library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.1.2 ✓ dplyr 1.0.6  
## ✓ tidyr 1.1.3 ✓ 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 0.1.3 ──

## ✓ broom 0.7.6 ✓ rsample 0.1.0   
## ✓ dials 0.0.9 ✓ tune 0.1.5   
## ✓ infer 0.5.4 ✓ workflows 0.2.2   
## ✓ modeldata 0.1.0 ✓ workflowsets 0.0.2   
## ✓ parsnip 0.1.5 ✓ yardstick 0.0.8   
## ✓ recipes 0.1.16

## ── 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()  
## • Use tidymodels\_prefer() to resolve common conflicts.

library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(readr)  
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()  
## )

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

## 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

How many rows of data are in each set (training and testing)? training -12163 testing - 5216

## Task 3

bike\_recipe = recipe(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, train)  
  
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   
## -427.33 -62.08 -9.82 51.84 503.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -81.6699 6.9466 -11.757 < 2e-16 \*\*\*  
## seasonSpring 27.4972 6.3951 4.300 1.72e-05 \*\*\*  
## seasonSummer 18.7645 7.5881 2.473 0.01342 \*   
## seasonFall 62.5367 6.4533 9.691 < 2e-16 \*\*\*  
## mnthFeb -0.5997 5.1373 -0.117 0.90707   
## mnthMar 3.0778 5.7904 0.532 0.59506   
## mnthApr -1.3130 8.6231 -0.152 0.87898   
## mnthMay -2.6894 9.2230 -0.292 0.77060   
## mnthJun -15.8125 9.4879 -1.667 0.09562 .   
## mnthJul -40.2300 10.6077 -3.793 0.00015 \*\*\*  
## mnthAug -16.4993 10.3574 -1.593 0.11119   
## mnthSep 3.9859 9.2187 0.432 0.66548   
## mnthOct -3.0817 8.5334 -0.361 0.71800   
## mnthNov -14.7632 8.2403 -1.792 0.07322 .   
## mnthDec -16.2734 6.5606 -2.480 0.01313 \*   
## hr1 -20.7836 6.9908 -2.973 0.00295 \*\*   
## hr2 -29.0673 6.9980 -4.154 3.29e-05 \*\*\*  
## hr3 -41.4592 7.0968 -5.842 5.29e-09 \*\*\*  
## hr4 -41.2506 7.0386 -5.861 4.73e-09 \*\*\*  
## hr5 -27.2665 6.9794 -3.907 9.41e-05 \*\*\*  
## hr6 31.8318 7.0125 4.539 5.70e-06 \*\*\*  
## hr7 164.5446 7.0278 23.413 < 2e-16 \*\*\*  
## hr8 305.3583 6.9782 43.759 < 2e-16 \*\*\*  
## hr9 163.9524 7.0096 23.390 < 2e-16 \*\*\*  
## hr10 105.9395 6.9986 15.137 < 2e-16 \*\*\*  
## hr11 138.1987 6.9861 19.782 < 2e-16 \*\*\*  
## hr12 179.5246 6.9799 25.720 < 2e-16 \*\*\*  
## hr13 177.5739 7.0533 25.176 < 2e-16 \*\*\*  
## hr14 152.0364 7.1106 21.382 < 2e-16 \*\*\*  
## hr15 170.3496 7.0967 24.004 < 2e-16 \*\*\*  
## hr16 229.1493 7.1110 32.225 < 2e-16 \*\*\*  
## hr17 384.6252 7.0221 54.774 < 2e-16 \*\*\*  
## hr18 342.3854 7.0387 48.643 < 2e-16 \*\*\*  
## hr19 236.7980 7.0437 33.618 < 2e-16 \*\*\*  
## hr20 158.1195 7.0488 22.432 < 2e-16 \*\*\*  
## hr21 107.9022 6.9453 15.536 < 2e-16 \*\*\*  
## hr22 72.0674 6.9890 10.312 < 2e-16 \*\*\*  
## hr23 31.3404 7.0004 4.477 7.64e-06 \*\*\*  
## holidayHoliday -25.5839 6.3712 -4.016 5.97e-05 \*\*\*  
## weekdaySunday -12.8572 3.7603 -3.419 0.00063 \*\*\*  
## weekdayMonday -8.6638 3.8974 -2.223 0.02623 \*   
## weekdayTuesday -6.7687 3.8295 -1.768 0.07716 .   
## weekdayWednesday -3.6852 3.8010 -0.970 0.33231   
## weekdayThursday -3.1739 3.8047 -0.834 0.40418   
## weekdayFriday 0.5683 3.7761 0.151 0.88036   
## temp 293.4586 12.1953 24.063 < 2e-16 \*\*\*  
## weathersitMisty -19.7902 2.3715 -8.345 < 2e-16 \*\*\*  
## weathersitLightPrecip -92.1438 3.8276 -24.073 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -78.2430 64.7522 -1.208 0.22694   
## ---  
## 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

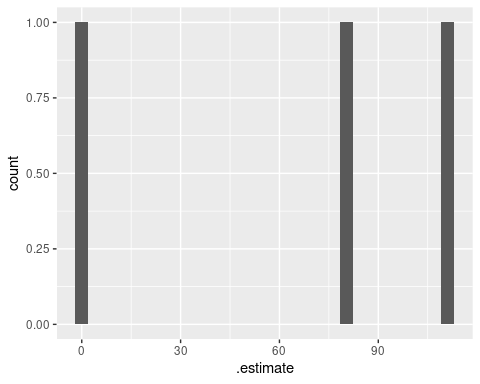
## Task 3 (Cont.)

Looking at the adjusted R-squared they seem to be just about the same and this model seems to be a fit, however running a test on the test set might help us see if this model is overfitting because of the amount of variables.

## Task 4

predict\_train <- lm\_fit %>% predict(train) %>% bind\_cols(train) %>% metrics(truth = count, estimate = .pred)  
  
ggplot(predict\_train, aes(x=.estimate))+  
 geom\_histogram()

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

 ## Task 4 Cont.

The model is overfitting because the r squared does not seem to change.

## Task 5

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

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -368.03 -62.17 -9.76 52.74 481.34   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -95.1958 10.6201 -8.964 < 2e-16 \*\*\*  
## seasonSpring 53.8146 9.5688 5.624 1.96e-08 \*\*\*  
## seasonSummer 45.9841 11.3109 4.065 4.87e-05 \*\*\*  
## seasonFall 70.1242 9.4945 7.386 1.76e-13 \*\*\*  
## mnthFeb 6.1502 7.8024 0.788 0.430594   
## mnthMar 11.6852 8.7203 1.340 0.180307   
## mnthApr -14.9273 12.8628 -1.161 0.245897   
## mnthMay -11.9519 13.7271 -0.871 0.383970   
## mnthJun -21.0722 14.0380 -1.501 0.133395   
## mnthJul -42.1616 15.9378 -2.645 0.008185 \*\*   
## mnthAug -30.3137 15.3587 -1.974 0.048467 \*   
## mnthSep 7.5058 13.6437 0.550 0.582255   
## mnthOct -3.5339 12.7437 -0.277 0.781555   
## mnthNov -23.5373 12.2306 -1.924 0.054352 .   
## mnthDec -10.9877 9.6034 -1.144 0.252617   
## hr1 -11.0652 10.7015 -1.034 0.301190   
## hr2 -21.8114 10.7908 -2.021 0.043300 \*   
## hr3 -29.5513 10.6958 -2.763 0.005749 \*\*   
## hr4 -40.4201 10.9208 -3.701 0.000217 \*\*\*  
## hr5 -19.3389 10.8915 -1.776 0.075859 .   
## hr6 37.0529 10.6798 3.469 0.000526 \*\*\*  
## hr7 180.4572 10.5846 17.049 < 2e-16 \*\*\*  
## hr8 323.7825 10.7211 30.201 < 2e-16 \*\*\*  
## hr9 166.8218 10.6382 15.681 < 2e-16 \*\*\*  
## hr10 125.4148 10.6944 11.727 < 2e-16 \*\*\*  
## hr11 141.5961 10.8453 13.056 < 2e-16 \*\*\*  
## hr12 182.2567 10.9941 16.578 < 2e-16 \*\*\*  
## hr13 173.0712 10.8264 15.986 < 2e-16 \*\*\*  
## hr14 179.6893 10.7555 16.707 < 2e-16 \*\*\*  
## hr15 169.4692 10.8306 15.647 < 2e-16 \*\*\*  
## hr16 237.3228 10.7529 22.070 < 2e-16 \*\*\*  
## hr17 384.3357 10.9501 35.099 < 2e-16 \*\*\*  
## hr18 375.0453 10.7877 34.766 < 2e-16 \*\*\*  
## hr19 252.8421 10.6451 23.752 < 2e-16 \*\*\*  
## hr20 168.1077 10.5605 15.919 < 2e-16 \*\*\*  
## hr21 116.4460 10.8600 10.722 < 2e-16 \*\*\*  
## hr22 72.4276 10.6795 6.782 1.32e-11 \*\*\*  
## hr23 38.4756 10.6167 3.624 0.000293 \*\*\*  
## holidayHoliday -27.2605 9.8003 -2.782 0.005429 \*\*   
## weekdaySunday -23.1150 5.7831 -3.997 6.50e-05 \*\*\*  
## weekdayMonday -4.6715 5.8982 -0.792 0.428385   
## weekdayTuesday -5.4022 5.6764 -0.952 0.341291   
## weekdayWednesday -3.9790 5.7330 -0.694 0.487675   
## weekdayThursday -0.2763 5.7232 -0.048 0.961493   
## weekdayFriday 5.1670 5.7521 0.898 0.369085   
## temp 273.3579 18.8030 14.538 < 2e-16 \*\*\*  
## weathersitMisty -18.3856 3.6236 -5.074 4.04e-07 \*\*\*  
## weathersitLightPrecip -87.1837 5.6581 -15.409 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 110.1 on 5168 degrees of freedom  
## Multiple R-squared: 0.632, Adjusted R-squared: 0.6287   
## F-statistic: 188.9 on 47 and 5168 DF, p-value: < 2.2e-16

## Task 5 Cont.

After running a linear regression on the test model, we see that the r squared value is 62.87 which is a little higher but does not change what we are seeing in both models for train and test.