Module3\_Assign1

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

## Rows: 17,379  
## Columns: 16  
## $ instant <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...  
## $ dteday <date> 2011-01-01, 2011-01-01, 2011-01-01, 2011-01-01, 2011-01...  
## $ season <fct> Winter, Winter, Winter, Winter, Winter, Winter, Winter, ...  
## $ mnth <fct> Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, J...  
## $ hr <fct> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...  
## $ holiday <fct> NotHoliday, NotHoliday, NotHoliday, NotHoliday, NotHolid...  
## $ weekday <fct> Saturday, Saturday, Saturday, Saturday, Saturday, Saturd...  
## $ workingday <fct> NotWorkingDay, NotWorkingDay, NotWorkingDay, NotWorkingD...  
## $ weathersit <fct> NoPrecip, NoPrecip, NoPrecip, NoPrecip, NoPrecip, Misty,...  
## $ temp <dbl> 0.24, 0.22, 0.22, 0.24, 0.24, 0.24, 0.22, 0.20, 0.24, 0....  
## $ atemp <dbl> 0.2879, 0.2727, 0.2727, 0.2879, 0.2879, 0.2576, 0.2727, ...  
## $ hum <dbl> 0.81, 0.80, 0.80, 0.75, 0.75, 0.75, 0.80, 0.86, 0.75, 0....  
## $ windspeed <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0896, 0.0000, ...  
## $ casual <dbl> 3, 8, 5, 3, 0, 0, 2, 1, 1, 8, 12, 26, 29, 47, 35, 40, 41...  
## $ registered <dbl> 13, 32, 27, 10, 1, 1, 0, 2, 7, 6, 24, 30, 55, 47, 71, 70...  
## $ count <dbl> 16, 40, 32, 13, 1, 1, 2, 3, 8, 14, 36, 56, 84, 94, 106, ...

summary(bike)

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

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

### Task 2

Test set contains 4,343 observations. Training set contains 13,036 observations.

### Task 3

biketrain\_recipe = recipe(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, train) %>%  
 step\_dummy(all\_nominal())  
  
lm\_model =   
 linear\_reg() %>%   
 set\_engine("lm")   
  
lm\_wflow =   
 workflow() %>%   
 add\_model(lm\_model) %>%   
 add\_recipe(biketrain\_recipe)  
  
lm\_fit = fit(lm\_wflow, train)  
  
summary(lm\_fit$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -413.11 -61.65 -10.20 52.16 493.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -85.3666 6.7480 -12.651 < 2e-16 \*\*\*  
## temp 289.3663 11.7834 24.557 < 2e-16 \*\*\*  
## season\_Spring 36.4869 6.0975 5.984 2.24e-09 \*\*\*  
## season\_Summer 31.2531 7.2079 4.336 1.46e-05 \*\*\*  
## season\_Fall 62.3139 6.1169 10.187 < 2e-16 \*\*\*  
## mnth\_Feb 1.1497 4.9544 0.232 0.816495   
## mnth\_Mar 6.1995 5.6065 1.106 0.268842   
## mnth\_Apr -8.0391 8.2890 -0.970 0.332141   
## mnth\_May -7.9182 8.8342 -0.896 0.370100   
## mnth\_Jun -24.0790 9.0886 -2.649 0.008074 \*\*   
## mnth\_Jul -46.7354 10.1901 -4.586 4.55e-06 \*\*\*  
## mnth\_Aug -27.7663 9.8890 -2.808 0.004996 \*\*   
## mnth\_Sep 0.3262 8.7760 0.037 0.970353   
## mnth\_Oct -2.8693 8.1729 -0.351 0.725537   
## mnth\_Nov -17.9871 7.8670 -2.286 0.022247 \*   
## mnth\_Dec -12.6331 6.2366 -2.026 0.042824 \*   
## hr\_X1 -17.6394 6.7829 -2.601 0.009318 \*\*   
## hr\_X2 -24.7408 6.7860 -3.646 0.000268 \*\*\*  
## hr\_X3 -36.3172 6.7857 -5.352 8.85e-08 \*\*\*  
## hr\_X4 -39.8317 6.8741 -5.794 7.01e-09 \*\*\*  
## hr\_X5 -23.5341 6.8326 -3.444 0.000574 \*\*\*  
## hr\_X6 34.9075 6.7378 5.181 2.24e-07 \*\*\*  
## hr\_X7 170.4187 6.7576 25.219 < 2e-16 \*\*\*  
## hr\_X8 310.2081 6.7874 45.703 < 2e-16 \*\*\*  
## hr\_X9 167.5555 6.6896 25.047 < 2e-16 \*\*\*  
## hr\_X10 112.2824 6.7742 16.575 < 2e-16 \*\*\*  
## hr\_X11 139.9731 6.7959 20.597 < 2e-16 \*\*\*  
## hr\_X12 180.4694 6.8816 26.225 < 2e-16 \*\*\*  
## hr\_X13 182.6847 6.8514 26.664 < 2e-16 \*\*\*  
## hr\_X14 163.6753 6.8350 23.947 < 2e-16 \*\*\*  
## hr\_X15 168.7255 6.8956 24.469 < 2e-16 \*\*\*  
## hr\_X16 228.8081 6.8944 33.187 < 2e-16 \*\*\*  
## hr\_X17 380.6338 6.8048 55.936 < 2e-16 \*\*\*  
## hr\_X18 355.7561 6.8635 51.833 < 2e-16 \*\*\*  
## hr\_X19 244.4088 6.7834 36.031 < 2e-16 \*\*\*  
## hr\_X20 160.9975 6.8198 23.607 < 2e-16 \*\*\*  
## hr\_X21 110.3631 6.7372 16.381 < 2e-16 \*\*\*  
## hr\_X22 73.3439 6.7251 10.906 < 2e-16 \*\*\*  
## hr\_X23 34.8460 6.7667 5.150 2.65e-07 \*\*\*  
## holiday\_Holiday -27.9348 6.1853 -4.516 6.35e-06 \*\*\*  
## weekday\_Sunday -18.8199 3.6318 -5.182 2.23e-07 \*\*\*  
## weekday\_Monday -7.3613 3.7407 -1.968 0.049102 \*   
## weekday\_Tuesday -6.4130 3.6577 -1.753 0.079581 .   
## weekday\_Wednesday -3.1261 3.6405 -0.859 0.390522   
## weekday\_Thursday -3.2735 3.6387 -0.900 0.368340   
## weekday\_Friday 1.2942 3.6295 0.357 0.721414   
## weathersit\_Misty -20.7095 2.2908 -9.040 < 2e-16 \*\*\*  
## weathersit\_LightPrecip -92.1176 3.6283 -25.389 < 2e-16 \*\*\*  
## weathersit\_HeavyPrecip -41.2406 78.7048 -0.524 0.600294   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 111.1 on 12987 degrees of freedom  
## Multiple R-squared: 0.6243, Adjusted R-squared: 0.6229   
## F-statistic: 449.6 on 48 and 12987 DF, p-value: < 2.2e-16

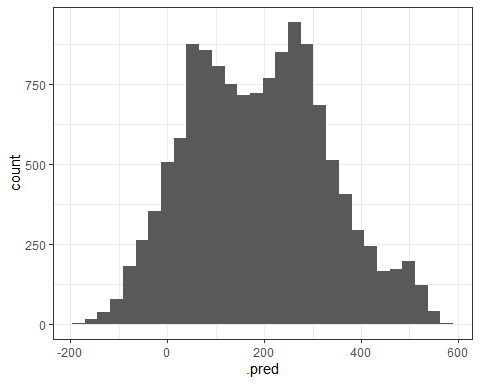
The model has an adjusted R squared of 62.29%. The model likely has multicollinarity with the predictors being highly correlated.

### Task 4

predict\_train = predict(lm\_fit,train)

ggplot(predict\_train, aes(x=.pred)) + geom\_histogram() + theme\_bw()

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



The distribution in the histogram above seems slightly skewed, perhaps indicating that their are outliers present in the predicted values. Outliers in the predicted values would indicate outliers in the data frame values as well.

#### Task 5

biketrain\_recipe = recipe(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, test) %>%  
 step\_dummy(all\_nominal())  
  
lm\_model\_test =   
 linear\_reg() %>%   
 set\_engine("lm")   
  
lm\_wflow\_test =   
 workflow() %>%   
 add\_model(lm\_model\_test) %>%   
 add\_recipe(biketrain\_recipe)  
  
lm\_fit\_test = fit(lm\_wflow\_test, test)  
  
summary(lm\_fit\_test$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -376.81 -64.23 -8.34 51.71 508.12   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.613e+01 1.147e+01 -7.509 7.21e-14 \*\*\*  
## temp 2.847e+02 2.064e+01 13.790 < 2e-16 \*\*\*  
## season\_Spring 3.305e+01 1.085e+01 3.046 0.002336 \*\*   
## season\_Summer 1.215e+01 1.297e+01 0.937 0.348702   
## season\_Fall 7.492e+01 1.092e+01 6.862 7.77e-12 \*\*\*  
## mnth\_Feb 2.273e+00 8.598e+00 0.264 0.791508   
## mnth\_Mar 1.628e+00 9.477e+00 0.172 0.863627   
## mnth\_Apr -6.372e-01 1.427e+01 -0.045 0.964376   
## mnth\_May 1.567e-02 1.539e+01 0.001 0.999188   
## mnth\_Jun -4.441e-01 1.569e+01 -0.028 0.977426   
## mnth\_Jul -2.448e+01 1.770e+01 -1.383 0.166616   
## mnth\_Aug 3.349e-04 1.734e+01 0.000 0.999985   
## mnth\_Sep 1.808e+01 1.552e+01 1.165 0.243947   
## mnth\_Oct -9.386e+00 1.428e+01 -0.657 0.511000   
## mnth\_Nov -2.110e+01 1.380e+01 -1.529 0.126337   
## mnth\_Dec -2.388e+01 1.095e+01 -2.182 0.029163 \*   
## hr\_X1 -1.904e+01 1.161e+01 -1.639 0.101229   
## hr\_X2 -3.220e+01 1.174e+01 -2.742 0.006126 \*\*   
## hr\_X3 -4.273e+01 1.211e+01 -3.528 0.000423 \*\*\*  
## hr\_X4 -4.311e+01 1.163e+01 -3.706 0.000213 \*\*\*  
## hr\_X5 -2.923e+01 1.155e+01 -2.530 0.011447 \*   
## hr\_X6 2.951e+01 1.195e+01 2.470 0.013566 \*   
## hr\_X7 1.682e+02 1.175e+01 14.309 < 2e-16 \*\*\*  
## hr\_X8 3.124e+02 1.154e+01 27.085 < 2e-16 \*\*\*  
## hr\_X9 1.553e+02 1.215e+01 12.781 < 2e-16 \*\*\*  
## hr\_X10 1.094e+02 1.168e+01 9.371 < 2e-16 \*\*\*  
## hr\_X11 1.377e+02 1.171e+01 11.754 < 2e-16 \*\*\*  
## hr\_X12 1.790e+02 1.144e+01 15.646 < 2e-16 \*\*\*  
## hr\_X13 1.567e+02 1.172e+01 13.371 < 2e-16 \*\*\*  
## hr\_X14 1.504e+02 1.194e+01 12.595 < 2e-16 \*\*\*  
## hr\_X15 1.730e+02 1.169e+01 14.791 < 2e-16 \*\*\*  
## hr\_X16 2.385e+02 1.163e+01 20.511 < 2e-16 \*\*\*  
## hr\_X17 3.973e+02 1.197e+01 33.185 < 2e-16 \*\*\*  
## hr\_X18 3.422e+02 1.155e+01 29.637 < 2e-16 \*\*\*  
## hr\_X19 2.331e+02 1.178e+01 19.795 < 2e-16 \*\*\*  
## hr\_X20 1.619e+02 1.150e+01 14.083 < 2e-16 \*\*\*  
## hr\_X21 1.116e+02 1.185e+01 9.417 < 2e-16 \*\*\*  
## hr\_X22 7.054e+01 1.187e+01 5.941 3.06e-09 \*\*\*  
## hr\_X23 2.936e+01 1.163e+01 2.524 0.011644 \*   
## holiday\_Holiday -2.114e+01 1.060e+01 -1.994 0.046205 \*   
## weekday\_Sunday -6.474e+00 6.354e+00 -1.019 0.308326   
## weekday\_Monday -8.466e+00 6.582e+00 -1.286 0.198386   
## weekday\_Tuesday -6.537e+00 6.404e+00 -1.021 0.307413   
## weekday\_Wednesday -5.344e+00 6.450e+00 -0.829 0.407383   
## weekday\_Thursday 4.956e-01 6.446e+00 0.077 0.938719   
## weekday\_Friday 3.428e+00 6.400e+00 0.536 0.592250   
## weathersit\_Misty -1.605e+01 3.974e+00 -4.040 5.44e-05 \*\*\*  
## weathersit\_LightPrecip -8.625e+01 6.552e+00 -13.163 < 2e-16 \*\*\*  
## weathersit\_HeavyPrecip -1.406e+02 1.128e+02 -1.246 0.212935   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 112.3 on 4294 degrees of freedom  
## Multiple R-squared: 0.6271, Adjusted R-squared: 0.623   
## F-statistic: 150.5 on 48 and 4294 DF, p-value: < 2.2e-16

The R-Squared value of the test data frame is nearly identical to the adjusted R-Squared of the testing set at .623.