library(tidyverse)

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

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.1.1 v forcats 0.5.1

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

library(tidymodels)

## Registered S3 method overwritten by 'tune':  
## method from   
## required\_pkgs.model\_spec parsnip

## -- Attaching packages -------------------------------------- tidymodels 0.1.4 --

## v broom 0.7.11 v rsample 0.1.1   
## v dials 0.0.10 v tune 0.1.6   
## v infer 1.0.0 v workflows 0.2.4   
## v modeldata 0.1.1 v workflowsets 0.1.0   
## v parsnip 0.1.7 v yardstick 0.0.9   
## v recipes 0.1.17

## -- Conflicts ----------------------------------------- tidymodels\_conflicts() --  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()  
## \* Learn how to get started at https://www.tidymodels.org/start/

library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.1-3

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggcorrplot)  
library(MASS)

##   
## Attaching package: 'MASS'

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

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':  
##   
## some

#install.packages("lubridate")  
library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

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.

summary(bike)

## instant dteday season mnth   
## Min. : 1 Length:17379 Length:17379 Length:17379   
## 1st Qu.: 4346 Class :character Class :character Class :character   
## Median : 8690 Mode :character Mode :character Mode :character   
## Mean : 8690   
## 3rd Qu.:13034   
## Max. :17379   
## hr holiday weekday workingday   
## Min. : 0.00 Length:17379 Length:17379 Length:17379   
## 1st Qu.: 6.00 Class :character Class :character Class :character   
## Median :12.00 Mode :character Mode :character Mode :character   
## Mean :11.55   
## 3rd Qu.:18.00   
## Max. :23.00   
## weathersit temp atemp hum   
## Length:17379 Min. :0.020 Min. :0.0000 Min. :0.0000   
## Class :character 1st Qu.:0.340 1st Qu.:0.3333 1st Qu.:0.4800   
## Mode :character Median :0.500 Median :0.4848 Median :0.6300   
## Mean :0.497 Mean :0.4758 Mean :0.6272   
## 3rd Qu.:0.660 3rd Qu.:0.6212 3rd Qu.:0.7800   
## Max. :1.000 Max. :1.0000 Max. :1.0000   
## windspeed casual registered count   
## Min. :0.0000 Min. : 0.00 Min. : 0.0 Min. : 1.0   
## 1st Qu.:0.1045 1st Qu.: 4.00 1st Qu.: 34.0 1st Qu.: 40.0   
## Median :0.1940 Median : 17.00 Median :115.0 Median :142.0   
## Mean :0.1901 Mean : 35.68 Mean :153.8 Mean :189.5   
## 3rd Qu.:0.2537 3rd Qu.: 48.00 3rd Qu.:220.0 3rd Qu.:281.0   
## Max. :0.8507 Max. :367.00 Max. :886.0 Max. :977.0

str(bike)

## spec\_tbl\_df [17,379 x 16] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ instant : num [1:17379] 1 2 3 4 5 6 7 8 9 10 ...  
## $ dteday : chr [1:17379] "1/1/2011" "1/1/2011" "1/1/2011" "1/1/2011" ...  
## $ season : chr [1:17379] "Winter" "Winter" "Winter" "Winter" ...  
## $ mnth : chr [1:17379] "Jan" "Jan" "Jan" "Jan" ...  
## $ hr : num [1:17379] 0 1 2 3 4 5 6 7 8 9 ...  
## $ holiday : chr [1:17379] "NotHoliday" "NotHoliday" "NotHoliday" "NotHoliday" ...  
## $ weekday : chr [1:17379] "Saturday" "Saturday" "Saturday" "Saturday" ...  
## $ workingday: chr [1:17379] "NotWorkingDay" "NotWorkingDay" "NotWorkingDay" "NotWorkingDay" ...  
## $ weathersit: chr [1:17379] "NoPrecip" "NoPrecip" "NoPrecip" "NoPrecip" ...  
## $ temp : num [1:17379] 0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...  
## $ atemp : num [1:17379] 0.288 0.273 0.273 0.288 0.288 ...  
## $ hum : num [1:17379] 0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...  
## $ windspeed : num [1:17379] 0 0 0 0 0 0.0896 0 0 0 0 ...  
## $ casual : num [1:17379] 3 8 5 3 0 0 2 1 1 8 ...  
## $ registered: num [1:17379] 13 32 27 10 1 1 0 2 7 6 ...  
## $ count : num [1:17379] 16 40 32 13 1 1 2 3 8 14 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. instant = col\_double(),  
## .. dteday = col\_character(),  
## .. season = col\_character(),  
## .. mnth = col\_character(),  
## .. hr = col\_double(),  
## .. holiday = col\_character(),  
## .. weekday = col\_character(),  
## .. workingday = col\_character(),  
## .. weathersit = col\_character(),  
## .. temp = col\_double(),  
## .. atemp = col\_double(),  
## .. hum = col\_double(),  
## .. windspeed = col\_double(),  
## .. casual = col\_double(),  
## .. registered = col\_double(),  
## .. count = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

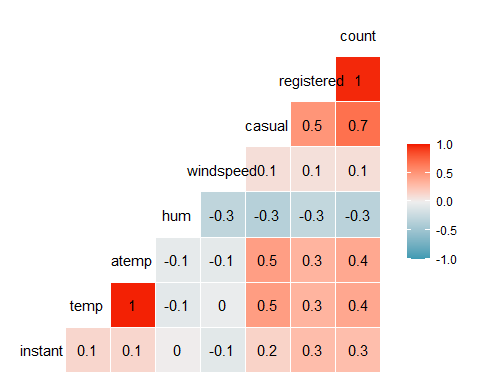
bike = bike %>% mutate(dteday = mdy(dteday)) #mdy is a lubridate package function  
bike = bike %>% mutate(season = as\_factor(season))  
bike = bike %>% mutate(mnth = as\_factor(mnth))  
bike = bike %>% mutate(holiday = as\_factor(holiday))  
bike = bike %>% mutate(weekday = as\_factor(weekday))  
bike = bike %>% mutate(workingday = as\_factor(workingday))  
bike = bike %>% mutate(weathersit = as\_factor(weathersit))  
bike = bike %>% mutate(hr = as\_factor(hr))  
str(bike)

## spec\_tbl\_df [17,379 x 16] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ instant : num [1:17379] 1 2 3 4 5 6 7 8 9 10 ...  
## $ dteday : Date[1:17379], format: "2011-01-01" "2011-01-01" ...  
## $ season : Factor w/ 4 levels "Winter","Spring",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ mnth : Factor w/ 12 levels "Jan","Feb","Mar",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ hr : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ holiday : Factor w/ 2 levels "NotHoliday","Holiday": 1 1 1 1 1 1 1 1 1 1 ...  
## $ weekday : Factor w/ 7 levels "Saturday","Sunday",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ workingday: Factor w/ 2 levels "NotWorkingDay",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ weathersit: Factor w/ 4 levels "NoPrecip","Misty",..: 1 1 1 1 1 2 1 1 1 1 ...  
## $ temp : num [1:17379] 0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...  
## $ atemp : num [1:17379] 0.288 0.273 0.273 0.288 0.288 ...  
## $ hum : num [1:17379] 0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...  
## $ windspeed : num [1:17379] 0 0 0 0 0 0.0896 0 0 0 0 ...  
## $ casual : num [1:17379] 3 8 5 3 0 0 2 1 1 8 ...  
## $ registered: num [1:17379] 13 32 27 10 1 1 0 2 7 6 ...  
## $ count : num [1:17379] 16 40 32 13 1 1 2 3 8 14 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. instant = col\_double(),  
## .. dteday = col\_character(),  
## .. season = col\_character(),  
## .. mnth = col\_character(),  
## .. hr = col\_double(),  
## .. holiday = col\_character(),  
## .. weekday = col\_character(),  
## .. workingday = col\_character(),  
## .. weathersit = col\_character(),  
## .. temp = col\_double(),  
## .. atemp = col\_double(),  
## .. hum = col\_double(),  
## .. windspeed = col\_double(),  
## .. casual = col\_double(),  
## .. registered = col\_double(),  
## .. count = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

Why do we convert the “hr” variable into factor? Why not just leave as numbers? Hr, which stand for Hour, essentially acts as a categorical variable since it is describing the hour given, not a random number with no upper/lower limit threshold. Moreover, using hours in the 0 to 23 form provides discontunuity since it should loop around from 23->0, but numbers would operate on a pure scale. In this case, it stands to make this variable a factor. Granted, hours could be formatted to be read normally as a number, but we can play it safe and label it as a factor (since it could be read as a coefficient of a per hour of day).

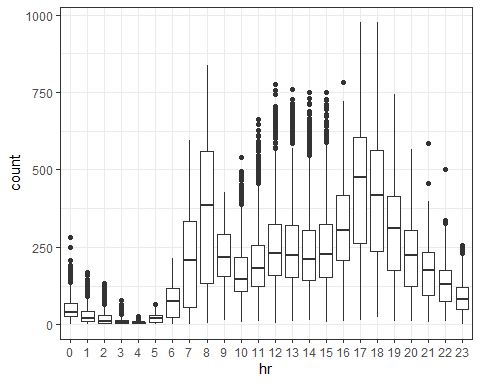
ggcorr(bike, label = TRUE)

## Warning in ggcorr(bike, label = TRUE): data in column(s) 'dteday', 'season',  
## 'mnth', 'hr', 'holiday', 'weekday', 'workingday', 'weathersit' are not numeric  
## and were ignored

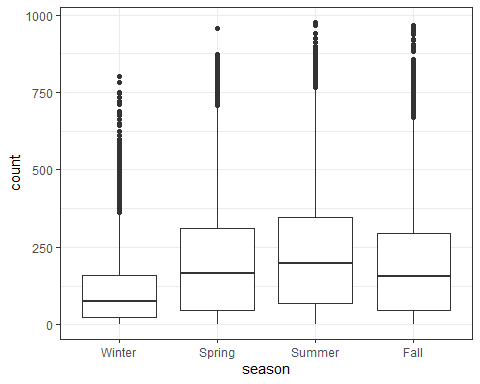


Which of the quantitative variables appears to be best correlated with “count”? atemp and temp appear to be the two most positively correlated with the count variable.

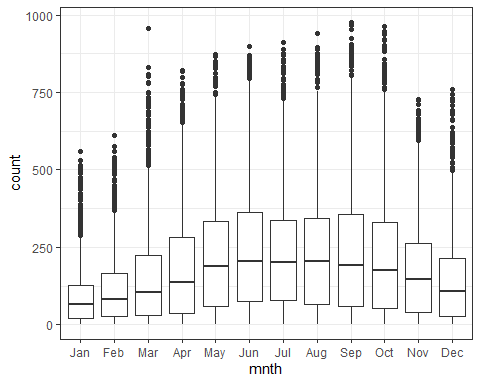
ggplot(bike,aes(x=hr,y=count)) + geom\_boxplot() + theme\_bw()



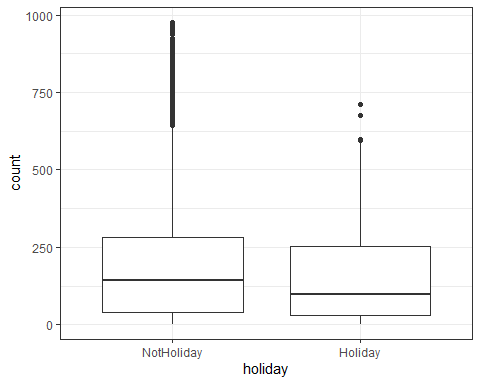
#ggplot(bike,aes(x=dteday,y=count)) + geom\_boxplot() + theme\_bw() doesn't seem to allow the var. to boxplot(?)  
ggplot(bike,aes(x=season,y=count)) + geom\_boxplot() + theme\_bw()



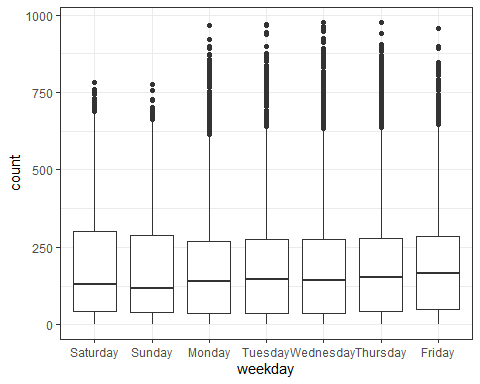
ggplot(bike,aes(x=mnth,y=count)) + geom\_boxplot() + theme\_bw()



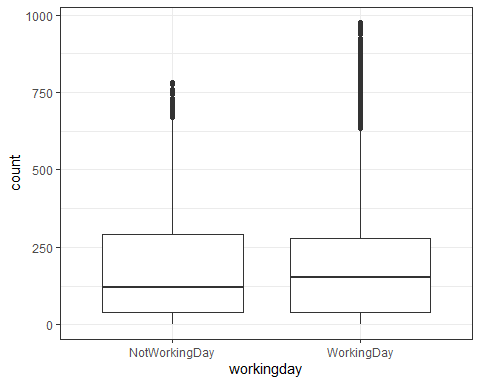
ggplot(bike,aes(x=holiday,y=count)) + geom\_boxplot() + theme\_bw()



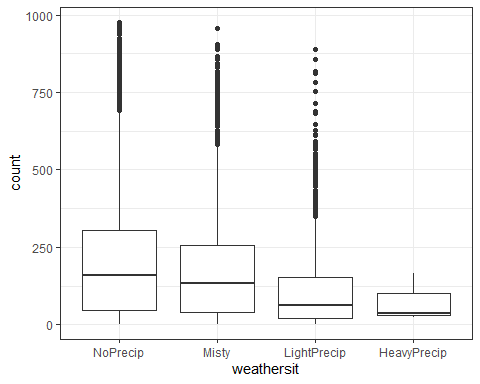
ggplot(bike,aes(x=weekday,y=count)) + geom\_boxplot() + theme\_bw()



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



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



Excluding HR, which we have established already obviously affects count, next we have season. Season appears to affect count, as we see a distribution that slightly bell curves during the Spring and Summer. The mnth variable has that same effect, the month might intuitively also affect count, the warmer months being more likely to rent out bikes. With holiday, the count has a small affect, seeing a greater dispersion on non-holidays, which makes sense. Who is renting a bike on Easter/Yom Kippur, etc.? Weekday doesn’t seem to show a marked difference, with only small differences in averages throughout the week. Workingday falls under the same idea as holiday- you may be more likely to rent a bike on a working day (need based industry?) versus a non-working day (leisure). Lastly, weathersit has shown a marked impact- seeing a greater quantity and dispersion with no rain vs very few bikes during dayswith heavier rainfall. To recap: weathersit, mnth, and season have marked affect, holiday and workingday have a much smaller affect (which may not materialize into a strong corr), and weekday does not seem to affect count.

bike\_recipe = recipe(count ~ temp , bike)   
   
  
lm\_model =   
 linear\_reg() %>%  
 set\_engine("lm")  
  
lm\_wflow =   
 workflow() %>%  
 add\_model(lm\_model) %>%  
 add\_recipe(bike\_recipe)  
  
lm\_fit = fit(lm\_wflow, bike)

summary(lm\_fit$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -291.37 -110.23 -32.86 76.77 744.76   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0356 3.4827 -0.01 0.992   
## temp 381.2949 6.5344 58.35 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 165.9 on 17377 degrees of freedom  
## Multiple R-squared: 0.1638, Adjusted R-squared: 0.1638   
## F-statistic: 3405 on 1 and 17377 DF, p-value: < 2.2e-16

Looking at this model, the R-squared (adjusted) being only .1638 is not particularly high, and does not instill much confidence in this being a quality model. The p-value is also <.05, which is the same issue as above. Despite there being a correlation in the corrplot from before, there may be some diagnosis we need to complete going forward.

bike\_recipe2 = recipe(count ~ temp + atemp + hr + season, bike) %>%  
 step\_dummy(all\_nominal())  
  
  
lm\_model = #give the model type a name  
 linear\_reg() %>%  
 set\_engine("lm")  
  
lm\_wflow2 =   
 workflow() %>%  
 add\_model(lm\_model) %>%  
 add\_recipe(bike\_recipe2)  
  
lm\_fit2 = fit(lm\_wflow2, bike)

summary(lm\_fit2$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -454.89 -61.67 -8.90 53.06 529.55   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -108.245 5.095 -21.247 < 2e-16 \*\*\*  
## temp 192.203 30.451 6.312 2.82e-10 \*\*\*  
## atemp 109.869 32.771 3.353 0.000802 \*\*\*  
## hr\_X1 -18.092 6.013 -3.009 0.002628 \*\*   
## hr\_X2 -27.437 6.033 -4.548 5.46e-06 \*\*\*  
## hr\_X3 -38.570 6.073 -6.351 2.20e-10 \*\*\*  
## hr\_X4 -42.220 6.075 -6.950 3.79e-12 \*\*\*  
## hr\_X5 -25.094 6.033 -4.159 3.21e-05 \*\*\*  
## hr\_X6 32.390 6.017 5.383 7.43e-08 \*\*\*  
## hr\_X7 166.508 6.011 27.700 < 2e-16 \*\*\*  
## hr\_X8 308.781 6.008 51.398 < 2e-16 \*\*\*  
## hr\_X9 162.876 6.007 27.113 < 2e-16 \*\*\*  
## hr\_X10 110.517 6.013 18.379 < 2e-16 \*\*\*  
## hr\_X11 138.631 6.023 23.015 < 2e-16 \*\*\*  
## hr\_X12 178.746 6.033 29.626 < 2e-16 \*\*\*  
## hr\_X13 174.886 6.044 28.937 < 2e-16 \*\*\*  
## hr\_X14 159.231 6.054 26.302 < 2e-16 \*\*\*  
## hr\_X15 168.570 6.058 27.825 < 2e-16 \*\*\*  
## hr\_X16 230.509 6.054 38.078 < 2e-16 \*\*\*  
## hr\_X17 382.988 6.044 63.370 < 2e-16 \*\*\*  
## hr\_X18 350.725 6.035 58.114 < 2e-16 \*\*\*  
## hr\_X19 241.465 6.022 40.095 < 2e-16 \*\*\*  
## hr\_X20 160.310 6.014 26.657 < 2e-16 \*\*\*  
## hr\_X21 110.328 6.009 18.360 < 2e-16 \*\*\*  
## hr\_X22 72.519 6.006 12.074 < 2e-16 \*\*\*  
## hr\_X23 31.764 6.005 5.290 1.24e-07 \*\*\*  
## season\_Spring 28.186 3.136 8.986 < 2e-16 \*\*\*  
## season\_Summer 9.870 4.051 2.436 0.014849 \*   
## season\_Fall 53.558 2.683 19.963 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 114.5 on 17350 degrees of freedom  
## Multiple R-squared: 0.6023, Adjusted R-squared: 0.6016   
## F-statistic: 938.3 on 28 and 17350 DF, p-value: < 2.2e-16

Looking at these variables, it is exciting to see a higher adjusted R-Sqaured of 0.6016. With the exception of some of the hours (1-5) having a negative slope, most of the slope values are to be expected (positive, rather large). Multicollinearity may be to blame for the negative slope values where we would intuitively think/imagine it would be positive! Utilizing the VIF function shows that temp, atemp especially are probable to be apart of the multicollinearity problem we are discussing. This many variables are also going to obviously increase our DoF. Multiple R-Squared being 0.6023 is to be expected, makes the confidence of the quality of this model slightly more comfortable for me personally.

car::vif(lm\_fit2$fit$fit$fit)

## temp atemp hr\_X1 hr\_X2 hr\_X3   
## 45.582928 42.049980 1.914264 1.903937 1.882879   
## hr\_X4 hr\_X5 hr\_X6 hr\_X7 hr\_X8   
## 1.883955 1.909213 1.919295 1.920325 1.918181   
## hr\_X9 hr\_X10 hr\_X11 hr\_X12 hr\_X13   
## 1.917966 1.921682 1.928237 1.937211 1.946342   
## hr\_X14 hr\_X15 hr\_X16 hr\_X17 hr\_X18   
## 1.952979 1.955748 1.955329 1.948879 1.938308   
## hr\_X19 hr\_X20 hr\_X21 hr\_X22 hr\_X23   
## 1.930098 1.924657 1.921540 1.919889 1.919082   
## season\_Spring season\_Summer season\_Fall   
## 2.469635 4.173240 1.758173

allmod = lm(count ~ temp + atemp + hr + season + mnth + holiday + weekday + workingday + weathersit + hum + windspeed, bike)  
summary(allmod)

##   
## Call:  
## lm(formula = count ~ temp + atemp + hr + season + mnth + holiday +   
## weekday + workingday + weathersit + hum + windspeed, data = bike)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -401.19 -61.46 -9.25 51.13 478.91   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12.7473 7.0760 -1.801 0.071642 .   
## temp 195.1921 31.9175 6.116 9.83e-10 \*\*\*  
## atemp 103.9531 33.1547 3.135 0.001719 \*\*   
## hr1 -16.4977 5.7874 -2.851 0.004369 \*\*   
## hr2 -24.4135 5.8076 -4.204 2.64e-05 \*\*\*  
## hr3 -34.3153 5.8494 -5.866 4.53e-09 \*\*\*  
## hr4 -36.0087 5.8549 -6.150 7.91e-10 \*\*\*  
## hr5 -19.7552 5.8171 -3.396 0.000685 \*\*\*  
## hr6 38.7841 5.8019 6.685 2.38e-11 \*\*\*  
## hr7 172.8931 5.7906 29.857 < 2e-16 \*\*\*  
## hr8 311.6446 5.7839 53.882 < 2e-16 \*\*\*  
## hr9 161.5087 5.7898 27.895 < 2e-16 \*\*\*  
## hr10 104.2107 5.8133 17.926 < 2e-16 \*\*\*  
## hr11 127.0021 5.8555 21.689 < 2e-16 \*\*\*  
## hr12 164.1738 5.9046 27.804 < 2e-16 \*\*\*  
## hr13 157.7321 5.9447 26.533 < 2e-16 \*\*\*  
## hr14 141.0138 5.9777 23.590 < 2e-16 \*\*\*  
## hr15 150.1886 5.9891 25.077 < 2e-16 \*\*\*  
## hr16 212.6712 5.9768 35.583 < 2e-16 \*\*\*  
## hr17 367.4906 5.9425 61.841 < 2e-16 \*\*\*  
## hr18 337.0098 5.9042 57.080 < 2e-16 \*\*\*  
## hr19 230.4248 5.8499 39.390 < 2e-16 \*\*\*  
## hr20 152.5053 5.8186 26.210 < 2e-16 \*\*\*  
## hr21 104.5791 5.7953 18.045 < 2e-16 \*\*\*  
## hr22 68.8433 5.7847 11.901 < 2e-16 \*\*\*  
## hr23 30.9933 5.7800 5.362 8.33e-08 \*\*\*  
## seasonSpring 39.4906 5.2578 7.511 6.16e-14 \*\*\*  
## seasonSummer 28.8578 6.2248 4.636 3.58e-06 \*\*\*  
## seasonFall 66.0479 5.2858 12.495 < 2e-16 \*\*\*  
## mnthFeb -0.4609 4.2442 -0.109 0.913527   
## mnthMar 4.4203 4.7679 0.927 0.353884   
## mnthApr -9.1995 7.0836 -1.299 0.194061   
## mnthMay 0.5152 7.5761 0.068 0.945784   
## mnthJun -20.7340 7.7826 -2.664 0.007726 \*\*   
## mnthJul -40.7689 8.7337 -4.668 3.06e-06 \*\*\*  
## mnthAug -15.6046 8.5184 -1.832 0.066988 .   
## mnthSep 15.5058 7.5733 2.047 0.040630 \*   
## mnthOct 3.9416 7.0161 0.562 0.574259   
## mnthNov -15.5826 6.7536 -2.307 0.021049 \*   
## mnthDec -9.3163 5.3638 -1.737 0.082422 .   
## holidayHoliday -26.3866 5.2846 -4.993 6.00e-07 \*\*\*  
## weekdaySunday -14.7269 3.1158 -4.727 2.30e-06 \*\*\*  
## weekdayMonday -6.6593 3.2160 -2.071 0.038401 \*   
## weekdayTuesday -5.4421 3.1389 -1.734 0.082980 .   
## weekdayWednesday -2.4169 3.1339 -0.771 0.440578   
## weekdayThursday -3.7139 3.1316 -1.186 0.235669   
## weekdayFriday 0.9168 3.1234 0.294 0.769126   
## workingdayWorkingDay NA NA NA NA   
## weathersitMisty -6.3713 2.0774 -3.067 0.002166 \*\*   
## weathersitLightPrecip -60.3711 3.5025 -17.236 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -33.1693 63.7637 -0.520 0.602937   
## hum -114.9024 5.9797 -19.216 < 2e-16 \*\*\*  
## windspeed -43.2351 7.6305 -5.666 1.48e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 110.2 on 17327 degrees of freedom  
## Multiple R-squared: 0.6323, Adjusted R-squared: 0.6312   
## F-statistic: 584.2 on 51 and 17327 DF, p-value: < 2.2e-16

emptymod = lm(count ~1, bike)   
summary(emptymod)

##   
## Call:  
## lm(formula = count ~ 1, data = bike)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -188.46 -149.46 -47.46 91.54 787.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 189.463 1.376 137.7 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 181.4 on 17378 degrees of freedom

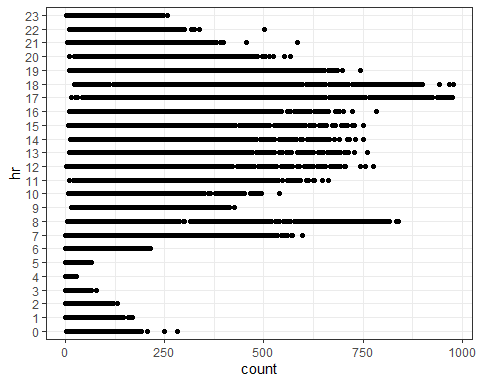
forwardmod = stepAIC(emptymod, direction = "forward", scope = list(upper=allmod,lower=emptymod))

## Start: AIC=180764.7  
## count ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + hr 23 286734681 285026910 168713  
## + temp 1 93677759 478083832 177657  
## + atemp 1 91907421 479854170 177721  
## + hum 1 59618351 512143240 178853  
## + mnth 11 42909976 528851615 179431  
## + season 3 37729358 534032233 179584  
## + weathersit 3 12285030 559476561 180393  
## + windspeed 1 4970060 566791531 180615  
## + holiday 1 546889 571214702 180750  
## + workingday 1 524387 571237204 180751  
## + weekday 6 687929 571073662 180756  
## <none> 571761591 180765  
##   
## Step: AIC=168712.5  
## count ~ hr  
##   
## Df Sum of Sq RSS AIC  
## + atemp 1 50518941 234507969 165324  
## + temp 1 50101685 234925225 165355  
## + mnth 11 44822160 240204750 165761  
## + season 3 39619754 245407156 166117  
## + weathersit 3 13766672 271260238 167858  
## + hum 1 4924310 280102600 168412  
## + windspeed 1 1476211 283550699 168624  
## + holiday 1 561784 284465126 168680  
## + weekday 6 719530 284307380 168681  
## + workingday 1 485366 284541544 168685  
## <none> 285026910 168713  
##   
## Step: AIC=165324  
## count ~ hr + atemp  
##   
## Df Sum of Sq RSS AIC  
## + weathersit 3 9227265 225280704 164632  
## + hum 1 7008684 227499285 164799  
## + season 3 6580442 227927527 164835  
## + mnth 11 5854560 228653409 164907  
## + weekday 6 607638 233900331 165291  
## + holiday 1 274006 234233963 165306  
## + temp 1 152153 234355816 165315  
## + windspeed 1 120557 234387412 165317  
## + workingday 1 90170 234417799 165319  
## <none> 234507969 165324  
##   
## Step: AIC=164632.4  
## count ~ hr + atemp + weathersit  
##   
## Df Sum of Sq RSS AIC  
## + season 3 7110590 218170114 164081  
## + mnth 11 6748849 218531855 164126  
## + hum 1 2218225 223062479 164462  
## + weekday 6 650146 224630558 164594  
## + holiday 1 361521 224919183 164606  
## + temp 1 210449 225070254 164618  
## + workingday 1 193080 225087624 164619  
## + windspeed 1 46236 225234467 164631  
## <none> 225280704 164632  
##   
## Step: AIC=164081  
## count ~ hr + atemp + weathersit + season  
##   
## Df Sum of Sq RSS AIC  
## + hum 1 3719099 214451015 163784  
## + mnth 11 1906375 216263739 163950  
## + temp 1 617921 217552193 164034  
## + weekday 6 671540 217498574 164039  
## + holiday 1 352599 217817515 164055  
## + workingday 1 189381 217980734 164068  
## <none> 218170114 164081  
## + windspeed 1 6 218170108 164083  
##   
## Step: AIC=163784.2  
## count ~ hr + atemp + weathersit + season + hum  
##   
## Df Sum of Sq RSS AIC  
## + mnth 11 2749042 211701972 163582  
## + weekday 6 533038 213917977 163753  
## + holiday 1 352089 214098926 163758  
## + temp 1 343320 214107695 163758  
## + windspeed 1 198909 214252106 163770  
## + workingday 1 172809 214278206 163772  
## <none> 214451015 163784  
##   
## Step: AIC=163581.9  
## count ~ hr + atemp + weathersit + season + hum + mnth  
##   
## Df Sum of Sq RSS AIC  
## + weekday 6 509911 211192062 163552  
## + holiday 1 311930 211390043 163558  
## + temp 1 296592 211405381 163560  
## + windspeed 1 231216 211470756 163565  
## + workingday 1 157064 211544909 163571  
## <none> 211701972 163582  
##   
## Step: AIC=163552  
## count ~ hr + atemp + weathersit + season + hum + mnth + weekday  
##   
## Df Sum of Sq RSS AIC  
## + holiday 1 261389 210930673 163533  
## + workingday 1 261389 210930673 163533  
## + temp 1 259160 210932901 163533  
## + windspeed 1 227960 210964101 163535  
## <none> 211192062 163552  
##   
## Step: AIC=163532.5  
## count ~ hr + atemp + weathersit + season + hum + mnth + weekday +   
## holiday  
##   
## Df Sum of Sq RSS AIC  
## + temp 1 291507 210639166 163510  
## + windspeed 1 227262 210703411 163516  
## <none> 210930673 163533  
##   
## Step: AIC=163510.5  
## count ~ hr + atemp + weathersit + season + hum + mnth + weekday +   
## holiday + temp  
##   
## Df Sum of Sq RSS AIC  
## + windspeed 1 389568 210249597 163480  
## <none> 210639166 163510  
##   
## Step: AIC=163480.3  
## count ~ hr + atemp + weathersit + season + hum + mnth + weekday +   
## holiday + temp + windspeed  
##   
## Df Sum of Sq RSS AIC  
## <none> 210249597 163480

summary(forwardmod)

##   
## Call:  
## lm(formula = count ~ hr + atemp + weathersit + season + hum +   
## mnth + weekday + holiday + temp + windspeed, data = bike)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -401.19 -61.46 -9.25 51.13 478.91   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12.7473 7.0760 -1.801 0.071642 .   
## hr1 -16.4977 5.7874 -2.851 0.004369 \*\*   
## hr2 -24.4135 5.8076 -4.204 2.64e-05 \*\*\*  
## hr3 -34.3153 5.8494 -5.866 4.53e-09 \*\*\*  
## hr4 -36.0087 5.8549 -6.150 7.91e-10 \*\*\*  
## hr5 -19.7552 5.8171 -3.396 0.000685 \*\*\*  
## hr6 38.7841 5.8019 6.685 2.38e-11 \*\*\*  
## hr7 172.8931 5.7906 29.857 < 2e-16 \*\*\*  
## hr8 311.6446 5.7839 53.882 < 2e-16 \*\*\*  
## hr9 161.5087 5.7898 27.895 < 2e-16 \*\*\*  
## hr10 104.2107 5.8133 17.926 < 2e-16 \*\*\*  
## hr11 127.0021 5.8555 21.689 < 2e-16 \*\*\*  
## hr12 164.1738 5.9046 27.804 < 2e-16 \*\*\*  
## hr13 157.7321 5.9447 26.533 < 2e-16 \*\*\*  
## hr14 141.0138 5.9777 23.590 < 2e-16 \*\*\*  
## hr15 150.1886 5.9891 25.077 < 2e-16 \*\*\*  
## hr16 212.6712 5.9768 35.583 < 2e-16 \*\*\*  
## hr17 367.4906 5.9425 61.841 < 2e-16 \*\*\*  
## hr18 337.0098 5.9042 57.080 < 2e-16 \*\*\*  
## hr19 230.4248 5.8499 39.390 < 2e-16 \*\*\*  
## hr20 152.5053 5.8186 26.210 < 2e-16 \*\*\*  
## hr21 104.5791 5.7953 18.045 < 2e-16 \*\*\*  
## hr22 68.8433 5.7847 11.901 < 2e-16 \*\*\*  
## hr23 30.9933 5.7800 5.362 8.33e-08 \*\*\*  
## atemp 103.9531 33.1547 3.135 0.001719 \*\*   
## weathersitMisty -6.3713 2.0774 -3.067 0.002166 \*\*   
## weathersitLightPrecip -60.3711 3.5025 -17.236 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -33.1693 63.7637 -0.520 0.602937   
## seasonSpring 39.4906 5.2578 7.511 6.16e-14 \*\*\*  
## seasonSummer 28.8578 6.2248 4.636 3.58e-06 \*\*\*  
## seasonFall 66.0479 5.2858 12.495 < 2e-16 \*\*\*  
## hum -114.9024 5.9797 -19.216 < 2e-16 \*\*\*  
## mnthFeb -0.4609 4.2442 -0.109 0.913527   
## mnthMar 4.4203 4.7679 0.927 0.353884   
## mnthApr -9.1995 7.0836 -1.299 0.194061   
## mnthMay 0.5152 7.5761 0.068 0.945784   
## mnthJun -20.7340 7.7826 -2.664 0.007726 \*\*   
## mnthJul -40.7689 8.7337 -4.668 3.06e-06 \*\*\*  
## mnthAug -15.6046 8.5184 -1.832 0.066988 .   
## mnthSep 15.5058 7.5733 2.047 0.040630 \*   
## mnthOct 3.9416 7.0161 0.562 0.574259   
## mnthNov -15.5826 6.7536 -2.307 0.021049 \*   
## mnthDec -9.3163 5.3638 -1.737 0.082422 .   
## weekdaySunday -14.7269 3.1158 -4.727 2.30e-06 \*\*\*  
## weekdayMonday -6.6593 3.2160 -2.071 0.038401 \*   
## weekdayTuesday -5.4421 3.1389 -1.734 0.082980 .   
## weekdayWednesday -2.4169 3.1339 -0.771 0.440578   
## weekdayThursday -3.7139 3.1316 -1.186 0.235669   
## weekdayFriday 0.9168 3.1234 0.294 0.769126   
## holidayHoliday -26.3866 5.2846 -4.993 6.00e-07 \*\*\*  
## temp 195.1921 31.9175 6.116 9.83e-10 \*\*\*  
## windspeed -43.2351 7.6305 -5.666 1.48e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 110.2 on 17327 degrees of freedom  
## Multiple R-squared: 0.6323, Adjusted R-squared: 0.6312   
## F-statistic: 584.2 on 51 and 17327 DF, p-value: < 2.2e-16

ggplot(bike, aes(x=count, y= hr)) + geom\_point() + theme\_bw()



bike\_recipe3 = recipe(count ~ hr + atemp + weathersit + season + hum +   
 mnth + weekday + holiday + temp + windspeed, data = bike) %>%  
 step\_dummy(all\_nominal()) %>%  
 step\_rm(hum, windspeed)  
  
  
lm\_model = #give the model type a name  
 linear\_reg() %>%  
 set\_engine("lm")  
  
lm\_wflow3 =   
 workflow() %>%  
 add\_model(lm\_model) %>%  
 add\_recipe(bike\_recipe3)  
  
lm\_fit3 = fit(lm\_wflow3, bike)

summary(lm\_fit3$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -419.95 -62.15 -9.81 51.86 496.22   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -87.380 5.870 -14.885 < 2e-16 \*\*\*  
## atemp 78.844 32.281 2.442 0.01460 \*   
## temp 216.066 31.121 6.943 3.98e-12 \*\*\*  
## hr\_X1 -17.876 5.848 -3.056 0.00224 \*\*   
## hr\_X2 -26.890 5.868 -4.583 4.62e-06 \*\*\*  
## hr\_X3 -37.846 5.908 -6.406 1.54e-10 \*\*\*  
## hr\_X4 -41.131 5.911 -6.958 3.57e-12 \*\*\*  
## hr\_X5 -24.934 5.872 -4.246 2.19e-05 \*\*\*  
## hr\_X6 33.478 5.857 5.716 1.11e-08 \*\*\*  
## hr\_X7 169.409 5.849 28.962 < 2e-16 \*\*\*  
## hr\_X8 310.667 5.844 53.160 < 2e-16 \*\*\*  
## hr\_X9 164.717 5.844 28.184 < 2e-16 \*\*\*  
## hr\_X10 111.894 5.853 19.117 < 2e-16 \*\*\*  
## hr\_X11 139.416 5.871 23.748 < 2e-16 \*\*\*  
## hr\_X12 180.485 5.890 30.643 < 2e-16 \*\*\*  
## hr\_X13 176.432 5.908 29.862 < 2e-16 \*\*\*  
## hr\_X14 160.800 5.926 27.135 < 2e-16 \*\*\*  
## hr\_X15 170.325 5.934 28.706 < 2e-16 \*\*\*  
## hr\_X16 231.935 5.929 39.122 < 2e-16 \*\*\*  
## hr\_X17 385.056 5.911 65.142 < 2e-16 \*\*\*  
## hr\_X18 352.364 5.893 59.790 < 2e-16 \*\*\*  
## hr\_X19 241.801 5.870 41.191 < 2e-16 \*\*\*  
## hr\_X20 161.224 5.857 27.527 < 2e-16 \*\*\*  
## hr\_X21 110.339 5.847 18.871 < 2e-16 \*\*\*  
## hr\_X22 72.389 5.842 12.390 < 2e-16 \*\*\*  
## hr\_X23 33.223 5.840 5.688 1.30e-08 \*\*\*  
## weathersit\_Misty -19.357 1.981 -9.773 < 2e-16 \*\*\*  
## weathersit\_LightPrecip -90.434 3.171 -28.520 < 2e-16 \*\*\*  
## weathersit\_HeavyPrecip -78.228 64.398 -1.215 0.22447   
## season\_Spring 35.220 5.307 6.636 3.31e-11 \*\*\*  
## season\_Summer 26.909 6.288 4.279 1.88e-05 \*\*\*  
## season\_Fall 64.727 5.332 12.140 < 2e-16 \*\*\*  
## mnth\_Feb 1.079 4.288 0.252 0.80130   
## mnth\_Mar 5.045 4.817 1.047 0.29494   
## mnth\_Apr -6.072 7.151 -0.849 0.39585   
## mnth\_May -5.481 7.647 -0.717 0.47351   
## mnth\_Jun -17.112 7.859 -2.177 0.02948 \*   
## mnth\_Jul -40.701 8.818 -4.616 3.94e-06 \*\*\*  
## mnth\_Aug -19.706 8.594 -2.293 0.02186 \*   
## mnth\_Sep 4.982 7.626 0.653 0.51356   
## mnth\_Oct -3.997 7.079 -0.565 0.57229   
## mnth\_Nov -18.461 6.822 -2.706 0.00682 \*\*   
## mnth\_Dec -15.458 5.411 -2.857 0.00429 \*\*   
## weekday\_Sunday -16.015 3.148 -5.087 3.67e-07 \*\*\*  
## weekday\_Monday -8.030 3.249 -2.471 0.01347 \*   
## weekday\_Tuesday -6.632 3.172 -2.091 0.03653 \*   
## weekday\_Wednesday -3.772 3.165 -1.192 0.23332   
## weekday\_Thursday -2.437 3.164 -0.770 0.44121   
## weekday\_Friday 1.862 3.155 0.590 0.55498   
## holiday\_Holiday -25.537 5.340 -4.782 1.75e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 111.3 on 17329 degrees of freedom  
## Multiple R-squared: 0.6244, Adjusted R-squared: 0.6233   
## F-statistic: 587.8 on 49 and 17329 DF, p-value: < 2.2e-16

car::vif(lm\_fit3$fit$fit$fit)

## atemp temp hr\_X1   
## 43.150859 50.352358 1.914858   
## hr\_X2 hr\_X3 hr\_X4   
## 1.904483 1.884314 1.886095   
## hr\_X5 hr\_X6 hr\_X7   
## 1.912460 1.922919 1.923048   
## hr\_X8 hr\_X9 hr\_X10   
## 1.919485 1.919670 1.925481   
## hr\_X11 hr\_X12 hr\_X13   
## 1.937128 1.952335 1.967139   
## hr\_X14 hr\_X15 hr\_X16   
## 1.978911 1.983959 1.983232   
## hr\_X17 hr\_X18 hr\_X19   
## 1.971559 1.954657 1.939311   
## hr\_X20 hr\_X21 hr\_X22   
## 1.930486 1.924020 1.920926   
## hr\_X23 weathersit\_Misty weathersit\_LightPrecip   
## 1.919603 1.062179 1.057155   
## weathersit\_HeavyPrecip season\_Spring season\_Summer   
## 1.003646 7.478051 10.632325   
## season\_Fall mnth\_Feb mnth\_Mar   
## 7.343002 1.835944 2.523863   
## mnth\_Apr mnth\_May mnth\_Jun   
## 5.438602 6.419604 6.581995   
## mnth\_Jul mnth\_Aug mnth\_Sep   
## 8.535348 8.044539 6.185239   
## mnth\_Oct mnth\_Nov mnth\_Dec   
## 5.376261 4.950186 3.204742   
## weekday\_Sunday weekday\_Monday weekday\_Tuesday   
## 1.712389 1.810506 1.709827   
## weekday\_Wednesday weekday\_Thursday weekday\_Friday   
## 1.715641 1.712306 1.711313   
## holiday\_Holiday   
## 1.117447

After running a forward stepwise variable method, I was able to find the most improved model based on that report. From that report, I was able to further eliminate hum and windspeed- those two variables seem to have been affected by multicollinearity but also did not appear to make much sense (sanity check!) for the overall report as it relates to the affect of count of bikes rented. From this, we can see an adjusted R-sqaure of 0.6233 and a multiple R-squared of 0.6244. Moreover, some variables still appear to be affected by multicollinearity with holiday, weekday, and weathersit. I thought that there effect would still provide some insight so I did not reduce the query any further. With the number of predictors involved, the report is complex. I was unable to create an upper limit variable (error message from ~., too many variables with <2 factors apparently) initially, but went through thereafter and selected all the non-categorical variables to juxtapose the empty variable for the forward stepwise. From there, I created a model to provide the above table to analyze our results. Truth be told, I feel uncomfortable with this material still. I feel the readings all touch on the same points the lectures do, but I am confident that I have missed something crucial in this last part of the assignment. Nonetheless, this appears to be the “best” model by the technical definition and the multicollinearity can be explained and identified easily (with a VIF for backup, of course).