Teaching Scores Data Analysis

2024-06-22

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(moderndive)
library(skimr)
```

Selecting a subset of variables

```
evals_ch5 <- evals %>%
  select(ID, score, bty_avg, age)
```

Previewing the raw data

```
glimpse(evals_ch5)
```

```
## Rows: 463
## Columns: 4
## $ ID      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,~
## $ score   <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, 4.~
## $ bty_avg <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, 3.333,~
## $ age     <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, 40~
```

Selecting a random sample of 5 rows to preview

```
evals_ch5 %>%
  sample_n(size = 5)
```

```
## # A tibble: 5 x 4
##      ID score bty_avg   age
##   <int> <dbl>   <dbl> <int>
## 1   415   4.9    6.83    54
## 2    96   4.1    4.33    48
## 3   164   4.4    4.33    63
## 4   183   3      4.33    47
## 5    36   3.4    4       51
```

Computing summary statistics for beauty score and teaching score

```
evals_ch5 %>%
  summarize(mean_bty_avg = mean(bty_avg), mean_score = mean(score),
            median_bty_avg = median(bty_avg), median_score = median(score))
```

```
## # A tibble: 1 x 4
##   mean_bty_avg mean_score median_bty_avg median_score
##          <dbl>      <dbl>          <dbl>        <dbl>
## 1         4.42       4.17           4.33          4.3
```

```
evals_ch5 %>%
  select(score, bty_avg) %>%
  skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	463
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
score	0	1	4.17	0.54	2.30	3.80	4.30	4.6	5.00	
bty_avg	0	1	4.42	1.53	1.67	3.17	4.33	5.5	8.17	

Correlation coefficient of beauty score and teaching score

```
evals_ch5 %>%
  get_correlation(formula = score ~ bty_avg)
```

```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.187
```

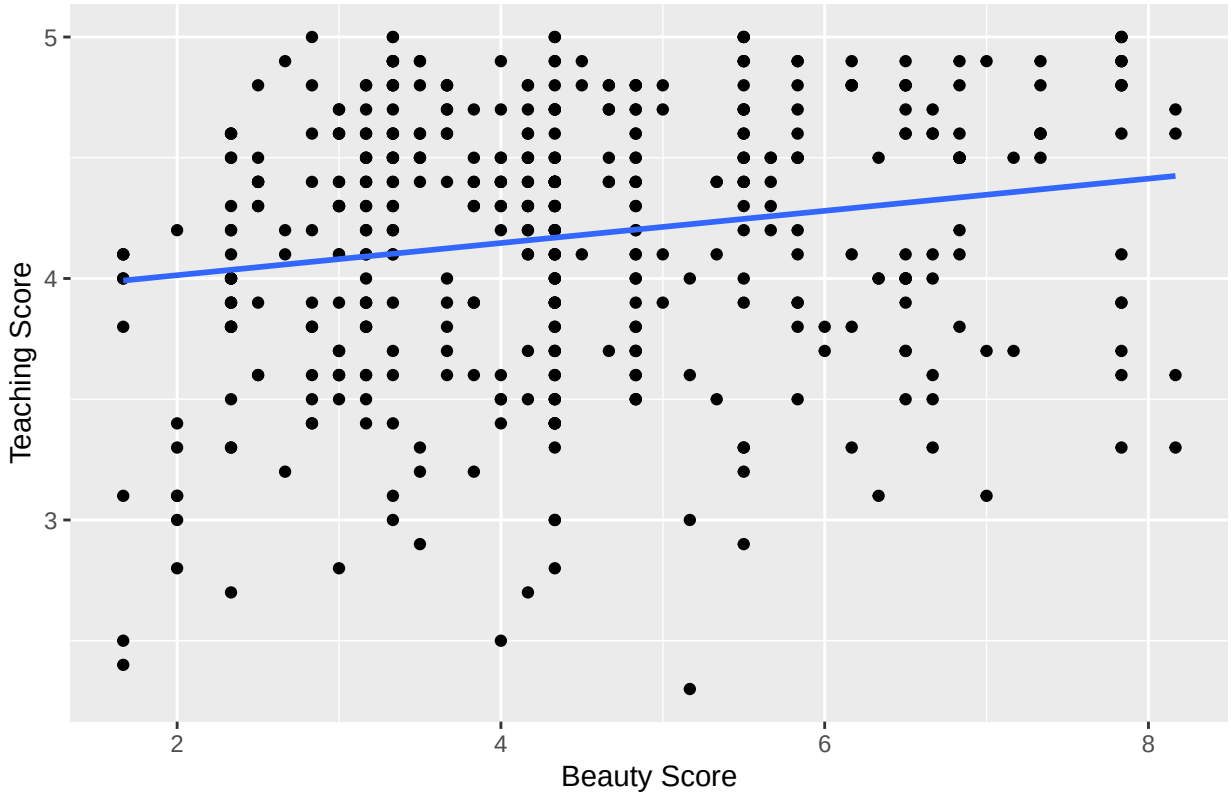
Data visualization: Scatterplot of the relationship between teaching and beauty scores

The positive slope of the regression line demonstrates a positive relationship, as instructors have a higher beauty score they also receive higher evaluations.

```
ggplot(evals_ch5, aes(x = bty_avg, y = score)) +
  geom_point() +
  labs(x = "Beauty Score",
       y = "Teaching Score",
       title = "Scatterplot of relationship between teaching and beauty scores.") +
  geom_smooth(method = "lm", se = FALSE)
```

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

Scatterplot of relationship between teaching and beauty scores.



Fit regression model

```
score_model <- lm(score ~ bty_avg, data = evals_ch5)
```

Get regression table

For every increase of 1 unit of bty_avg there is an associated increase of, on average, 0.067 units of score.

```
get_regression_table(score_model)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    3.88      0.076     51.0        0    3.73     4.03
## 2 bty_avg      0.067     0.016      4.09       0    0.035    0.099
```

Dataframe of the residuals of the model for teaching and beauty scores

```
regression_points <- get_regression_points(score_model)
regression_points
```

```
## # A tibble: 463 x 5
##       ID score bty_avg score_hat residual
##    <int> <dbl>   <dbl>     <dbl>    <dbl>
##  1     1   4.7    5         4.21    0.486
##  2     2   4.1    5         4.21   -0.114
##  3     3   3.9    5         4.21   -0.314
##  4     4   4.8    5         4.21    0.586
##  5     5   4.6    3         4.08    0.52
##  6     6   4.3    3         4.08    0.22
```

```
##  7     7   2.8    3         4.08   -1.28
##  8     8   4.1    3.33      4.10   -0.002
##  9     9   3.4    3.33      4.10   -0.702
## 10    10   4.5    3.17      4.09    0.409
## # i 453 more rows
```

Computing summary statistics for age and teaching score

```
evals_ch5 %>%
  summarize(mean_age = mean(age), mean_score = mean(score),
            median_age = median(age), median_score = median(score))
```

```
## # A tibble: 1 x 4
##   mean_age mean_score median_age median_score
##      <dbl>      <dbl>      <int>        <dbl>
## 1     48.4       4.17         48          4.3
```

```
evals_ch5 %>%
  select(age, score) %>%
  skim()
```

Table 3: Data summary

Name	Piped data
Number of rows	463
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	48.37	9.80	29.0	42.0	48.0	57.0	73	
score	0	1	4.17	0.54	2.3	3.8	4.3	4.6	5	

Correlation coefficient of age and teaching score

```
evals_ch5 %>%
  get_correlation(formula = age ~ score)
```

```
## # A tibble: 1 x 1
##      cor
##    <dbl>
## 1 -0.107
```

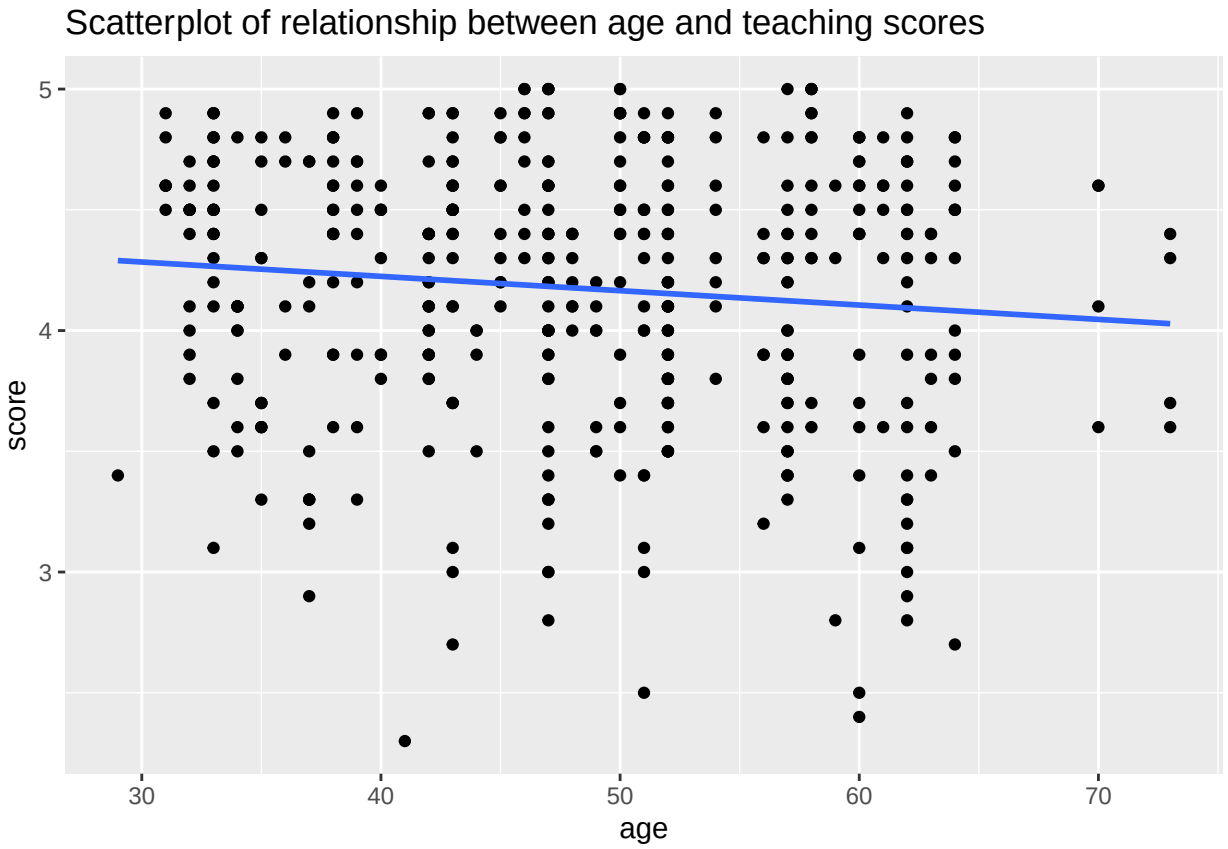
Data visualization: Scatterplot of the relationship between age and teaching scores

The negative slope of the regression line demonstrates a negative relationship, as the age decreases they also receive lower evaluations.

```
ggplot(evals_ch5, aes(x = age, y = score)) +
  geom_point() +
```

```
  labs(x = "age",
       y = "score",
       title = "Scatterplot of relationship between age and teaching scores") +
  geom_smooth(method = "lm",se = FALSE)
```

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



Fit regression model

```
age_model <- lm(score ~ age, data = evals_ch5)
```

Get regression table

For every increase of 1 unit of age there is an associated increase of, on average, -0.006 units of score.

```
get_regression_table(age_model)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    4.46      0.127     35.2    0        4.21     4.71
## 2 age         -0.006     0.003     -2.31   0.021   -0.011   -0.001
```

Dataframe of the residuals of the model for age and teaching scores

```
regression_points2 <- get_regression_points(age_model)
regression_points2
```

```
## # A tibble: 463 x 5
##       ID score   age score_hat residual
```

```
##    <int> <dbl> <int>     <dbl>    <dbl>
##  1     1   4.7    36      4.25    0.452
##  2     2   4.1    36      4.25   -0.148
##  3     3   3.9    36      4.25   -0.348
##  4     4   4.8    36      4.25    0.552
##  5     5   4.6    59      4.11    0.488
##  6     6   4.3    59      4.11    0.188
##  7     7   2.8    59      4.11   -1.31
##  8     8   4.1    51      4.16   -0.059
##  9     9   3.4    51      4.16   -0.759
## 10    10   4.5    40      4.22    0.276
## # i 453 more rows
```