Stats 101C - Homework 2

Luke Villanueva

Summer 2023

Homework 2 Requirements

You will submit two files.

The files you submit will be:

1. `101C_HW_02_First_Last.Rmd` Take the provided R Markdown file and make the necessary edits so that it generates the requested output.
2. `101C_HW_02_First_Last.pdf` Your output file. This must be a PDF. This is the primary file that will be graded. **Make sure all requested output is visible in the output file.**

Academic Integrity

At the top of your R markdown file, be sure to include the following statement after modifying it with your name.

"By including this statement, I, **Luke Villanueva**, declare that all of the work in this assignment is my own original work. At no time did I look at the code of other students nor did I search for code solutions online. I understand that plagiarism on any single part of this assignment will result in a 0 for the entire assignment and that I will be referred to the dean of students."

Reading:

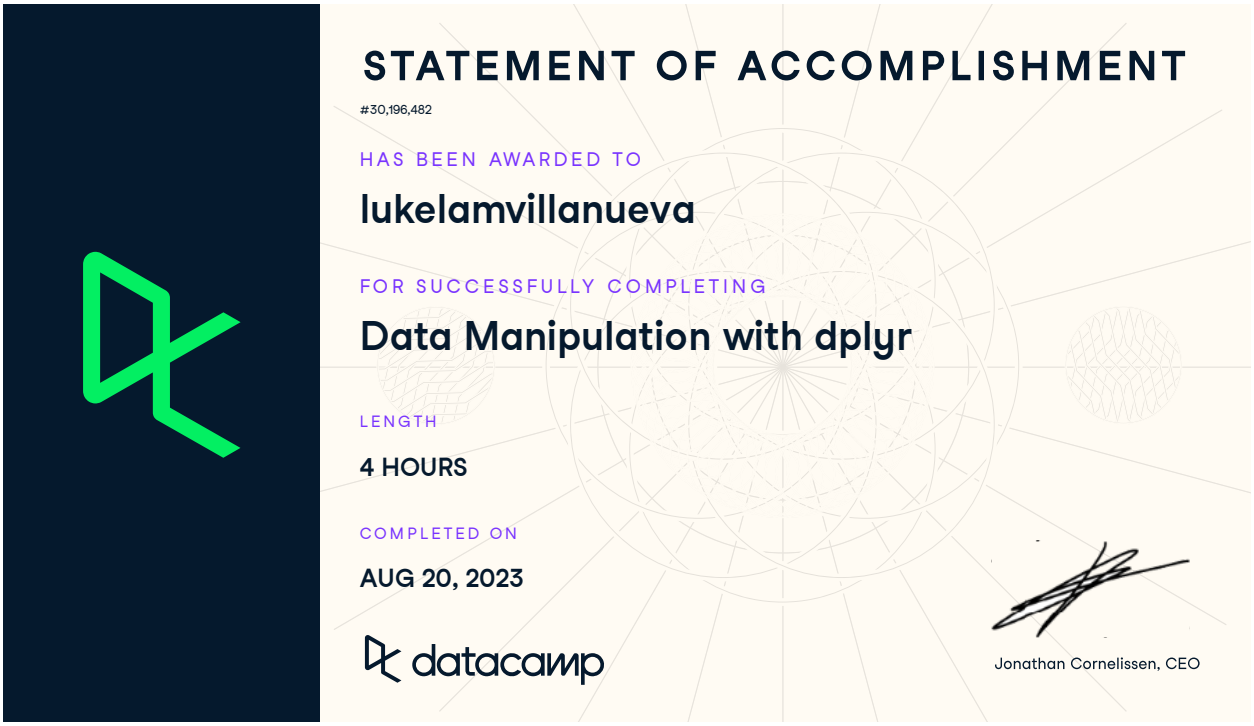
- Read: Introduction to Statistical Learning with R: Chapter 3
- Read: Tidy Modeling with R: Chapter 6 Fitting Models with parsnip
- Read: Tidy Modeling with R: Chapter 7 A Model Workflow
- Read: parsnip documentation: linear_regression() https://parsnip.tidymodels.org/reference/linear_reg.html
- Read: parsnip documentation: fit() https://parsnip.tidymodels.org/reference/fit.html

DataCamp Homework Part 1 (25 pts)

- Course: Data Manipulation with dplyr
- https://app.datacamp.com/learn/courses/data-manipulation-with-dplyr

Include certificate of completion here (30 pts):

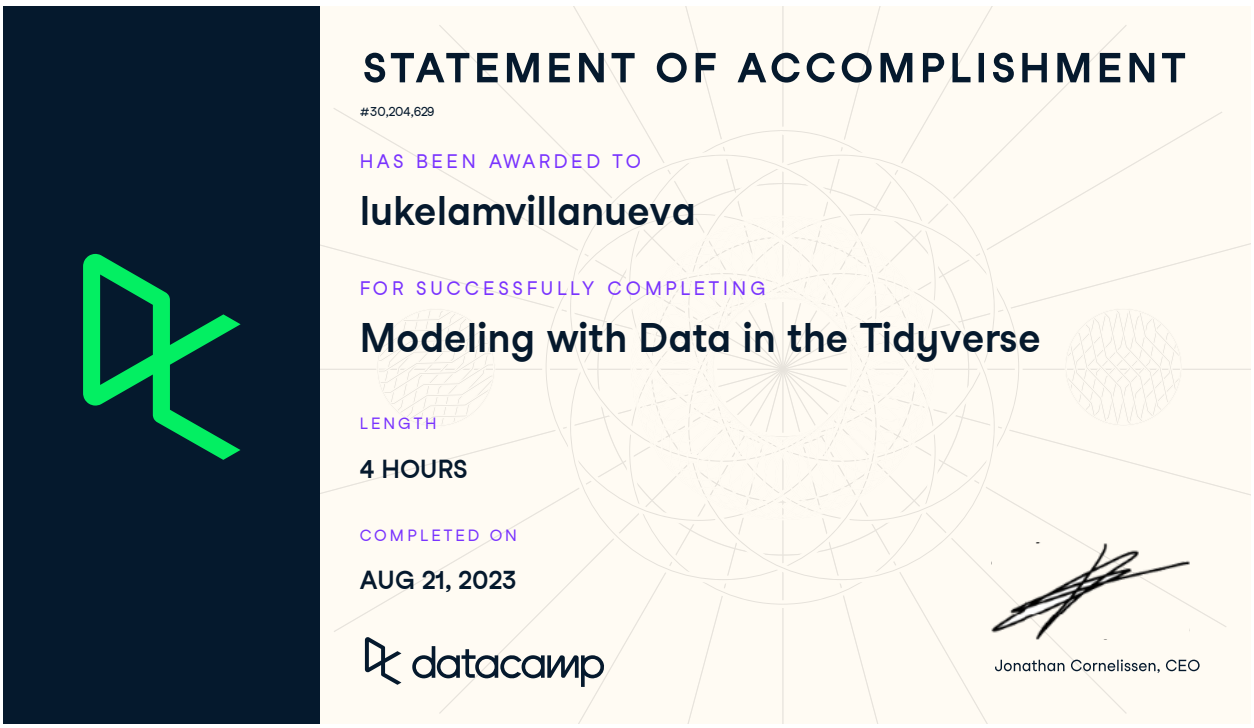
```
include_graphics("certificate.pdf")
```



DataCamp Homework Part 2 (25 pts)

- Course: Modeling with Data in the Tidyverse
- https://app.datacamp.com/learn/courses/modeling-with-data-in-the-tidyverse

```
include_graphics("certificate2.pdf")
```



ISLR Chapter 3 Applied Exercises

The following questions are based on exercises from ISLR Chapter 3, but I have modified them. You can refer to the original questions in the chapter text if some of the questions are confusing because of missing context.

Exercise 8 (modified to use tidymodels)

```
library(ISLR)
library(tidymodels)
data(Auto)
```

step 0. Use `tidymodels` and `rsample` to split Auto into a training set (prop = 0.80) and test set. Use `set.seed(101)` before using `initial_split()`. Stratify on mpg. Report the dimensions of the training and test sets.

a. Use `tidymodels` to fit a simple linear regression model to the training data with mpg as the response and horsepower as the predictor. Use the engine `lm`. Once you fit the model, print the model summary.

If you call `summary()` on the parsnip `model_fit` object, it will print a list summary. To get the traditional `summary()` output associated with lm, use `extract_fit_engine()` along with `summary()`.

```
set.seed(101)

split <- initial_split(Auto, prop = 0.8, strata = mpg)

trainAuto <- training(split)
```

```r
testAuto <- testing(split)

print(paste0("Training: ", nrow(trainAuto)))
```

```
## [1] "Training: 312"
```

```r
print(paste0("Testing: ", nrow(testAuto)))
```

```
## [1] "Testing: 80"
```

```r
lm_model <- linear_reg() %>% set_engine("lm")

lm_fit <- lm_model %>% fit(mpg ~ horsepower, data = trainAuto)

lm_fit %>% extract_fit_engine() %>% summary() %>% print()
```

```
##
## Call:
## stats::lm(formula = mpg ~ horsepower, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6159  -3.1739  -0.4144   2.5778  16.8674
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.102631   0.793211   50.56   <2e-16 ***
## horsepower  -0.159538   0.007168  -22.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.891 on 310 degrees of freedom
## Multiple R-squared:  0.6151, Adjusted R-squared:  0.6139
## F-statistic: 495.4 on 1 and 310 DF,  p-value: < 2.2e-16
```

Answer the following questions based on the fit model:

i. Is there a relationship between the predictor and the response?

- Answer: Yes, there is a relationship because the slope of horsepower has a p-value of close to 0. Meaning, it is unlikely that the slope is obtained through random sampling (i.e. the slope is statistically significant). Also, the R-squared value is generally high as well. In addition, the F-stat's p-value is close to 0, meaning the predictors generally have a statistically significant relationship.

ii. How strong is the relationship between the predictor and the response?

- Answer: The relationship is generally strong because the R-squared value is about 0.62. Meaning, around 62% of the total variance is explained by the predictor's variance. This is a generally strong relationship.

iii. Is the relationship between the predictor and the response positive or negative?

- Answer: The relationship is negative because there is an average of a 0.1595 decrease in mpg as horsepower increases. Meaning, this is a negative relationship.

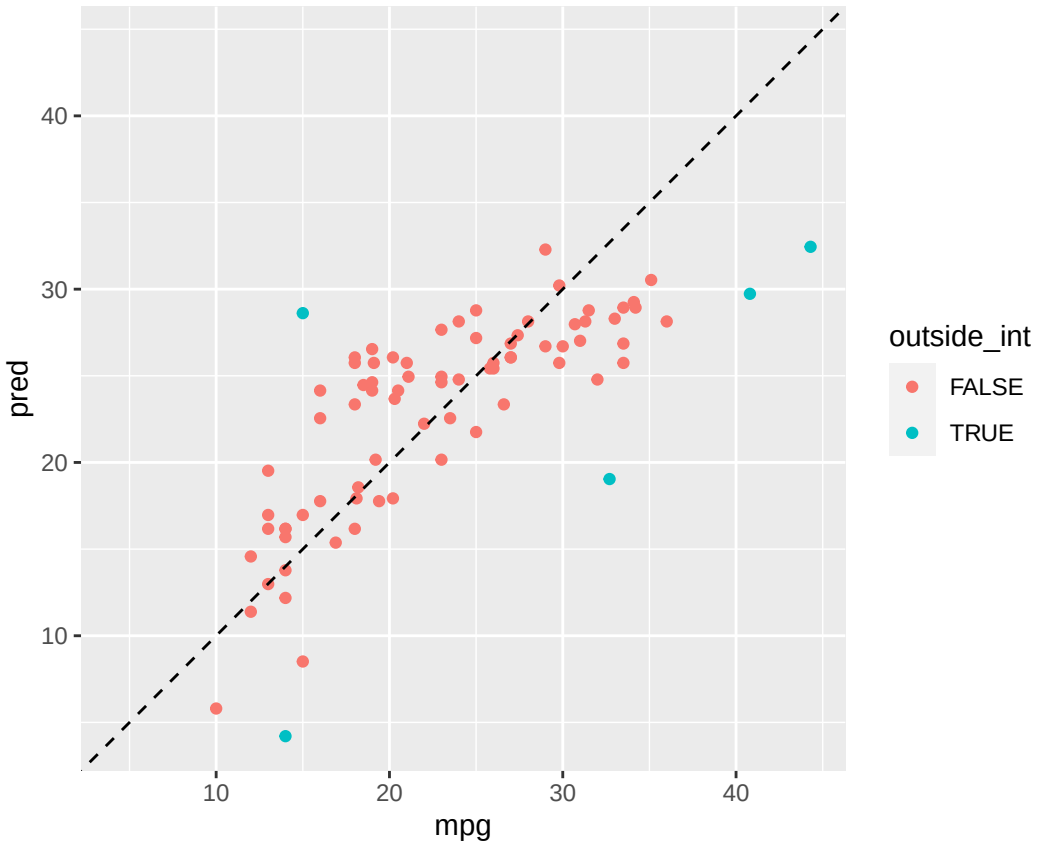
iv. Make predictions on the test set. Create a new data frame that contains the actual mpg, the prediction made by the model, as well as the lower and upper bounds of a 95% prediction interval. Add another column indicating if the actual mpg value is outside the bounds of the prediction interval. Identify which observations failed to make successful prediction intervals.

```
pred <- lm_fit %>% predict(new_data = testAuto)
pred_int <- lm_fit %>% predict(new_data = testAuto, type = "pred_int", level = 0.95)

res <- testAuto %>% select(mpg) %>%
  mutate(pred = pred %>% pull(.pred), pred_lower = pred_int %>% pull(.pred_lower), pred_upper = pred_in
  mutate(outside_int = if_else(mpg < pred_lower | mpg > pred_upper, TRUE, FALSE))

# identify the failed prediction intervals

res %>% filter(outside_int == TRUE)
```

```
##      mpg      pred pred_lower pred_upper outside_int
## 14   14.0  4.206516  -5.583179   13.99621        TRUE
## 155 15.0 28.615874  18.966432   38.26532        TRUE
## 325 40.8 29.732642  20.078121   39.38716        TRUE
## 326 44.3 32.444793  22.773754   42.11583        TRUE
## 334 32.7 19.043577   9.396243   28.69091        TRUE
```

b. Use ggplot and create a plot for the test set with actual mpg on the x-axis and the predicted mpg on the y-axis. Add a geom_abline with a slope of 1 and intercept of 0 (also use lty = 2 to make it dotted) - this line represents where the predictions would be if they were 100% accurate. Add the option `coord_obs_pred()`. Color the observations that failed to make successful prediction intervals a different color from the other observations.

```
ggplot(res, aes(x = mpg, y = pred, color = outside_int)) + geom_point() + geom_abline(slope = 1, interc
```



Exercise 9 (modified to use tidymodels)

This exercise involves the use of multiple linear regression on the Auto data set.

Skip part (a). You can make the plot for your own benefit, but don't include it in your HW solutions.

(b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable from cor() which is qualitative.

```
Auto %>% select(-name) %>% cor()
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
```

```
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

(c) Use tidymodels with the lm engine to create a multiple linear regression with mpg as the response and all other variables except name as the predictors. Print the summary and answer the following questions:

```
mlm_fit <- linear_reg() %>% set_engine("lm") %>% fit(mpg ~ . - name, data = trainAuto)

mlm_fit %>% extract_fit_engine() %>% summary()
```

```
##
## Call:
## stats::lm(formula = mpg ~ . - name, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6870 -2.1372 -0.2185  1.8711 12.8691
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.642e+01  4.989e+00  -3.290  0.00112 **
## cylinders    -6.321e-01  3.491e-01  -1.811  0.07119 .
## displacement  2.695e-02  8.249e-03   3.267  0.00121 **
## horsepower   -1.318e-02  1.467e-02  -0.899  0.36960
## weight       -7.337e-03  7.452e-04  -9.846  < 2e-16 ***
## acceleration  9.477e-02  1.071e-01   0.885  0.37671
## year          7.606e-01  5.527e-02  13.762  < 2e-16 ***
## origin        1.384e+00  2.968e-01   4.662 4.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.192 on 304 degrees of freedom
## Multiple R-squared:  0.8393, Adjusted R-squared:  0.8356
## F-statistic: 226.8 on 7 and 304 DF,  p-value: < 2.2e-16
```

i. Is there a relationship between the predictors and the response?

- Answer: There is a relationship between the predictors and the response because the summary provides that the p-values of the slopes of some variables are statistically significant. However, some are not significant, so not all the predictors have a relationship with the response variable. Also, the r-squared value is 0.83, which means there is a strong correlation between the predictors and the response. In addition, the F-stat's p-value is close to 0, meaning the predictors generally have a statistically significant relationship.

ii. Which predictors appear to have a statistically significant relationship to the response?

- Answer: Based on a 5% significance level, the significant predictors are displacement, weight, year, and origin.

iii. What does the coefficient for the year variable suggest?

- Answer: The coefficient for year suggests that if all other variables were to be controlled, then on average, as year increases by 1, mpg increases by around 0.76.

skip (d)

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
autoCols <- colnames(Auto)[-c(1,9)]

recipeObj <- recipe(mpg ~ ., data = Auto) %>% step_rm(name) %>% step_interact(~all_of(autoCols):all_of(a

wf <- workflow() %>% add_recipe(recipeObj) %>% add_model(lm_model)

wf_fit <- fit(wf, data = Auto)

summary(extract_fit_engine(wf_fit))
```

```
##
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.548e+01  5.314e+01   0.668  0.50475
## cylinders                    6.989e+00  8.248e+00   0.847  0.39738
## displacement                -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                   5.034e-01  3.470e-01   1.451  0.14769
## weight                       4.133e-03  1.759e-02   0.235  0.81442
## acceleration                -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                         6.974e-01  6.097e-01   1.144  0.25340
## origin                      -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders_x_displacement    -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders_x_horsepower       1.161e-02  2.420e-02   0.480  0.63157
## cylinders_x_weight           3.575e-04  8.955e-04   0.399  0.69000
## cylinders_x_acceleration     2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders_x_year            -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders_x_origin           4.022e-01  4.926e-01   0.816  0.41482
## displacement_x_horsepower   -8.491e-05  2.885e-04  -0.294  0.76867
## displacement_x_weight        2.472e-05  1.470e-05   1.682  0.09342 .
## displacement_x_acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement_x_year          5.934e-03  2.391e-03   2.482  0.01352 *
## displacement_x_origin        2.398e-02  1.947e-02   1.232  0.21875
## horsepower_x_weight         -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower_x_acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower_x_year           -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower_x_origin          2.233e-03  2.930e-02   0.076  0.93931
```

```
## weight_x_acceleration        2.346e-04  2.289e-04   1.025  0.30596
## weight_x_year               -2.245e-04  2.127e-04  -1.056  0.29182
## weight_x_origin             -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration_x_year          5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration_x_origin        4.583e-01  1.567e-01   2.926  0.00365 **
## year_x_origin                1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

- Answer: Based on a 5% significance level, the interactions that are statistically significant would be displacement:year, acceleration:year, and acceleration:origin.

Skip part (f)