# Daniel Rauscher

## Module 3 Assignment 1

bike <- 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()  
## )

#view(bike)  
  
bike = bike %>% mutate(dteday = mdy(dteday))   
bike <- bike %>% mutate(season = as\_factor(season)) %>% mutate(holiday = as\_factor(holiday)) %>% mutate(mnth = as\_factor(mnth)) %>% mutate(weekday = as\_factor(weekday)) %>% mutate(workingday = as\_factor(workingday)) %>% mutate(weathersit = as\_factor(weathersit)) %>% mutate(hr = as\_factor(hr))

## Task 1

set.seed(1234)  
bike\_split = initial\_split(bike, prob = 0.70, strata = count)  
train = training(bike\_split)  
test = testing(bike\_split)

## Task 2

The training set has 13036 rows and the testing set has 4343 rows.

## Task 3

The adjusted R squared value for this model is pretty good with a value of 0.6284. When using this data set in the previous module I noticed a non linear relationship between temp and count so I added a spline. Some of the variables (hr, temp) appear to be more strongly correlated with count than others (weekday).

bike\_recipe <- recipe(count ~ temp + season + mnth + hr + holiday + weekday + weathersit, train) %>%  
 step\_ns(temp, deg\_free = 7) %>%  
 step\_dummy(all\_nominal())  
  
  
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)

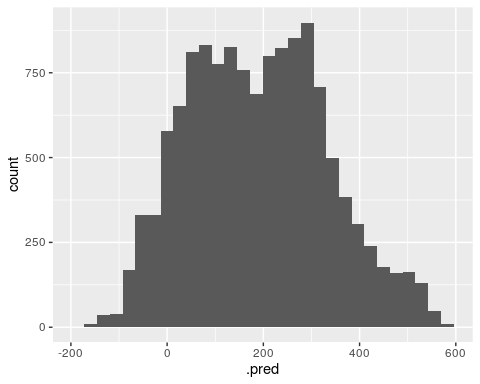
##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -423.91 -63.45 -8.32 52.19 506.69   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -25.6976 12.4902 -2.057 0.039661 \*   
## temp\_ns\_1 28.3751 11.7015 2.425 0.015322 \*   
## temp\_ns\_2 47.5015 13.9124 3.414 0.000641 \*\*\*  
## temp\_ns\_3 102.1670 13.2493 7.711 1.31e-14 \*\*\*  
## temp\_ns\_4 118.8624 13.6274 8.722 < 2e-16 \*\*\*  
## temp\_ns\_5 237.2920 10.5429 22.507 < 2e-16 \*\*\*  
## temp\_ns\_6 142.1846 27.8364 5.108 3.29e-07 \*\*\*  
## temp\_ns\_7 88.4387 15.3282 5.770 8.08e-09 \*\*\*  
## season\_Spring 35.4563 5.2747 6.722 1.85e-11 \*\*\*  
## season\_Summer 24.4407 6.2791 3.892 9.96e-05 \*\*\*  
## season\_Fall 67.2075 5.3075 12.663 < 2e-16 \*\*\*  
## mnth\_Feb 8.4886 4.3727 1.941 0.052242 .   
## mnth\_Mar 14.2901 4.9716 2.874 0.004053 \*\*   
## mnth\_Apr 1.7969 7.2349 0.248 0.803852   
## mnth\_May -8.8029 7.6848 -1.145 0.252020   
## mnth\_Jun -26.1202 7.8817 -3.314 0.000921 \*\*\*  
## mnth\_Jul -46.6526 8.8078 -5.297 1.19e-07 \*\*\*  
## mnth\_Aug -32.8637 8.6322 -3.807 0.000141 \*\*\*  
## mnth\_Sep -0.6782 7.7790 -0.087 0.930527   
## mnth\_Oct 0.8767 7.1862 0.122 0.902897   
## mnth\_Nov -8.1136 6.9571 -1.166 0.243536   
## mnth\_Dec -5.5983 5.5583 -1.007 0.313861   
## hr\_X1 -17.5768 5.8085 -3.026 0.002481 \*\*   
## hr\_X2 -26.1774 5.8278 -4.492 7.11e-06 \*\*\*  
## hr\_X3 -36.7077 5.8686 -6.255 4.07e-10 \*\*\*  
## hr\_X4 -39.6130 5.8716 -6.747 1.56e-11 \*\*\*  
## hr\_X5 -23.5173 5.8334 -4.032 5.57e-05 \*\*\*  
## hr\_X6 34.4767 5.8185 5.925 3.18e-09 \*\*\*  
## hr\_X7 169.2841 5.8113 29.130 < 2e-16 \*\*\*  
## hr\_X8 309.5784 5.8069 53.312 < 2e-16 \*\*\*  
## hr\_X9 163.2259 5.8092 28.098 < 2e-16 \*\*\*  
## hr\_X10 110.6124 5.8214 19.001 < 2e-16 \*\*\*  
## hr\_X11 138.7539 5.8419 23.752 < 2e-16 \*\*\*  
## hr\_X12 180.7016 5.8649 30.811 < 2e-16 \*\*\*  
## hr\_X13 177.0252 5.8863 30.074 < 2e-16 \*\*\*  
## hr\_X14 161.6875 5.9066 27.374 < 2e-16 \*\*\*  
## hr\_X15 171.1364 5.9135 28.940 < 2e-16 \*\*\*  
## hr\_X16 232.0385 5.9049 39.296 < 2e-16 \*\*\*  
## hr\_X17 384.3982 5.8855 65.313 < 2e-16 \*\*\*  
## hr\_X18 351.2594 5.8651 59.890 < 2e-16 \*\*\*  
## hr\_X19 240.2191 5.8369 41.155 < 2e-16 \*\*\*  
## hr\_X20 159.8983 5.8211 27.469 < 2e-16 \*\*\*  
## hr\_X21 109.3355 5.8091 18.822 < 2e-16 \*\*\*  
## hr\_X22 71.5463 5.8036 12.328 < 2e-16 \*\*\*  
## hr\_X23 32.8492 5.8005 5.663 1.51e-08 \*\*\*  
## holiday\_Holiday -26.6861 5.3152 -5.021 5.20e-07 \*\*\*  
## weekday\_Sunday -16.2815 3.1271 -5.207 1.95e-07 \*\*\*  
## weekday\_Monday -9.8589 3.2321 -3.050 0.002289 \*\*   
## weekday\_Tuesday -7.3802 3.1521 -2.341 0.019223 \*   
## weekday\_Wednesday -5.2476 3.1455 -1.668 0.095276 .   
## weekday\_Thursday -2.7656 3.1430 -0.880 0.378925   
## weekday\_Friday 2.1799 3.1328 0.696 0.486534   
## weathersit\_Misty -19.2494 1.9765 -9.739 < 2e-16 \*\*\*  
## weathersit\_LightPrecip -88.7345 3.1596 -28.084 < 2e-16 \*\*\*  
## weathersit\_HeavyPrecip -82.6549 63.9624 -1.292 0.196291   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 110.6 on 17324 degrees of freedom  
## Multiple R-squared: 0.6296, Adjusted R-squared: 0.6284   
## F-statistic: 545.3 on 54 and 17324 DF, p-value: < 2.2e-16

## Task 4

The distribution of the predictions forms several peaks. One peak is roughly 70-80, another roughly 125 and the other is roughly 300.

testdata = data.frame(train)  
predict\_train <- predict(lm\_fit, new\_data = testdata)  
ggplot(predict\_train, aes(x=.pred)) +   
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Task 5

The adjusted r square value for the test data (0.6288) is very close to the training data’s adjusted r square value of 0.6284. Based on this we can say this model is not over fitting the data.

lm\_fit %>% predict(test) %>% bind\_cols(test) %>% metrics(truth = count, estimate = .pred)

## # A tibble: 3 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 rmse standard 111.   
## 2 rsq standard 0.629  
## 3 mae standard 81.1