# Celeste Matute

## Module 2 Assignment 2

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.0.5 ✓ dplyr 1.0.3  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.1

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

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 0.1.2 ──

## ✓ broom 0.7.3 ✓ recipes 0.1.15  
## ✓ dials 0.0.9 ✓ rsample 0.0.8   
## ✓ infer 0.5.4 ✓ tune 0.1.2   
## ✓ modeldata 0.1.0 ✓ workflows 0.2.1   
## ✓ parsnip 0.1.5 ✓ 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(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

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

## Loaded glmnet 4.1

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(leaps)   
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(splines)  
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\_cleaned <- read\_csv("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\_cleaned

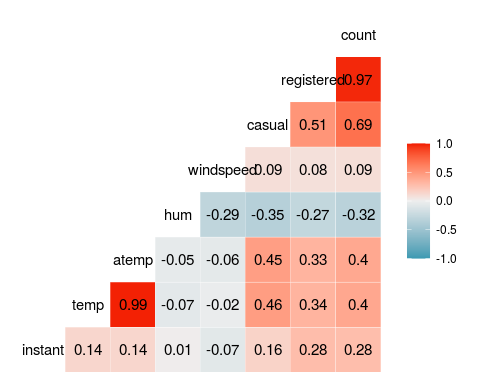
Conversions

bike = bike %>% mutate(dteday = mdy(dteday))  
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))

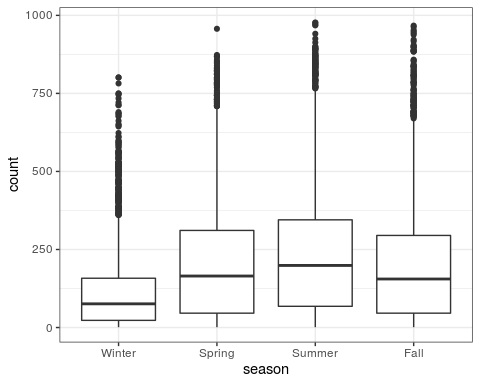
Correlations

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

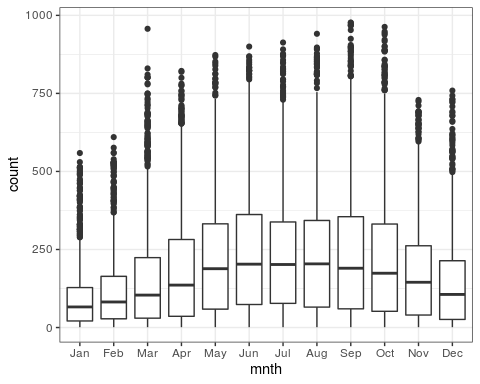
## Warning in ggcorr(bike, label = "TRUE", label\_round = 2): data in column(s)  
## 'dteday', 'season', 'mnth', 'hr', 'holiday', 'weekday', 'workingday',  
## 'weathersit' are not numeric and were ignored



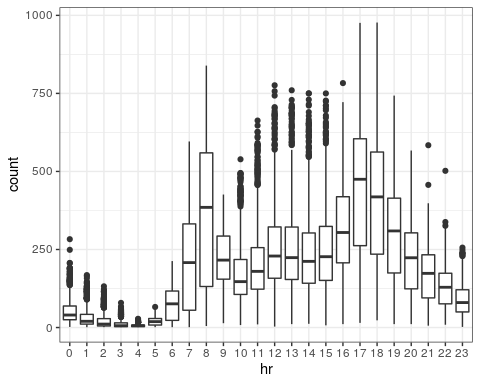
ggplot(bike,aes(x=season,y=count)) + geom\_boxplot() + theme\_bw()



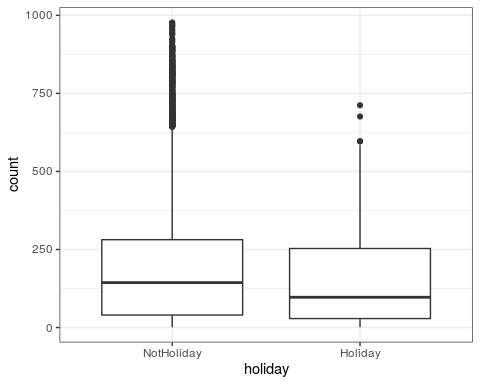
ggplot(bike,aes(x=mnth,y=count)) + geom\_boxplot() + theme\_bw()



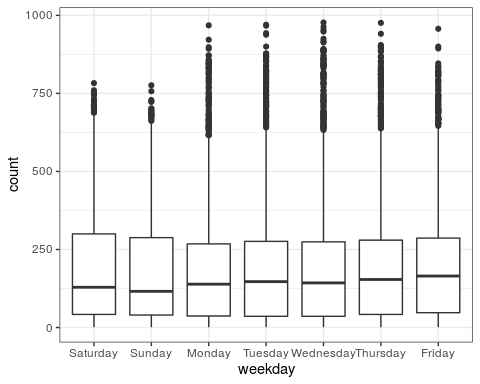
ggplot(bike,aes(x=hr,y=count)) + geom\_boxplot() + theme\_bw()



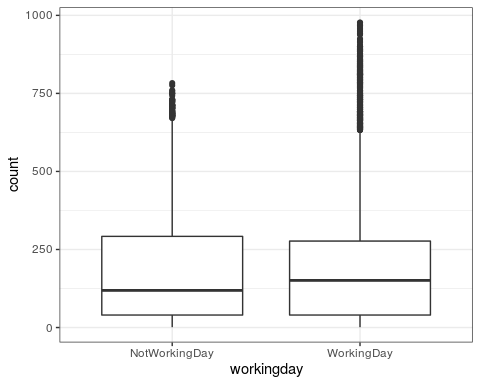
ggplot(bike,aes(x=holiday,y=count)) + geom\_boxplot() + theme\_bw()



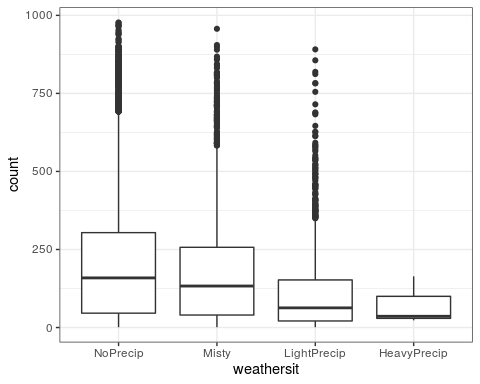
ggplot(bike,aes(x=weekday,y=count)) + geom\_boxplot() + theme\_bw()



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



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



The boxplots showed that the categories most effected were the month of the year, the day of the week, the weather, the season and hour of the day. The most correlated variable with a count was the temperature.

Correlation Model

bike\_recipe = recipe(count ~ hr , 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)

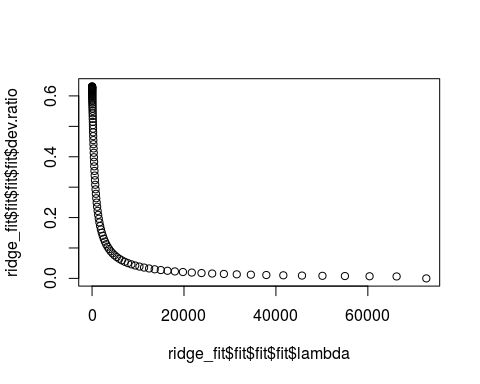
##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -446.45 -60.99 -6.01 50.10 551.49   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 53.898 4.756 11.332 < 2e-16 \*\*\*  
## hr1 -20.522 6.731 -3.049 0.002300 \*\*   
## hr2 -31.028 6.752 -4.595 4.35e-06 \*\*\*  
## hr3 -42.171 6.796 -6.205 5.58e-10 \*\*\*  
## hr4 -47.545 6.796 -6.996 2.73e-12 \*\*\*  
## hr5 -34.008 6.747 -5.040 4.70e-07 \*\*\*  
## hr6 22.146 6.729 3.291 0.000999 \*\*\*  
## hr7 158.167 6.724 23.523 < 2e-16 \*\*\*  
## hr8 305.113 6.724 45.377 < 2e-16 \*\*\*  
## hr9 165.411 6.724 24.600 < 2e-16 \*\*\*  
## hr10 119.770 6.724 17.812 < 2e-16 \*\*\*  
## hr11 154.245 6.724 22.939 < 2e-16 \*\*\*  
## hr12 199.418 6.722 29.668 < 2e-16 \*\*\*  
## hr13 199.763 6.719 29.729 < 2e-16 \*\*\*  
## hr14 187.051 6.719 27.838 < 2e-16 \*\*\*  
## hr15 197.335 6.719 29.368 < 2e-16 \*\*\*  
## hr16 258.085 6.717 38.422 < 2e-16 \*\*\*  
## hr17 407.554 6.717 60.674 < 2e-16 \*\*\*  
## hr18 371.613 6.722 55.286 < 2e-16 \*\*\*  
## hr19 257.625 6.722 38.327 < 2e-16 \*\*\*  
## hr20 172.132 6.722 25.608 < 2e-16 \*\*\*  
## hr21 118.416 6.722 17.617 < 2e-16 \*\*\*  
## hr22 77.437 6.722 11.520 < 2e-16 \*\*\*  
## hr23 33.933 6.722 5.048 4.50e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 128.2 on 17355 degrees of freedom  
## Multiple R-squared: 0.5015, Adjusted R-squared: 0.5008   
## F-statistic: 759.1 on 23 and 17355 DF, p-value: < 2.2e-16

The R squared value is at .5 and that is compared to the other variables where their R squared value was around .02 and .07. All the hours were also significant, showing this model was a good predictor.

Ridge Model

bike\_recipe2 = recipe(count ~ ., bike) %>%  
 step\_ns(temp, deg\_free = 4) %>%  
 step\_rm("instant", "dteday", "registered", "casual") %>%  
 step\_dummy(all\_nominal()) %>%   
 step\_center(all\_predictors()) %>%   
 step\_scale(all\_predictors())   
   
   
ridge\_model =   
 linear\_reg(mixture = 0 ) %>%   
 set\_engine("glmnet")   
  
ridge\_wflow =   
 workflow() %>%   
 add\_model(ridge\_model) %>%   
 add\_recipe(bike\_recipe2)  
  
ridge\_fit = fit(ridge\_wflow, bike)

plot(ridge\_fit$fit$fit$fit$lambda,ridge\_fit$fit$fit$fit$dev.ratio)



ridge\_fit %>%  
 pull\_workflow\_fit() %>%  
 pluck("fit")

##   
## Call: glmnet::glmnet(x = maybe\_matrix(x), y = y, family = "gaussian", alpha = ~0)   
##   
## Df %Dev Lambda  
## 1 55 0.00 72720  
## 2 55 0.64 66260  
## 3 55 0.70 60370  
## 4 55 0.77 55010  
## 5 55 0.85 50120  
## 6 55 0.93 45670  
## 7 55 1.02 41610  
## 8 55 1.11 37920  
## 9 55 1.22 34550  
## 10 55 1.34 31480  
## 11 55 1.46 28680  
## 12 55 1.60 26130  
## 13 55 1.75 23810  
## 14 55 1.92 21700  
## 15 55 2.10 19770  
## 16 55 2.29 18010  
## 17 55 2.51 16410  
## 18 55 2.74 14960  
## 19 55 2.99 13630  
## 20 55 3.27 12420  
## 21 55 3.56 11310  
## 22 55 3.89 10310  
## 23 55 4.24 9392  
## 24 55 4.62 8558  
## 25 55 5.03 7798  
## 26 55 5.47 7105  
## 27 55 5.95 6474  
## 28 55 6.46 5899  
## 29 55 7.01 5375  
## 30 55 7.61 4897  
## 31 55 8.24 4462  
## 32 55 8.92 4066  
## 33 55 9.65 3705  
## 34 55 10.42 3375  
## 35 55 11.24 3076  
## 36 55 12.11 2802  
## 37 55 13.03 2553  
## 38 55 14.00 2327  
## 39 55 15.02 2120  
## 40 55 16.10 1932  
## 41 55 17.22 1760  
## 42 55 18.40 1604  
## 43 55 19.62 1461  
## 44 55 20.88 1331  
## 45 55 22.19 1213  
## 46 55 23.55 1105  
## 47 55 24.94 1007  
## 48 55 26.36 918  
## 49 55 27.81 836  
## 50 55 29.29 762  
## 51 55 30.78 694  
## 52 55 32.29 632  
## 53 55 33.81 576  
## 54 55 35.33 525  
## 55 55 36.85 478  
## 56 55 38.35 436  
## 57 55 39.84 397  
## 58 55 41.30 362  
## 59 55 42.73 330  
## 60 55 44.12 300  
## 61 55 45.47 274  
## 62 55 46.78 250  
## 63 55 48.03 227  
## 64 55 49.22 207  
## 65 55 50.36 189  
## 66 55 51.43 172  
## 67 55 52.44 157  
## 68 55 53.39 143  
## 69 55 54.27 130  
## 70 55 55.09 118  
## 71 55 55.85 108  
## 72 55 56.54 98  
## 73 55 57.18 90  
## 74 55 57.77 82  
## 75 55 58.30 74  
## 76 55 58.78 68  
## 77 55 59.22 62  
## 78 55 59.61 56  
## 79 55 59.97 51  
## 80 55 60.30 47  
## 81 55 60.59 43  
## 82 55 60.85 39  
## 83 55 61.09 35  
## 84 55 61.31 32  
## 85 55 61.51 29  
## 86 55 61.68 27  
## 87 55 61.85 24  
## 88 55 62.00 22  
## 89 55 62.14 20  
## 90 55 62.26 18  
## 91 55 62.38 17  
## 92 55 62.49 15  
## 93 55 62.59 14  
## 94 55 62.68 13  
## 95 55 62.77 12  
## 96 55 62.86 11  
## 97 55 62.93 10  
## 98 55 63.00 9  
## 99 55 63.07 8  
## 100 55 63.13 7

The R squared value was 63 because we used the lambada of 9.

ridge\_fit %>%  
 pull\_workflow\_fit() %>%  
 pluck("fit") %>%   
 coef(s = 9)

## 56 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 189.46308763  
## atemp 29.37584344  
## hum -24.99608563  
## windspeed -3.48078739  
## temp\_ns\_1 -0.08483601  
## temp\_ns\_2 30.41663952  
## temp\_ns\_3 0.82250180  
## temp\_ns\_4 3.80461416  
## season\_Spring 10.96037896  
## season\_Summer 4.37276040  
## season\_Fall 22.14569936  
## mnth\_Feb 1.00209711  
## mnth\_Mar 4.35927571  
## mnth\_Apr 2.70013232  
## mnth\_May 2.92853725  
## mnth\_Jun -3.34284365  
## mnth\_Jul -6.55019381  
## mnth\_Aug -1.43534527  
## mnth\_Sep 7.00876668  
## mnth\_Oct 6.38376985  
## mnth\_Nov 2.70006180  
## mnth\_Dec 2.69937151  
## hr\_X1 -15.61264033  
## hr\_X2 -16.93978317  
## hr\_X3 -18.52190949  
## hr\_X4 -18.78177449  
## hr\_X5 -15.97665441  
## hr\_X6 -4.96458753  
## hr\_X7 20.47353589  
## hr\_X8 46.84357422  
## hr\_X9 18.10448033  
## hr\_X10 7.22391274  
## hr\_X11 11.61848185  
## hr\_X12 18.76005181  
## hr\_X13 17.54925690  
## hr\_X14 14.32962681  
## hr\_X15 16.06355368  
## hr\_X16 27.98943852  
## hr\_X17 57.60591917  
## hr\_X18 51.66924741  
## hr\_X19 31.25521086  
## hr\_X20 16.38080028  
## hr\_X21 7.26730423  
## hr\_X22 0.50692131  
## hr\_X23 -6.66553274  
## holiday\_Holiday -4.45136357  
## weekday\_Sunday -4.67930451  
## weekday\_Monday -2.14409069  
## weekday\_Tuesday -1.21396320  
## weekday\_Wednesday -0.20951399  
## weekday\_Thursday -0.60123974  
## weekday\_Friday 1.25315857  
## workingday\_WorkingDay -0.54338480  
## weathersit\_Misty -1.67747995  
## weathersit\_LightPrecip -14.33129994  
## weathersit\_HeavyPrecip -0.36852770

LASSO Method

bike\_recipe3 = recipe(count ~ ., bike) %>%  
 step\_ns(temp, deg\_free = 4) %>%  
 step\_rm("instant", "dteday", "registered", "casual") %>%  
 step\_dummy(all\_nominal()) %>%   
 step\_center(all\_predictors()) %>%   
 step\_scale(all\_predictors())   
lasso\_model =   
 linear\_reg(mixture = 1 ) %>%   
 set\_engine("glmnet")   
lasso\_wflow =   
 workflow() %>%   
 add\_model(lasso\_model) %>%   
 add\_recipe(bike\_recipe3)  
lasso\_fit = fit(lasso\_wflow, bike)

lasso\_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 72.720  
## 2 1 2.73 66.260  
## 3 1 4.99 60.370  
## 4 3 7.91 55.010  
## 5 3 11.96 50.120  
## 6 4 15.82 45.670  
## 7 4 19.53 41.610  
## 8 5 22.97 37.920  
## 9 5 26.57 34.550  
## 10 5 29.55 31.480  
## 11 7 32.35 28.680  
## 12 10 35.31 26.130  
## 13 11 38.41 23.810  
## 14 11 41.11 21.700  
## 15 12 43.43 19.770  
## 16 14 45.48 18.010  
## 17 15 47.53 16.410  
## 18 15 49.33 14.960  
## 19 16 50.92 13.630  
## 20 17 52.27 12.420  
## 21 18 53.49 11.310  
## 22 18 54.54 10.310  
## 23 21 55.43 9.392  
## 24 22 56.28 8.558  
## 25 23 57.04 7.798  
## 26 25 57.73 7.105  
## 27 27 58.37 6.474  
## 28 27 58.95 5.899  
## 29 28 59.44 5.375  
## 30 28 59.89 4.897  
## 31 30 60.27 4.462  
## 32 31 60.61 4.066  
## 33 32 60.93 3.705  
## 34 33 61.21 3.375  
## 35 34 61.44 3.076  
## 36 34 61.64 2.802  
## 37 36 61.82 2.553  
## 38 36 62.11 2.327  
## 39 39 62.32 2.120  
## 40 39 62.48 1.932  
## 41 39 62.63 1.760  
## 42 39 62.74 1.604  
## 43 39 62.84 1.461  
## 44 40 62.93 1.331  
## 45 41 63.04 1.213  
## 46 41 63.14 1.105  
## 47 41 63.22 1.007  
## 48 41 63.28 0.918  
## 49 41 63.32 0.836  
## 50 42 63.36 0.762  
## 51 43 63.39 0.694  
## 52 45 63.42 0.632  
## 53 45 63.46 0.576  
## 54 46 63.49 0.525  
## 55 46 63.52 0.478  
## 56 49 63.54 0.436  
## 57 50 63.56 0.397  
## 58 49 63.59 0.362  
## 59 50 63.61 0.330  
## 60 50 63.62 0.300  
## 61 50 63.64 0.274  
## 62 50 63.65 0.250  
## 63 51 63.66 0.227  
## 64 52 63.67 0.207  
## 65 52 63.67 0.189  
## 66 52 63.68 0.172  
## 67 52 63.69 0.157  
## 68 52 63.69 0.143  
## 69 52 63.69 0.130  
## 70 52 63.70 0.118  
## 71 52 63.70 0.108  
## 72 52 63.70 0.098  
## 73 52 63.70 0.090  
## 74 52 63.71 0.082  
## 75 53 63.71 0.074  
## 76 53 63.71 0.068  
## 77 53 63.71 0.062  
## 78 53 63.71 0.056  
## 79 53 63.71 0.051  
## 80 54 63.71 0.047

The lambada of 1.213 is an R squared value of 63.04.

lasso\_fit %>%  
 pull\_workflow\_fit() %>%  
 pluck("fit") %>%   
 coef(s = 1.213)

## 56 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 189.46308763  
## atemp 34.80158293  
## hum -25.67971384  
## windspeed -2.11331651  
## temp\_ns\_1 .   
## temp\_ns\_2 28.78585749  
## temp\_ns\_3 .   
## temp\_ns\_4 .   
## season\_Spring 10.22668109  
## season\_Summer .   
## season\_Fall 23.33828716  
## mnth\_Feb .   
## mnth\_Mar 1.44917298  
## mnth\_Apr .   
## mnth\_May 0.06585405  
## mnth\_Jun -3.20960172  
## mnth\_Jul -4.82082074  
## mnth\_Aug .   
## mnth\_Sep 5.82501070  
## mnth\_Oct 2.52910746  
## mnth\_Nov .   
## mnth\_Dec .   
## hr\_X1 -12.94944250  
## hr\_X2 -14.32530658  
## hr\_X3 -15.98211930  
## hr\_X4 -16.26690158  
## hr\_X5 -13.26412408  
## hr\_X6 -1.73490846  
## hr\_X7 22.55618264  
## hr\_X8 50.11142285  
## hr\_X9 19.90799218  
## hr\_X10 8.45791360  
## hr\_X11 13.00129885  
## hr\_X12 20.43199065  
## hr\_X13 19.15279106  
## hr\_X14 15.75999149  
## hr\_X15 17.58061776  
## hr\_X16 30.08815325  
## hr\_X17 61.13644865  
## hr\_X18 54.92582487  
## hr\_X19 33.59198228  
## hr\_X20 18.06500924  
## hr\_X21 8.56186251  
## hr\_X22 1.51725229  
## hr\_X23 -3.64244896  
## holiday\_Holiday -3.32721007  
## weekday\_Sunday -2.78467975  
## weekday\_Monday -0.90922702  
## weekday\_Tuesday .   
## weekday\_Wednesday .   
## weekday\_Thursday .   
## weekday\_Friday 0.86009924  
## workingday\_WorkingDay .   
## weathersit\_Misty -0.04349270  
## weathersit\_LightPrecip -13.26572359  
## weathersit\_HeavyPrecip .

I liked the Lasso method more then the ridge. With the lasso method it is smoother and most of the coefficients are zero which helps differentiate which ones are more important then others.