# Module 2 - Mmultiple Linear Regression Assignment

## Katia Mottert

## Task 1

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.4 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

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

library(tidymodels)

## -- Attaching packages -------------------------------------- tidymodels 0.1.2 --

## v broom 0.7.3 v recipes 0.1.15  
## v dials 0.0.9 v rsample 0.0.8   
## v infer 0.5.4 v tune 0.1.2   
## v modeldata 0.1.0 v workflows 0.2.1   
## v parsnip 0.1.5 v yardstick 0.0.7

## -- 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()

library(GGally)

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

library(MASS)

##   
## Attaching package: 'MASS'

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

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

library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

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

## Loaded glmnet 4.1

library(ggcorrplot)  
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

bike = bike\_cleaned <- read\_csv("~/Courses/BAN502/Module2/MultLinearRegressionAssign/bike\_cleaned.csv")

##   
## -- Column specification --------------------------------------------------------  
## 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()  
## )

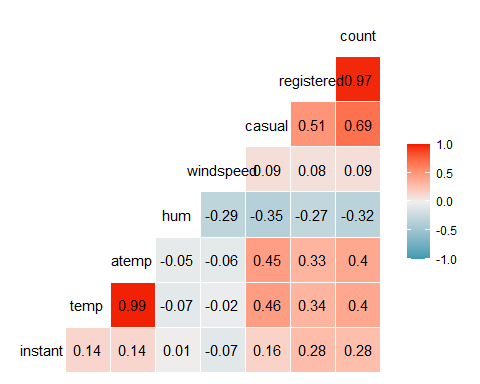
bike = bike %>%  
 mutate(dteday = mdy(dteday))  
  
bike = bike %>%  
 mutate\_if(is.character,as.factor)  
  
bike$hr <- factor(bike$hr)

. We change HR to a factor because a number will return a single coefficient. It could have been interpreted as a slope.

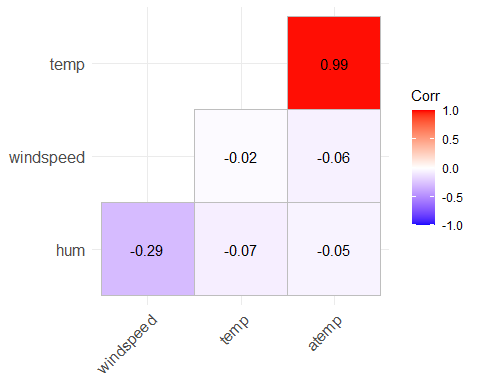
## Task 2

ggcorr(bike,label = "True", label\_round = 2)

## Warning in ggcorr(bike, label = "True", label\_round = 2): data in column(s)  
## 'dteday', 'season', 'mnth', 'hr', 'holiday', 'weekday', 'workingday',  
## 'weathersit' are not numeric and were ignored

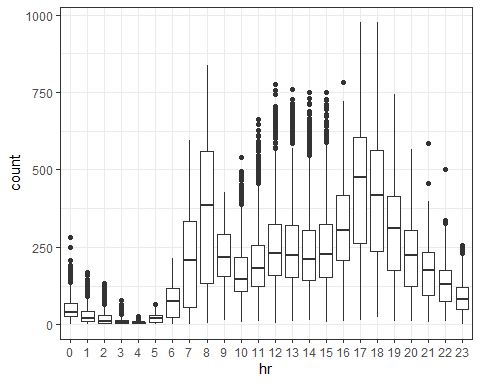


corr= round(cor(bike[,10:13,16]),2)  
ggcorrplot(corr,hc.order = TRUE,type="lower", lab = TRUE)

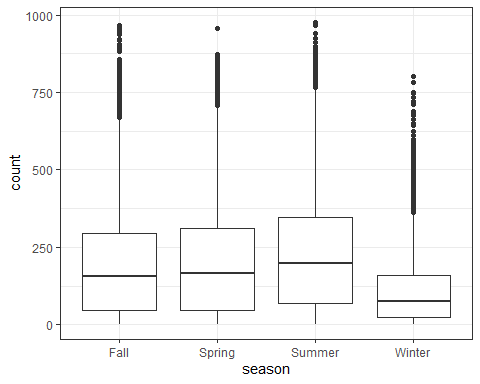
 . Temp and atemp seem to be highly correlated to count.

## Task 3

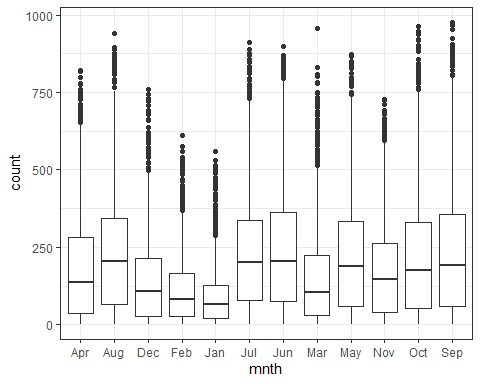
ggplot(bike,aes(x=hr,y=count)) +  
 geom\_boxplot() + theme\_bw()



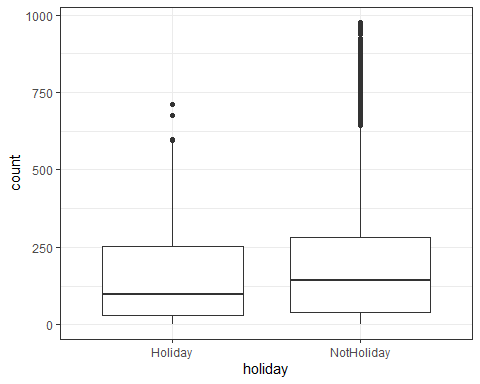
ggplot(bike,aes(x=season,y=count)) +  
 geom\_boxplot() + theme\_bw()



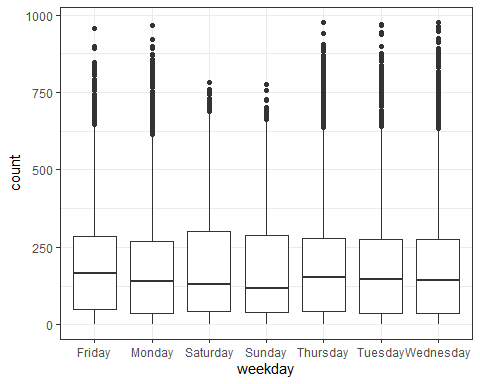
ggplot(bike,aes(x=mnth,y=count)) +  
 geom\_boxplot() + theme\_bw()



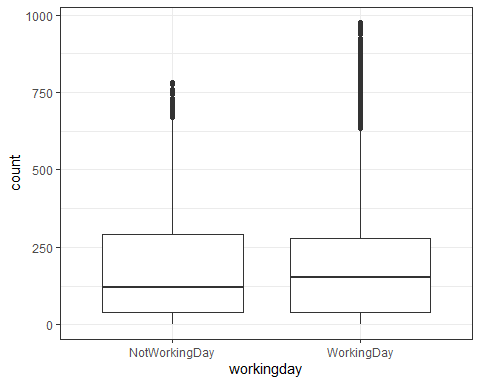
ggplot(bike,aes(x=holiday,y=count)) +  
 geom\_boxplot() + theme\_bw()



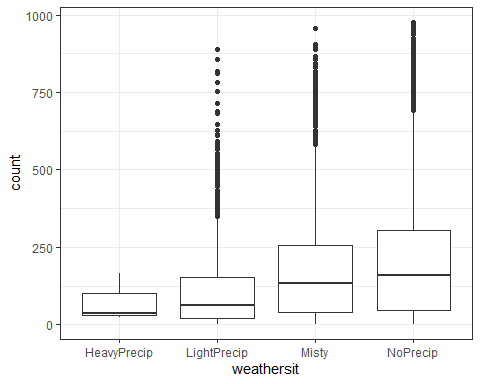
ggplot(bike,aes(x=weekday,y=count)) +  
 geom\_boxplot() + theme\_bw()



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



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

 . Hour and month (and thus season and weathersit as well) seem to impact the count greatly. It makes sense that the time of day and better weather will impact the use of bike sharing. If its cold and rainy, obviously people will choose a rideshare or other forms of public transportation. It will be the same thing if people use it for work or tourism in DC (impacted by the holiday, working day, and weekday).

## Task 4

bike\_recipe= recipe(count ~ mnth, 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   
## -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 \*\*\*  
## mnthAug 50.837 6.468 7.860 4.08e-15 \*\*\*  
## mnthDec -44.958 6.459 -6.960 3.53e-12 \*\*\*  
## mnthFeb -74.396 6.626 -11.228 < 2e-16 \*\*\*  
## mnthJan -92.836 6.519 -14.240 < 2e-16 \*\*\*  
## mnthJul 44.559 6.454 6.904 5.23e-12 \*\*\*  
## mnthJun 53.254 6.507 8.184 2.92e-16 \*\*\*  
## mnthMar -31.850 6.470 -4.923 8.62e-07 \*\*\*  
## mnthMay 35.646 6.454 5.523 3.38e-08 \*\*\*  
## mnthNov -9.926 6.510 -1.525 0.127   
## mnthOct 34.898 6.494 5.373 7.82e-08 \*\*\*  
## mnthSep 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

. hr seemed obvious, so I chose the month to evaluate if this impacts the count. There are a few negative values, which don’t really make sense, however, almost all (except November) seem significant.

## Task 5

bikeR\_recipe = recipe(count ~.,bike)%>%  
 step\_rm(instant,dteday,registered,casual)%>%  
 step\_dummy(all\_nominal())%>%  
 step\_center(all\_predictors())%>%  
 step\_scale(all\_predictors())  
  
bikeR\_model =  
 linear\_reg(mixture = 0)%>%  
 set\_engine('glmnet')  
  
bikeR\_wflow <-  
 workflow()%>%  
 add\_model(bikeR\_model)%>%  
 add\_recipe(bikeR\_recipe)  
  
bikeR\_fit <-  
 fit(bikeR\_wflow,bike)  
  
bikeR\_fit %>%  
 pull\_workflow\_fit() %>%  
 pluck("fit")

##   
## Call: glmnet::glmnet(x = maybe\_matrix(x), y = y, family = "gaussian", alpha = ~0)   
##   
## Df %Dev Lambda  
## 1 52 0.00 73420  
## 2 52 0.61 66900  
## 3 52 0.67 60950  
## 4 52 0.74 55540  
## 5 52 0.81 50600  
## 6 52 0.88 46110  
## 7 52 0.97 42010  
## 8 52 1.06 38280  
## 9 52 1.16 34880  
## 10 52 1.27 31780  
## 11 52 1.39 28960  
## 12 52 1.53 26390  
## 13 52 1.67 24040  
## 14 52 1.83 21910  
## 15 52 2.00 19960  
## 16 52 2.19 18190  
## 17 52 2.40 16570  
## 18 52 2.62 15100  
## 19 52 2.86 13760  
## 20 52 3.13 12540  
## 21 52 3.41 11420  
## 22 52 3.72 10410  
## 23 52 4.06 9482  
## 24 52 4.43 8640  
## 25 52 4.83 7872  
## 26 52 5.26 7173  
## 27 52 5.72 6536  
## 28 52 6.22 5955  
## 29 52 6.76 5426  
## 30 52 7.34 4944  
## 31 52 7.96 4505  
## 32 52 8.62 4105  
## 33 52 9.33 3740  
## 34 52 10.09 3408  
## 35 52 10.90 3105  
## 36 52 11.76 2829  
## 37 52 12.67 2578  
## 38 52 13.63 2349  
## 39 52 14.65 2140  
## 40 52 15.72 1950  
## 41 52 16.83 1777  
## 42 52 18.01 1619  
## 43 52 19.23 1475  
## 44 52 20.49 1344  
## 45 52 21.81 1225  
## 46 52 23.16 1116  
## 47 52 24.56 1017  
## 48 52 25.98 926  
## 49 52 27.44 844  
## 50 52 28.93 769  
## 51 52 30.43 701  
## 52 52 31.95 639  
## 53 52 33.48 582  
## 54 52 35.01 530  
## 55 52 36.53 483  
## 56 52 38.04 440  
## 57 52 39.54 401  
## 58 52 41.01 365  
## 59 52 42.44 333  
## 60 52 43.84 303  
## 61 52 45.20 276  
## 62 52 46.51 252  
## 63 52 47.77 230  
## 64 52 48.96 209  
## 65 52 50.10 190  
## 66 52 51.18 174  
## 67 52 52.19 158  
## 68 52 53.14 144  
## 69 52 54.02 131  
## 70 52 54.83 120  
## 71 52 55.59 109  
## 72 52 56.28 99  
## 73 52 56.91 91  
## 74 52 57.49 82  
## 75 52 58.01 75  
## 76 52 58.48 68  
## 77 52 58.91 62  
## 78 52 59.30 57  
## 79 52 59.64 52  
## 80 52 59.96 47  
## 81 52 60.24 43  
## 82 52 60.49 39  
## 83 52 60.72 36  
## 84 52 60.93 33  
## 85 52 61.11 30  
## 86 52 61.28 27  
## 87 52 61.44 25  
## 88 52 61.58 22  
## 89 52 61.71 20  
## 90 52 61.83 19  
## 91 52 61.95 17  
## 92 52 62.05 15  
## 93 52 62.14 14  
## 94 52 62.23 13  
## 95 52 62.32 12  
## 96 52 62.40 11  
## 97 52 62.47 10  
## 98 52 62.54 9  
## 99 52 62.60 8  
## 100 52 62.66 7

bikeR\_fit %>%  
 pull\_workflow\_fit() %>%  
 pluck("fit") %>%  
 coef(s = 30)

## 53 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 189.4630876  
## temp 24.7608664  
## atemp 24.3401839  
## hum -24.4757351  
## windspeed -1.8375628  
## season\_Spring -1.9696496  
## season\_Summer -5.9637409  
## season\_Winter -14.8899184  
## mnth\_Aug -0.3121922  
## mnth\_Dec 0.9058286  
## mnth\_Feb -2.3854903  
## mnth\_Jan -2.6820662  
## mnth\_Jul -6.2059740  
## mnth\_Jun -1.8125541  
## mnth\_Mar 0.2075416  
## mnth\_May 2.8178202  
## mnth\_Nov 2.2777050  
## mnth\_Oct 8.0569833  
## mnth\_Sep 7.9205987  
## hr\_X1 -19.4424332  
## hr\_X2 -20.6748056  
## hr\_X3 -22.0851546  
## hr\_X4 -22.4146131  
## hr\_X5 -19.9327607  
## hr\_X6 -9.9921297  
## hr\_X7 13.1158327  
## hr\_X8 37.0862793  
## hr\_X9 11.1162855  
## hr\_X10 1.2248065  
## hr\_X11 5.1875749  
## hr\_X12 11.5928052  
## hr\_X13 10.4957075  
## hr\_X14 7.5819635  
## hr\_X15 9.1676935  
## hr\_X16 19.9787338  
## hr\_X17 46.7759457  
## hr\_X18 41.4520209  
## hr\_X19 23.0862085  
## hr\_X20 9.6319765  
## hr\_X21 1.4118615  
## hr\_X22 -4.7497397  
## hr\_X23 -11.2911010  
## holiday\_NotHoliday 3.3545925  
## weekday\_Monday -1.6789317  
## weekday\_Saturday 1.5208355  
## weekday\_Sunday -2.7367975  
## weekday\_Thursday -0.7025073  
## weekday\_Tuesday -1.1073966  
## weekday\_Wednesday -0.2854113  
## workingday\_WorkingDay 2.1271880  
## weathersit\_LightPrecip -11.1467697  
## weathersit\_Misty 2.1617532  
## weathersit\_NoPrecip 4.4239482

. Quite a few values are close to zero but varies greatly by category. for example the hour of 10 is only 1.22, but at 17 is at 46.77. This is similar to the days of the week with some positive and almost a mirror of negative values other days of the week.

## Task 6

bikeL\_recipe = recipe(count ~., bike)%>%  
 step\_rm(instant,dteday,registered,casual)%>%  
 step\_dummy(all\_nominal())%>%  
 step\_center(all\_predictors())%>%  
 step\_scale(all\_predictors())  
  
bikeL\_model =  
 linear\_reg(mixture = 1)%>%  
 set\_engine('glmnet')  
  
bikeL\_wflow <-  
 workflow()%>%  
 add\_model(bikeL\_model)%>%  
 add\_recipe(bikeL\_recipe)  
  
bikeL\_fit <-  
 fit(bikeL\_wflow,bike)  
  
bikeL\_fit %>%  
 pull\_workflow\_fit() %>%  
 pluck("fit")

##   
## Call: glmnet::glmnet(x = maybe\_matrix(x), y = y, family = "gaussian", alpha = ~1)   
##   
## Df %Dev Lambda  
## 1 0 0.00 73.420  
## 2 1 2.78 66.900  
## 3 1 5.09 60.950  
## 4 3 7.60 55.540  
## 5 3 11.69 50.600  
## 6 4 15.44 46.110  
## 7 4 19.18 42.010  
## 8 6 22.56 38.280  
## 9 6 26.23 34.880  
## 10 6 29.28 31.780  
## 11 8 32.06 28.960  
## 12 11 34.97 26.390  
## 13 12 38.11 24.040  
## 14 12 40.86 21.910  
## 15 14 43.28 19.960  
## 16 14 45.50 18.190  
## 17 15 47.37 16.570  
## 18 15 49.03 15.100  
## 19 16 50.55 13.760  
## 20 16 51.81 12.540  
## 21 18 52.98 11.420  
## 22 19 54.01 10.410  
## 23 21 54.90 9.482  
## 24 24 55.78 8.640  
## 25 25 56.58 7.872  
## 26 26 57.29 7.173  
## 27 27 57.91 6.536  
## 28 27 58.47 5.955  
## 29 28 58.95 5.426  
## 30 28 59.38 4.944  
## 31 29 59.74 4.505  
## 32 31 60.09 4.105  
## 33 32 60.41 3.740  
## 34 32 60.69 3.408  
## 35 32 60.92 3.105  
## 36 33 61.11 2.829  
## 37 36 61.30 2.578  
## 38 37 61.60 2.349  
## 39 36 61.82 2.140  
## 40 36 61.98 1.950  
## 41 38 62.13 1.777  
## 42 39 62.25 1.619  
## 43 40 62.36 1.475  
## 44 41 62.46 1.344  
## 45 42 62.58 1.225  
## 46 42 62.69 1.116  
## 47 42 62.77 1.017  
## 48 41 62.84 0.926  
## 49 42 62.89 0.844  
## 50 42 62.92 0.769  
## 51 42 62.96 0.701  
## 52 42 62.98 0.639  
## 53 42 63.01 0.582  
## 54 42 63.04 0.530  
## 55 42 63.05 0.483  
## 56 43 63.07 0.440  
## 57 44 63.09 0.401  
## 58 45 63.11 0.365  
## 59 45 63.13 0.333  
## 60 45 63.14 0.303  
## 61 46 63.15 0.276  
## 62 49 63.16 0.252  
## 63 49 63.17 0.230  
## 64 49 63.18 0.209  
## 65 49 63.19 0.190  
## 66 49 63.19 0.174  
## 67 49 63.20 0.158  
## 68 49 63.20 0.144  
## 69 49 63.21 0.131  
## 70 48 63.21 0.120  
## 71 48 63.21 0.109  
## 72 48 63.21 0.099  
## 73 48 63.22 0.091  
## 74 49 63.22 0.082  
## 75 49 63.22 0.075  
## 76 49 63.22 0.068  
## 77 49 63.22 0.062  
## 78 49 63.22 0.057  
## 79 50 63.22 0.052  
## 80 50 63.22 0.047  
## 81 50 63.22 0.043

bikeL\_fit %>%  
 pull\_workflow\_fit() %>%  
 pluck("fit") %>%  
 coef(s = 1.47)

## 53 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 189.46308763  
## temp 24.55166926  
## atemp 26.65529779  
## hum -24.27671944  
## windspeed -2.33079874  
## season\_Spring -1.61412811  
## season\_Summer -8.14608608  
## season\_Winter -18.96903547  
## mnth\_Aug .   
## mnth\_Dec .   
## mnth\_Feb .   
## mnth\_Jan .   
## mnth\_Jul -6.34475077  
## mnth\_Jun -2.00980531  
## mnth\_Mar .   
## mnth\_May 0.13167129  
## mnth\_Nov 0.42421788  
## mnth\_Oct 6.70346179  
## mnth\_Sep 7.20257966  
## hr\_X1 -14.49563506  
## hr\_X2 -15.96085652  
## hr\_X3 -17.66992086  
## hr\_X4 -18.04371782  
## hr\_X5 -15.07411236  
## hr\_X6 -3.52288677  
## hr\_X7 20.43426785  
## hr\_X8 48.16732123  
## hr\_X9 18.04394534  
## hr\_X10 6.56603060  
## hr\_X11 11.11812666  
## hr\_X12 18.52178222  
## hr\_X13 17.25188711  
## hr\_X14 13.87838425  
## hr\_X15 15.72630591  
## hr\_X16 28.28540460  
## hr\_X17 59.35045036  
## hr\_X18 53.17631471  
## hr\_X19 31.91120445  
## hr\_X20 16.35485695  
## hr\_X21 6.82296800  
## hr\_X22 0.01562789  
## hr\_X23 -5.05887291  
## holiday\_NotHoliday 3.23327166  
## weekday\_Monday -0.19776334  
## weekday\_Saturday .   
## weekday\_Sunday -2.59733649  
## weekday\_Thursday .   
## weekday\_Tuesday .   
## weekday\_Wednesday .   
## workingday\_WorkingDay .   
## weathersit\_LightPrecip -13.54967501  
## weathersit\_Misty .   
## weathersit\_NoPrecip 0.71943647

. Lasso is helpful to see which variables fall out. Like the days of the week mentioned above, easy to weed those out as not really helpful when evaluating relationship.