Lab 09 - Modeling course evaluations, Pt. 1

Tensorflow 2.0

11/05/2020

Load packages and data

```r
library(tidyverse)
library(broom)
```

```r
evals <- read_csv("data/evals-mod.csv")
```

Exercise 1

At first, lets add the average attractiveness score of professors in the dataframe.

```r
evals <- evals %>%
  mutate(bty_avg = rowMeans(select(., bty_f1lower:bty_m2upper)))

evals
```

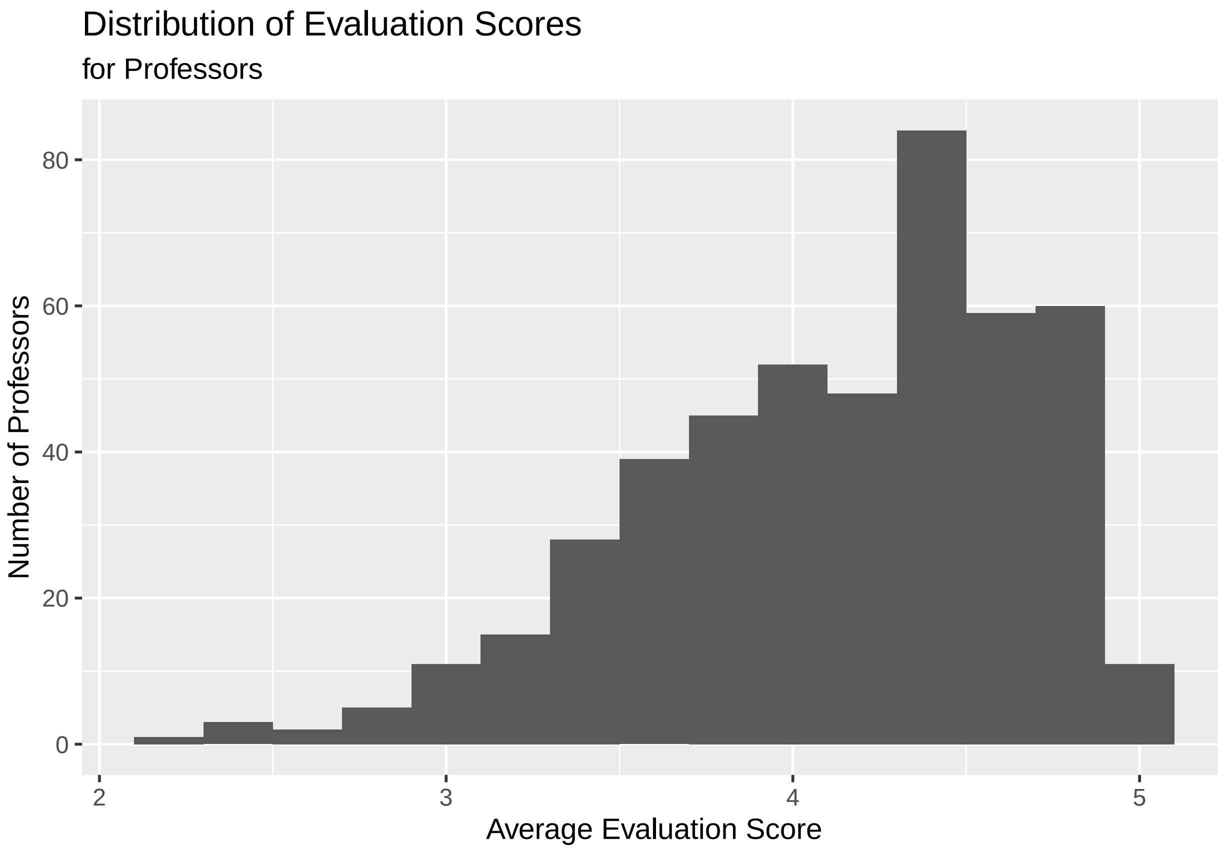
```
## # A tibble: 463 x 19
##    score rank  ethnicity gender language   age cls_perc_eval cls_did_eval
##    <dbl> <chr> <chr>     <chr>  <chr>    <dbl>         <dbl>        <dbl>
##  1   4.7 tenu~ minority  female english     36          55.8           24
##  2   4.1 tenu~ minority  female english     36          68.8           86
##  3   3.9 tenu~ minority  female english     36          60.8           76
##  4   4.8 tenu~ minority  female english     36          62.6           77
##  5   4.6 tenu~ not mino~ male   english     59          85             17
##  6   4.3 tenu~ not mino~ male   english     59          87.5           35
##  7   2.8 tenu~ not mino~ male   english     59          88.6           39
##  8   4.1 tenu~ not mino~ male   english     51         100             55
##  9   3.4 tenu~ not mino~ male   english     51          56.9          111
## 10   4.5 tenu~ not mino~ female english     40          87.0           40
## # ... with 453 more rows, and 11 more variables: cls_students <dbl>,
## #   cls_level <chr>, cls_profs <chr>, cls_credits <chr>, bty_f1lower <dbl>,
## #   bty_f1upper <dbl>, bty_f2upper <dbl>, bty_m1lower <dbl>, bty_m1upper <dbl>,
## #   bty_m2upper <dbl>, bty_avg <dbl>
```

Exercise 2

Now, lets visualize the distribution of the score.

```r
evals %>%
  ggplot(mapping = aes(x = score)) +
  geom_histogram(binwidth = 0.2) +
  labs(
```

```
    x = "Average Evaluation Score",
    y = "Number of Professors",
    title = "Distribution of Evaluation Scores",
    subtitle = "for Professors"
  )
```

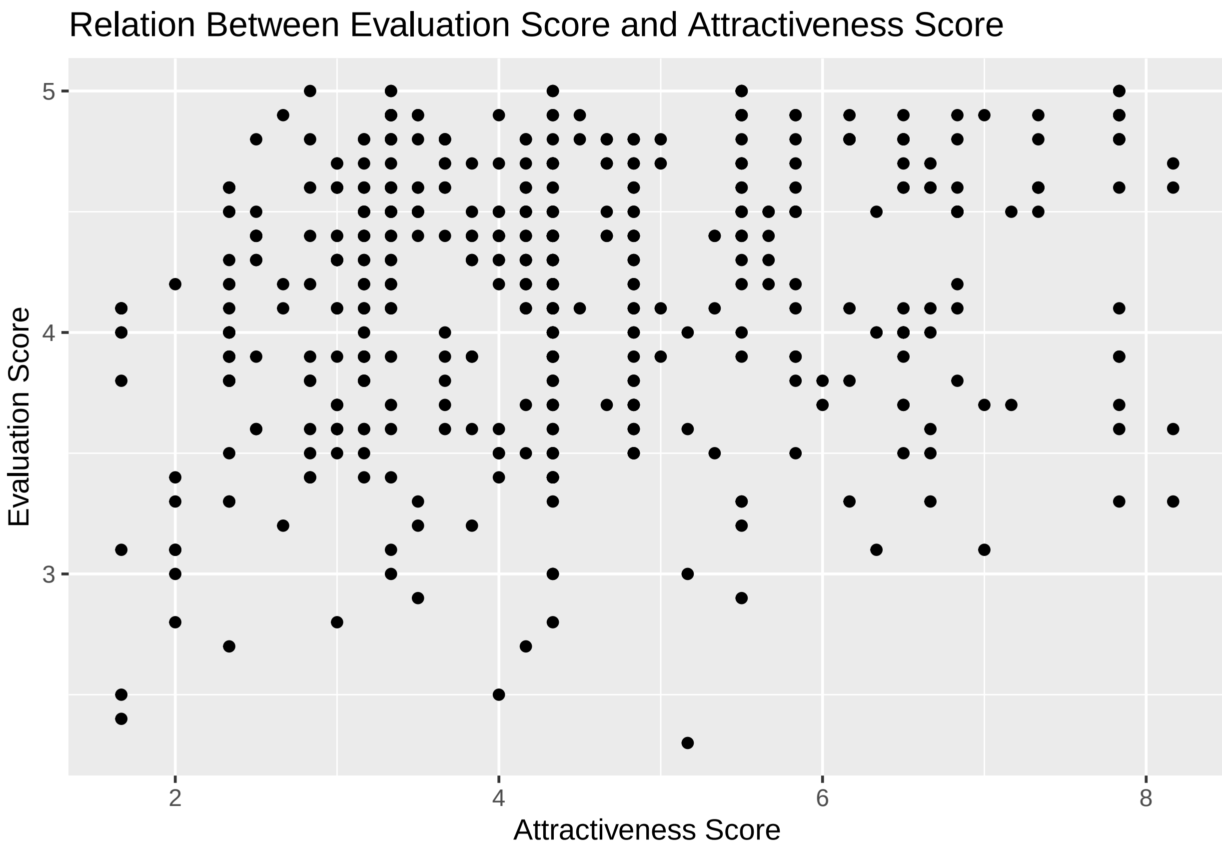


The distribution is left skewed and unimodal. It appears that majority of students had given higher ratings for the professors (4 and 5) and only a few students had given lower rating(less than 3) for the professors. This was as expected because at such a prestigious institution like UT Austin, majority of professors will be well-qualified and have great teaching skills and hence more likely to be appreciated by their students.

Exercise 3

Now, lets visualize the relation between average evaluation score and and average attractiveness score.

```
evals %>%
  ggplot(mapping = aes(x = bty_avg, y = score)) +
  geom_point() +
  labs(
    x = "Attractiveness Score",
    y = "Evaluation Score",
    title = "Relation Between Evaluation Score and Attractiveness Score"
  )
```

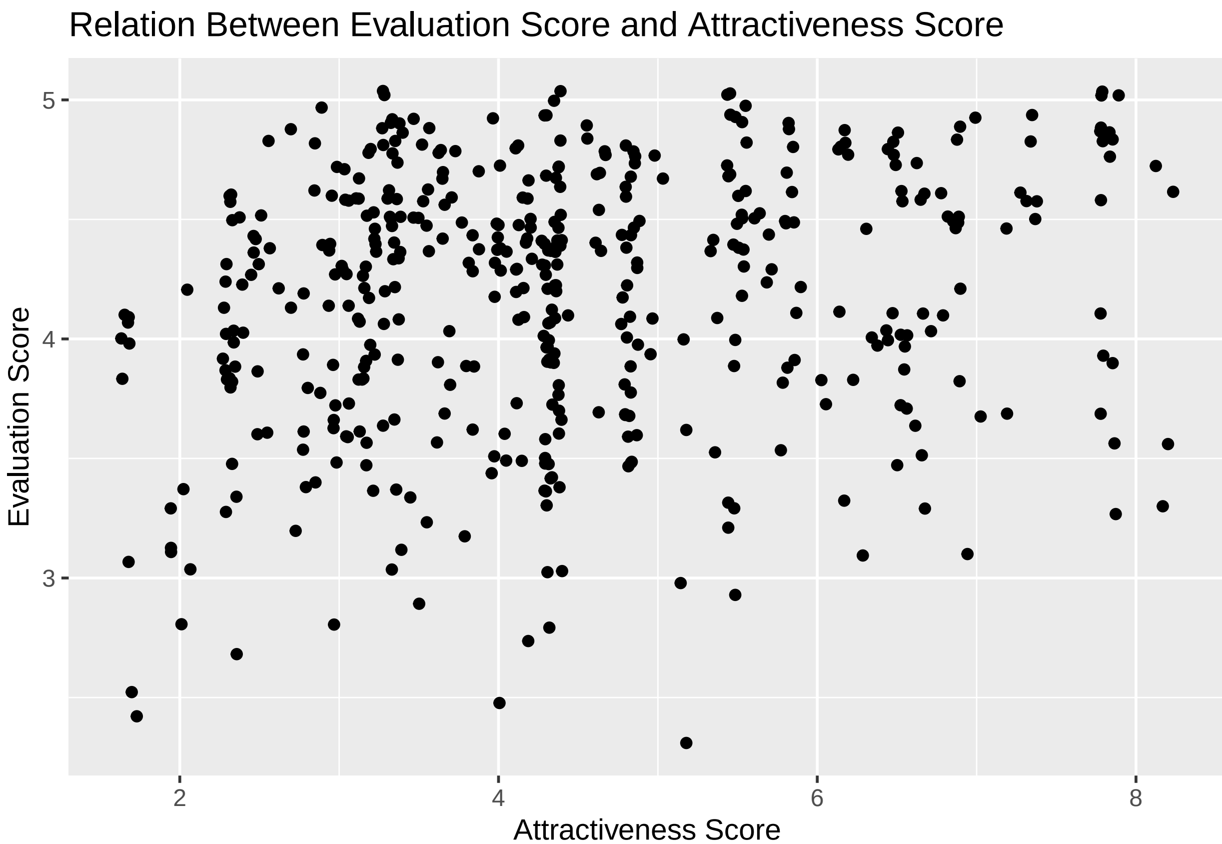


It appears that when the attractiveness scores are higher, the professors are more likely to be given higher evaluation score as it can be seen in the graph that as the average attractiveness score increases, there are fewer professors that have average evaluation score less than 3.5. In other words, the attractiveness score and evaluation score appear to be positively correlated.

Exercise 4

Now, lets use jitter plot to visualize the relation between average evaluation score and average attractiveness score.

```
evals %>%
  ggplot(mapping = aes(x = bty_avg, y = score)) +
  geom_jitter() +
  labs(
    x = "Attractiveness Score",
    y = "Evaluation Score",
    title = "Relation Between Evaluation Score and Attractiveness Score"
  )
```



The advantage of jitter plot over scatter plot is that it adds randomness to the graph as a result of which over-plotted points can be seen as separate from each other. This helps us to visualize the concentration of points more effectively.

The scatter plot showed that we had far fewer number of ratings for the professors as there were only few points in the graph. But it appears from the jitter plot that the number of ratings are far too greater than initially seen. The randomness added in the jitter plot has also given detailed insight into the distribution of evaluation score for any given average attractiveness score of the professors as now, we can visualize the number of professors with different evaluation score for each given average attractiveness score.

Exercise 5

Now, lets fit a model to predict average evaluation score based on average attractiveness score.

```
m_bty <- lm(score ~ bty_avg, data = evals)
tidy(m_bty)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   3.88      0.0761     51.0  1.56e-191
## 2 bty_avg       0.0666    0.0163      4.09 5.08e-  5
```

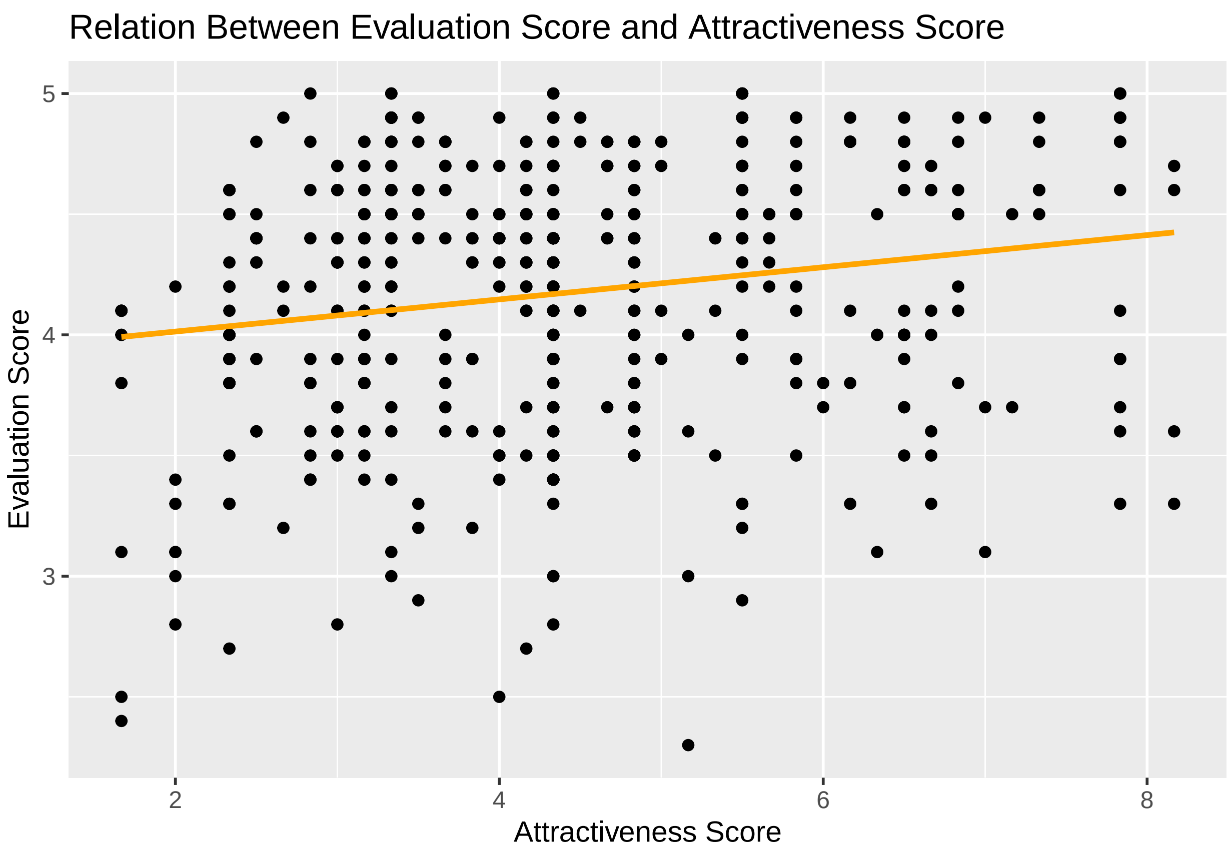
The linear model can be written as:
$\hat{y} = b_0 + b_1 x$
$sc\hat{o}re = 3.88 + 0.067 * btyscore$

Exercise 6

Now, lets visualize the relation between average evaluation score and average attractiveness score with an added regression line.

```
evals %>%
  ggplot(mapping = aes(x = bty_avg, y = score)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "orange") +
  labs(
    x = "Attractiveness Score",
    y = "Evaluation Score",
    title = "Relation Between Evaluation Score and Attractiveness Score"
  )
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Exercise 7

The slope of the model suggests that the average evaluation score of a professor increases by 0.067 on average for each unit increase in the average attractiveness score holding everything else constant.

Exercise 8

The intercept of the model suggests that professors with average attractiveness score of 0 will have average evaluation score of 3.88 on average. It doesn't make sense in this context because attractiveness score can only range in between 1 and 10 inclusive hence there can be no average attractiveness score that is 0.

Exercise 9

The coefficient of determination, R^2, of the model is 0.0350232. It means that 3.5023217 percent of the variability in the value of average evaluation score can be explained by the average attractiveness score.

Exercise 10

Now, lets fit a model to predict average evaluation score based on the gender of the professor.

```
m_gen <- lm(score ~ gender, data = evals)
tidy(m_gen)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    4.09     0.0387    106.   0
## 2 gendermale     0.142    0.0508      2.78 0.00558
```

The equation of regression line for the model above is
$\hat{y} = b_0 + b_1 x$
$\hat{score} = 4.09 + 0.14 * gender$, where gender takes value 1 if it is male and 0 if it is female.

The slope above suggests that male professors will have an average evaluation score that is higher than the average evaluation scores of female by 0.14 on average holding everything else constant. The y-intercept suggests that professors who are female will have an average evaluation score of 4.09 on average.

Exercise 11

The equation of the regression line corresponding to the male professors can be written as:
$\hat{score} = 4.09 + 0.14 * gender$
$\hat{score} = 4.09 + 0.14 * 1$
$\hat{score} = 4.23$

The equation of the regression line corresponding to the female professors can be written as:
$\hat{score} = 4.09 + 0.14 * gender$
$\hat{score} = 4.09 + 0.14 * 0$
$\hat{score} = 4.09$

Exercise 12

Now, lets fit a model to predict average evaluation score based on the rank of the professor.

```
m_rank <- lm(score ~ rank, data = evals)
tidy(m_rank)
```

```
## # A tibble: 3 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)          4.28     0.0537     79.9  1.02e-271
## 2 ranktenure track    -0.130    0.0748     -1.73 8.37e-  2
## 3 ranktenured         -0.145    0.0636     -2.28 2.28e-  2
```

The equation of the linear model that predicts the average professor evaluation score based on the rank of the professor can be written as:
$\hat{y} = b_0 + b_1x_1 + b_2x_2$
$\hat{score} = 4.28 - 0.130 * rank_{tenure\,track} - 0.145 * rank_{tenured}$

The equation above shows that on average professors who are in teaching track will have average evaluation score of 4.28. Those professors who are on a tenure track will have average score that is less than a professor on teaching track by 0.130 on average holding everything else constant. Similarly, those professors who are tenured will have average score that is less than a professor on teaching track by 0.145 on average holding everything else constant.

Exercise 13

Now, adding a new variable called `rank_leveled`.

```
evals <- evals %>%
  mutate(rank_leveled = fct_relevel(rank, "tenure track"))

evals
```

```
## # A tibble: 463 x 20
##    score rank  ethnicity gender language   age cls_perc_eval cls_did_eval
##    <dbl> <chr> <chr>     <chr>  <chr>    <dbl>         <dbl>        <dbl>
##  1   4.7 tenu~ minority  female english     36          55.8           24
##  2   4.1 tenu~ minority  female english     36          68.8           86
##  3   3.9 tenu~ minority  female english     36          60.8           76
##  4   4.8 tenu~ minority  female english     36          62.6           77
##  5   4.6 tenu~ not mino~ male   english     59          85             17
##  6   4.3 tenu~ not mino~ male   english     59          87.5           35
##  7   2.8 tenu~ not mino~ male   english     59          88.6           39
##  8   4.1 tenu~ not mino~ male   english     51         100             55
##  9   3.4 tenu~ not mino~ male   english     51          56.9          111
## 10   4.5 tenu~ not mino~ female english     40          87.0           40
## # ... with 453 more rows, and 12 more variables: cls_students <dbl>,
## #   cls_level <chr>, cls_profs <chr>, cls_credits <chr>, bty_f1lower <dbl>,
## #   bty_f1upper <dbl>, bty_f2upper <dbl>, bty_m1lower <dbl>, bty_m1upper <dbl>,
## #   bty_m2upper <dbl>, bty_avg <dbl>, rank_leveled <fct>
```

Exercise 14

Finally, fitting a model to predict average evaluation score based on the leveled rank of the professor.

```
m_rank_leveled <- lm(score ~ rank_leveled, data = evals)
tidy(m_rank_leveled)
```

```
## # A tibble: 3 x 5
##   term                 estimate std.error statistic   p.value
##   <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)            4.15      0.0521    79.7   2.58e-271
## 2 rank_leveledteaching   0.130     0.0748     1.73  8.37e-  2
## 3 rank_leveledtenured   -0.0155    0.0623    -0.249 8.04e-  1
```

The equation of the linear model that predicts the average professor evaluation score based on the leveled rank can be written as:

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$
$\hat{score} = 4.15 + 0.130 * rankleveled_{teaching} - 0.016 * rankleveled_{tenured}$

This model will produce the same prediction as the model we fitted in exercise 12 because releveling of the rank will only change the interpretation of individual predictors in reference to the intercept.In other words, this releveling will simply help us view other levels against "tenure track" as the reference level without changing the average evaluation score for any given instance of an observation.

The coefficient of determinant, R^2, of the model is 0.0116289. This means that 1.1628942 percent of the variability in the value of average evaluation score can be explained by the rank of the professor.