### Claire Alt

### Module 2 Assignment 1

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(lubridate)

## Warning: package 'lubridate' was built under R version 4.1.2

##   
## Attaching package: 'lubridate'

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

library(readr)  
bike <- read\_csv("bike\_cleaned.csv")

## Rows: 17379 Columns: 16

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): dteday, season, mnth, holiday, weekday, workingday, weathersit  
## dbl (9): instant, hr, temp, atemp, hum, windspeed, casual, registered, count

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

bike = bike %>% mutate(dteday = mdy(dteday))   
bike = bike %>% mutate\_if(is.character,as.factor)

cor(bike$hum, bike$count)

## [1] -0.3229107

cor(bike$atemp, bike$count)

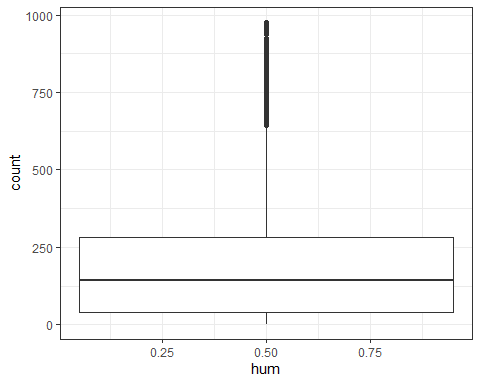
## [1] 0.4009293

cor(bike$temp, bike$count)

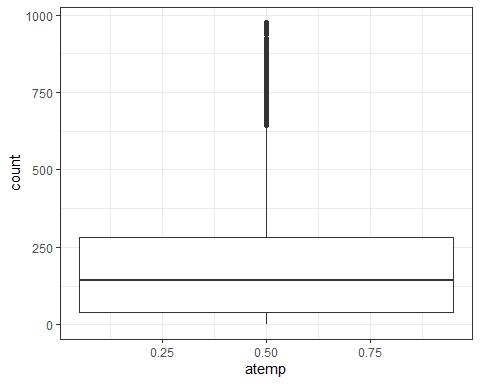
## [1] 0.4047723

Task 2 answer: After computing the correlation with the quantitative variables, it appears that the best correlation is the temp to the count.

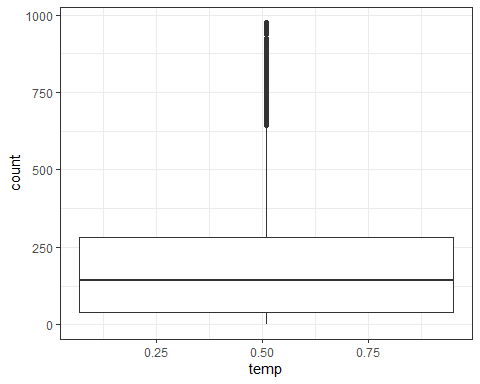
ggplot(bike,aes(x=hum,y=count)) + geom\_boxplot(aes(group = 1)) + theme\_bw()



ggplot(bike,aes(x=atemp,y=count)) + geom\_boxplot(aes(group = 1)) + theme\_bw()

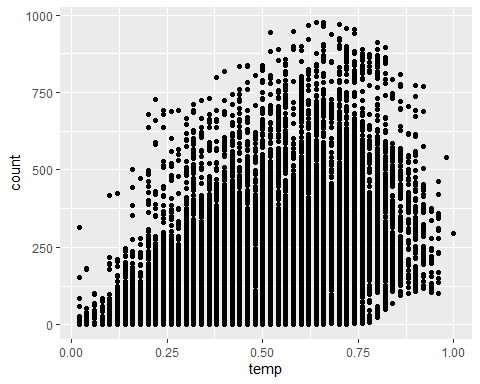


ggplot(bike,aes(x=temp,y=count)) + geom\_boxplot(aes(group = 1)) + theme\_bw()



Task 3: I believe that all of the variable affect the Count variable in some way, but the temp affects it the most. Simply because, with the colder or warmer it is outside - the more or fewer bikes people will be renting.

ggplot(bike,aes(x=temp,y=count)) +  
 geom\_point()

 Task 4: I used a scatter plot to plot the regression/correlation between Count and Temp. As you can see, there is an increase, as the temperature goes up, in the count of bikes being rented. Of course, if the temperature gets too hot, people will likely want ot stay inside - hence the decrease of bikes being rented.

model1 <- lm(count ~ temp + atemp + hum + windspeed, data = bike)  
summary (model1)

##   
## Call:  
## lm(formula = count ~ temp + atemp + hum + windspeed, data = bike)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -338.41 -101.49 -33.08 65.18 704.70   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 161.807 6.506 24.872 < 2e-16 \*\*\*  
## temp 85.576 40.751 2.100 0.0357 \*   
## atemp 314.343 45.714 6.876 6.35e-12 \*\*\*  
## hum -275.180 6.466 -42.560 < 2e-16 \*\*\*  
## windspeed 42.979 10.451 4.112 3.93e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 156.7 on 17374 degrees of freedom  
## Multiple R-squared: 0.2534, Adjusted R-squared: 0.2532   
## F-statistic: 1474 on 4 and 17374 DF, p-value: < 2.2e-16

Task 5 and Task 6: It seems that the most significant here is the Hum variable, to count. However, it does not seem that temp is that significant - so I was wrong in my original prediction. Referring to the multicollinearity of the model, it seems that Hum and Windspeed are the most inter correlated of the variables, to count.

model1 <- lm(count ~ atemp + hum + windspeed, data = bike)  
summary (model1)

##   
## Call:  
## lm(formula = count ~ atemp + hum + windspeed, data = bike)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -348.79 -101.66 -33.04 65.67 703.38   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 158.695 6.335 25.049 < 2e-16 \*\*\*  
## atemp 409.225 6.952 58.864 < 2e-16 \*\*\*  
## hum -275.863 6.458 -42.716 < 2e-16 \*\*\*  
## windspeed 47.860 10.191 4.697 2.67e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 156.8 on 17375 degrees of freedom  
## Multiple R-squared: 0.2532, Adjusted R-squared: 0.2531   
## F-statistic: 1964 on 3 and 17375 DF, p-value: < 2.2e-16

Task 7: I got rid of the “temp” variable, because it did not show much significance. Due to this change, the model shows a higher significance and coorelation in the three variables: Atemp, Hum, and Windspeed.