HW 3

2023-01-12

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(moderndive)
theme_set(theme_minimal())

# now let's load the data from the bayesrules packages
data(bikes, package = "bayesrules")
glimpse(bikes)
```

```
## Rows: 500
## Columns: 13
## $ date         <date> 2011-01-01, 2011-01-03, 2011-01-04, 2011-01-05, 2011-01-0~
## $ season       <fct> winter, winter, winter, winter, winter, winter, winter, wi~
## $ year         <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011~
## $ month        <fct> Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan~
## $ day_of_week  <fct> Sat, Mon, Tue, Wed, Fri, Sat, Mon, Tue, Wed, Thu, Fri, Sat~
## $ weekend      <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALS~
## $ holiday      <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, yes, n~
## $ temp_actual  <dbl> 57.39952, 46.49166, 46.76000, 48.74943, 46.50332, 44.17700~
## $ temp_feel    <dbl> 64.72625, 49.04645, 51.09098, 52.63430, 50.79551, 46.60286~
## $ humidity     <dbl> 80.5833, 43.7273, 59.0435, 43.6957, 49.8696, 53.5833, 48.2~
## $ windspeed    <dbl> 10.749882, 16.636703, 10.739832, 12.522300, 11.304642, 17.~
## $ weather_cat  <fct> categ2, categ1, categ1, categ1, categ2, categ2, categ1, ca~
## $ rides        <int> 654, 1229, 1454, 1518, 1362, 891, 1280, 1220, 1137, 1368, ~
```

```
bikes <- tibble(bikes)
```

Question 0

1. How many rows are in the dataset?
There are 500 rows

2. What does each row represent?
Each row represents a case
3. What dates does the dataset cover?

```
sort(bikes$date)
```

```
##   [1] "2011-01-01" "2011-01-03" "2011-01-04" "2011-01-05" "2011-01-07"
##   [6] "2011-01-08" "2011-01-10" "2011-01-11" "2011-01-12" "2011-01-13"
##  [11] "2011-01-14" "2011-01-15" "2011-01-16" "2011-01-17" "2011-01-18"
##  [16] "2011-01-19" "2011-01-20" "2011-01-21" "2011-01-25" "2011-01-26"
##  [21] "2011-01-27" "2011-01-29" "2011-01-30" "2011-01-31" "2011-02-01"
##  [26] "2011-02-02" "2011-02-03" "2011-02-05" "2011-02-06" "2011-02-07"
##  [31] "2011-02-08" "2011-02-10" "2011-02-11" "2011-02-12" "2011-02-13"
##  [36] "2011-02-14" "2011-02-15" "2011-02-16" "2011-02-17" "2011-02-18"
##  [41] "2011-02-19" "2011-02-20" "2011-02-21" "2011-02-22" "2011-02-23"
##  [46] "2011-02-24" "2011-02-25" "2011-02-26" "2011-02-27" "2011-03-01"
##  [51] "2011-03-03" "2011-03-04" "2011-03-05" "2011-03-07" "2011-03-08"
##  [56] "2011-03-09" "2011-03-10" "2011-03-11" "2011-03-12" "2011-03-13"
##  [61] "2011-03-14" "2011-03-15" "2011-03-16" "2011-03-17" "2011-03-18"
##  [66] "2011-03-19" "2011-03-20" "2011-03-21" "2011-03-23" "2011-03-24"
##  [71] "2011-03-26" "2011-03-27" "2011-03-28" "2011-03-29" "2011-03-30"
##  [76] "2011-03-31" "2011-04-01" "2011-04-02" "2011-04-03" "2011-04-04"
##  [81] "2011-04-05" "2011-04-07" "2011-04-08" "2011-04-09" "2011-04-10"
##  [86] "2011-04-11" "2011-04-12" "2011-04-13" "2011-04-14" "2011-04-15"
##  [91] "2011-04-16" "2011-04-17" "2011-04-18" "2011-04-19" "2011-04-20"
##  [96] "2011-04-21" "2011-04-22" "2011-04-23" "2011-04-24" "2011-04-25"
## [101] "2011-04-26" "2011-04-27" "2011-04-28" "2011-04-29" "2011-04-30"
## [106] "2011-05-01" "2011-05-02" "2011-05-03" "2011-05-04" "2011-05-05"
## [111] "2011-05-06" "2011-05-07" "2011-05-08" "2011-05-10" "2011-05-11"
## [116] "2011-05-12" "2011-05-13" "2011-05-14" "2011-05-15" "2011-05-16"
## [121] "2011-05-17" "2011-05-18" "2011-05-19" "2011-05-20" "2011-05-21"
## [126] "2011-05-22" "2011-05-23" "2011-05-24" "2011-06-03" "2011-06-04"
## [131] "2011-06-13" "2011-06-14" "2011-06-16" "2011-06-17" "2011-06-20"
## [136] "2011-08-23" "2011-08-29" "2011-08-30" "2011-09-02" "2011-09-06"
## [141] "2011-09-07" "2011-09-08" "2011-09-09" "2011-09-10" "2011-09-11"
## [146] "2011-09-12" "2011-09-13" "2011-09-16" "2011-09-17" "2011-09-18"
## [151] "2011-09-19" "2011-09-20" "2011-09-21" "2011-09-22" "2011-09-23"
## [156] "2011-09-24" "2011-09-25" "2011-09-26" "2011-09-27" "2011-09-28"
## [161] "2011-09-29" "2011-09-30" "2011-10-01" "2011-10-03" "2011-10-04"
## [166] "2011-10-05" "2011-10-07" "2011-10-08" "2011-10-09" "2011-10-10"
## [171] "2011-10-11" "2011-10-12" "2011-10-13" "2011-10-14" "2011-10-15"
## [176] "2011-10-16" "2011-10-17" "2011-10-18" "2011-10-20" "2011-10-21"
## [181] "2011-10-22" "2011-10-23" "2011-10-24" "2011-10-25" "2011-10-26"
## [186] "2011-10-27" "2011-10-28" "2011-10-29" "2011-10-31" "2011-11-01"
## [191] "2011-11-02" "2011-11-03" "2011-11-04" "2011-11-05" "2011-11-06"
## [196] "2011-11-07" "2011-11-08" "2011-11-09" "2011-11-10" "2011-11-11"
## [201] "2011-11-12" "2011-11-13" "2011-11-15" "2011-11-16" "2011-11-17"
## [206] "2011-11-18" "2011-11-19" "2011-11-20" "2011-11-21" "2011-11-22"
## [211] "2011-11-23" "2011-11-25" "2011-11-26" "2011-11-27" "2011-11-28"
## [216] "2011-11-29" "2011-11-30" "2011-12-01" "2011-12-02" "2011-12-03"
## [221] "2011-12-04" "2011-12-05" "2011-12-06" "2011-12-07" "2011-12-08"
## [226] "2011-12-09" "2011-12-11" "2011-12-13" "2011-12-15" "2011-12-16"
## [231] "2011-12-17" "2011-12-18" "2011-12-19" "2011-12-20" "2011-12-21"
```

```
## [236] "2011-12-22" "2011-12-23" "2011-12-24" "2011-12-26" "2011-12-27"
## [241] "2011-12-28" "2011-12-30" "2011-12-31" "2012-01-01" "2012-01-02"
## [246] "2012-01-05" "2012-01-06" "2012-01-08" "2012-01-10" "2012-01-11"
## [251] "2012-01-12" "2012-01-13" "2012-01-14" "2012-01-15" "2012-01-16"
## [256] "2012-01-17" "2012-01-18" "2012-01-19" "2012-01-20" "2012-01-21"
## [261] "2012-01-22" "2012-01-23" "2012-01-24" "2012-01-25" "2012-01-26"
## [266] "2012-01-28" "2012-01-29" "2012-01-30" "2012-01-31" "2012-02-01"
## [271] "2012-02-02" "2012-02-03" "2012-02-04" "2012-02-05" "2012-02-06"
## [276] "2012-02-07" "2012-02-08" "2012-02-09" "2012-02-10" "2012-02-11"
## [281] "2012-02-13" "2012-02-14" "2012-02-15" "2012-02-16" "2012-02-17"
## [286] "2012-02-18" "2012-02-19" "2012-02-20" "2012-02-21" "2012-02-22"
## [291] "2012-02-23" "2012-02-24" "2012-02-25" "2012-02-27" "2012-02-28"
## [296] "2012-02-29" "2012-03-01" "2012-03-02" "2012-03-03" "2012-03-04"
## [301] "2012-03-05" "2012-03-06" "2012-03-07" "2012-03-08" "2012-03-09"
## [306] "2012-03-10" "2012-03-11" "2012-03-12" "2012-03-13" "2012-03-14"
## [311] "2012-03-15" "2012-03-16" "2012-03-17" "2012-03-18" "2012-03-19"
## [316] "2012-03-20" "2012-03-21" "2012-03-22" "2012-03-23" "2012-03-24"
## [321] "2012-03-25" "2012-03-26" "2012-03-27" "2012-03-28" "2012-03-29"
## [326] "2012-03-30" "2012-03-31" "2012-04-01" "2012-04-02" "2012-04-03"
## [331] "2012-04-05" "2012-04-06" "2012-04-07" "2012-04-08" "2012-04-09"
## [336] "2012-04-10" "2012-04-12" "2012-04-13" "2012-04-14" "2012-04-15"
## [341] "2012-04-18" "2012-04-19" "2012-04-20" "2012-04-21" "2012-04-22"
## [346] "2012-04-23" "2012-04-24" "2012-04-25" "2012-04-26" "2012-04-27"
## [351] "2012-04-28" "2012-04-29" "2012-04-30" "2012-05-01" "2012-05-02"
## [356] "2012-05-03" "2012-05-04" "2012-05-05" "2012-05-06" "2012-05-07"
## [361] "2012-05-08" "2012-05-09" "2012-05-10" "2012-05-12" "2012-05-13"
## [366] "2012-05-14" "2012-05-15" "2012-05-16" "2012-05-17" "2012-05-20"
## [371] "2012-05-21" "2012-05-22" "2012-05-23" "2012-05-24" "2012-06-01"
## [376] "2012-06-02" "2012-06-03" "2012-06-04" "2012-06-05" "2012-06-06"
## [381] "2012-06-07" "2012-06-08" "2012-06-12" "2012-06-15" "2012-06-16"
## [386] "2012-06-17" "2012-06-18" "2012-06-26" "2012-07-21" "2012-08-17"
## [391] "2012-08-19" "2012-08-20" "2012-08-21" "2012-08-26" "2012-09-09"
## [396] "2012-09-10" "2012-09-11" "2012-09-12" "2012-09-13" "2012-09-14"
## [401] "2012-09-15" "2012-09-18" "2012-09-19" "2012-09-20" "2012-09-21"
## [406] "2012-09-22" "2012-09-23" "2012-09-24" "2012-09-25" "2012-09-26"
## [411] "2012-09-27" "2012-09-28" "2012-09-29" "2012-09-30" "2012-10-01"
## [416] "2012-10-02" "2012-10-03" "2012-10-04" "2012-10-05" "2012-10-06"
## [421] "2012-10-09" "2012-10-10" "2012-10-11" "2012-10-12" "2012-10-13"
## [426] "2012-10-16" "2012-10-17" "2012-10-18" "2012-10-19" "2012-10-20"
## [431] "2012-10-21" "2012-10-22" "2012-10-23" "2012-10-24" "2012-10-25"
## [436] "2012-10-26" "2012-10-27" "2012-10-28" "2012-10-29" "2012-10-30"
## [441] "2012-10-31" "2012-11-01" "2012-11-02" "2012-11-03" "2012-11-04"
## [446] "2012-11-05" "2012-11-06" "2012-11-07" "2012-11-08" "2012-11-09"
## [451] "2012-11-10" "2012-11-11" "2012-11-12" "2012-11-13" "2012-11-14"
## [456] "2012-11-15" "2012-11-17" "2012-11-18" "2012-11-19" "2012-11-20"
## [461] "2012-11-21" "2012-11-22" "2012-11-23" "2012-11-24" "2012-11-25"
## [466] "2012-11-26" "2012-11-27" "2012-11-28" "2012-11-29" "2012-11-30"
## [471] "2012-12-01" "2012-12-02" "2012-12-03" "2012-12-04" "2012-12-05"
## [476] "2012-12-06" "2012-12-07" "2012-12-08" "2012-12-09" "2012-12-10"
## [481] "2012-12-11" "2012-12-12" "2012-12-13" "2012-12-14" "2012-12-15"
## [486] "2012-12-16" "2012-12-17" "2012-12-18" "2012-12-19" "2012-12-20"
## [491] "2012-12-21" "2012-12-22" "2012-12-23" "2012-12-24" "2012-12-25"
## [496] "2012-12-27" "2012-12-28" "2012-12-29" "2012-12-30" "2012-12-31"
```

The data set covers from 2011-01-01 to 2012-12-31.
4. What is the highest observed ridership in the dataset?

```
max(bikes$rides)
```

```
## [1] 6946
```

6946 rides

5. What was the highest wind speed recorded in the dataset

```
max(bikes$windspeed)
```

```
## [1] 34.00002
```

34.00002

Question 1

What is the correlation between number of rides and what the temperature feels like (in Fahrenheit)? What is the correlation between the number of rides and wind speed (miles per hour)?

```
library(boot)
corr(bikes %>% select(temp_feel, rides))
```

```
## [1] 0.5824898
```

```
corr(bikes %>% select(rides, windspeed))
```

```
## [1] -0.1949352
```

There is a strong, positive correlation (0.58) between rides and the temperature feels like (in Fahrenheit).There is a weak, negative correlation (-0.19) between number of rides and wind speed (miles per hour).

Question 2

Using the approximation that a mile is equal to 1.61 kilometers, convert wind speed to kilometers per hour. Call the new variable wind_kph and add it to the bikes data frame. What is the correlation between wind speed in MPH and wind speed in KPH? Explain why in enough detail that I know you understand.

```
bikes <- bikes %>%
  mutate(wind_kph = windspeed * 1.61)

corr(bikes %>% select(wind_kph, windspeed))
```

```
## [1] 1
```

The correlation is 1, meaning there is a perfect correlation. The valued are related in an identical manner (perfectly) because the values for each are the same just converted to different units (for example, when you have 1 mile its perfectly correlated to 1.61 km because this is the same distance just represented differently).

Question 3

Estimate two simple regressions.

```
# regression where windspeed mph predicts rides
mph_fit <- lm(rides ~ windspeed, data = bikes)

# regression where windspeed km predicts rides
kph_fit <- lm(rides ~ wind_kph, data = bikes)
```

Use `get_regression_table()` or `broom::tidy()` to disply the results. f any coefficients are the same between models, explain why. If any coefficients are different between models, explain why. Make sure to give me enough detail to convince me you understand.

```
library(broom)
tidy(mph_fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   4205.      177.      23.8  5.99e-84
## 2 windspeed      -55.5      12.5     -4.44 1.13e- 5
```

```
tidy(kph_fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   4205.     177.       23.8  5.99e-84
## 2 wind_kph       -34.5      7.78     -4.44 1.13e- 5
```

The intercept and the estimate is the same for both models. This is because the only change in the model is the units, which has no effect on the rate of change or the intercept. This is because the change that occurred is perfectly correlated. When ran through the regression equation it will lead to the same results regardless of the changed numerical input, because they are proportional to each other.

Question 4

Using the models from above, tell me what the predicted ridership would be if the wind is blowing at 20 KPH. What would the predicted ridership be if the wind is blowing at 20 MPH?

1. kph

$$4205.06 - 34.49 * 20$$

   The predicted ridership would be 3515.26.

2. mph

$$4205.06 - 34.49 * 20$$

   The predicted ridership would also be 3515.26.

Question 5

Now we're going to move to multiple regression. We will add in `temp_feel` to our model, converted to Celsius.

```
# first I will convert Fahrenheit to Celsius
bikes <- bikes %>%
  mutate(temp_c = temp_feel*5/9)

# now, create the model (additive)
kph_c_fit <- lm(rides ~ wind_kph + temp_c, data = bikes)
tidy(kph_c_fit)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -1604.     402.       -3.99 7.58e- 5
## 2 wind_kph       -19.8      6.46     -3.07 2.24e- 3
## 3 temp_c         143.       9.24     15.5  1.65e-44
```

INTERPRETATIONS.

1. For a 1 degree increase in temperature, the model expects a 143.23 increase in ridership.
2. For a 1 kph increase in wind speed, the model expects a 19.84 decrease in ridership.
3. intercept interpretation: When temperature is at 0 degrees and there is 0 wind, the ridership is -1603.96.

Question 6

Ridership predictions
1.

$$-1603.96 - 19.84 * 15 + 143.23 * 25$$

1679.19 rides.
2.

$$-1603.96 - 19.84 * 5 + 143.23 * 15$$

445.29 rides
3.

$$-1603.96 - 19.84 * 40 + 143.23 * 10$$

The last one is more tricky I think. The answer is -965.26, but since this is a real-world example. I would say that this means there would be 0 rides.

Question 7

Let's add another predictor into the mix. Estimate a new model that uses weekend in addition to the predictors already in the model. Display the model results. Interpret the coefficient on weekend using a complete sentence.

```
# add in another predictor, weekend

complex_fit <- lm(rides ~ wind_kph + temp_c + weekend, data = bikes)
tidy(complex_fit)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -1280.     393.       -3.26 1.21e- 3
## 2 wind_kph       -20.4      6.26     -3.26 1.20e- 3
## 3 temp_c         140.       8.96     15.7  3.26e-45
## 4 weekendTRUE   -714.     122.       -5.83 1.02e- 8
```

When it is the weekend, the model expect a 713.58 decrease in ridership.

Question 8

If the temperature and the wind speed are average, what is the expected ridership for a weekend day? What is the expected ridership for a weekday? Show the code that gets you your answers.

```r
# first, let's find out what "average" means for windspeed and temp
mean(bikes$wind_kph)
```

```
## [1] 20.96062
```

```r
mean(bikes$temp_c)
```

```
## [1] 38.41317
```

The average wind speed is 20.96062 kph.
The average temperature in Celsius is 38.41317 degrees.

```r
# Now that we have the averages, we can create an equation and input these
# numbers in the correct places.

# FOR A WEEKDAY
-1280.10650 - 20.38598 * 20.96062 + 140.33863 * 38.41317 - 713.57504 * 0
```

```
## [1] 3683.442
```

```r
# we put a 0 because this is would only be a 1 if it was the weekend
# this value is for WEEKEND = TRUE

# FOR A WEEKEND
-1280.10650 - 20.38598 * 20.96062 + 140.33863 * 38.41317 - 713.57504 * 1
```

```
## [1] 2969.867
```

The predicted ridership for weekday with average temperature and windspeed is approximately 3638 rides.
The predicted ridership for a weekend with average temperature and windspeed is approximately 2970 rides.

Question 9

You can use `get_regression_points()` *or* `predict()` *to see how the model did at predicting each individual value of rides. Use one of these functions to find the date with the largest absolute residual. That is, find the day where the model is most wrong about predicted ridership. Why is the model so wrong about this day? (There is a correct answer here.)*

```r
q8 <- get_regression_points(complex_fit)

# find the highest residual
q8 %>%
  mutate(residual_ab = abs(residual)) %>%
  arrange(desc(residual_ab))
```

```
## # A tibble: 500 x 8
##       ID rides wind_kph temp_c weekend rides_hat residual residual_ab
##    <int> <int>    <dbl>  <dbl> <lgl>       <dbl>    <dbl>       <dbl>
##  1   439    20    38.6    39.7 FALSE       3510.   -3490.       3490.
##  2   390  5665    25.0    29.9 FALSE       2407.    3258.       3258.
##  3   142  1689    20.8    45.5 FALSE       4688.   -2999.       2999.
##  4   141  1878    10.5    45.0 FALSE       4820.   -2942.       2942.
##  5   423  6736    19.6    39.3 FALSE       3841.    2895.       2895.
##  6   224   655    28.7    37.8 FALSE       3438.   -2783.       2783.
##  7    57   577    28.2    37.1 FALSE       3345.   -2768.       2768.
##  8   422  6911    20.3    42.9 FALSE       4332.    2579.       2579.
##  9   155  2137     8.45   43.9 FALSE       4706.   -2569.       2569.
## 10   426  6612    19.7    40.8 FALSE       4048.    2564.       2564.
## # ... with 490 more rows
```

The day is 2012-10-29. I am not quite sure why the model is wrong, but I am wondering if it has to do with the limited information the model has on the weather conditions on that day. Scanning over the other information available in the data, the weather was a cat3 which means there was a high possibility of stormy weather which could have impacted the ridership.