Final Project - Group 14

2022-11-28

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()
```

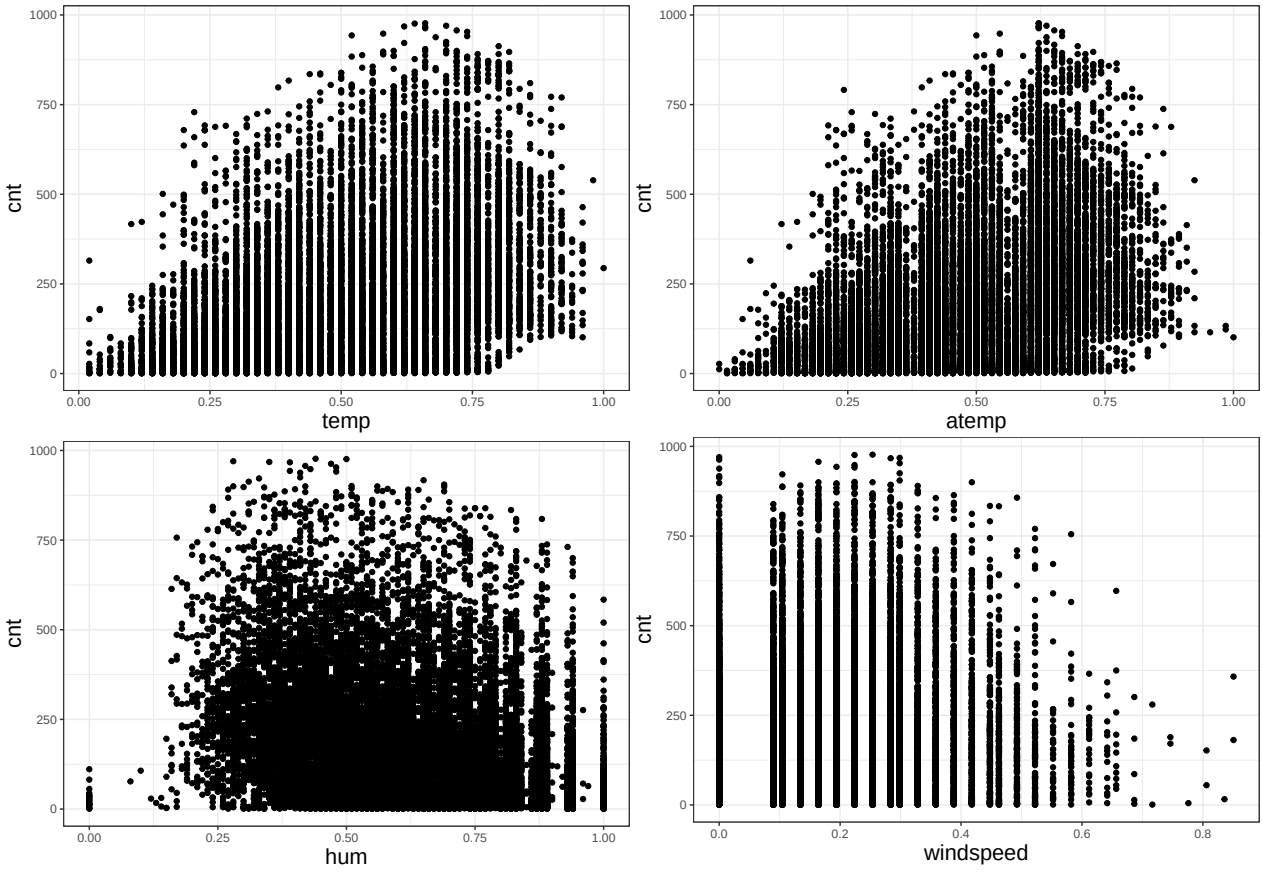
```
## [1] 0
```

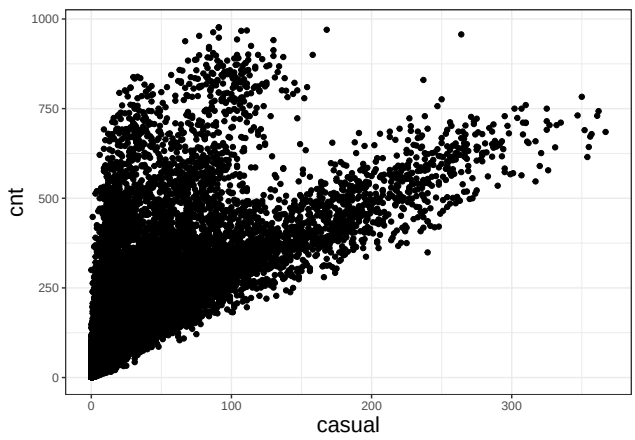
```
#Avoid changing original data
bs_hour=origin_data%>%mutate(dteday=as.Date(dteday))%>%select(-instant)
#Add one hourly identifiable column to identify every row
bs_hour=bs_hour%>%mutate(hourly_id=paste(as.character(dteday),as.character(hr)))%>%mutate(hourly_id=ymd
bs_hour=bs_hour[,c(1:15,17,16)]
bs_hour$windspeed=as.numeric(bs_hour$windspeed)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

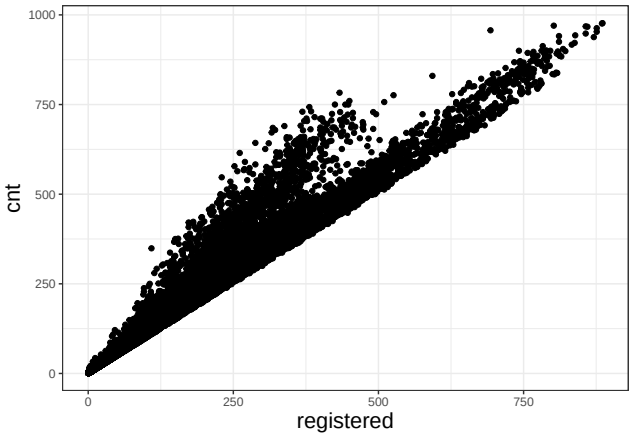
Scatter plots & Box plots for total counts.

```
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)))
}
```

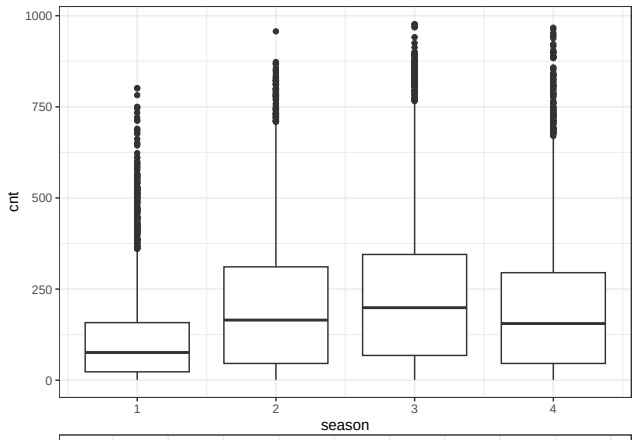
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation ideoms with `aes()`
```

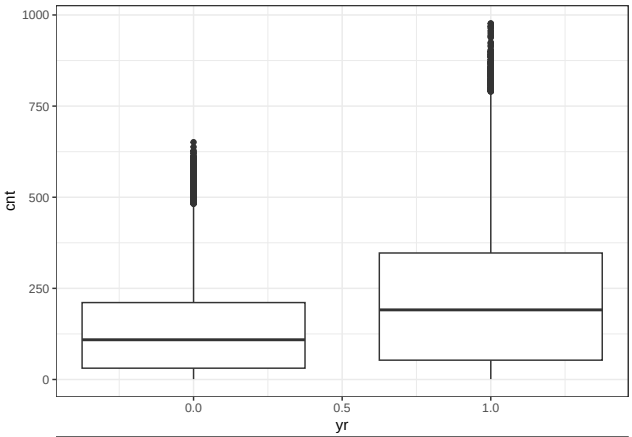


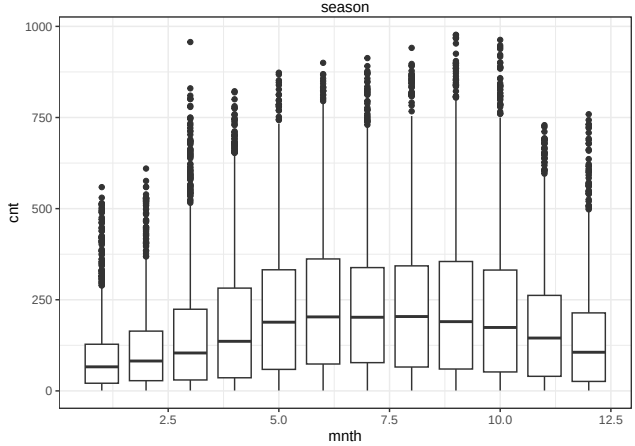


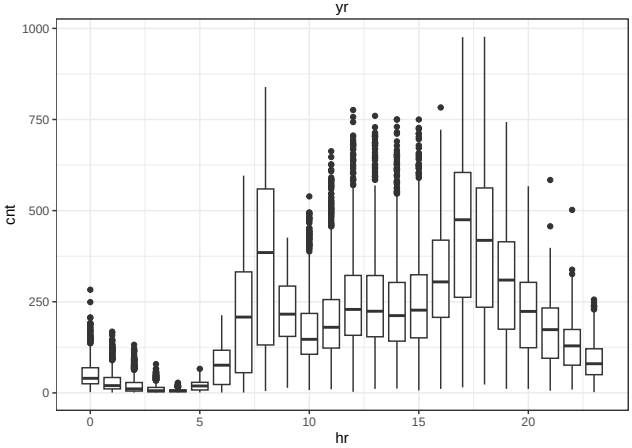


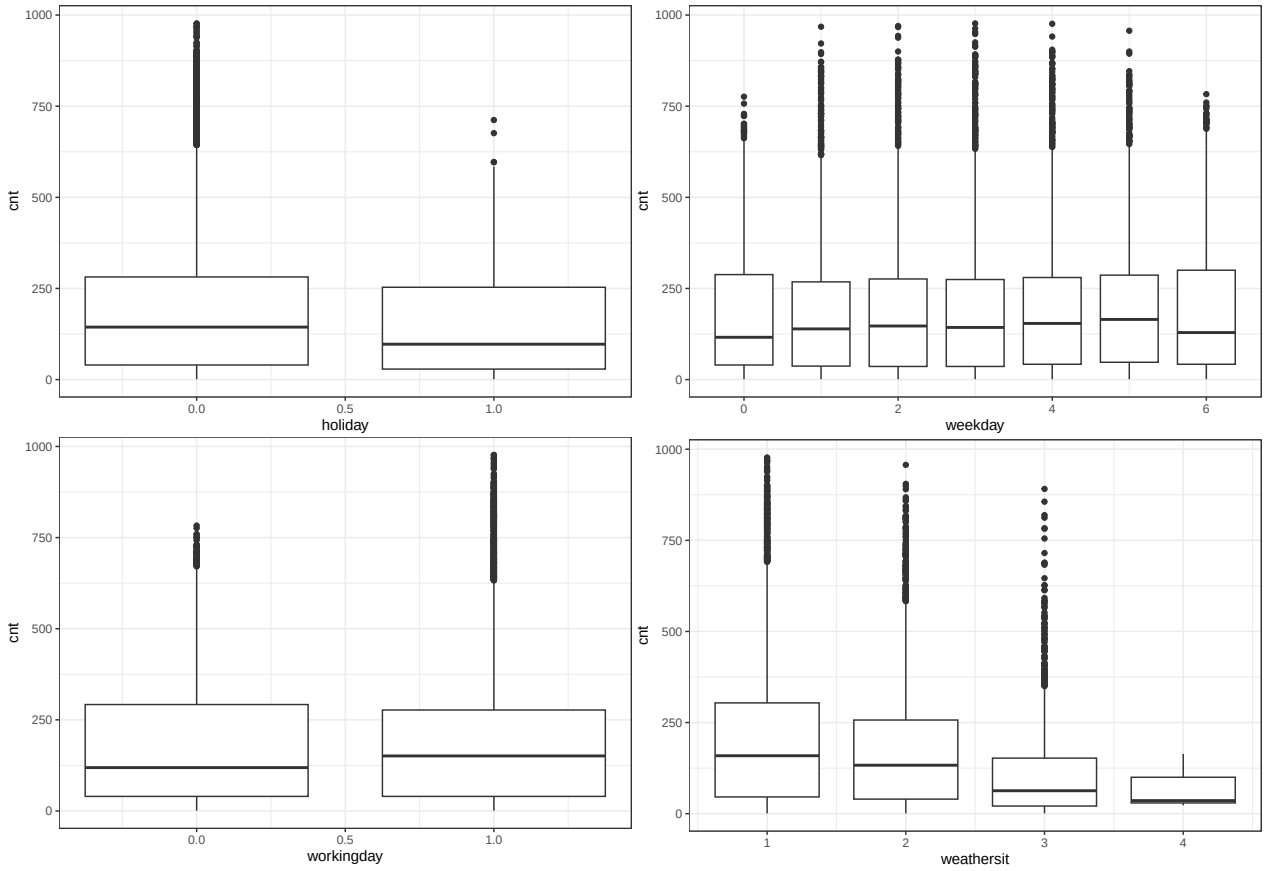

```
for (value in col_vec_box) {
    print(ggplot(bs_hour)+geom_boxplot(aes_string(value,'cnt',group=value))+theme_bw())+
            theme(axis.title.y=element_text(size=16),
        axis.title.x=element_text(size=16))
}
```



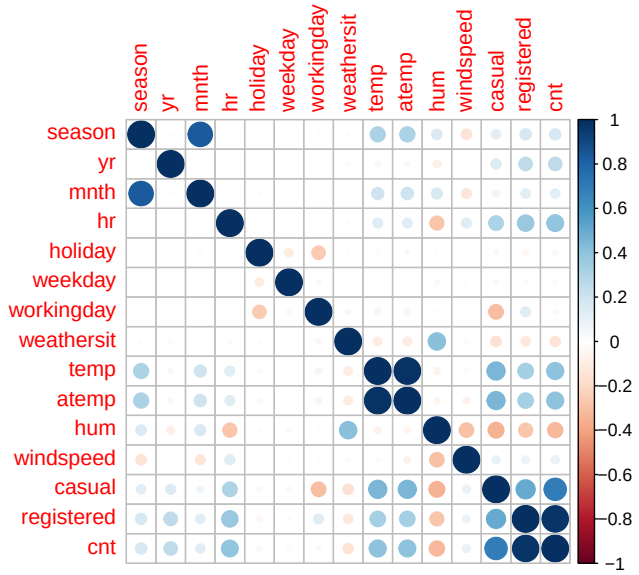









```
cor(bs_hour[, -c(1,16)]) %>%
  corrplot::corrplot()
```



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.

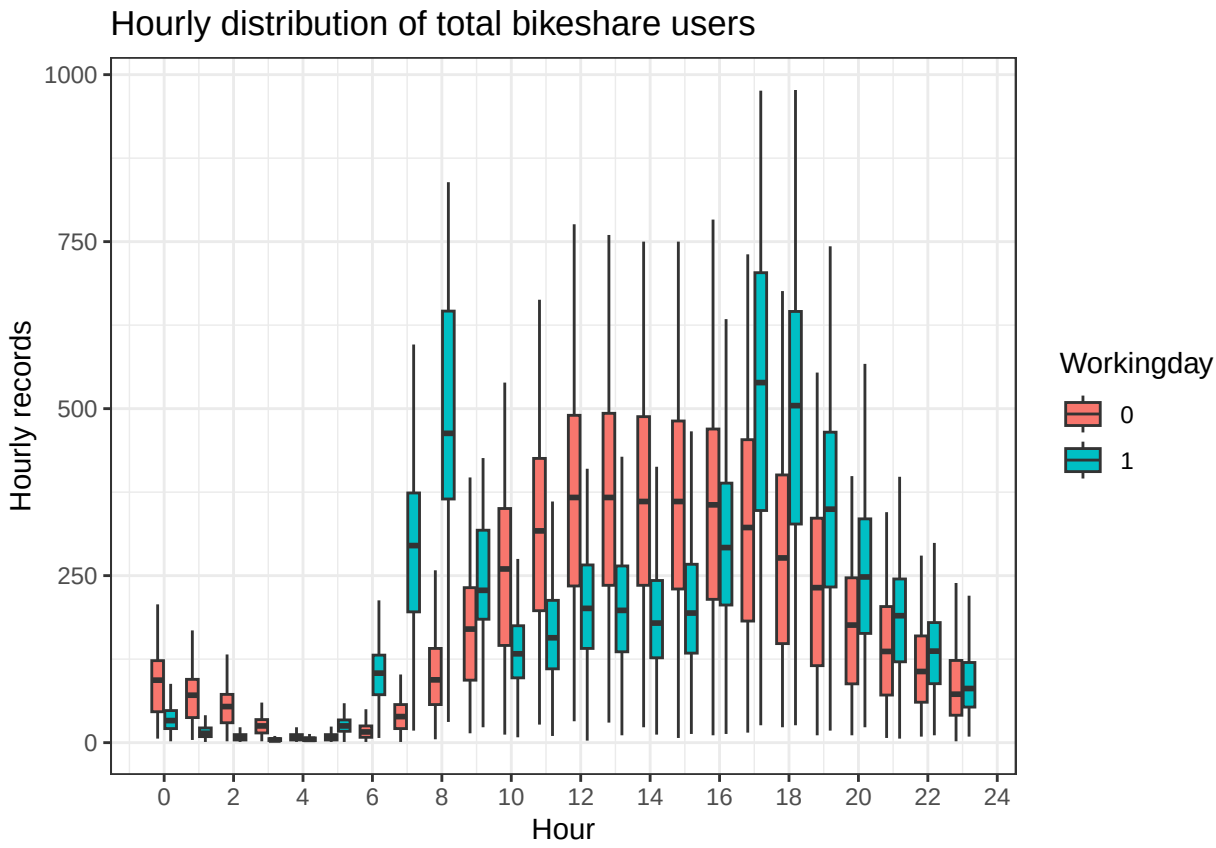
```
# col_vec_scatter=colnames(bs_hour)[10:13]
# for (value in col_vec_scatter) {
#     print(ggplot(bs_hour)+geom_point(aes_string(value,'registered'))+theme_bw()+
#             theme(axis.title.y=element_text(size=16),
#         axis.title.x=element_text(size=16)))
# }
# for (value in col_vec_box)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

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

Total bikeshare users

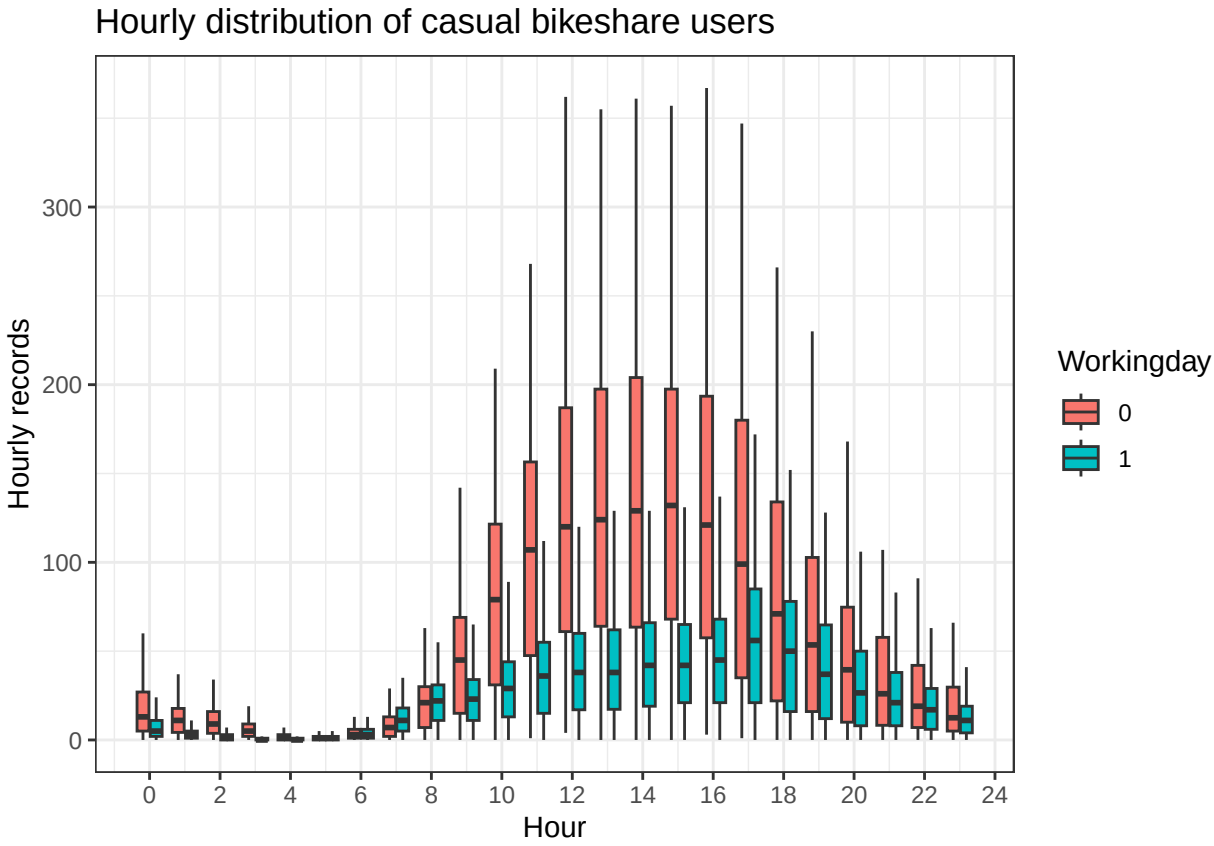
```
ggplot(bs_hour)+
  geom_boxplot(aes(hr,cnt,group=interaction(workingday,hr),fill=factor(workingday)),outlier.shape = NA)+
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Workingday',title='Hourly distribution of total bikeshare users')+
  scale_x_continuous(breaks=seq(0,24,2))
```



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.

Casual bikeshare users

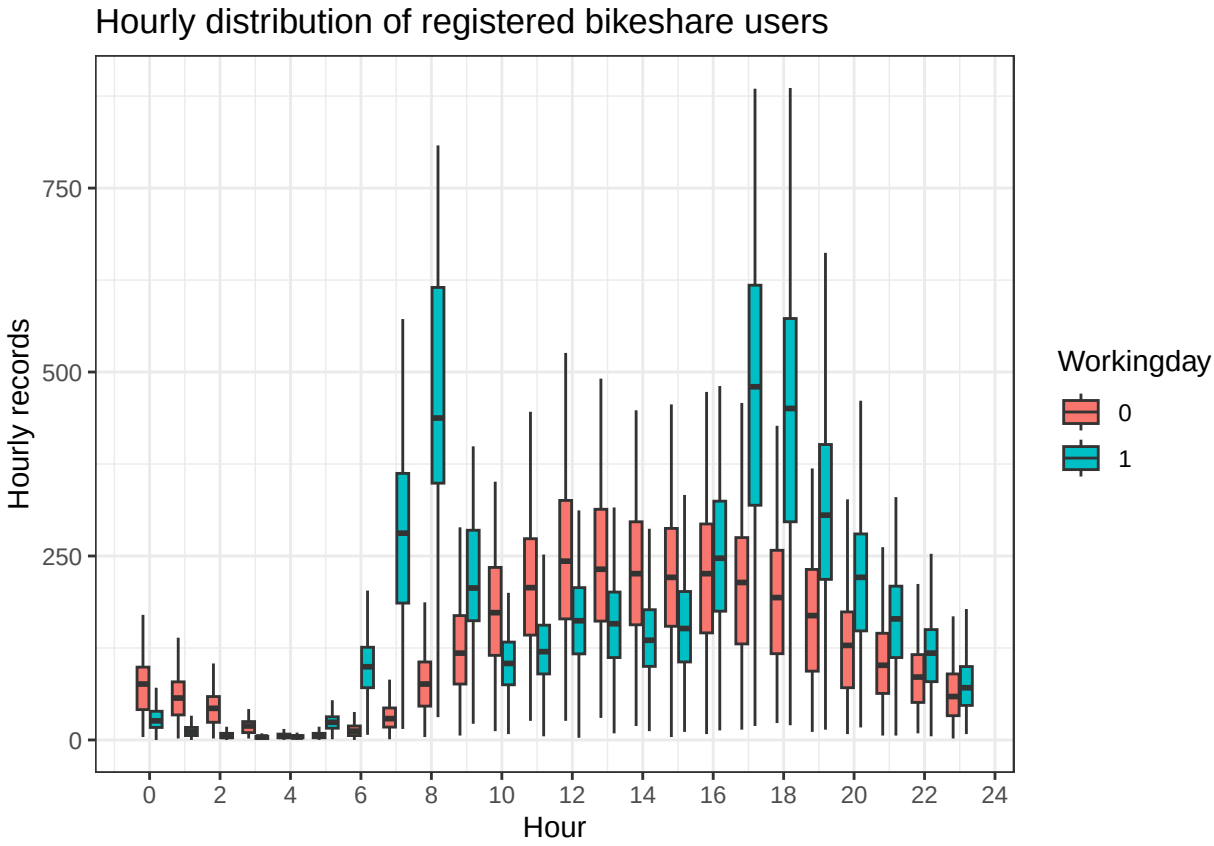
```
ggplot(bs_hour)+
  geom_boxplot(aes(hr,casual,group=interaction(workingday,hr),fill=factor(workingday)),outlier.shape = N
  theme_bw()+
  xlab('Hour')+ylab('Hourly records')+
  labs(fill='Workingday',title='Hourly distribution of casual bikeshare users')+
  scale_x_continuous(breaks=seq(0,24,2))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



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.

Registered bikeshare users

```
ggplot(bs_hour)+
  geom_boxplot(aes(hr,registered,group=interaction(workingday,hr),fill=factor(workingday)),outlier.shape
  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.