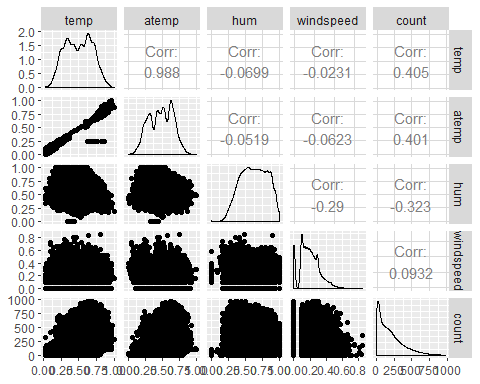
library(tidyverse)  
library(GGally)  
library(MASS)  
library(car)  
library(gridExtra)

bike = read\_csv("hour.csv")

## Parsed with column specification:  
## cols(  
## instant = col\_double(),  
## dteday = col\_date(format = ""),  
## season = col\_double(),  
## yr = col\_double(),  
## mnth = col\_double(),  
## hr = col\_double(),  
## holiday = col\_double(),  
## weekday = col\_double(),  
## workingday = col\_double(),  
## weathersit = col\_double(),  
## temp = col\_double(),  
## atemp = col\_double(),  
## hum = col\_double(),  
## windspeed = col\_double(),  
## casual = col\_double(),  
## registered = col\_double(),  
## count = col\_double()  
## )

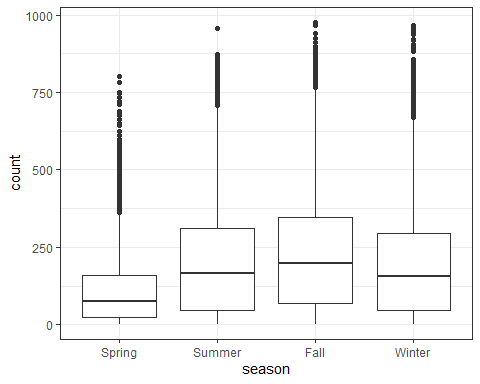
bike = bike %>%   
 mutate(season = as\_factor(as.character(season))) %>%  
 mutate(season = fct\_recode(season,  
 "Spring" = "1",  
 "Summer" = "2",  
 "Fall" = "3",  
 "Winter" = "4")) %>%  
 mutate(yr = as\_factor(as.character(yr))) %>%  
 mutate(mnth = as\_factor(as.character(mnth))) %>%  
 mutate(hr = as\_factor(as.character(hr))) %>%  
 mutate(holiday = as\_factor(as.character(holiday))) %>%  
 mutate(holiday = fct\_recode(holiday,  
 "NotHoliday" = "0",  
 "Holiday" = "1")) %>%  
 mutate(workingday = as\_factor(as.character(workingday))) %>%  
 mutate(workingday = fct\_recode(workingday,  
 "NotWorkingDay" = "0",  
 "WorkingDay" = "1")) %>%   
 mutate(weathersit = as\_factor(as.character(weathersit))) %>%  
 mutate(weathersit = fct\_recode(weathersit,  
 "NoPrecip" = "1",  
 "Misty" = "2",  
 "LightPrecip" = "3",  
 "HeavyPrecip" = "4")) %>%  
 mutate(weekday = as\_factor(as.character(weekday))) %>%  
 mutate(weekday = fct\_recode(weekday,  
 "Sunday" = "0",  
 "Monday" = "1",  
 "Tuesday" = "2",  
 "Wednesday" = "3",  
 "Thursday" = "4",  
 "Friday" = "5",  
 "Saturday" = "6"))

ggpairs(bike, columns=c("temp","atemp","hum","windspeed","count"))

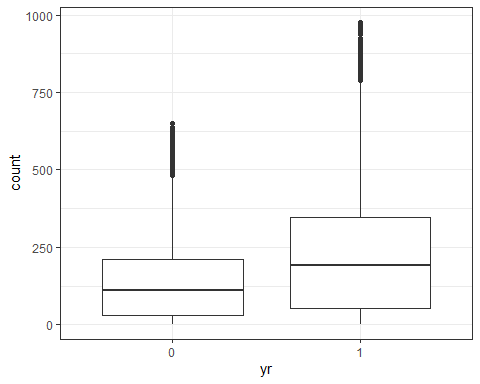


“Temp” seems to be the quantitative variable that is best correlated with “Count”. It produces a correlation of 0.405, narrowly beating out “atemp”.

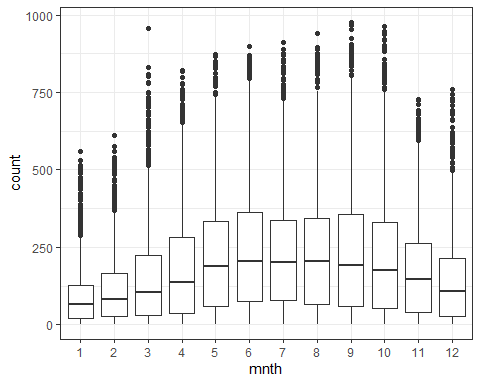
ggplot(bike, aes(season, count)) +  
 geom\_boxplot() +  
 theme\_bw()



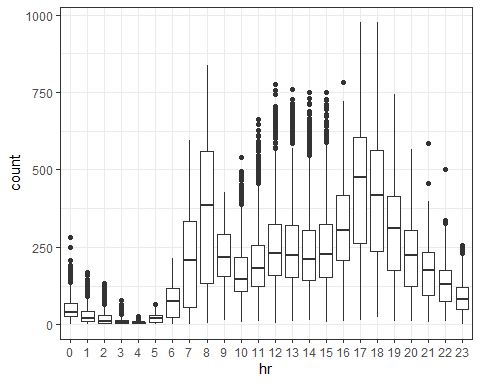
ggplot(bike, aes(yr, count)) +  
 geom\_boxplot() +  
 theme\_bw()



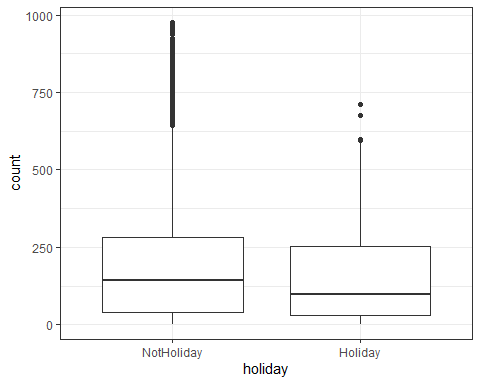
ggplot(bike, aes(mnth, count)) +  
 geom\_boxplot() +  
 theme\_bw()



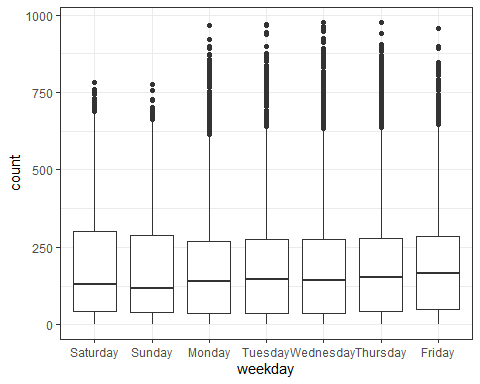
ggplot(bike, aes(hr, count)) +  
 geom\_boxplot() +  
 theme\_bw()



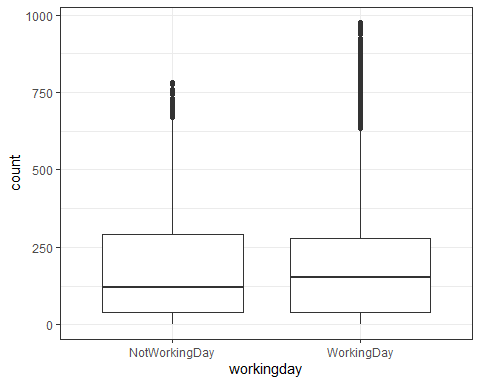
ggplot(bike, aes(holiday, count)) +  
 geom\_boxplot() +  
 theme\_bw()



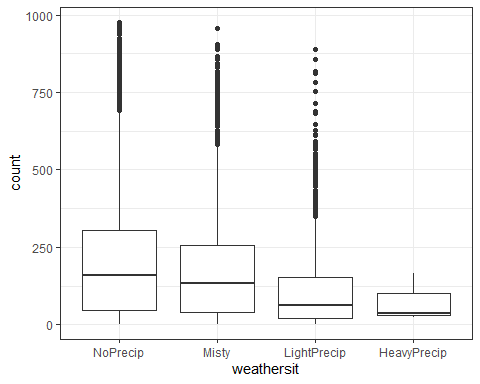
ggplot(bike, aes(weekday, count)) +  
 geom\_boxplot() +  
 theme\_bw()



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



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



The box and whisker plot is designed to show variance among the interquartile range and median of the y-value (count) affected by the x-values (variables). So, we can conclude that the variables that have box and whisker plots with very little change in y-value do not affect the count much. Conversely, if there is a bunch of variance among the interquartile range and median of a plot, we can conclude that that particular variable has an affect on the count, thus, there is a correlation.

With that being said, the box and whisker plots for yr, holiday, weekday, and workingday all show very little variance and therefore do not appear to affect count. However, the season, mnth, hr, and weathersit all have considerable changes in y-values (specifically hr). Logically, this makes sense as count totals would obviously be affected by the hour of the day, the season and thus the month, and the type of weather, since this is an outdoor activity.

bike2 = bike %>% dplyr::select(-c(instant, dteday, registered, casual))  
allmod = lm(count ~., bike2)  
emptymod = lm(count ~1, bike2)

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

## 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  
## + yr 1 35876722 535884870 179641  
## + 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  
## + yr 1 36875130 248151780 166307  
## + 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  
## + yr 1 33463769 201044200 162650  
## + 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=162650.2  
## count ~ hr + atemp + yr  
##   
## Df Sum of Sq RSS AIC  
## + weathersit 3 8408358 192635842 161914  
## + season 3 7190305 193853896 162023  
## + mnth 11 6486062 194558138 162102  
## + hum 1 4341837 196702363 162273  
## + weekday 6 641648 200402552 162607  
## + holiday 1 324763 200719438 162624  
## + windspeed 1 109311 200934889 162643  
## + workingday 1 106404 200937797 162643  
## + temp 1 91735 200952465 162644  
## <none> 201044200 162650  
##   
## Step: AIC=161913.7  
## count ~ hr + atemp + yr + weathersit  
##   
## Df Sum of Sq RSS AIC  
## + season 3 7771024 184864818 161204  
## + mnth 11 7464989 185170852 161249  
## + hum 1 805099 191830743 161843  
## + weekday 6 686172 191949670 161864  
## + holiday 1 413536 192222305 161878  
## + workingday 1 212428 192423414 161897  
## + temp 1 134482 192501360 161904  
## + windspeed 1 44407 192591435 161912  
## <none> 192635842 161914  
##   
## Step: AIC=161204.1  
## count ~ hr + atemp + yr + weathersit + season  
##   
## Df Sum of Sq RSS AIC  
## + mnth 11 2051323 182813495 161032  
## + hum 1 1810161 183054657 161035  
## + weekday 6 704303 184160515 161150  
## + holiday 1 392702 184472116 161169  
## + temp 1 352584 184512234 161173  
## + workingday 1 214973 184649845 161186  
## <none> 184864818 161204  
## + windspeed 1 158 184864660 161206  
##   
## Step: AIC=161032.2  
## count ~ hr + atemp + yr + weathersit + season + mnth  
##   
## Df Sum of Sq RSS AIC  
## + hum 1 2356411 180457084 160809  
## + weekday 6 692672 182120823 160978  
## + holiday 1 312321 182501174 161004  
## + temp 1 233052 182580443 161012  
## + workingday 1 203953 182609542 161015  
## <none> 182813495 161032  
## + windspeed 1 68 182813428 161034  
##   
## Step: AIC=160808.7  
## count ~ hr + atemp + yr + weathersit + season + mnth + hum  
##   
## Df Sum of Sq RSS AIC  
## + weekday 6 581105 179875980 160765  
## + holiday 1 322997 180134087 160780  
## + workingday 1 194139 180262945 160792  
## + windspeed 1 114287 180342797 160800  
## + temp 1 100025 180357059 160801  
## <none> 180457084 160809  
##   
## Step: AIC=160764.7  
## count ~ hr + atemp + yr + weathersit + season + mnth + hum +   
## weekday  
##   
## Df Sum of Sq RSS AIC  
## + holiday 1 274717 179601263 160740  
## + workingday 1 274717 179601263 160740  
## + windspeed 1 112085 179763895 160756  
## + temp 1 77171 179798809 160759  
## <none> 179875980 160765  
##   
## Step: AIC=160740.1  
## count ~ hr + atemp + yr + weathersit + season + mnth + hum +   
## weekday + holiday  
##   
## Df Sum of Sq RSS AIC  
## + windspeed 1 111562 179489701 160731  
## + temp 1 95460 179505803 160733  
## <none> 179601263 160740  
##   
## Step: AIC=160731.3  
## count ~ hr + atemp + yr + weathersit + season + mnth + hum +   
## weekday + holiday + windspeed  
##   
## Df Sum of Sq RSS AIC  
## + temp 1 160954 179328746 160718  
## <none> 179489701 160731  
##   
## Step: AIC=160717.7  
## count ~ hr + atemp + yr + weathersit + season + mnth + hum +   
## weekday + holiday + windspeed + temp  
##   
## Df Sum of Sq RSS AIC  
## <none> 179328746 160718

summary(forwardmod)

##   
## Call:  
## lm(formula = count ~ hr + atemp + yr + weathersit + season +   
## mnth + hum + weekday + holiday + windspeed + temp, data = bike2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -393.87 -60.66 -7.96 51.31 439.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -67.542 6.612 -10.216 < 2e-16 \*\*\*  
## hr1 -17.294 5.345 -3.236 0.00122 \*\*   
## hr2 -26.369 5.364 -4.916 8.91e-07 \*\*\*  
## hr3 -37.112 5.403 -6.869 6.67e-12 \*\*\*  
## hr4 -40.263 5.408 -7.445 1.01e-13 \*\*\*  
## hr5 -23.501 5.373 -4.374 1.23e-05 \*\*\*  
## hr6 35.393 5.359 6.605 4.10e-11 \*\*\*  
## hr7 170.418 5.348 31.864 < 2e-16 \*\*\*  
## hr8 310.801 5.342 58.183 < 2e-16 \*\*\*  
## hr9 163.101 5.347 30.501 < 2e-16 \*\*\*  
## hr10 108.444 5.370 20.196 < 2e-16 \*\*\*  
## hr11 133.843 5.409 24.742 < 2e-16 \*\*\*  
## hr12 173.142 5.456 31.735 < 2e-16 \*\*\*  
## hr13 168.102 5.494 30.600 < 2e-16 \*\*\*  
## hr14 152.249 5.525 27.558 < 2e-16 \*\*\*  
## hr15 161.707 5.535 29.213 < 2e-16 \*\*\*  
## hr16 223.834 5.524 40.522 < 2e-16 \*\*\*  
## hr17 377.535 5.491 68.750 < 2e-16 \*\*\*  
## hr18 345.587 5.455 63.350 < 2e-16 \*\*\*  
## hr19 236.919 5.404 43.841 < 2e-16 \*\*\*  
## hr20 157.293 5.375 29.266 < 2e-16 \*\*\*  
## hr21 107.840 5.353 20.147 < 2e-16 \*\*\*  
## hr22 70.907 5.343 13.272 < 2e-16 \*\*\*  
## hr23 32.112 5.338 6.015 1.83e-09 \*\*\*  
## atemp 127.975 30.624 4.179 2.94e-05 \*\*\*  
## yr1 85.431 1.563 54.658 < 2e-16 \*\*\*  
## weathersitMisty -10.409 1.920 -5.421 6.00e-08 \*\*\*  
## weathersitLightPrecip -65.189 3.236 -20.145 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -62.580 58.893 -1.063 0.28797   
## seasonSummer 38.178 4.856 7.862 4.00e-15 \*\*\*  
## seasonFall 32.055 5.749 5.575 2.51e-08 \*\*\*  
## seasonWinter 67.994 4.882 13.928 < 2e-16 \*\*\*  
## mnth2 3.426 3.920 0.874 0.38219   
## mnth3 14.299 4.407 3.244 0.00118 \*\*   
## mnth4 6.230 6.548 0.951 0.34144   
## mnth5 20.657 7.007 2.948 0.00320 \*\*   
## mnth6 6.238 7.205 0.866 0.38662   
## mnth7 -13.269 8.082 -1.642 0.10065   
## mnth8 7.897 7.879 1.002 0.31622   
## mnth9 32.269 7.001 4.609 4.07e-06 \*\*\*  
## mnth10 15.843 6.483 2.444 0.01455 \*   
## mnth11 -9.840 6.238 -1.577 0.11474   
## mnth12 -6.256 4.954 -1.263 0.20672   
## hum -82.802 5.554 -14.909 < 2e-16 \*\*\*  
## weekdaySunday -16.089 2.878 -5.591 2.30e-08 \*\*\*  
## weekdayMonday -6.814 2.970 -2.294 0.02180 \*   
## weekdayTuesday -5.240 2.899 -1.807 0.07071 .   
## weekdayWednesday -2.464 2.894 -0.851 0.39469   
## weekdayThursday -2.940 2.892 -1.016 0.30947   
## weekdayFriday 1.356 2.885 0.470 0.63823   
## holidayHoliday -26.228 4.881 -5.374 7.81e-08 \*\*\*  
## windspeed -29.167 7.052 -4.136 3.55e-05 \*\*\*  
## temp 116.384 29.513 3.943 8.06e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 101.7 on 17326 degrees of freedom  
## Multiple R-squared: 0.6864, Adjusted R-squared: 0.6854   
## F-statistic: 729.1 on 52 and 17326 DF, p-value: < 2.2e-16

There are 11 variables included in my forward model - hr, atemp, yr, weathersit, season, mnth, hum, weekday, holiday, windspeed, and temp. There is an adjusted R-squared value of about .69 which is okay, but could be higher. Furthermore, there are multiple negative coefficients among multiple variables, so I’m thinking there is definitely some multicollinearity going on. This may explain why a variables like weekday and holiday, ones whose box and whisker plots did not show a ton of variation, are included in the model. Perhaps the model will change once multicollinearity is accounted for.

backmod = stepAIC(allmod, direction="backward", trace=TRUE)

## Start: AIC=160717.7  
## count ~ season + yr + mnth + hr + holiday + weekday + workingday +   
## weathersit + temp + atemp + hum + windspeed  
##   
##   
## Step: AIC=160717.7  
## count ~ season + yr + mnth + hr + holiday + weekday + weathersit +   
## temp + atemp + hum + windspeed  
##   
## Df Sum of Sq RSS AIC  
## <none> 179328746 160718  
## - temp 1 160954 179489701 160731  
## - windspeed 1 177057 179505803 160733  
## - atemp 1 180751 179509498 160733  
## - holiday 1 298893 179627639 160745  
## - weekday 6 498795 179827541 160754  
## - mnth 11 2426171 181754917 160929  
## - hum 1 2300667 181629413 160937  
## - season 3 2398467 181727213 160943  
## - weathersit 3 4208731 183537478 161115  
## - yr 1 30920851 210249597 163480  
## - hr 23 196741474 376070220 173542

summary(backmod)

##   
## Call:  
## lm(formula = count ~ season + yr + mnth + hr + holiday + weekday +   
## weathersit + temp + atemp + hum + windspeed, data = bike2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -393.87 -60.66 -7.96 51.31 439.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -67.542 6.612 -10.216 < 2e-16 \*\*\*  
## seasonSummer 38.178 4.856 7.862 4.00e-15 \*\*\*  
## seasonFall 32.055 5.749 5.575 2.51e-08 \*\*\*  
## seasonWinter 67.994 4.882 13.928 < 2e-16 \*\*\*  
## yr1 85.431 1.563 54.658 < 2e-16 \*\*\*  
## mnth2 3.426 3.920 0.874 0.38219   
## mnth3 14.299 4.407 3.244 0.00118 \*\*   
## mnth4 6.230 6.548 0.951 0.34144   
## mnth5 20.657 7.007 2.948 0.00320 \*\*   
## mnth6 6.238 7.205 0.866 0.38662   
## mnth7 -13.269 8.082 -1.642 0.10065   
## mnth8 7.897 7.879 1.002 0.31622   
## mnth9 32.269 7.001 4.609 4.07e-06 \*\*\*  
## mnth10 15.843 6.483 2.444 0.01455 \*   
## mnth11 -9.840 6.238 -1.577 0.11474   
## mnth12 -6.256 4.954 -1.263 0.20672   
## hr1 -17.294 5.345 -3.236 0.00122 \*\*   
## hr2 -26.369 5.364 -4.916 8.91e-07 \*\*\*  
## hr3 -37.112 5.403 -6.869 6.67e-12 \*\*\*  
## hr4 -40.263 5.408 -7.445 1.01e-13 \*\*\*  
## hr5 -23.501 5.373 -4.374 1.23e-05 \*\*\*  
## hr6 35.393 5.359 6.605 4.10e-11 \*\*\*  
## hr7 170.418 5.348 31.864 < 2e-16 \*\*\*  
## hr8 310.801 5.342 58.183 < 2e-16 \*\*\*  
## hr9 163.101 5.347 30.501 < 2e-16 \*\*\*  
## hr10 108.444 5.370 20.196 < 2e-16 \*\*\*  
## hr11 133.843 5.409 24.742 < 2e-16 \*\*\*  
## hr12 173.142 5.456 31.735 < 2e-16 \*\*\*  
## hr13 168.102 5.494 30.600 < 2e-16 \*\*\*  
## hr14 152.249 5.525 27.558 < 2e-16 \*\*\*  
## hr15 161.707 5.535 29.213 < 2e-16 \*\*\*  
## hr16 223.834 5.524 40.522 < 2e-16 \*\*\*  
## hr17 377.535 5.491 68.750 < 2e-16 \*\*\*  
## hr18 345.587 5.455 63.350 < 2e-16 \*\*\*  
## hr19 236.919 5.404 43.841 < 2e-16 \*\*\*  
## hr20 157.293 5.375 29.266 < 2e-16 \*\*\*  
## hr21 107.840 5.353 20.147 < 2e-16 \*\*\*  
## hr22 70.907 5.343 13.272 < 2e-16 \*\*\*  
## hr23 32.112 5.338 6.015 1.83e-09 \*\*\*  
## holidayHoliday -26.228 4.881 -5.374 7.81e-08 \*\*\*  
## weekdaySunday -16.089 2.878 -5.591 2.30e-08 \*\*\*  
## weekdayMonday -6.814 2.970 -2.294 0.02180 \*   
## weekdayTuesday -5.240 2.899 -1.807 0.07071 .   
## weekdayWednesday -2.464 2.894 -0.851 0.39469   
## weekdayThursday -2.940 2.892 -1.016 0.30947   
## weekdayFriday 1.356 2.885 0.470 0.63823   
## weathersitMisty -10.409 1.920 -5.421 6.00e-08 \*\*\*  
## weathersitLightPrecip -65.189 3.236 -20.145 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -62.580 58.893 -1.063 0.28797   
## temp 116.384 29.513 3.943 8.06e-05 \*\*\*  
## atemp 127.975 30.624 4.179 2.94e-05 \*\*\*  
## hum -82.802 5.554 -14.909 < 2e-16 \*\*\*  
## windspeed -29.167 7.052 -4.136 3.55e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 101.7 on 17326 degrees of freedom  
## Multiple R-squared: 0.6864, Adjusted R-squared: 0.6854   
## F-statistic: 729.1 on 52 and 17326 DF, p-value: < 2.2e-16

There are also 11 variables in my backward model, and all 11 are the same as my forward model. The adjusted R-squared value is also the same. Just by eyeballing it, from what I can tell, there isn’t a whole lot of difference between the two models in terms of results. The correlation values and coefficients seem the same in both models. The main noticeable difference is the list order of the variables - they seem to have filled themselves with each variable at different stages of the process. This seems logical given the difference in approach between the two models.

Task 6 - So, I think I may lose points for this. Not quite sure what happened, but I don’t get coefficients and p values for the model used to begin the backward stepwise approach. I don’t see anything listed as NA. I can easily see that workingday was removed in the next model, but RStudio is only showing the last model’s coefficients and p values.

That being said, thinking logically, workingday would be perfectly correlated with the weekday variable. This is confirmed when taking a look at the dataset.

I would have reservations in using this model. First and foremost, I feel like the R-squared value isn’t too high. I would like to get it a bit closer to 1. There is also evidence of multicollinearity. The model already took out workingday since it was perfectly correlated with weekday, but, thinking logically, I speculate there are additional variables that are more correlated to themselves than the count variable. For example, mnth and season may be heavily correlated, and weathersit and mnth may be as well. Overall, I would prefer looking into other model options to analyze this dataset multi-linearlly.