# BAN 502 - Model Validation Assignment

## Khayrayyah Haamid-Day

**Task 1**

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

**Task 2**

There are 12,163 rows in training and 5,216 in testing.

bike\_recipe = recipe(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, train) %>%  
 step\_dummy(all\_nominal\_predictors())  
  
bike\_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(bike\_model) %>%   
 add\_recipe(bike\_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   
## -427.33 -62.08 -9.82 51.84 503.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -123.7048 66.1177 -1.871 0.061372 .   
## temp 293.4586 12.1953 24.063 < 2e-16 \*\*\*  
## season\_Spring -35.0395 7.5737 -4.626 3.76e-06 \*\*\*  
## season\_Summer -43.7722 6.8705 -6.371 1.95e-10 \*\*\*  
## season\_Winter -62.5367 6.4533 -9.691 < 2e-16 \*\*\*  
## mnth\_Aug -15.1863 8.4226 -1.803 0.071405 .   
## mnth\_Dec -14.9604 8.4415 -1.772 0.076380 .   
## mnth\_Feb 0.7133 8.4470 0.084 0.932706   
## mnth\_Jan 1.3130 8.6231 0.152 0.878982   
## mnth\_Jul -38.9170 8.5386 -4.558 5.22e-06 \*\*\*  
## mnth\_Jun -14.4995 5.9791 -2.425 0.015321 \*   
## mnth\_Mar 4.3908 6.5373 0.672 0.501819   
## mnth\_May -1.3764 5.1503 -0.267 0.789277   
## mnth\_Nov -13.4502 9.2393 -1.456 0.145485   
## mnth\_Oct -1.7687 9.0406 -0.196 0.844894   
## mnth\_Sep 5.2989 7.9195 0.669 0.503449   
## hr\_X1 -20.7836 6.9908 -2.973 0.002955 \*\*   
## hr\_X2 -29.0673 6.9980 -4.154 3.29e-05 \*\*\*  
## hr\_X3 -41.4592 7.0968 -5.842 5.29e-09 \*\*\*  
## hr\_X4 -41.2506 7.0386 -5.861 4.73e-09 \*\*\*  
## hr\_X5 -27.2665 6.9794 -3.907 9.41e-05 \*\*\*  
## hr\_X6 31.8318 7.0125 4.539 5.70e-06 \*\*\*  
## hr\_X7 164.5446 7.0278 23.413 < 2e-16 \*\*\*  
## hr\_X8 305.3583 6.9782 43.759 < 2e-16 \*\*\*  
## hr\_X9 163.9524 7.0096 23.390 < 2e-16 \*\*\*  
## hr\_X10 105.9395 6.9986 15.137 < 2e-16 \*\*\*  
## hr\_X11 138.1987 6.9861 19.782 < 2e-16 \*\*\*  
## hr\_X12 179.5246 6.9799 25.720 < 2e-16 \*\*\*  
## hr\_X13 177.5739 7.0533 25.176 < 2e-16 \*\*\*  
## hr\_X14 152.0364 7.1106 21.382 < 2e-16 \*\*\*  
## hr\_X15 170.3496 7.0967 24.004 < 2e-16 \*\*\*  
## hr\_X16 229.1493 7.1110 32.225 < 2e-16 \*\*\*  
## hr\_X17 384.6252 7.0221 54.774 < 2e-16 \*\*\*  
## hr\_X18 342.3854 7.0387 48.643 < 2e-16 \*\*\*  
## hr\_X19 236.7980 7.0437 33.618 < 2e-16 \*\*\*  
## hr\_X20 158.1195 7.0488 22.432 < 2e-16 \*\*\*  
## hr\_X21 107.9022 6.9453 15.536 < 2e-16 \*\*\*  
## hr\_X22 72.0674 6.9890 10.312 < 2e-16 \*\*\*  
## hr\_X23 31.3404 7.0004 4.477 7.64e-06 \*\*\*  
## holiday\_NotHoliday 25.5839 6.3712 4.016 5.97e-05 \*\*\*  
## weekday\_Monday -9.2322 3.8759 -2.382 0.017238 \*   
## weekday\_Saturday -0.5683 3.7761 -0.151 0.880363   
## weekday\_Sunday -13.4256 3.7705 -3.561 0.000371 \*\*\*  
## weekday\_Thursday -3.7422 3.8044 -0.984 0.325297   
## weekday\_Tuesday -7.3370 3.8298 -1.916 0.055420 .   
## weekday\_Wednesday -4.2535 3.8010 -1.119 0.263137   
## weathersit\_LightPrecip -13.9008 64.8336 -0.214 0.830233   
## weathersit\_Misty 58.4528 64.7679 0.902 0.366811   
## weathersit\_NoPrecip 78.2430 64.7522 1.208 0.226938   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 111.8 on 12114 degrees of freedom  
## Multiple R-squared: 0.6224, Adjusted R-squared: 0.6209   
## F-statistic: 416.1 on 48 and 12114 DF, p-value: < 2.2e-16

This model has a good r-squared with several variables being significant though there are some variables that have a negative coefficient when they should be positive so there is some multicollinearity occuring.

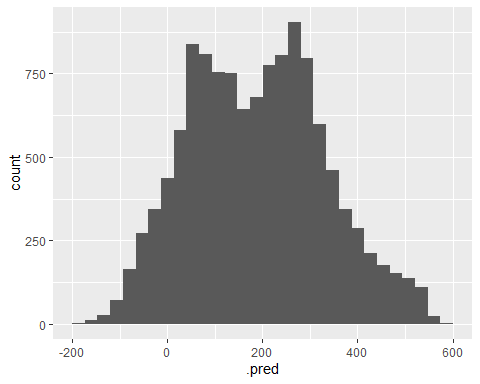
**Task 4**

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

## # A tibble: 6 x 1  
## .pred  
## <dbl>  
## 1 -37.9  
## 2 -46.2  
## 3 -52.7  
## 4 -52.5  
## 5 -58.3  
## 6 14.7

ggplot(predict\_train,aes(.pred))+  
 geom\_histogram()

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

 There is a normal distribution in this histogram.

**Task 5**

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 110.   
## 2 rsq standard 0.627  
## 3 mae standard 80.1

The R-Squared value on the testing set is very close to the R-Squared from the training model.