# Multiple Linear Regression Assignment - Mod 2

## Khayrayyah Haamid-Day

library(tidyverse)  
library(tidymodels)  
library(glmnet)  
library(GGally)  
library(ggcorrplot)  
library(MASS)  
library(car)  
library(lubridate)  
library(lmtest)

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

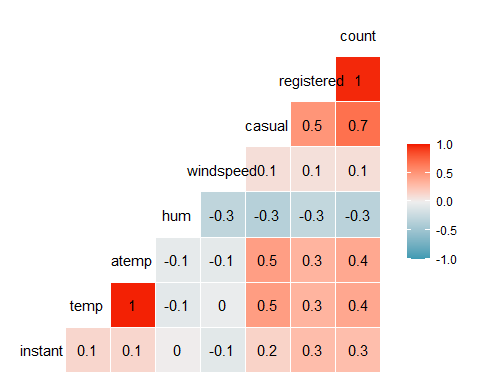
## Rows: 17379 Columns: 16

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

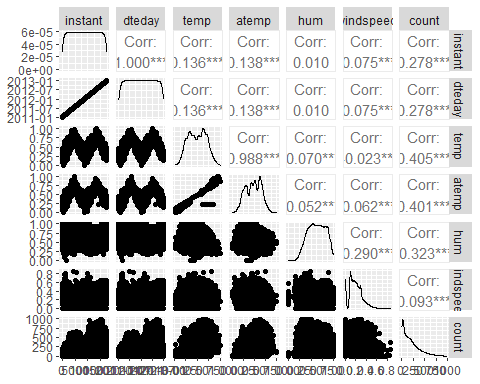
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

bike = bike %>% mutate(dteday = mdy(dteday)) %>%   
 mutate\_if(is.character, as.factor) %>%  
 mutate(hr = as.factor(hr)) #Task 1  
  
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



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



summary(bike)

## instant dteday season mnth   
## Min. : 1 Min. :2011-01-01 Fall :4232 Jul :1488   
## 1st Qu.: 4346 1st Qu.:2011-07-04 Spring:4409 May :1488   
## Median : 8690 Median :2012-01-02 Summer:4496 Dec :1483   
## Mean : 8690 Mean :2012-01-02 Winter:4242 Aug :1475   
## 3rd Qu.:13034 3rd Qu.:2012-07-02 Mar :1473   
## Max. :17379 Max. :2012-12-31 Oct :1451   
## (Other):8521   
## hr holiday weekday workingday   
## 16 : 730 Holiday : 500 Friday :2487 NotWorkingDay: 5514   
## 17 : 730 NotHoliday:16879 Monday :2479 WorkingDay :11865   
## 13 : 729 Saturday :2512   
## 14 : 729 Sunday :2502   
## 15 : 729 Thursday :2471   
## 12 : 728 Tuesday :2453   
## (Other):13004 Wednesday:2475   
## weathersit temp atemp hum   
## HeavyPrecip: 3 Min. :0.020 Min. :0.0000 Min. :0.0000   
## LightPrecip: 1419 1st Qu.:0.340 1st Qu.:0.3333 1st Qu.:0.4800   
## Misty : 4544 Median :0.500 Median :0.4848 Median :0.6300   
## NoPrecip :11413 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   
##

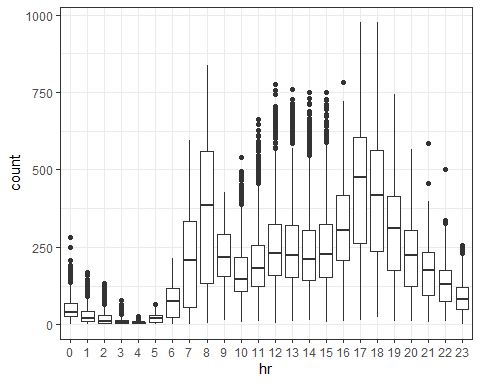
**Task 1 - Why do we convert the “hr” variable into factor? Why not just leave as numbers?**

We would want to convert hr to a factor so we can use it to compare with other categorical data.

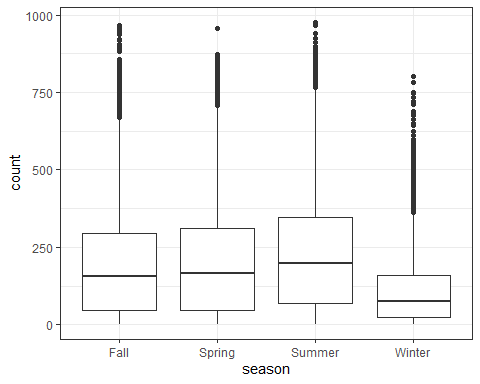
**Task 2 - Which of the quantitative variables appears to be best correlated with “count” (ignore the “registered” and “casual” variable as the sum of these two variables equals “count”)?**

The variables atemp and temp appear to be best correlated with the variable count.

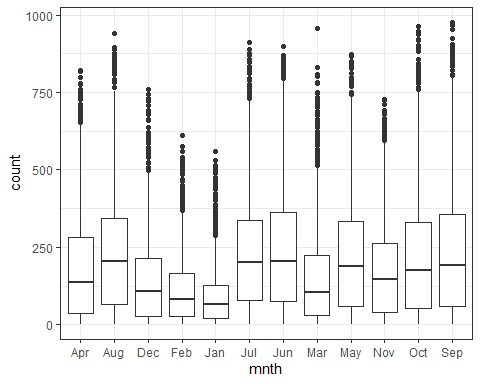
#Task 3  
ggplot(bike,aes(x=hr,y=count)) + geom\_boxplot() + theme\_bw()



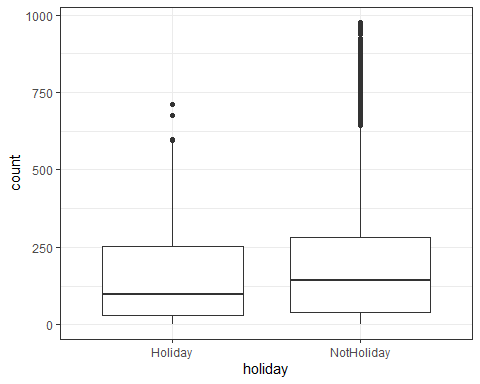
ggplot(bike,aes(x=season,y=count)) + geom\_boxplot() + theme\_bw()



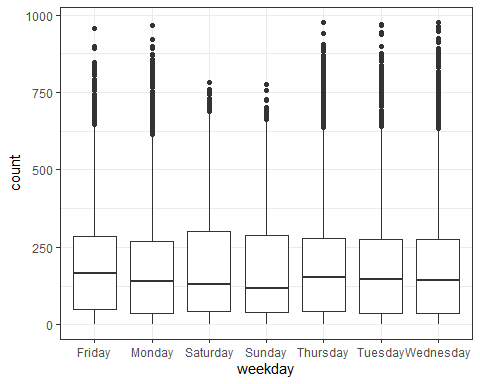
ggplot(bike,aes(x=mnth,y=count)) + geom\_boxplot() + theme\_bw()



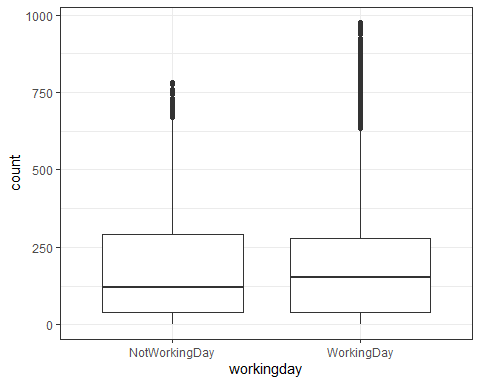
ggplot(bike,aes(x=holiday,y=count)) + geom\_boxplot() + theme\_bw()



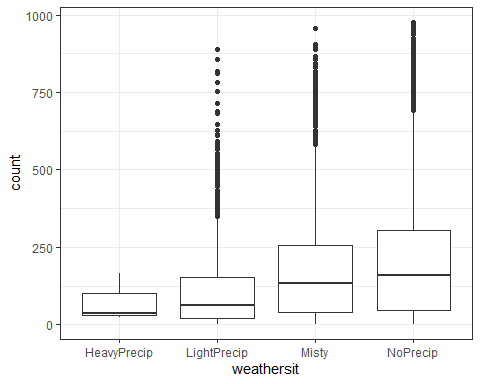
ggplot(bike,aes(x=weekday,y=count)) + geom\_boxplot() + theme\_bw()



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



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



**Task 3 - Which variables appear to affect “count”? Provide a brief explanation as to why you believe that each variable does or does not affect “count” (use your intuition to help you answer this question).**

* Season - This variable increases in count when the season is warmer and vastly decreases in winter when it would be a lot colder in DC as well as harder to bike with snowfall.
* Mnth - In months that are colder there are less people using the bike service in comparison to the warmer weather months. In particular, this service is used most during the hottest months of summer.
* Holiday - There is not too strong of a relationship between count and holiday. Usage of the bikes remains fairly similar whether its a holiday or not. This may be due to the fact that people are just as active outdoors whether its a holiday or not.
* Weekday - There is not a too strong of a relationship between count and weekday. People in DC use public transportation as a way to get around the city so they would still require the bikes no matter the day of the week.
* Workingday - Just like holiday and weekday, this is not too strong of a relationship with count and this is most likely due to the continued need to get around town regardless of the day.
* Weathersit - There is a strong relationship with count, this makes sense as people are less likely to use a bike when it is raining outside.

bikemodel = recipe(count ~ mnth, bike) %>%  
 step\_dummy(mnth)  
  
lm\_model = #give the model type a name   
 linear\_reg() %>% #specify that we are doing linear regression  
 set\_engine("lm") #specify the specify type of linear tool we want to use   
  
lm\_wflow =   
 workflow() %>%   
 add\_model(lm\_model) %>%   
 add\_recipe(bikemodel)  
  
lm\_fit1 = fit(lm\_wflow, bike)  
  
  
summary(lm\_fit1$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -239.77 -124.91 -37.42 85.72 801.59   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 187.261 4.603 40.679 < 2e-16 \*\*\*  
## mnth\_Aug 50.837 6.468 7.860 4.08e-15 \*\*\*  
## mnth\_Dec -44.958 6.459 -6.960 3.53e-12 \*\*\*  
## mnth\_Feb -74.396 6.626 -11.228 < 2e-16 \*\*\*  
## mnth\_Jan -92.836 6.519 -14.240 < 2e-16 \*\*\*  
## mnth\_Jul 44.559 6.454 6.904 5.23e-12 \*\*\*  
## mnth\_Jun 53.254 6.507 8.184 2.92e-16 \*\*\*  
## mnth\_Mar -31.850 6.470 -4.923 8.62e-07 \*\*\*  
## mnth\_May 35.646 6.454 5.523 3.38e-08 \*\*\*  
## mnth\_Nov -9.926 6.510 -1.525 0.127   
## mnth\_Oct 34.898 6.494 5.373 7.82e-08 \*\*\*  
## mnth\_Sep 53.512 6.510 8.220 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 174.5 on 17367 degrees of freedom  
## Multiple R-squared: 0.07505, Adjusted R-squared: 0.07446   
## F-statistic: 128.1 on 11 and 17367 DF, p-value: < 2.2e-16

**Task 4**

Though the R-Squared value is low, the p-values show that every variable except for November is significant. The negative coefficients show that as the month gets colder in DC, the amount of bikes being used decrease, while the opposite is true for the positive coefficients.

#Task 5  
bikemod = recipe(count ~., bike) %>%  
 step\_rm(instant,dteday,registered,casual)  
  
all\_model =   
 linear\_reg() %>%   
 set\_engine("lm")  
  
lm\_wflow =   
 workflow() %>%   
 add\_model(all\_model) %>%   
 add\_recipe(bikemod)  
  
lm\_fit2 = fit(lm\_wflow, bike)  
  
summary(lm\_fit2$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## 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) -14.5381 64.9466 -0.224 0.822879   
## seasonSpring -26.5573 6.2028 -4.282 1.87e-05 \*\*\*  
## seasonSummer -37.1901 5.5866 -6.657 2.88e-11 \*\*\*  
## seasonWinter -66.0479 5.2858 -12.495 < 2e-16 \*\*\*  
## mnthAug -6.4051 7.0307 -0.911 0.362305   
## mnthDec -0.1168 6.9419 -0.017 0.986575   
## mnthFeb 8.7386 6.9343 1.260 0.207610   
## mnthJan 9.1995 7.0836 1.299 0.194061   
## mnthJul -31.5694 7.1235 -4.432 9.40e-06 \*\*\*  
## mnthJun -11.5345 4.9339 -2.338 0.019409 \*   
## mnthMar 13.6198 5.3742 2.534 0.011276 \*   
## mnthMay 9.7147 4.3243 2.247 0.024680 \*   
## mnthNov -6.3831 7.5801 -0.842 0.399751   
## mnthOct 13.1412 7.4558 1.763 0.077997 .   
## mnthSep 24.7054 6.6167 3.734 0.000189 \*\*\*  
## 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 \*\*\*  
## holidayNotHoliday 26.3866 5.2846 4.993 6.00e-07 \*\*\*  
## weekdayMonday -7.5761 3.2058 -2.363 0.018127 \*   
## weekdaySaturday -0.9168 3.1234 -0.294 0.769126   
## weekdaySunday -15.6437 3.1311 -4.996 5.90e-07 \*\*\*  
## weekdayThursday -4.6307 3.1359 -1.477 0.139791   
## weekdayTuesday -6.3589 3.1461 -2.021 0.043276 \*   
## weekdayWednesday -3.3337 3.1382 -1.062 0.288116   
## workingdayWorkingDay NA NA NA NA   
## weathersitLightPrecip -27.2018 63.7810 -0.426 0.669758   
## weathersitMisty 26.7980 63.7544 0.420 0.674249   
## weathersitNoPrecip 33.1693 63.7637 0.520 0.602937   
## temp 195.1921 31.9175 6.116 9.83e-10 \*\*\*  
## atemp 103.9531 33.1547 3.135 0.001719 \*\*   
## 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

**Task 5**

One example of multicollinearity in this summary are the month variables, that show warmer weather months have negative coefficients which indicate which is implying that the warmer the weather is the less bikes being used. Previous information contradicts this. After attempting to run the vif function, there is an error that suggests there are variables that are linearly dependent upon each other.

allPredict = recipe(count ~.,bike) %>%  
 step\_rm(instant,dteday,registered,casual,windspeed,weekday,workingday,hum,weathersit,atemp,temp,season)  
  
all\_predict =   
 linear\_reg() %>%   
 set\_engine("lm")  
  
lm\_wflow =   
 workflow() %>%   
 add\_model(all\_predict) %>%   
 add\_recipe(allPredict)  
  
lm\_fit3 = fit(lm\_wflow, bike)  
summary(lm\_fit3$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -494.91 -60.83 -9.13 54.64 529.98   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.595 7.408 4.265 2.01e-05 \*\*\*  
## mnthAug 49.901 4.364 11.435 < 2e-16 \*\*\*  
## mnthDec -45.246 4.354 -10.391 < 2e-16 \*\*\*  
## mnthFeb -77.528 4.467 -17.358 < 2e-16 \*\*\*  
## mnthJan -96.065 4.396 -21.855 < 2e-16 \*\*\*  
## mnthJul 44.912 4.351 10.323 < 2e-16 \*\*\*  
## mnthJun 52.912 4.390 12.053 < 2e-16 \*\*\*  
## mnthMar -33.994 4.365 -7.787 7.23e-15 \*\*\*  
## mnthMay 35.999 4.351 8.274 < 2e-16 \*\*\*  
## mnthNov -9.193 4.392 -2.093 0.036368 \*   
## mnthOct 34.683 4.378 7.922 2.47e-15 \*\*\*  
## mnthSep 53.542 4.388 12.201 < 2e-16 \*\*\*  
## hr1 -20.403 6.178 -3.302 0.000961 \*\*\*  
## hr2 -31.605 6.198 -5.099 3.44e-07 \*\*\*  
## hr3 -44.720 6.238 -7.169 7.88e-13 \*\*\*  
## hr4 -50.144 6.238 -8.038 9.70e-16 \*\*\*  
## hr5 -34.790 6.193 -5.617 1.97e-08 \*\*\*  
## hr6 22.167 6.176 3.589 0.000333 \*\*\*  
## hr7 158.253 6.172 25.641 < 2e-16 \*\*\*  
## hr8 305.200 6.172 49.449 < 2e-16 \*\*\*  
## hr9 165.498 6.172 26.814 < 2e-16 \*\*\*  
## hr10 119.857 6.172 19.420 < 2e-16 \*\*\*  
## hr11 154.332 6.172 25.005 < 2e-16 \*\*\*  
## hr12 199.637 6.170 32.357 < 2e-16 \*\*\*  
## hr13 199.936 6.168 32.416 < 2e-16 \*\*\*  
## hr14 187.224 6.168 30.355 < 2e-16 \*\*\*  
## hr15 197.508 6.168 32.023 < 2e-16 \*\*\*  
## hr16 258.391 6.166 41.908 < 2e-16 \*\*\*  
## hr17 407.859 6.166 66.150 < 2e-16 \*\*\*  
## hr18 371.853 6.170 60.269 < 2e-16 \*\*\*  
## hr19 257.866 6.170 41.794 < 2e-16 \*\*\*  
## hr20 172.373 6.170 27.938 < 2e-16 \*\*\*  
## hr21 118.657 6.170 19.232 < 2e-16 \*\*\*  
## hr22 77.678 6.170 12.590 < 2e-16 \*\*\*  
## hr23 34.173 6.170 5.539 3.09e-08 \*\*\*  
## holidayNotHoliday 21.558 5.374 4.011 6.06e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 117.6 on 17343 degrees of freedom  
## Multiple R-squared: 0.5803, Adjusted R-squared: 0.5794   
## F-statistic: 685.1 on 35 and 17343 DF, p-value: < 2.2e-16

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

## GVIF Df GVIF^(1/(2\*Df))  
## mnth 1.013974 11 1.000631  
## hr 1.000371 23 1.000008  
## holiday 1.013622 1 1.006788

With this model there appears to be little multicollinearity, a moderate rsquared, and VIF of less than 4. I found that variables relating to the overall weather and season interacted with each other too much causing positive coefficients to become negative. With a bit of tweaking I found that the variable month, hr, and holiday do not interact with each other too strongly and positive coefficients remain so. There is only one variable in this model that is not as significant as the rest of the variables are.