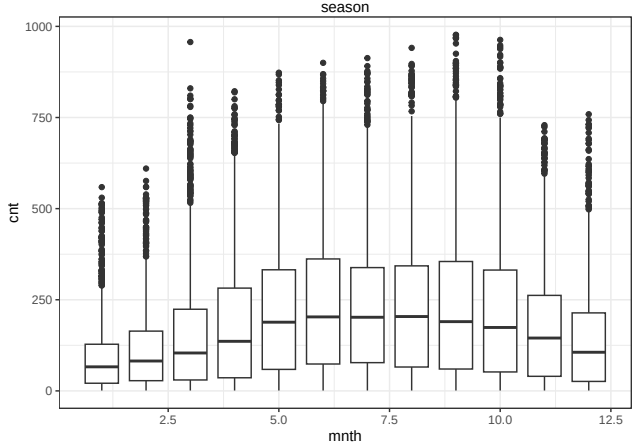


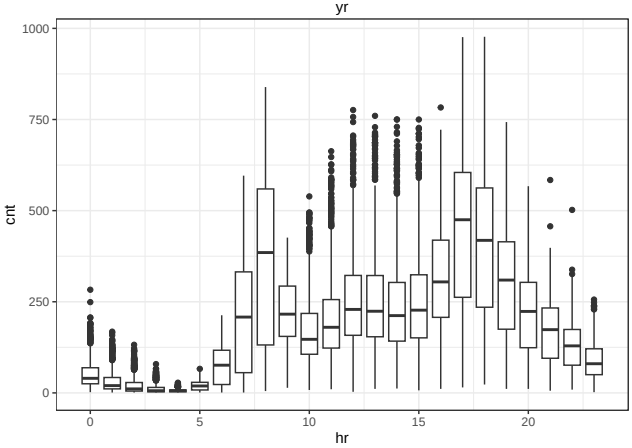


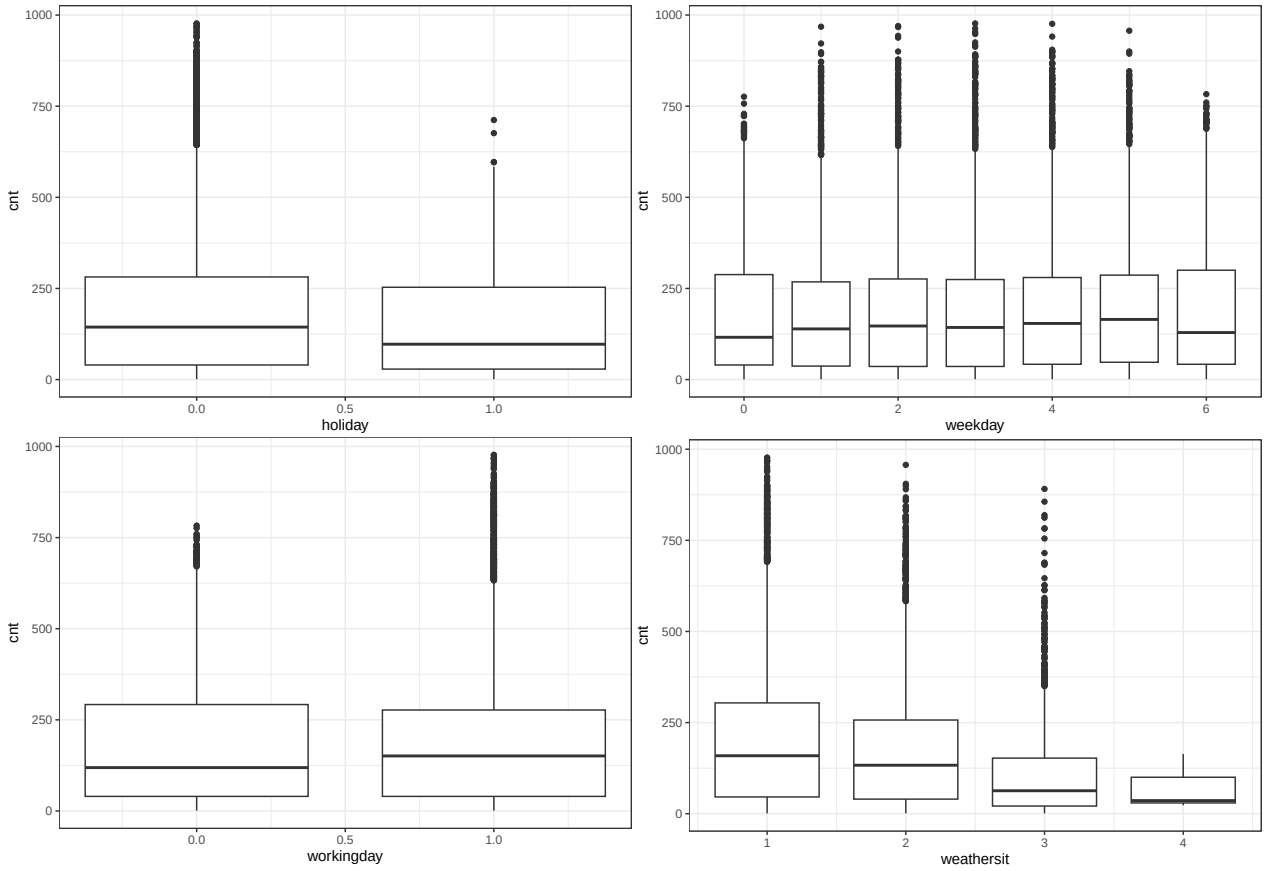

```
for (value in col_vec_box) {
    print(ggplot(bs_hour)+geom_boxplot(aes_string(value,'cnt',group=value))+theme_bw())+
            theme(axis.title.y=element_text(size=16),
        axis.title.x=element_text(size=16))
}
```



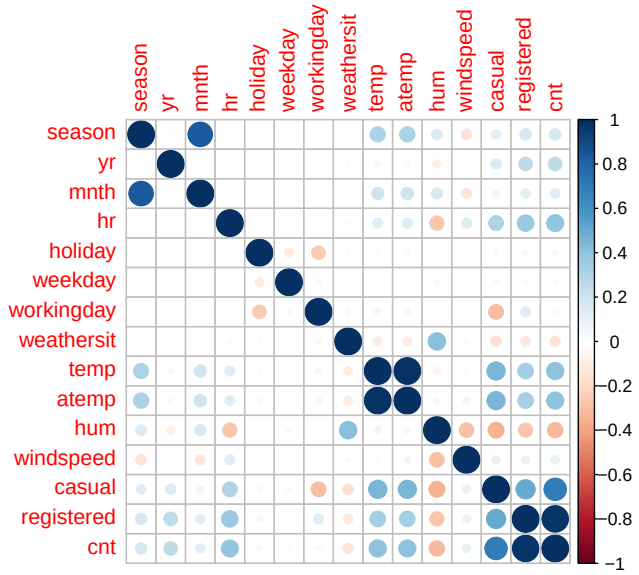









```
cor(bs_hour[, -c(1,16)]) %>%
  corrplot::corrplot()
```



Scatter plots & Box plots for casual user counts.

```
# col_vec_scatter=colnames(bs_hour)[10:13]
# for (value in col_vec_scatter) {
#     print(ggplot(bs_hour)+geom_point(aes_string(value,'casual'))+theme_bw()+
#             theme(axis.title.y=element_text(size=16),
#         axis.title.x=element_text(size=16)))
# }
# for (value in col_vec_box) {
#     print(ggplot(bs_hour)+geom_boxplot(aes_string(value,'casual',group=value))+theme_bw())+
#             theme(axis.title.y=element_text(size=16),
#         axis.title.x=element_text(size=16))
# }
```

Scatter plots & Box plots for registered user counts.

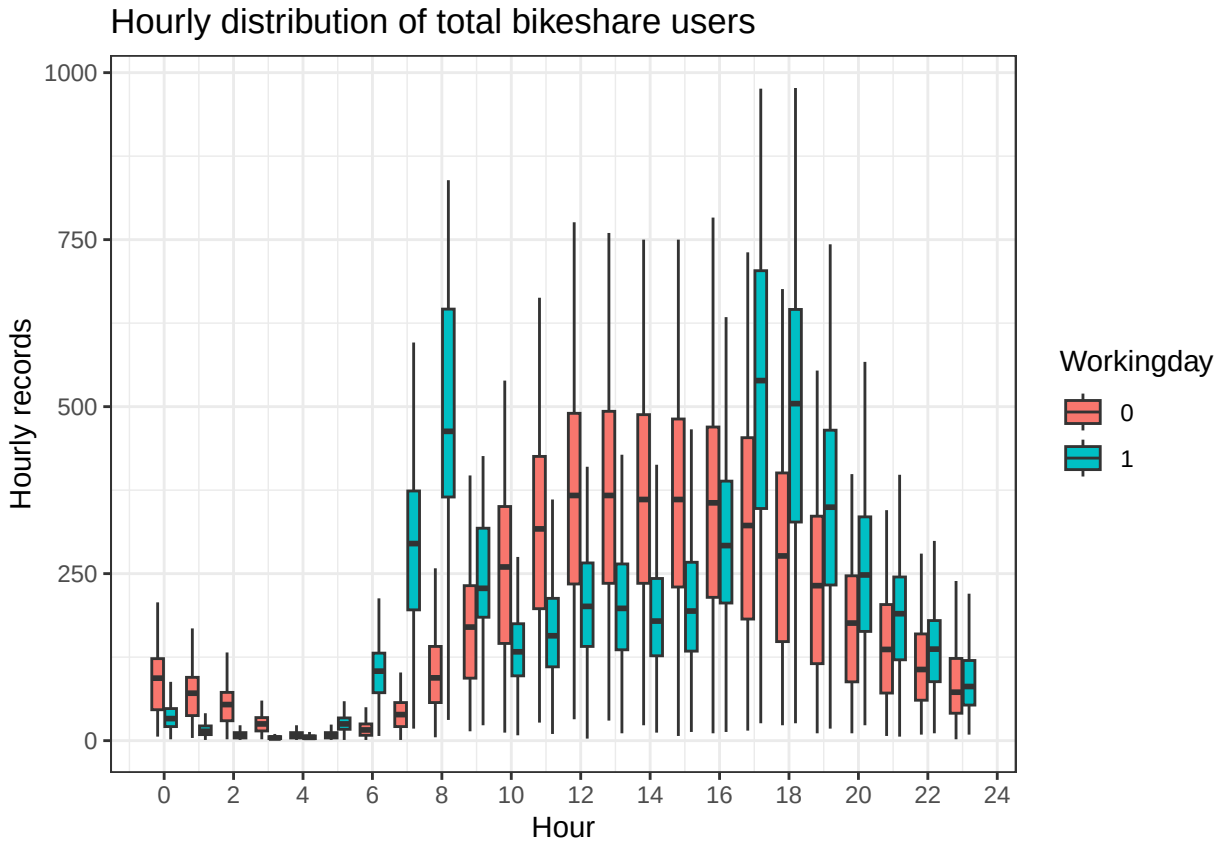
```
# col_vec_scatter=colnames(bs_hour)[10:13]
# for (value in col_vec_scatter) {
#     print(ggplot(bs_hour)+geom_point(aes_string(value,'registered'))+theme_bw()+
#             theme(axis.title.y=element_text(size=16),
#         axis.title.x=element_text(size=16)))
# }
# for (value in col_vec_box) {
#     print(ggplot(bs_hour)+geom_boxplot(aes_string(value,'registered',group=value))+theme_bw())+
#             theme(axis.title.y=element_text(size=16),
#         axis.title.x=element_text(size=16))
# }
```

2. Problems through the data and answering.

Question 1. Would the hourly distributions of bikeshare users on working days / non-working days different, and what about casual / registered users?

Total bikeshare users

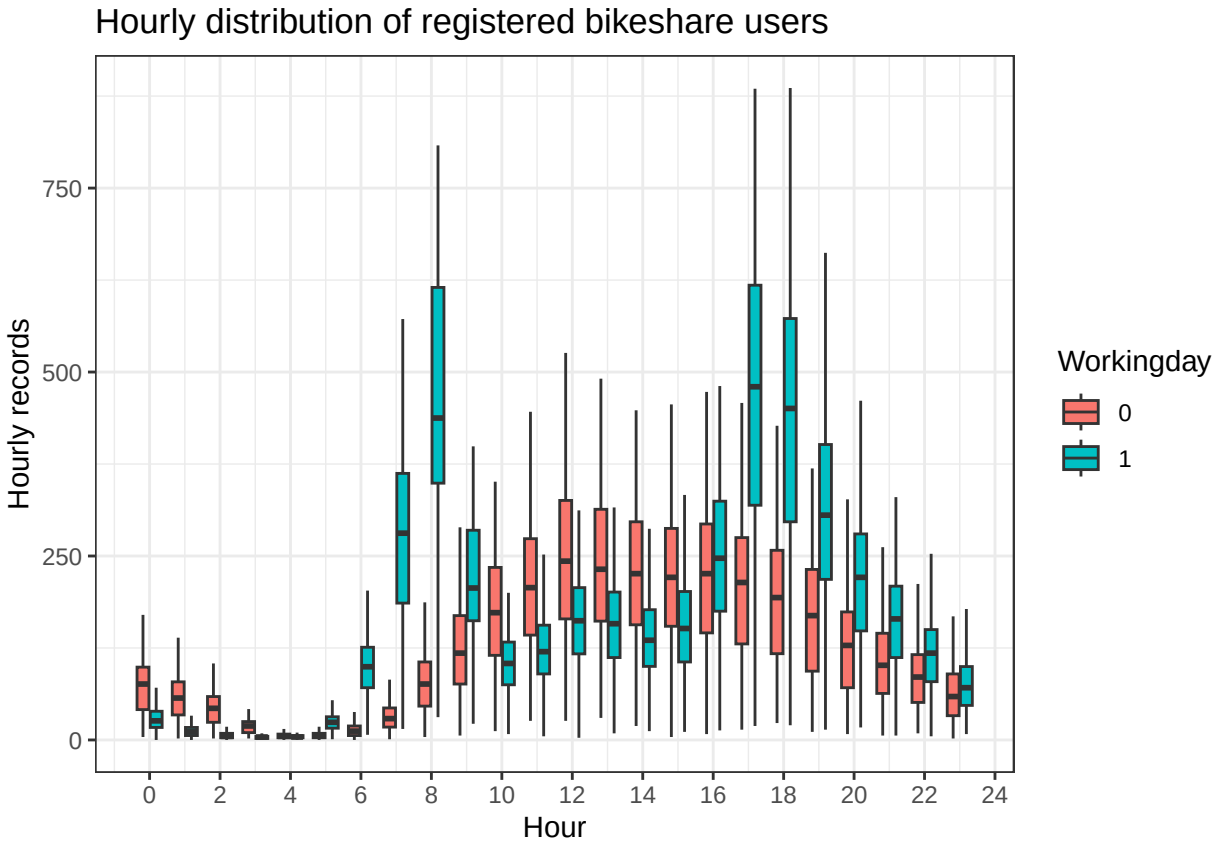
```
ggplot(bs_hour)+
  geom_boxplot(aes(hr,cnt,group=interaction(workingday,hr),fill=factor(workingday)),outlier.shape = NA)+
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Workingday',title='Hourly distribution of total bikeshare users')+
  scale_x_continuous(breaks=seq(0,24,2))
```



The answer is Yes.
Before 6am, there are a few bike share users for both working-day types while more people on non-working days tend to use bike share from 0am to 2am.
From 6am to 11pm, Two peaks of users are shown around 8am and 5pm on working days, which may reflect commuting during the rush hours. While on non-working days, we saw a smooth increasing then decreasing trend on bike share users.

Registered bikeshare users

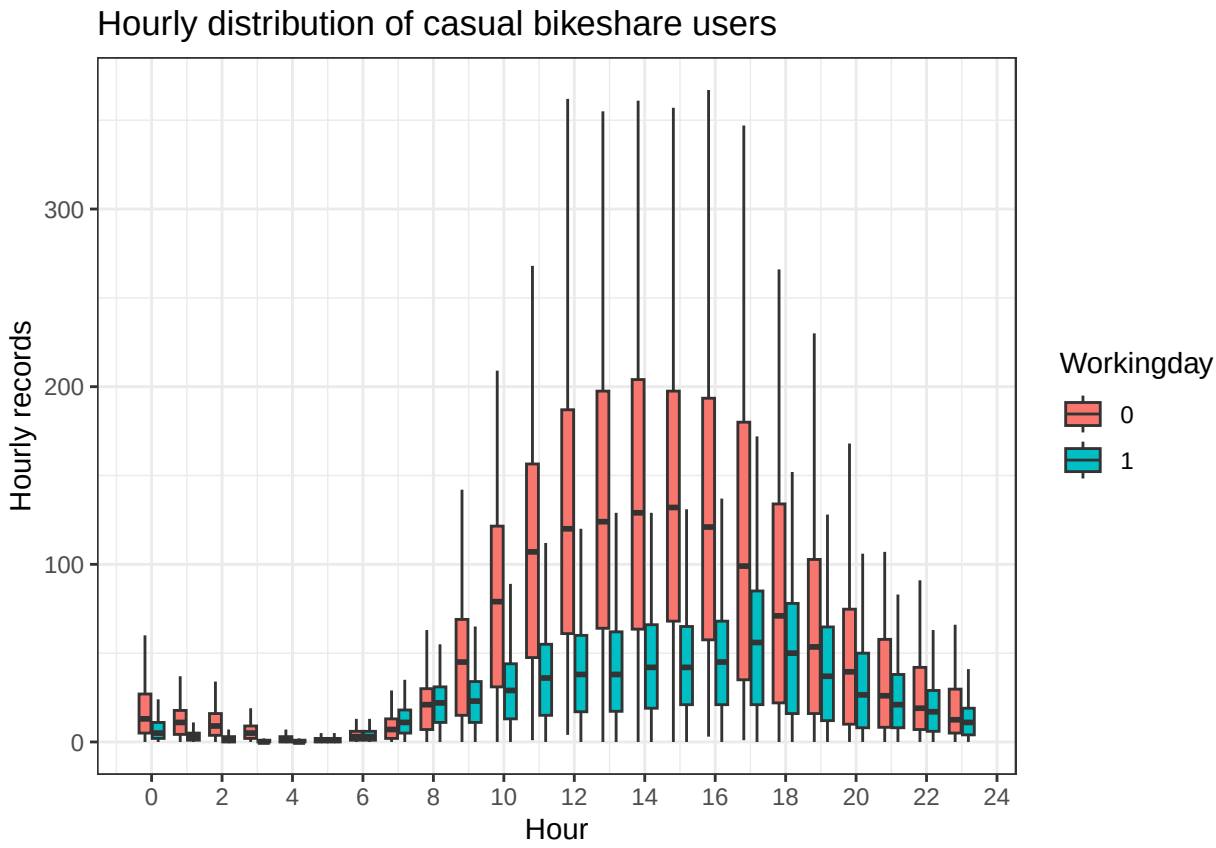
```
ggplot(bs_hour)+
  geom_boxplot(aes(hr,registered,group=interaction(workingday,hr),fill=factor(workingday)),outlier.shap
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Workingday',title='Hourly distribution of registered bikeshare users')+
  scale_x_continuous(breaks=seq(0,24,2))
```



The hourly distribution of registered users are quite like that of the total users.

Casual bikeshare users

```
ggplot(bs_hour)+
  geom_boxplot(aes(hr,casual,group=interaction(workingday,hr),fill=factor(workingday)),outlier.shape = N
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Workingday',title='Hourly distribution of casual bikeshare users')+
  scale_x_continuous(breaks=seq(0,24,2))
```

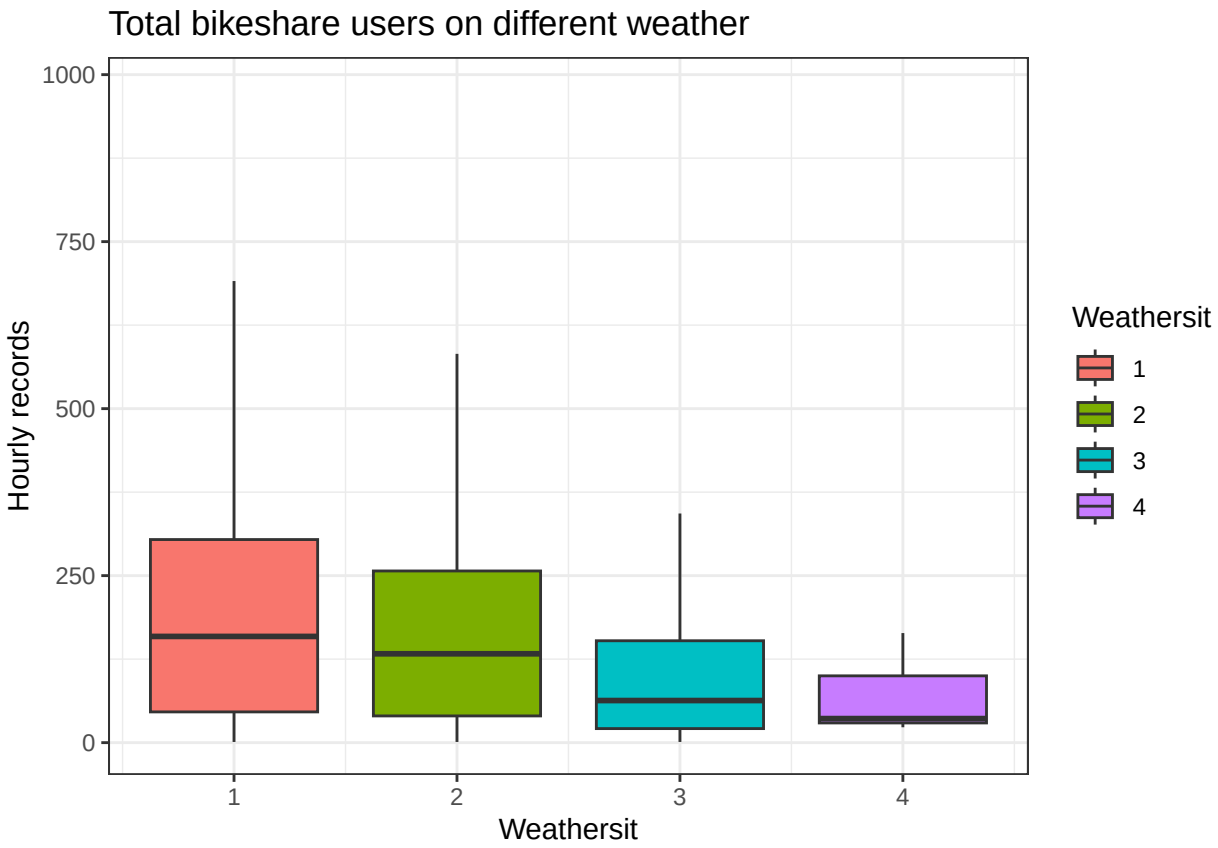


Casual users tend to use bikeshare more often on non-workding days while there is no strong evidence they would use bikeshare for commuting on rush hours.

Question 2. Would bikeshare users counts on different weather be different, and what about casual / registered users?

Total bikeshare users

```
ggplot(bs_hour)+
  geom_boxplot(aes(weathersit,cnt,group=factor(weathersit),fill=factor(weathersit)),outlier.shape = NA)+
  theme_bw()+
  xlab('Weathersit')+ylab('Hourly records')+
  labs(fill='Weathersit',title='Total bikeshare users on different weather')
```

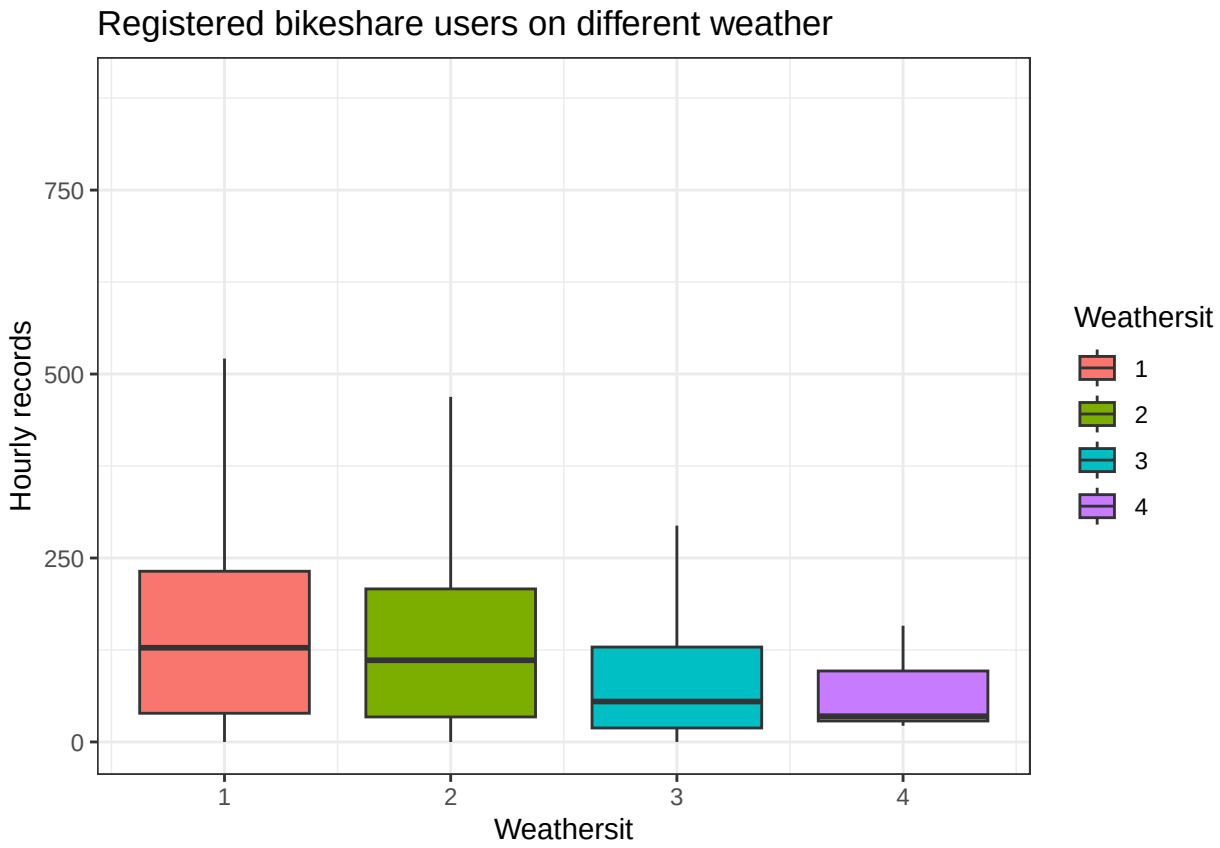


Here's the description for weather type 1 - 4: weathersit :
- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fogv

From these boxplots, it's obvious that from weather type 1 - 4, people are becoming more unwillingly to use bike share. This makes sense because commonly speaking, ordinary people would see weather type from 1 - 4 as weather getting worse. And for type 4, it's definitely bad weather.

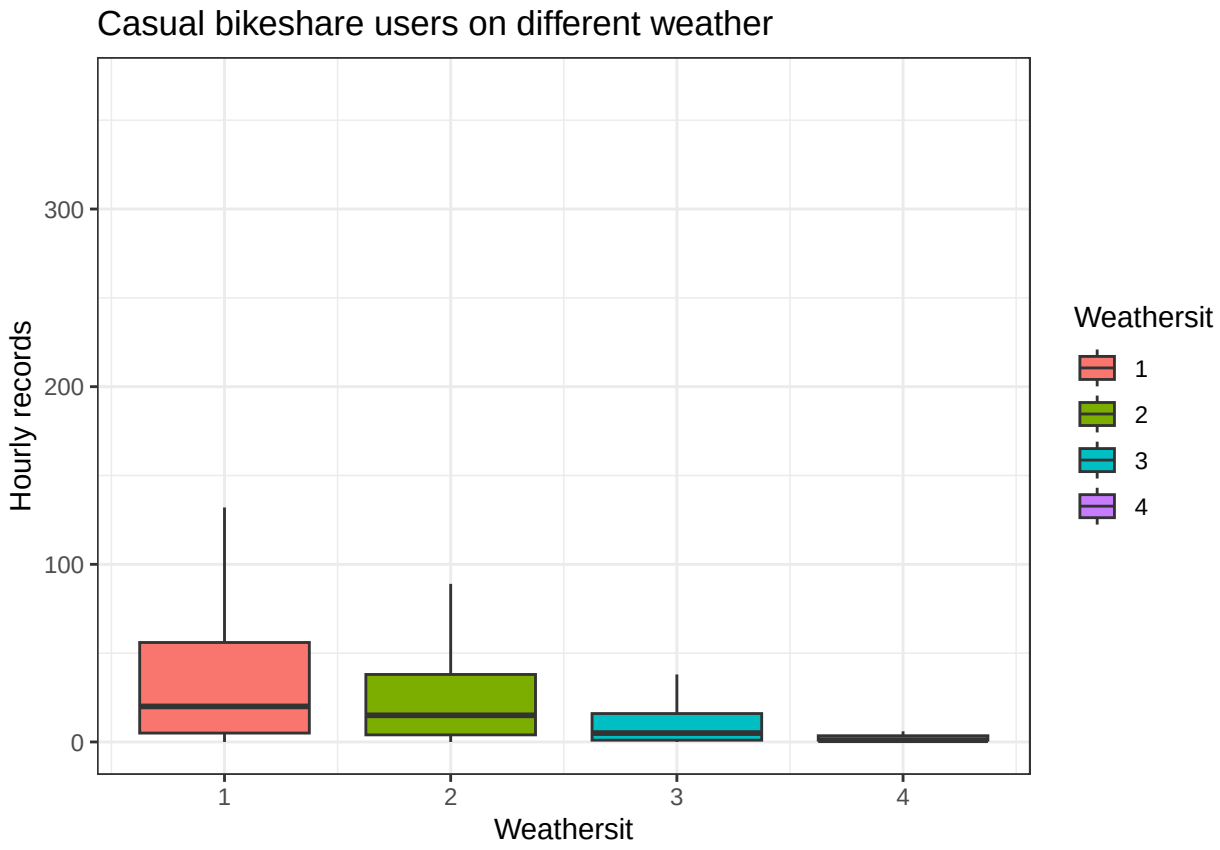
Registered bikeshare users

```
ggplot(bs_hour)+
  geom_boxplot(aes(weathersit,registered,group=factor(weathersit),fill=factor(weathersit)),outlier.shape
  theme_bw()+
  xlab('Weathersit')+ylab('Hourly records')+
  labs(fill='Weathersit',title='Registered bikeshare users on different weather')
```



Casual bikeshare users

```
ggplot(bs_hour)+
  geom_boxplot(aes(weathersit,casual,group=factor(weathersit),fill=factor(weathersit)),outlier.shape = N
  theme_bw()+
  xlab('Weathersit')+ylab('Hourly records')+
  labs(fill='Weathersit',title='Casual bikeshare users on different weather')
```



Question 3s. Digging deeper into rush hour patterns. Which would affect people's willingness of using bikeshare during rush hours?

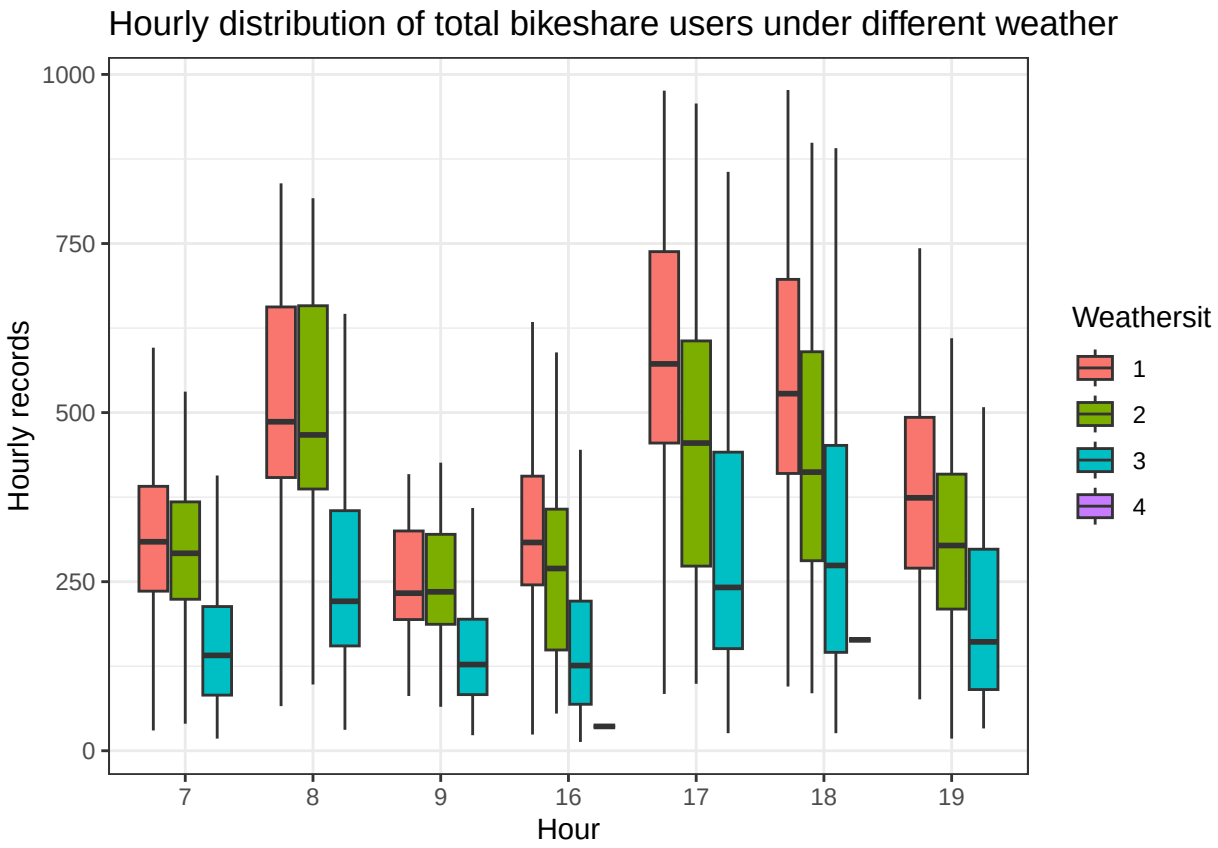
Question 3s - 1. Influence of weathertype

First let's find out those records during rush hours and likely to be considered as commuting.

```
bs_rush=bs_hour%>%filter(hr %in% c(7,8,9,16,17,18,19))%>%filter(workingday==1)
bs_rush$hr=as.factor(bs_rush$hr)
```

Total bikeshare users

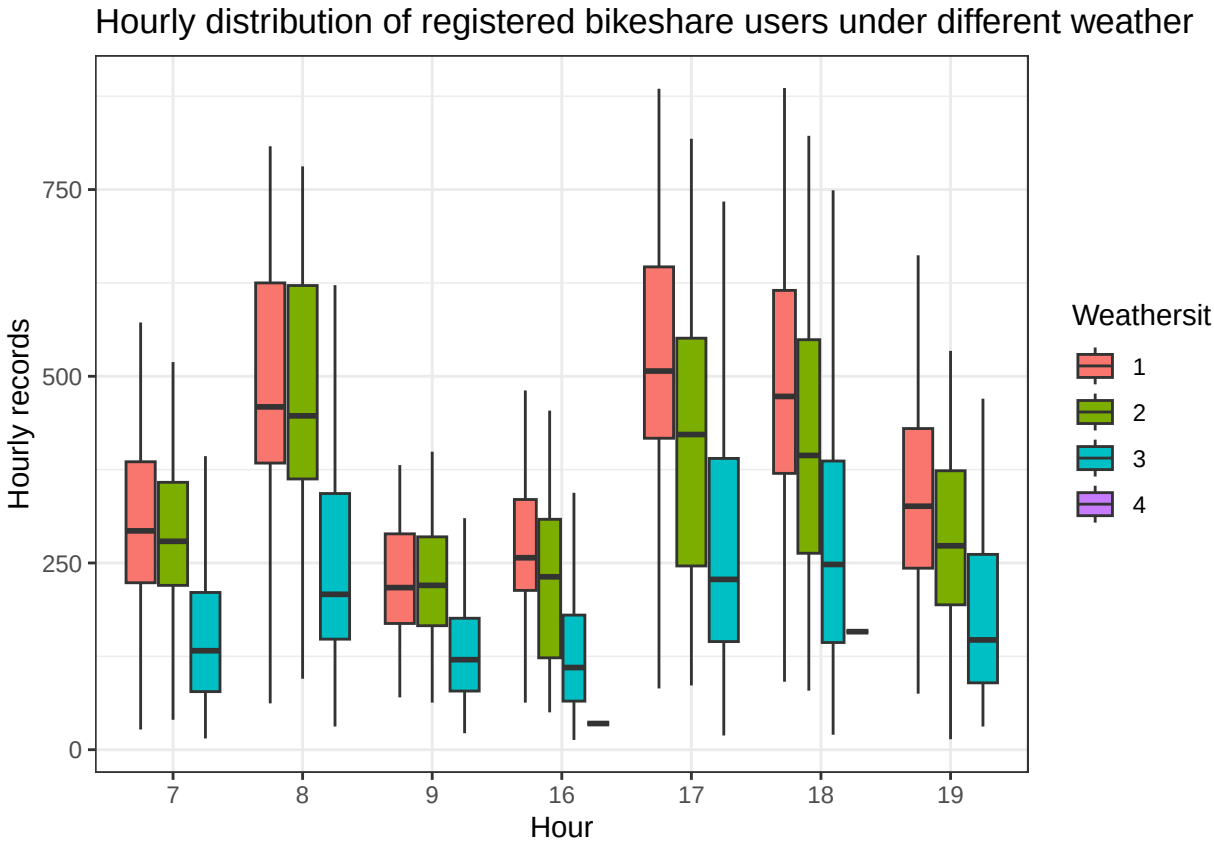
```
ggplot(bs_rush)+
  geom_boxplot(aes(hr,cnt,group=interaction(weathersit,hr),fill=factor(weathersit)),outlier.shape = NA)+
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Weathersit',title='Hourly distribution of total bikeshare users under different weather')
```



During the rush hours in the morning, weather type 1 and 2 seems have similar pattern, a little cloudy weather won't affect people's choice in the morning. Only when weather getting worse to type 3 or 4, people would not willing to ride a bike.
However during the evening rush hours, things are different. As the weather getting worse, even a little cloudy would make people not willing to ride a bike.
What about for casual or registered users?

Registered bikeshare users

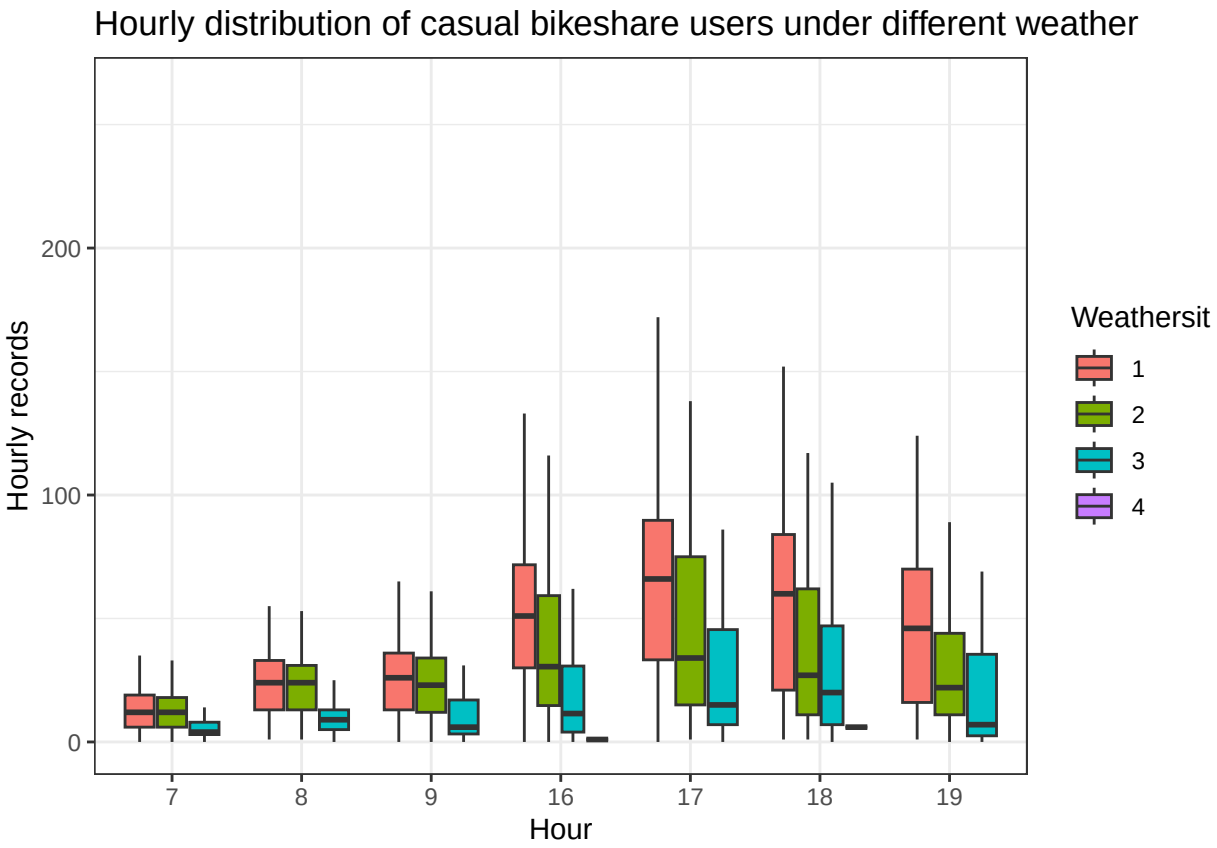
```
ggplot(bs_rush)+
  geom_boxplot(aes(hr,registered,group=interaction(weathersit,hr),fill=factor(weathersit)),outlier.shape
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Weathersit',title='Hourly distribution of registered bikeshare users under different weathe
```



Similar to total users.

Casual bikeshare users

```
ggplot(bs_rush)+
  geom_boxplot(aes(hr,casual,group=interaction(weathersit,hr),fill=factor(weathersit)),outlier.shape = N
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Weathersit',title='Hourly distribution of casual bikeshare users under different weather')
```



Weather influences are similar to total users. While as shown in previous hourly distribution of casual users, casual users are more willing to use bikehsare in the afternoon.

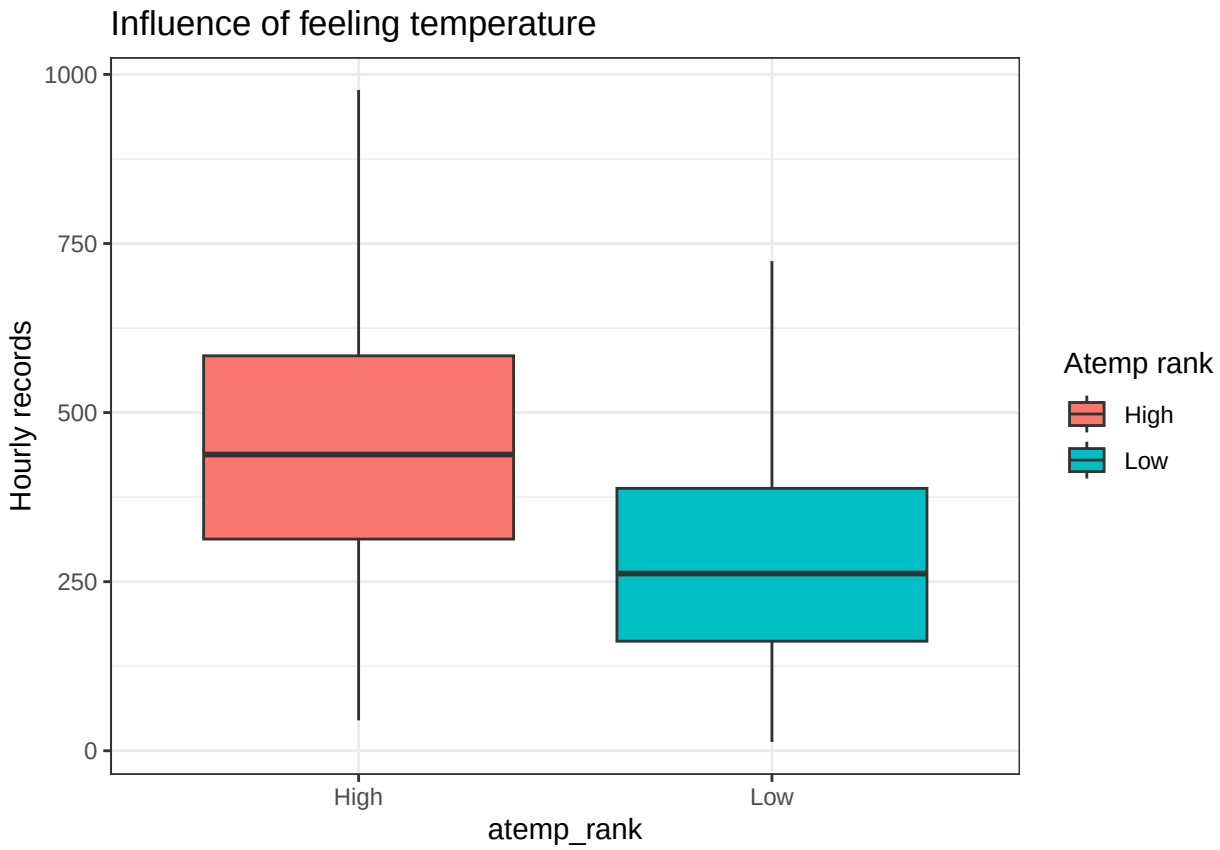
Question 3s - 2. Influence of feeling temperature, humidity and windspeed

In this dataset, all the values of meteorological data are normalized. Which makes us easier to classify them into two levels: high and low (or another medium level).

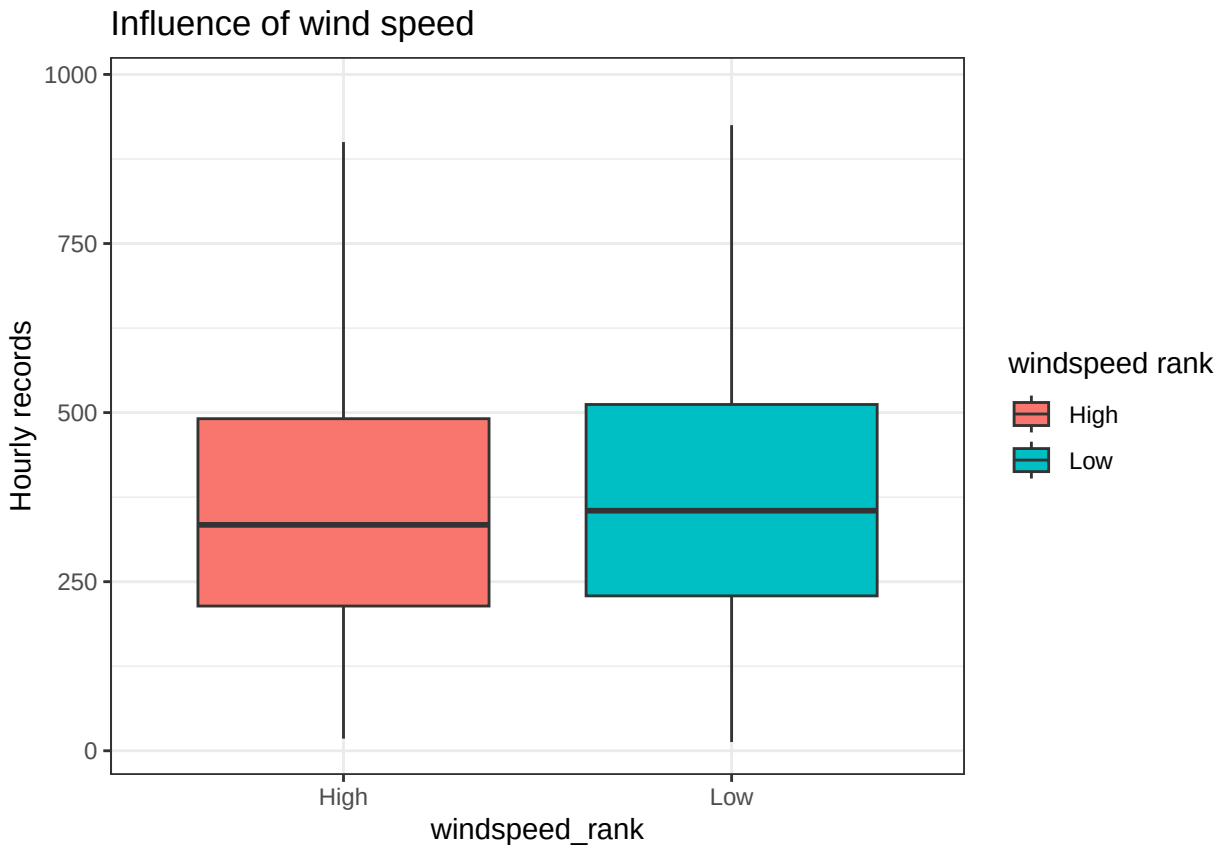
```
f=function(x,a){
  (max(x)-min(x))*a+min(x)
}
bs_rush$atemp_rank=ifelse(bs_rush$atemp>=f(bs_rush$atemp,0.5),'High','Low')%>%as.factor()
bs_rush$windspeed_rank=ifelse(bs_rush$windspeed>=f(bs_rush$windspeed,0.5),'High','Low')%>%as.factor()
bs_rush$humi_rank=ifelse(bs_rush$hum>=f(bs_rush$hum,0.5),'High','Low')%>%as.factor()
```

Total bikeshare users

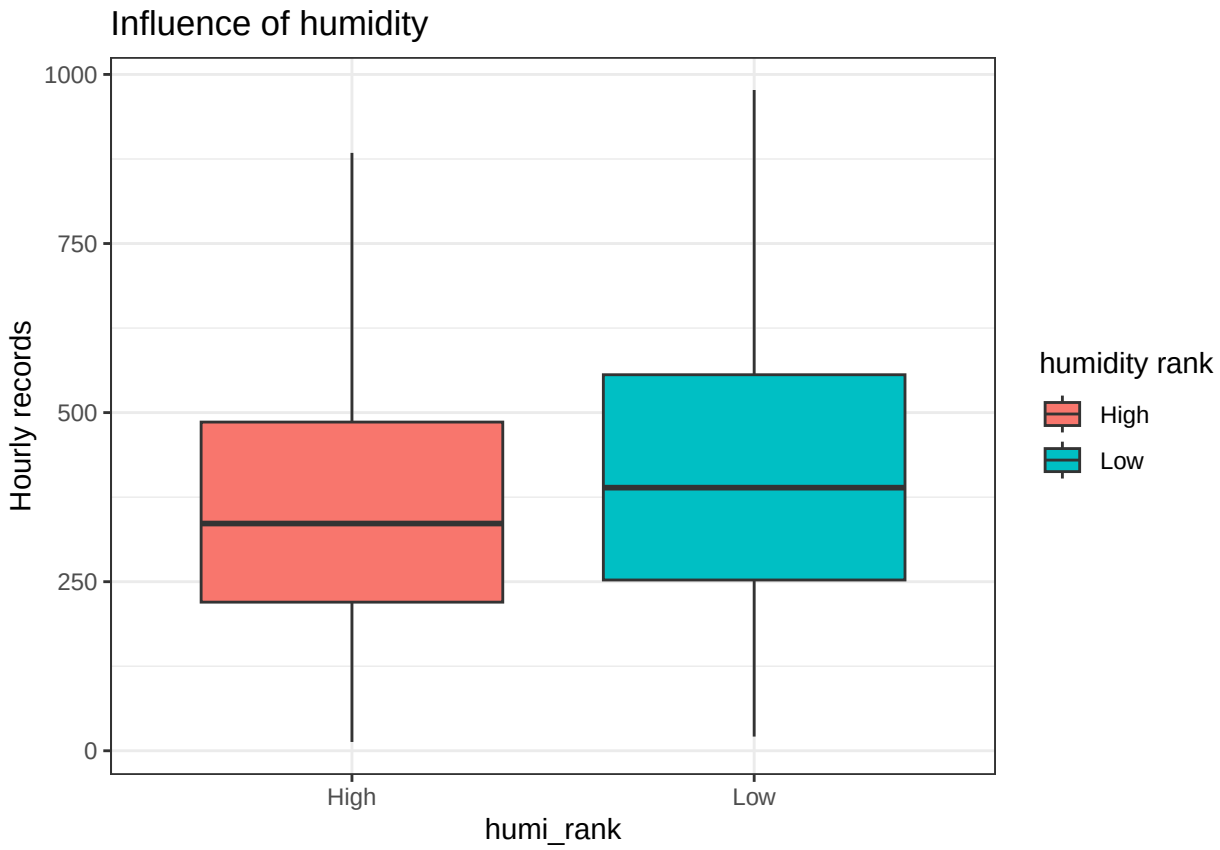
```
ggplot(bs_rush)+
  geom_boxplot(aes(atemp_rank,cnt,
                   group=atemp_rank,fill=factor(atemp_rank)),outlier.shape = NA)+
  theme_bw()+
  xlab('atemp_rank')+ylab('Hourly records')+
  labs(fill='Atemp rank',title='Influence of feeling temperature')#+guides(fill=guide_legend(ncol=2))
```




```
ggplot(bs_rush)+
  geom_boxplot(aes(windspeed_rank,cnt,
                   group=windspeed_rank,fill=factor(windspeed_rank)),outlier.shape = NA)+
  theme_bw()+
  xlab('windspeed_rank')+ylab('Hourly records')+
  labs(fill='windspeed rank',title='Influence of wind speed')#+guides(fill=guide_legend(ncol=2))
```




```
ggplot(bs_rush)+
  geom_boxplot(aes(humi_rank,cnt,
                   group=humi_rank,fill=factor(humi_rank)),outlier.shape = NA)+
  theme_bw()+
  xlab('humi_rank')+ylab('Hourly records')+
  labs(fill='humidity rank',title='Influence of humidity')#+guides(fill=guide_legend(ncol=2))
```



3. Random forest regression to predict hourly bikeshare user counts

```
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

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

## The following object is masked from 'package:ggplot2':
##
##     margin
```

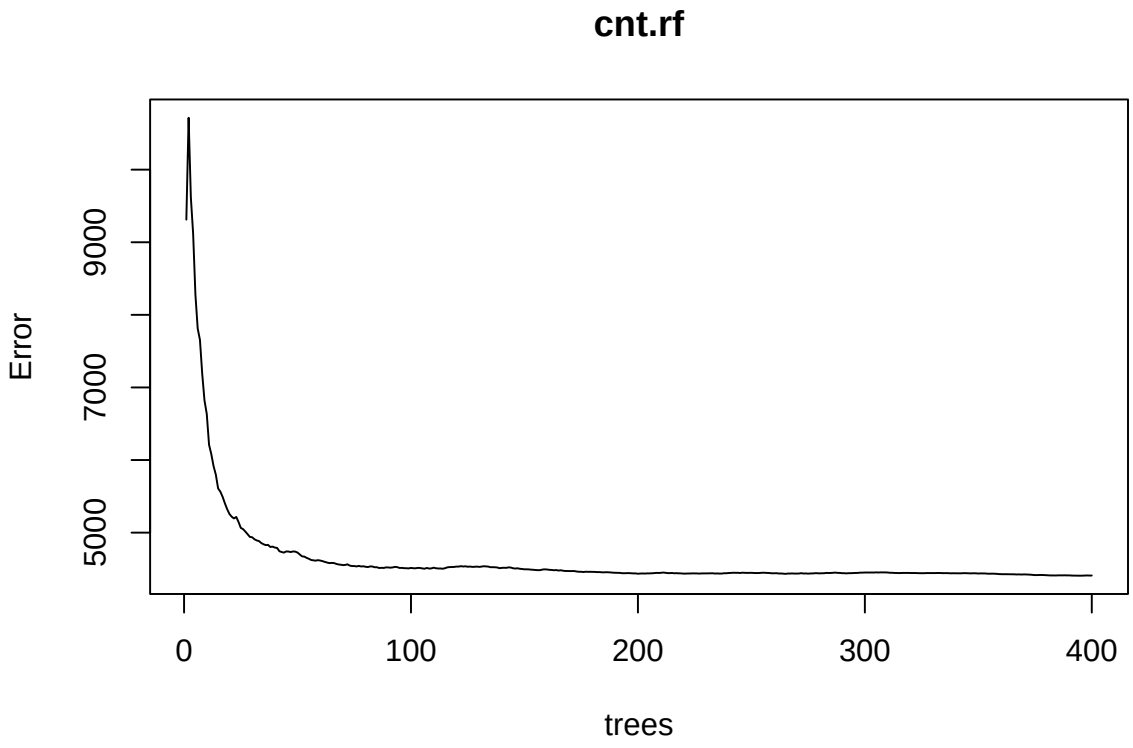
```
model_table=bs_rush[,-c(1,6,8,14,15,16)]
str(model_table)
```

```
## tibble [3,482 x 14] (S3: tbl_df/tbl/data.frame)
##  $ season        : num [1:3482] 1 1 1 1 1 1 1 1 1 1 ...
##  $ yr            : num [1:3482] 0 0 0 0 0 0 0 0 0 0 ...
##  $ mnth          : num [1:3482] 1 1 1 1 1 1 1 1 1 1 ...
##  $ hr            : Factor w/ 7 levels "7","8","9","16",..: 1 2 3 4 5 6 7 1 2 3 ...
##  $ weekday       : num [1:3482] 1 1 1 1 1 1 1 2 2 2 ...
##  $ weathersit    : num [1:3482] 1 1 1 1 1 1 1 1 1 1 ...
##  $ temp          : num [1:3482] 0.14 0.14 0.16 0.26 0.24 0.24 0.2 0.12 0.14 0.16 ...
##  $ atemp         : num [1:3482] 0.136 0.121 0.136 0.242 0.227 ...
##  $ hum           : num [1:3482] 0.5 0.5 0.43 0.3 0.3 0.32 0.47 0.74 0.69 0.64 ...
##  $ windspeed     : num [1:3482] 0.194 0.284 0.388 0.254 0.224 ...
##  $ cnt           : num [1:3482] 64 154 88 76 157 157 110 94 179 100 ...
##  $ atemp_rank    : Factor w/ 2 levels "High","Low": 2 2 2 2 2 2 2 2 2 2 ...
##  $ windspeed_rank: Factor w/ 2 levels "High","Low": 2 2 1 2 2 2 2 2 2 2 ...
##  $ humi_rank     : Factor w/ 2 levels "High","Low": 1 1 2 2 2 2 2 1 1 1 ...
```

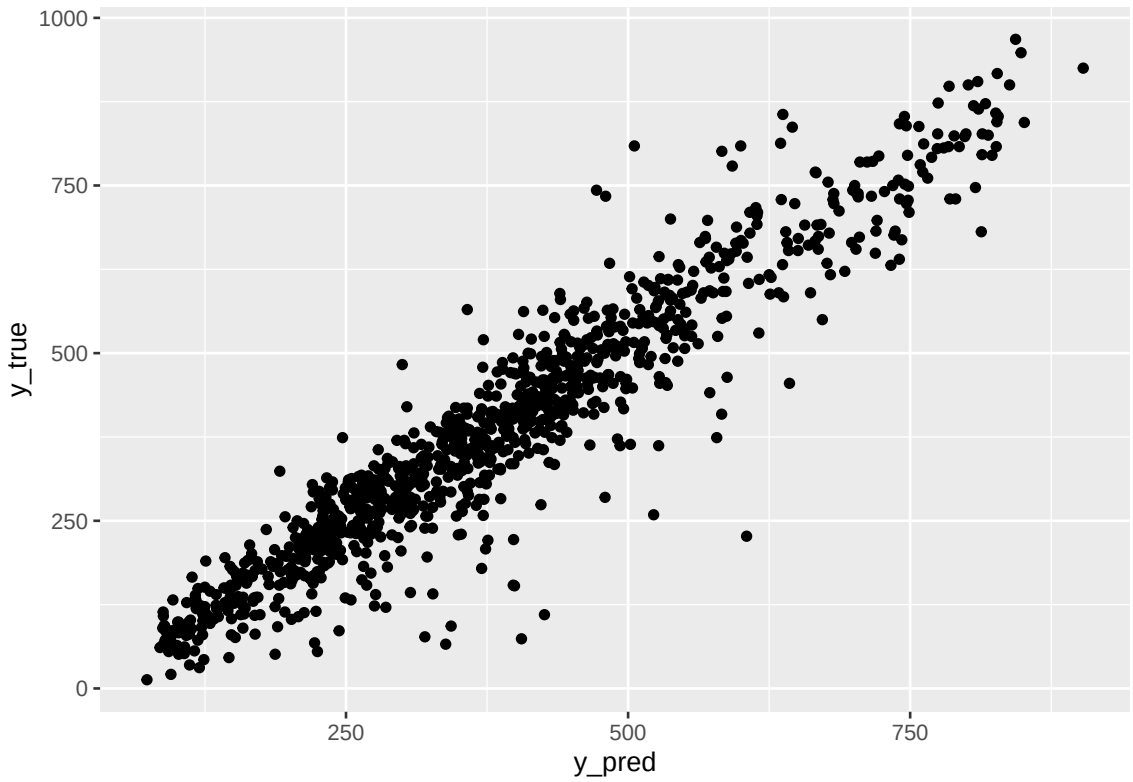
```
model_table$season=as.factor(model_table$season)
model_table$yr=as.factor(model_table$yr)
model_table$mnth=as.factor(model_table$mnth)
model_table$weekday=as.factor(model_table$weekday)
model_table$weathersit=as.factor(model_table$weathersit)

set.seed(1)
train_rows=sample(nrow(model_table),dim(model_table)[1]*0.7)
train_set=model_table[train_rows,]
test_set=model_table[-train_rows,]
test_x=test_set[,-11]
test_y=test_set[,11]
cnt.rf <- randomForest(cnt ~ ., data = train_set,
                       importance = TRUE,ntree=400,xtest=test_x,ytest=test_y$cnt,
                       keep.forest=TRUE)
print(cnt.rf)
```

```
##
## Call:
##  randomForest(formula = cnt ~ ., data = train_set, importance = TRUE,      ntree = 400, xtest = test_
##                Type of random forest: regression
##                      Number of trees: 400
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 4409.085
##                     % Var explained: 89.3
##                        Test set MSE: 4080.91
##                     % Var explained: 89.41
```

```
plot(cnt.rf)
```




```
y_pred=predict(cnt.rf,test_x)
table=cbind(y_pred,test_y$cnt)%>%as.data.frame()
colnames(table)=c('y_pred','y_true')
ggplot(table)+geom_point(aes(y_pred,y_true))
```



```r
#30 most important attributes
varImpPlot(cnt.rf, n.var = min(30, nrow(cnt.rf$importance)),
main = 'Top 30 - variable importance')
```