

Final Project - Group 14

2022-11-09

1. “Read data and some plots”

Read data and simple processing.

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

## 
## 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

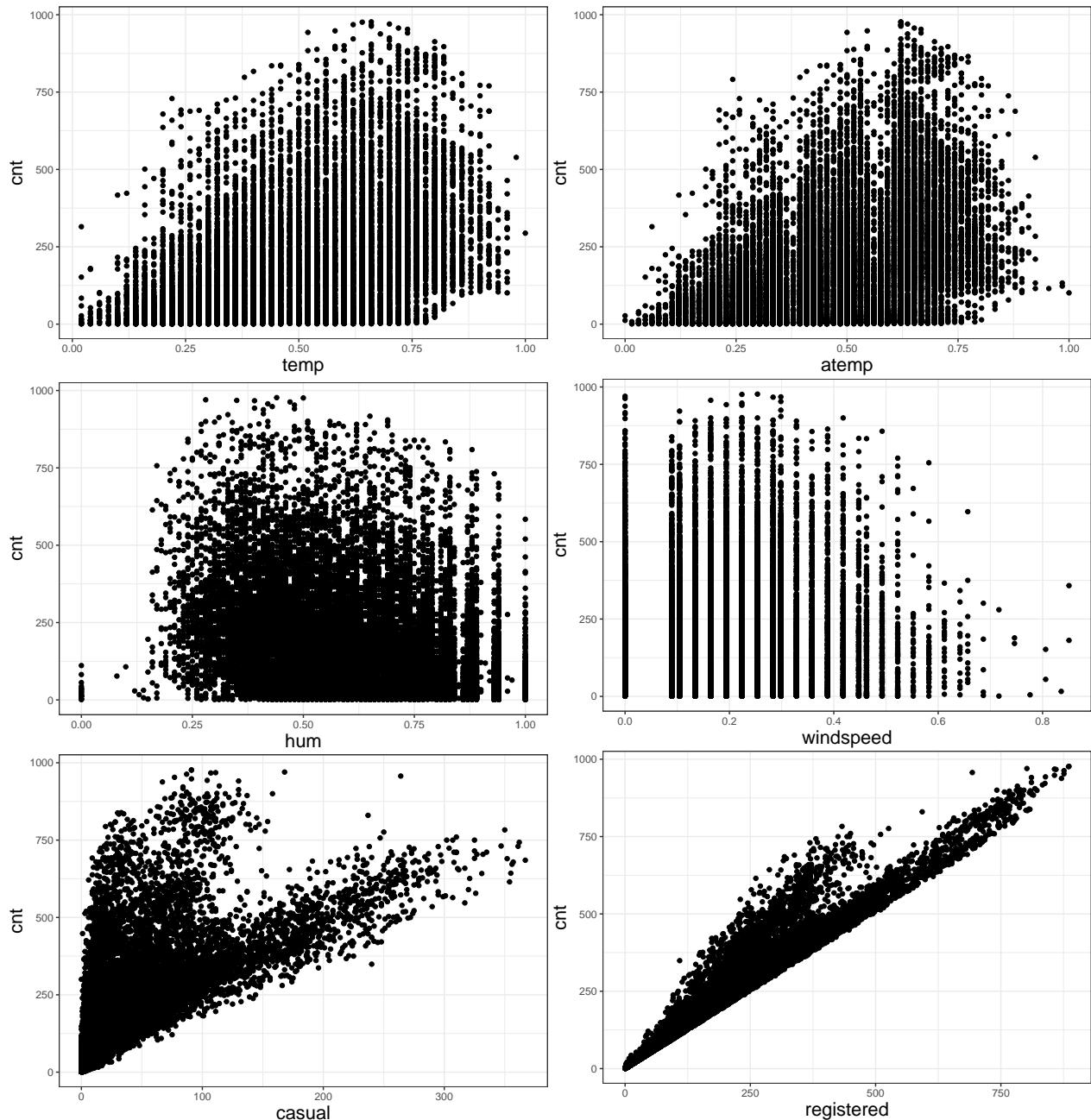
#Avoid changing original data
bs_hour=origin_data%>%mutate(dteday=as.Date(dteday))%>%select(-instant)
#Add one hourly identifiable column to identify every row
bs_hour=bs_hour%>%mutate(hourly_id=paste(as.character(dteday),as.character(hr)))%>%mutate(hourly_id=ymd(
bs_hour=bs_hour[,c(1:15,17,16)]
```

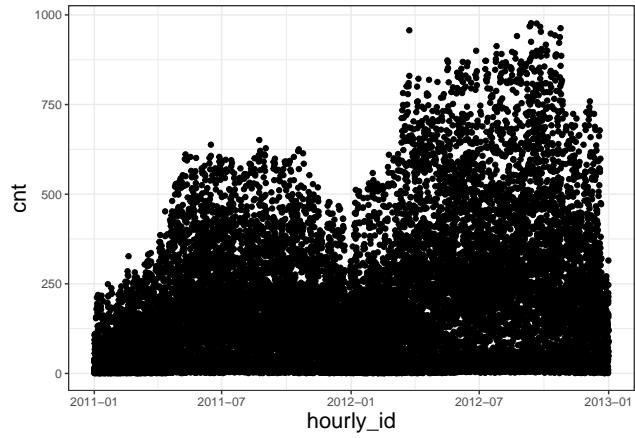
Scatter plots & Box plots.

```

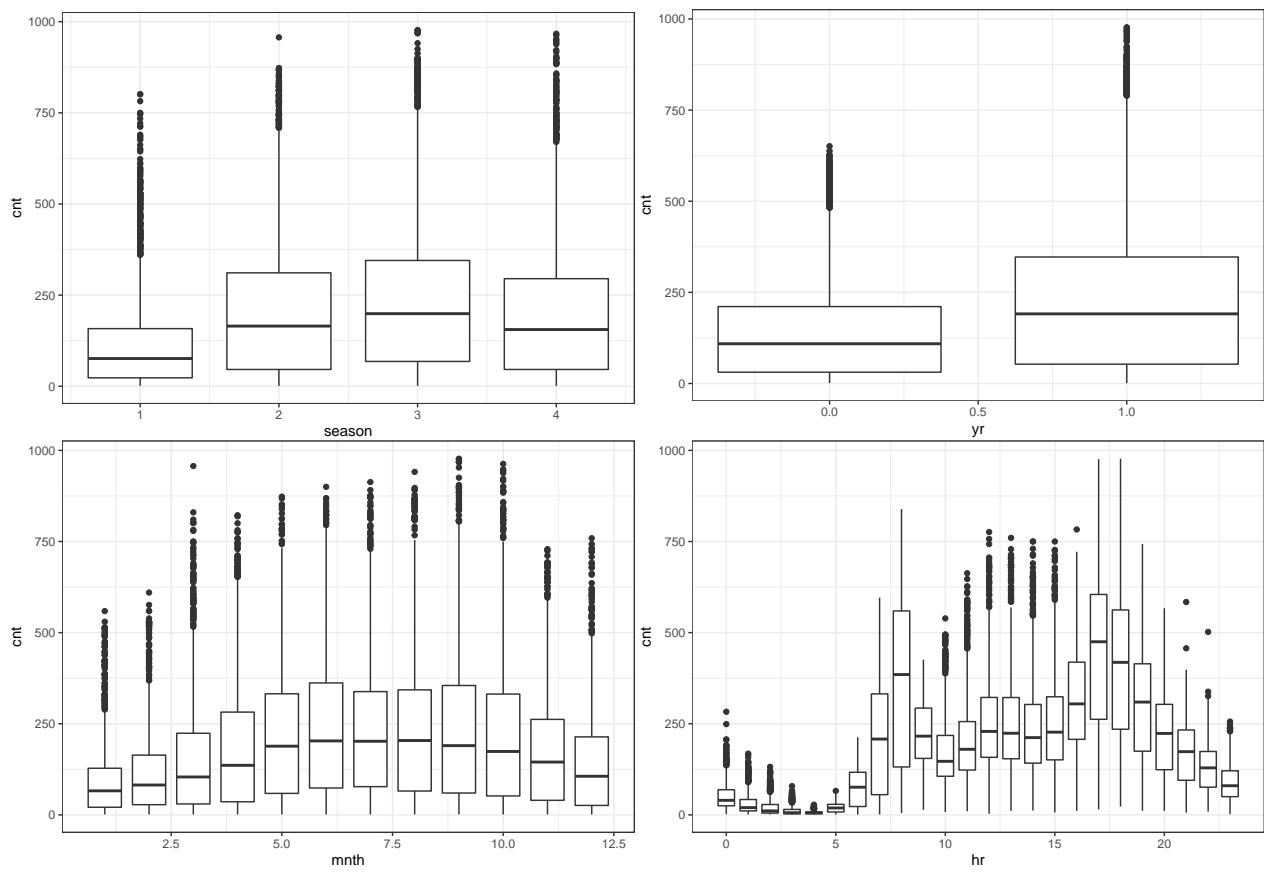
col_vec_scatter=colnames(bs_hour) [10:16]
col_vec_box=colnames(bs_hour) [2:9]
?xlab()
for (value in col_vec_scatter) {
  print(ggplot(bs_hour)+geom_point(aes_string(value, 'cnt'))+theme_bw()+
    theme(axis.title.y=element_text(size=16),
          axis.title.x=element_text(size=16)))
}

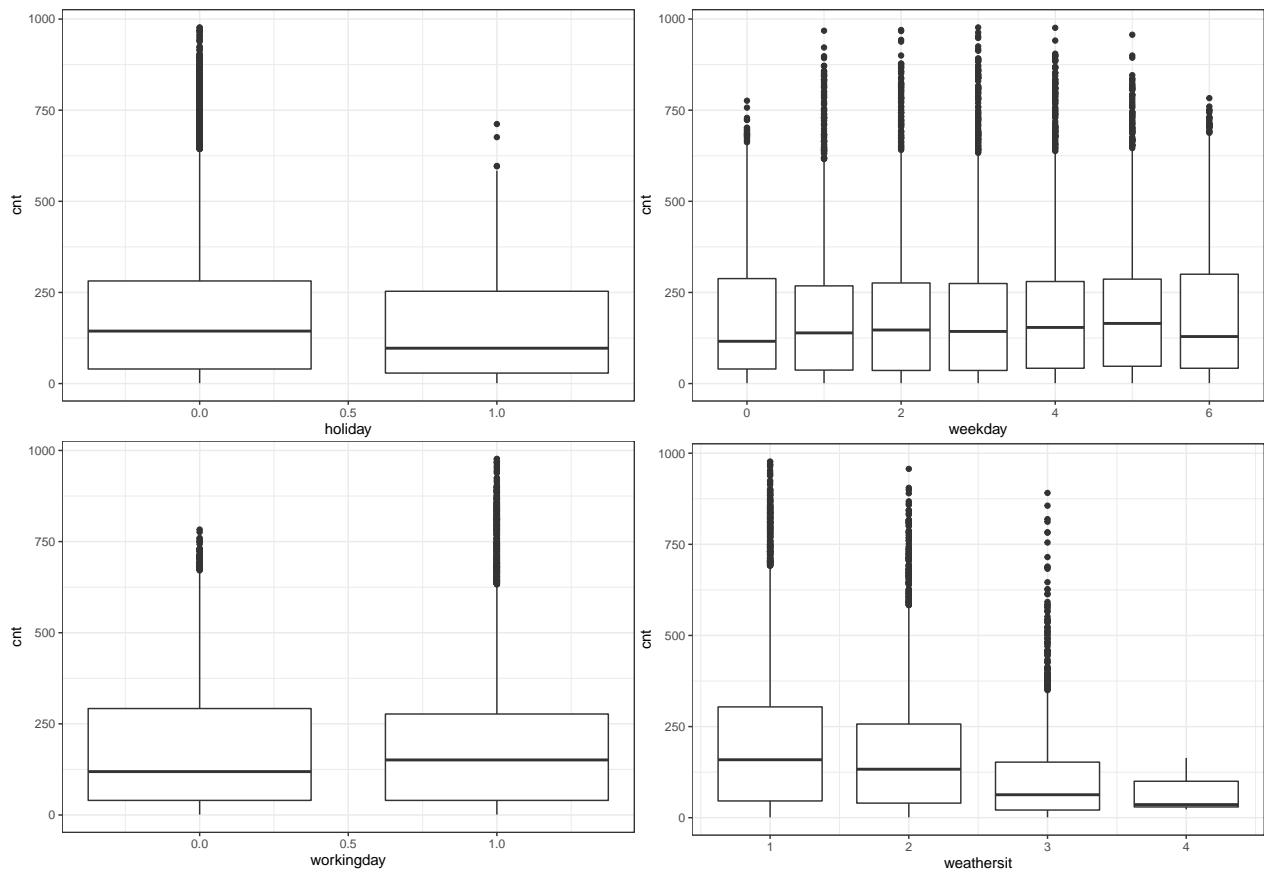
```





```
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()
```

