Stat 480 - Homework #4

Vahid Azizi

02/13/2020

Bike rentals in DC

1. Download the RMarkdown file with these homework instructions to use as a template for your work. Make sure to replace “Your Name” in the YAML with your name.
2. The data include daily bike rental counts (by members and casual users) of Capital Bikeshare in Washington, DC in 2011 and 2012 as well as weather information on these days. The original data sources are http://capitalbikeshare.com/system-data (http://capitalbikeshare.com/system-data) and http://www.freemeteo.com (http://www.freemeteo.com). Using the command below, read in the spotify data set into your R session.

```
bikes <- read.csv("https://raw.githubusercontent.com/Stat480-at-ISU/Stat480-at-ISU.github.io/master/homework/data/bikes.csv")
```

3. Recode the variable `holiday` to be logical variables with `0` as `FALSE` and `1` as `TRUE`.

```
bikes$holiday<-as.logical(bikes$holiday)
str(bikes)
```

```
## 'data.frame':    731 obs. of  15 variables:
##  $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ date      : Factor w/ 731 levels "2011-01-01","2011-01-02",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ season    : Factor w/ 4 levels "fall","spring",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ year      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ weekday   : int  7 1 2 3 4 5 6 7 1 2 ...
##  $ weather   : int  2 2 1 1 1 1 2 2 1 1 ...
##  $ temp      : num  0.344 0.363 0.196 0.2 0.227 ...
##  $ atemp     : num  0.364 0.354 0.189 0.212 0.229 ...
##  $ hum       : num  0.806 0.696 0.437 0.59 0.437 ...
##  $ windspeed : num  0.16 0.249 0.248 0.16 0.187 ...
##  $ casual    : int  331 131 120 108 82 88 148 68 54 41 ...
##  $ registered: int  654 670 1229 1454 1518 1518 1362 891 768 1280 ...
##  $ count     : int  985 801 1349 1562 1600 1606 1510 959 822 1321 ...
```

4. Create a variable `workingday` in that is `FALSE` if it is a holiday or the weekend (use `weekday` where 1 = Sunday, 2 = Monday, etc.). You may find De Morgan’s laws helpful here. Use ggplot to create a scatterplot comparing the number of registered bike rentals with the number of casual bike rentals. Map `workingday` to color. Interpret the result.

```
bikes["workingday"] <- TRUE
bikes$workingday<-replace(bikes$workingday, which(bikes$holiday==TRUE | bikes$weekday==1 | bikes
$weekday==2), FALSE)

ggplot(data=bikes, aes(x=casual, y=registered )) +
  geom_point(aes(color = workingday))
```

Apparently during the workdays there are more registered bike rentals compared to casual ones. Also, it can be seen during non-workingdays the number of casual bike rentals increases and on average the number of registered bike rentals decreases.

5. Recode the `year` variable so that the value 0 becomes 2011 and the value 1 becomes 2012.

```
bikes$year<-replace(bikes$year, which(bikes$year==0), 2011)
bikes$year<-replace(bikes$year, which(bikes$year==1), 2012)
head(bikes)
```

```
##   instant       date season year month holiday weekday weather     temp
## 1       1 2011-01-01 winter 2011     1   FALSE       7       2 0.344167
## 2       2 2011-01-02 winter 2011     1   FALSE       1       2 0.363478
## 3       3 2011-01-03 winter 2011     1   FALSE       2       1 0.196364
## 4       4 2011-01-04 winter 2011     1   FALSE       3       1 0.200000
## 5       5 2011-01-05 winter 2011     1   FALSE       4       1 0.226957
## 6       6 2011-01-06 winter 2011     1   FALSE       5       1 0.204348
##      atemp      hum windspeed casual registered count workingday
## 1 0.363625 0.805833 0.1604460    331        654   985       TRUE
## 2 0.353739 0.696087 0.2485390    131        670   801      FALSE
## 3 0.189405 0.437273 0.2483090    120       1229  1349      FALSE
## 4 0.212122 0.590435 0.1602960    108       1454  1562       TRUE
## 5 0.229270 0.436957 0.1869000     82       1518  1600       TRUE
## 6 0.233209 0.518261 0.0895652     88       1518  1606       TRUE
```

6. For each observation, verify that the variable `count` is equal to `casual` plus `registered`. You should be able to verify this without having to print out the columns. (Hint: one option is to use the function `any()`)

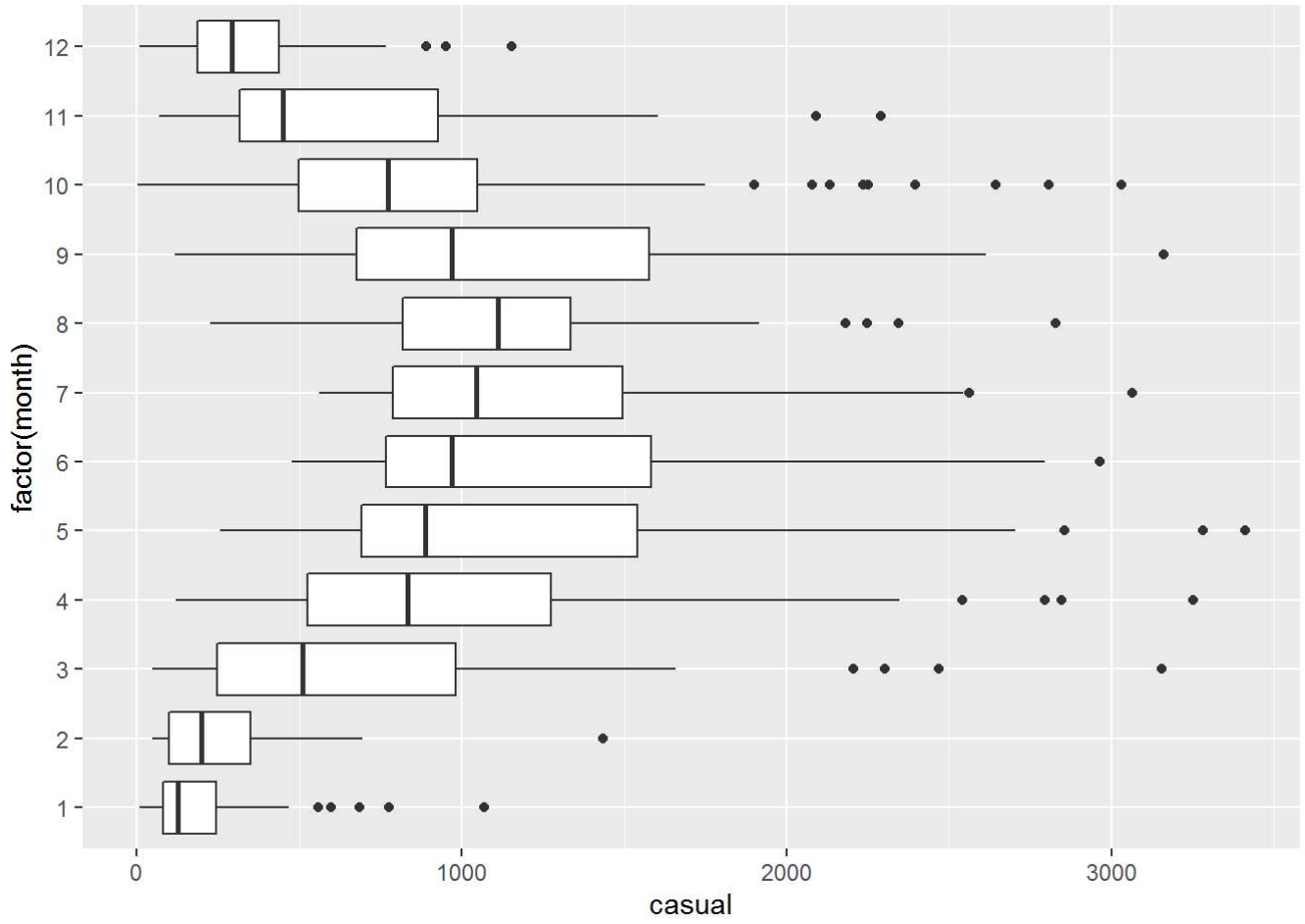
```
any(bikes$casual+bikes$registered!=bikes$count)
```

```
## [1] FALSE
```

This means that there is no observation that summation of ‘casual’ and ‘registered’ is different from its corresponding ‘count’.

7. How does the number of casual riders renting bikes compare across the months? Use ggplot2 to draw side-by-side boxplots of `casual` by `month`. Interpret the result.

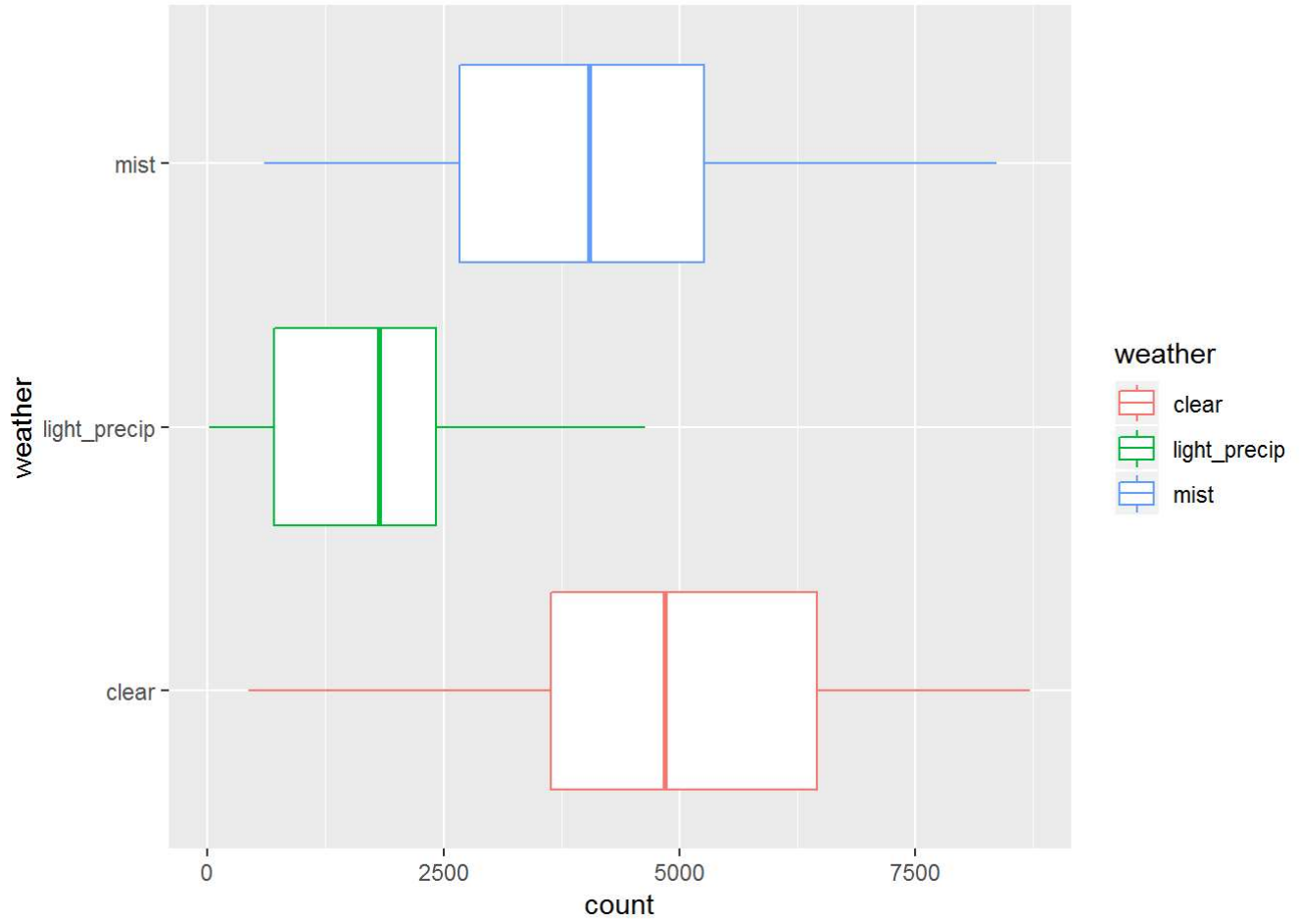
```
ggplot(data=bikes,aes(x=factor(month),y=casual))+
  geom_boxplot()+
  coord_flip()
```



Obviously, during the months that weather is good (months 5-9) on average the number of casual bike rentals is greater than the other months.

8. How does the number of rentals compare for different weather conditions? Recode the variable `weather` to be a factor with 1 - clear, 2 - mist, 3 - light_precip. Use ggplot2 to draw side-by-side boxplots of `count` by `weather` colored by `weather` . Interpret the result.

```
bikes$weather<-replace(bikes$weather, which(bikes$weather==1), "clear")
bikes$weather<-replace(bikes$weather, which(bikes$weather==2), "mist")
bikes$weather<-replace(bikes$weather, which(bikes$weather==3), "light_precip")
bikes$weather<-as.factor(bikes$weather)
ggplot(data=bikes,aes(x=weather,y=count))+
  geom_boxplot(aes(color = weather))+
  coord_flip()
```



It can be seen that in a good weather (here means clear weather) on average there is larger numer of bike rentals compared to bad weather (light-precip or mist). The least number of bike rentals belongs to category light_precip. It should be mention that clear weather on average is better than mist category but due to overlap in boxplots for these two categories we cannot say clear weather shows more number of bike rentals than mist category for sure. But certainly both categories do better than light-precip if we don't include outliers.