# Module 3 Assignment 1

# David Wells

#Load Libraries  
  
library("tidyverse")  
library("lubridate")  
library("tidymodels")

# Read data in   
  
bike <- read\_csv("C:/Users/wdavi/Documents/BAN 502/Module 3/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(sapply(bike, is\_character), as\_factor)  
  
bike <- bike %>%  
 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

# In the Train dataset there are 13036 rows. In the Test dataset there are 4343 rows.

### Task\_3

bike\_recipe = recipe(count ~., bike) %>%   
 step\_rm(instant,dteday,registered,casual) %>%  
 step\_dummy(all\_nominal()) %>%  
 step\_center(all\_predictors()) %>%   
 step\_scale(all\_predictors())  
  
lm\_model =   
 linear\_reg() %>%   
 set\_engine("lm")   
  
lm\_wflow =   
 workflow() %>%   
 add\_model(lm\_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   
## -392.60 -61.14 -9.77 50.67 466.08   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 188.9574 0.9618 196.472 < 2e-16 \*\*\*  
## temp 37.9837 7.1080 5.344 9.26e-08 \*\*\*  
## atemp 17.8510 6.5866 2.710 0.006733 \*\*   
## hum -23.1057 1.3370 -17.282 < 2e-16 \*\*\*  
## windspeed -4.6961 1.0719 -4.381 1.19e-05 \*\*\*  
## season\_Spring 17.7886 2.6243 6.778 1.27e-11 \*\*\*  
## season\_Summer 14.3534 3.1276 4.589 4.49e-06 \*\*\*  
## season\_Fall 27.0767 2.6065 10.388 < 2e-16 \*\*\*  
## mnth\_Feb -0.1546 1.3103 -0.118 0.906102   
## mnth\_Mar 1.7392 1.5402 1.129 0.258824   
## mnth\_Apr -3.1573 2.2348 -1.413 0.157735   
## mnth\_May -0.3196 2.4492 -0.131 0.896165   
## mnth\_Jun -7.1658 2.4824 -2.887 0.003900 \*\*   
## mnth\_Jul -12.6059 2.8230 -4.465 8.06e-06 \*\*\*  
## mnth\_Aug -5.9253 2.7663 -2.142 0.032212 \*   
## mnth\_Sep 3.5721 2.4151 1.479 0.139147   
## mnth\_Oct 1.7747 2.2444 0.791 0.429124   
## mnth\_Nov -3.8768 2.1392 -1.812 0.069968 .   
## mnth\_Dec -1.6939 1.7297 -0.979 0.327451   
## hr\_X1 -3.2601 1.3364 -2.439 0.014723 \*   
## hr\_X2 -4.4301 1.3360 -3.316 0.000916 \*\*\*  
## hr\_X3 -6.5005 1.3377 -4.859 1.19e-06 \*\*\*  
## hr\_X4 -6.8199 1.3239 -5.151 2.62e-07 \*\*\*  
## hr\_X5 -3.5370 1.3335 -2.652 0.008002 \*\*   
## hr\_X6 8.1488 1.3511 6.031 1.67e-09 \*\*\*  
## hr\_X7 35.0107 1.3449 26.031 < 2e-16 \*\*\*  
## hr\_X8 62.0305 1.3365 46.413 < 2e-16 \*\*\*  
## hr\_X9 33.4142 1.3568 24.627 < 2e-16 \*\*\*  
## hr\_X10 20.8594 1.3470 15.486 < 2e-16 \*\*\*  
## hr\_X11 25.3669 1.3577 18.684 < 2e-16 \*\*\*  
## hr\_X12 32.1029 1.3559 23.676 < 2e-16 \*\*\*  
## hr\_X13 32.4099 1.3764 23.547 < 2e-16 \*\*\*  
## hr\_X14 28.8219 1.3922 20.702 < 2e-16 \*\*\*  
## hr\_X15 29.3147 1.3801 21.242 < 2e-16 \*\*\*  
## hr\_X16 41.4926 1.3810 30.045 < 2e-16 \*\*\*  
## hr\_X17 73.1621 1.3859 52.790 < 2e-16 \*\*\*  
## hr\_X18 67.4278 1.3604 49.566 < 2e-16 \*\*\*  
## hr\_X19 46.6999 1.3568 34.420 < 2e-16 \*\*\*  
## hr\_X20 29.9751 1.3413 22.348 < 2e-16 \*\*\*  
## hr\_X21 20.9054 1.3500 15.485 < 2e-16 \*\*\*  
## hr\_X22 14.0398 1.3479 10.416 < 2e-16 \*\*\*  
## hr\_X23 6.4519 1.3394 4.817 1.47e-06 \*\*\*  
## holiday\_Holiday -4.6206 1.0157 -4.549 5.43e-06 \*\*\*  
## weekday\_Sunday -6.2957 1.2579 -5.005 5.66e-07 \*\*\*  
## weekday\_Monday -2.2712 1.2917 -1.758 0.078718 .   
## weekday\_Tuesday -1.8598 1.2563 -1.480 0.138785   
## weekday\_Wednesday -0.7795 1.2608 -0.618 0.536434   
## weekday\_Thursday -1.6518 1.2603 -1.311 0.190012   
## weekday\_Friday 0.1064 1.2596 0.084 0.932676   
## workingday\_WorkingDay NA NA NA NA   
## weathersit\_Misty -2.9873 1.0537 -2.835 0.004587 \*\*   
## weathersit\_LightPrecip -16.6032 1.1109 -14.946 < 2e-16 \*\*\*  
## weathersit\_HeavyPrecip 0.1190 0.9644 0.123 0.901767   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 109.8 on 12984 degrees of freedom  
## Multiple R-squared: 0.6329, Adjusted R-squared: 0.6315   
## F-statistic: 438.9 on 51 and 12984 DF, p-value: < 2.2e-16

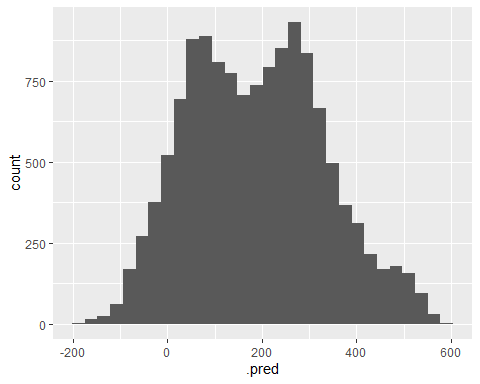
### Task\_ \_4

predict\_train = lm\_fit %>% predict(train)

## Warning in predict.lm(object = object$fit, newdata = new\_data, type =  
## "response"): prediction from a rank-deficient fit may be misleading

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

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



# Based upon my model, it appears that the distribution of predictions peak around 100 and again at around 300.THe data are not normally-distributed.  
  
lm\_fit %>% predict(train) %>% bind\_cols(train) %>% metrics(truth = count, estimate = .pred)

## Warning in predict.lm(object = object$fit, newdata = new\_data, type =  
## "response"): prediction from a rank-deficient fit may be misleading

## # A tibble: 3 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 rmse standard 110.   
## 2 rsq standard 0.633  
## 3 mae standard 79.2

### Task\_5

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

## Warning in predict.lm(object = object$fit, newdata = new\_data, type =  
## "response"): prediction from a rank-deficient fit may be misleading

## # A tibble: 3 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 rmse standard 111.   
## 2 rsq standard 0.629  
## 3 mae standard 80.5

# Applying the model on the testing set, we get an R-squared of 0.6292. This is close to the Adjusted R-squared of 0.6315 for the training dataset.