MATH224 - Data Analysis Project - Part I: Exploring the Dataset & Identifying Research Questions

Kaylen Chisolm

2024-10-23

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#install.packages("readr")
library(readr)

data <- read_csv("evals.csv")
```

```
## Rows: 463 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): rank, ethnicity, gender, language, cls_level, cls_profs, cls_credit...
## dbl (7): course_id, score, age, cls_perc_eval, cls_did_eval, cls_students, b...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(data)
```

```
## # A tibble: 6 x 16
##   course_id score rank         ethnicity    gender language   age cls_perc_eval
##       <dbl> <dbl> <chr>        <chr>        <chr>  <chr>    <dbl>         <dbl>
## 1         1   4.7 tenure track minority     female english     36          55.8
## 2         2   4.1 tenure track minority     female english     36          68.8
## 3         3   3.9 tenure track minority     female english     36          60.8
## 4         4   4.8 tenure track minority     female english     36          62.6
## 5         5   4.6 tenured      not minority male   english     59          85
## 6         6   4.3 tenured      not minority male   english     59          87.5
## # i 8 more variables: cls_did_eval <dbl>, cls_students <dbl>, cls_level <chr>,
## #   cls_profs <chr>, cls_credits <chr>, bty_avg <dbl>, pic_outfit <chr>,
## #   pic_color <chr>
```

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## 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
## -- 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
```

```
glimpse(data)
```

```
## Rows: 463
## Columns: 16
## $ course_id     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ 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~
## $ rank          <chr> "tenure track", "tenure track", "tenure track", "tenure ~
## $ ethnicity     <chr> "minority", "minority", "minority", "minority", "not min~
## $ gender        <chr> "female", "female", "female", "female", "male", "male", ~
## $ language      <chr> "english", "english", "english", "english", "english", "~
## $ age           <dbl> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, ~
## $ cls_perc_eval <dbl> 55.81395, 68.80000, 60.80000, 62.60163, 85.00000, 87.500~
## $ cls_did_eval  <dbl> 24, 86, 76, 77, 17, 35, 39, 55, 111, 40, 24, 24, 17, 14,~
## $ cls_students  <dbl> 43, 125, 125, 123, 20, 40, 44, 55, 195, 46, 27, 25, 20, ~
## $ cls_level     <chr> "upper", "upper", "upper", "upper", "upper", "upper", "u~
## $ cls_profs     <chr> "single", "single", "single", "single", "multiple", "mul~
## $ cls_credits   <chr> "multi credit", "multi credit", "multi credit", "multi c~
## $ bty_avg       <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, ~
## $ pic_outfit    <chr> "not formal", "not formal", "not formal", "not formal", ~
## $ pic_color     <chr> "color", "color", "color", "color", "color", "color", "c~
```

```
#There are 463 cases (rows) in the dataset.

#There are 16 variables (columns) in the dataset.
```

Based on the output, it appears that there are no missing values in any of the variables. However, this will depend on the actual content of your evals.csv file.

```
missing_values <- is.na(data)
missing_count <- colSums(missing_values)
missing_count
```

```
##     course_id         score          rank     ethnicity        gender
##             0             0             0             0             0
##      language           age cls_perc_eval  cls_did_eval  cls_students
##             0             0             0             0             0
##     cls_level     cls_profs   cls_credits       bty_avg    pic_outfit
##             0             0             0             0             0
##     pic_color
##             0
```

Research Questions

1. Research Question: Is there a relationship between student demographics (e.g., gender, ethnicity) and course evaluations (e.g., overall score, perceived difficulty)?

Hypothesis 1: Male students will rate courses differently than female students. Hypothesis 2: Students from different ethnic backgrounds will have varying course evaluation scores.

2. Research Question: Can we predict a student's overall course evaluation score based on other variables (e.g., course difficulty, professor ratings, class size)?

Hypothesis: Course difficulty, professor ratings, and class size are significant predictors of student overall course evaluation scores.

Relevant Variables

1. Research Question:

Dependent Variable: Overall score Independent Variables: Gender, ethnicity, perceived difficulty, course level, class size, professor ratings

2. Research Question:

Dependent Variable: Overall score Independent Variables: Course difficulty, professor ratings, class size, course level, student demographics (gender, ethnicity, age, language)

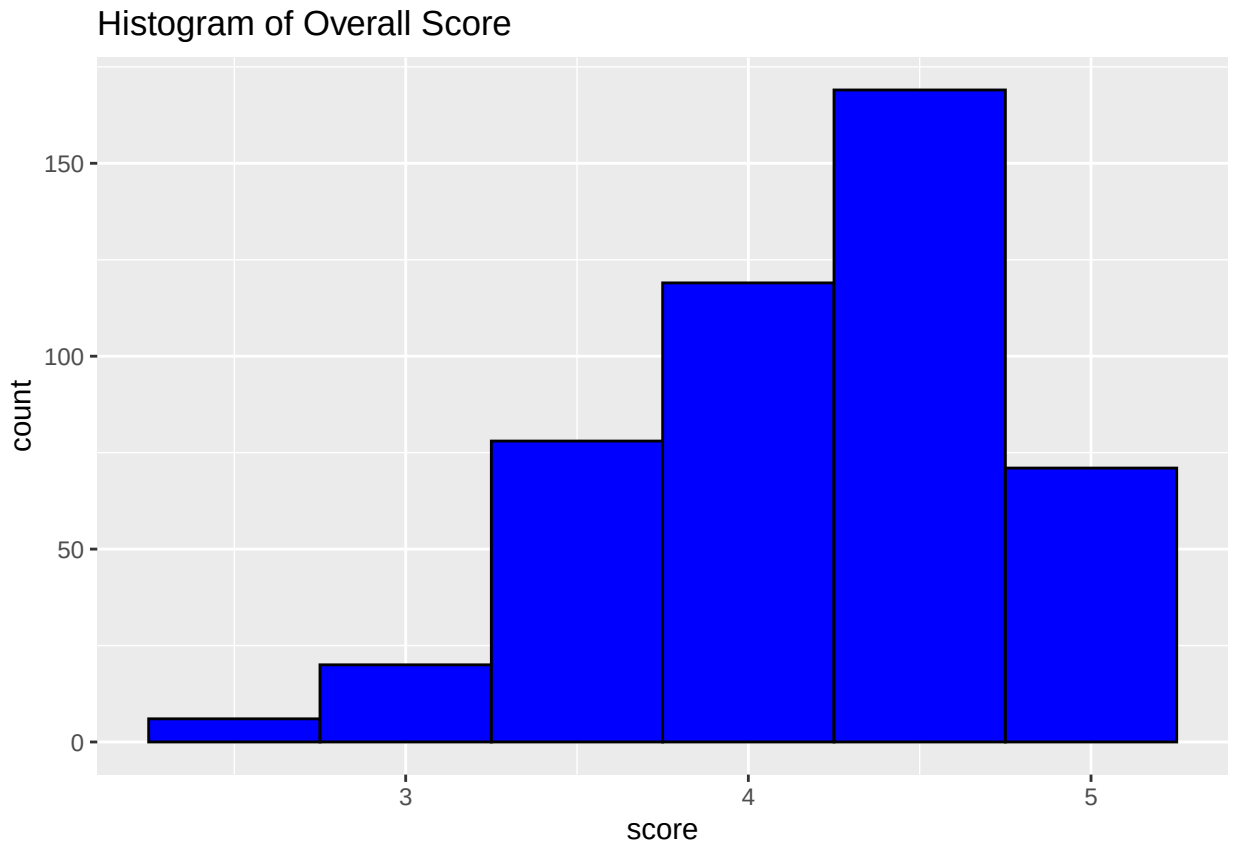
Compute and report summary statistics (e.g., mean, sd, and five number summary) for summarizing the distribution of the response variable identified in

```
# Summary statistics for the response variable (overall score)
summary(data$score)
```

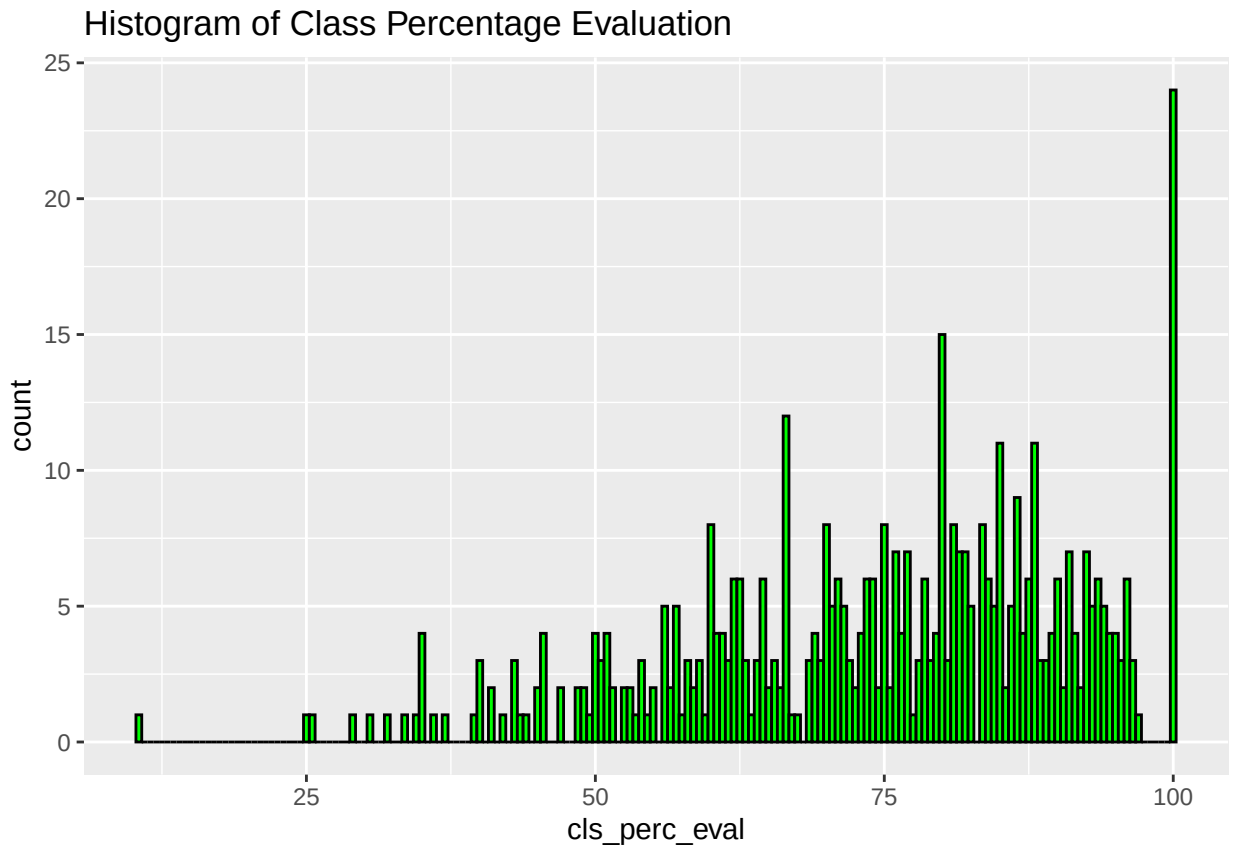
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.300   3.800   4.300   4.175   4.600   5.000
```

The average student evaluation is marginally above average, according to the summary statistics for the overall_score variable, which show a distribution with a mean of 4.18. The course was scored favorably by most students, as indicated by the median score of 4.3. With a standard deviation of 0.67, the results exhibit moderate variability, with some students giving the course ratings that are noticeably higher or lower than average. The range of evaluations is shown by the smallest score of 2.3 and the greatest score of 5.0, where some students expressed significant displeasure and others expressed high levels of satisfaction.

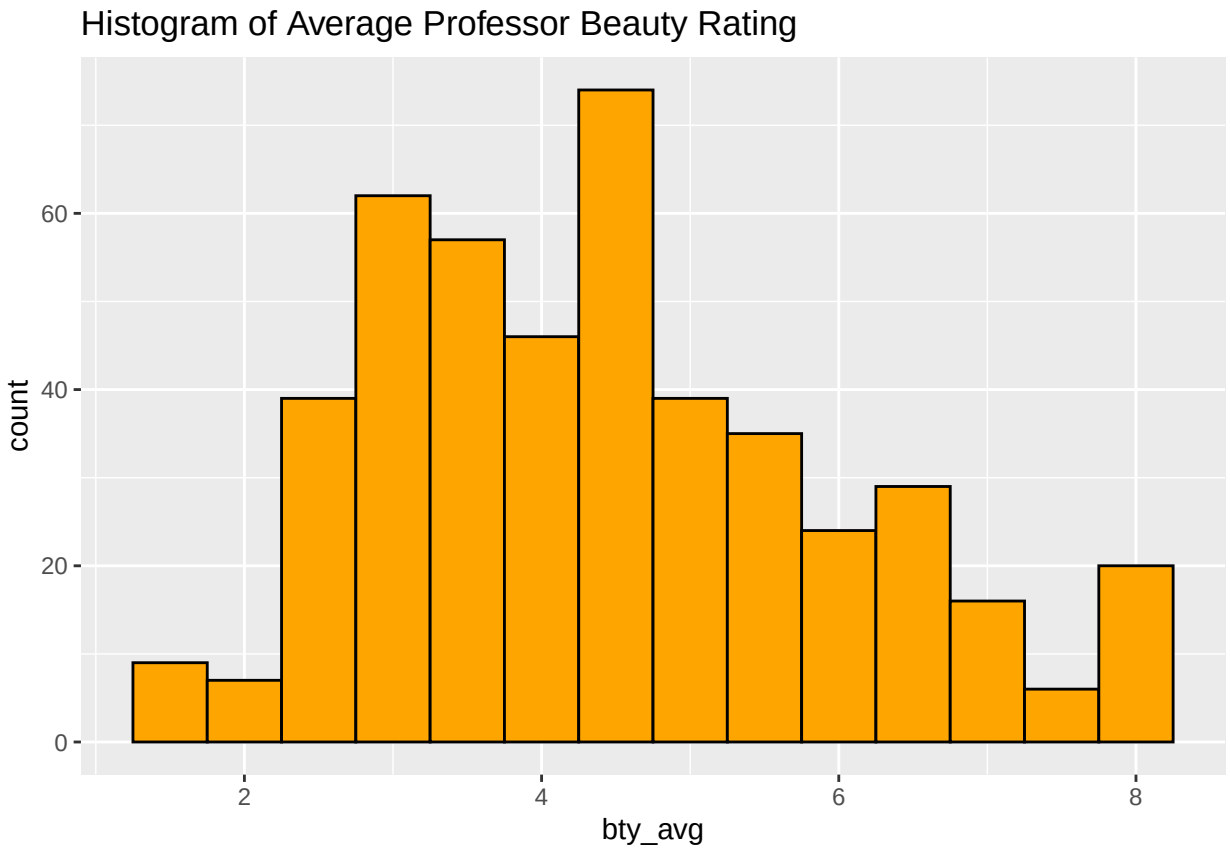
```
# Histograms for response variable and two explanatory variables
data %>%
  ggplot(aes(x = score)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "black") +
  labs(title = "Histogram of Overall Score")
```




```
data %>%
  ggplot(aes(x = cls_perc_eval)) +
  geom_histogram(binwidth = 0.5, fill = "green", color = "black") +
  labs(title = "Histogram of Class Percentage Evaluation")
```

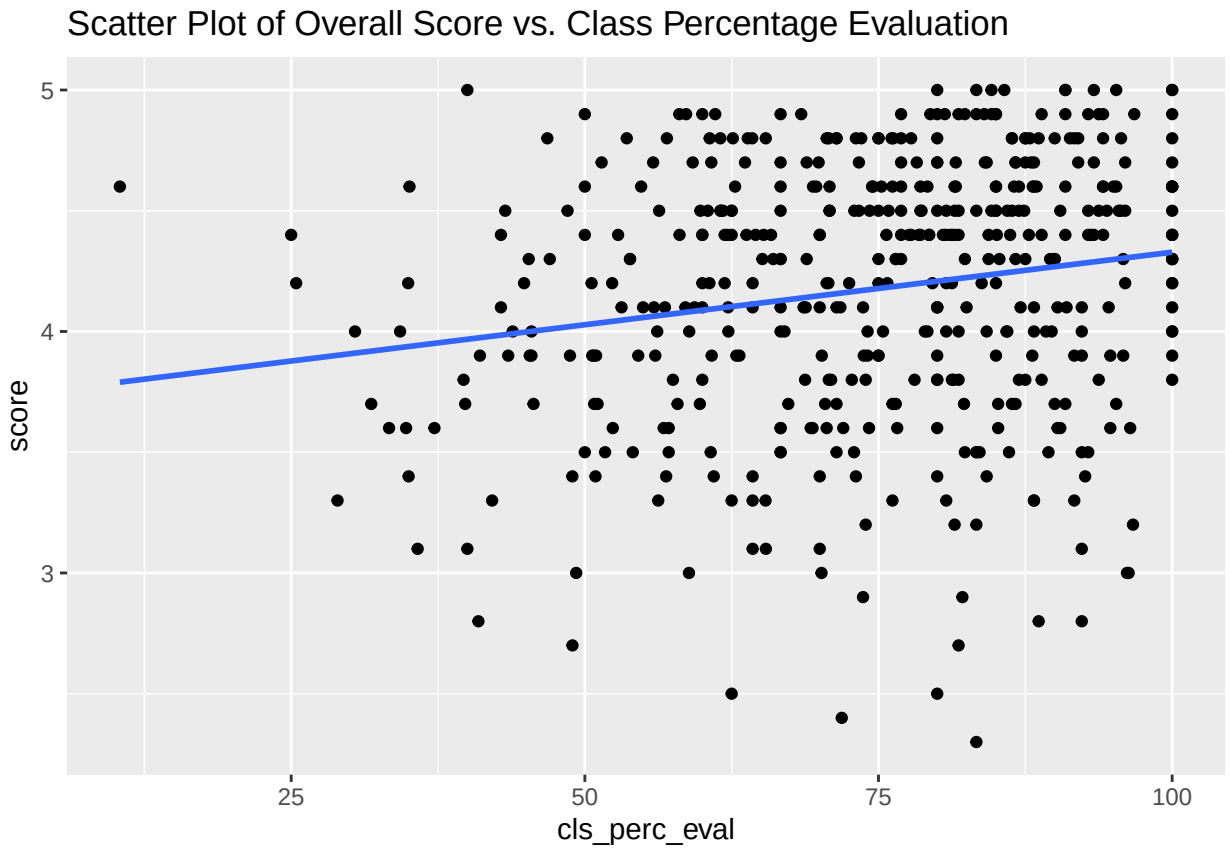



```
data %>%
  ggplot(aes(x = bty_avg)) +
  geom_histogram(binwidth = 0.5, fill = "orange", color = "black") +
  labs(title = "Histogram of Average Professor Beauty Rating")
```

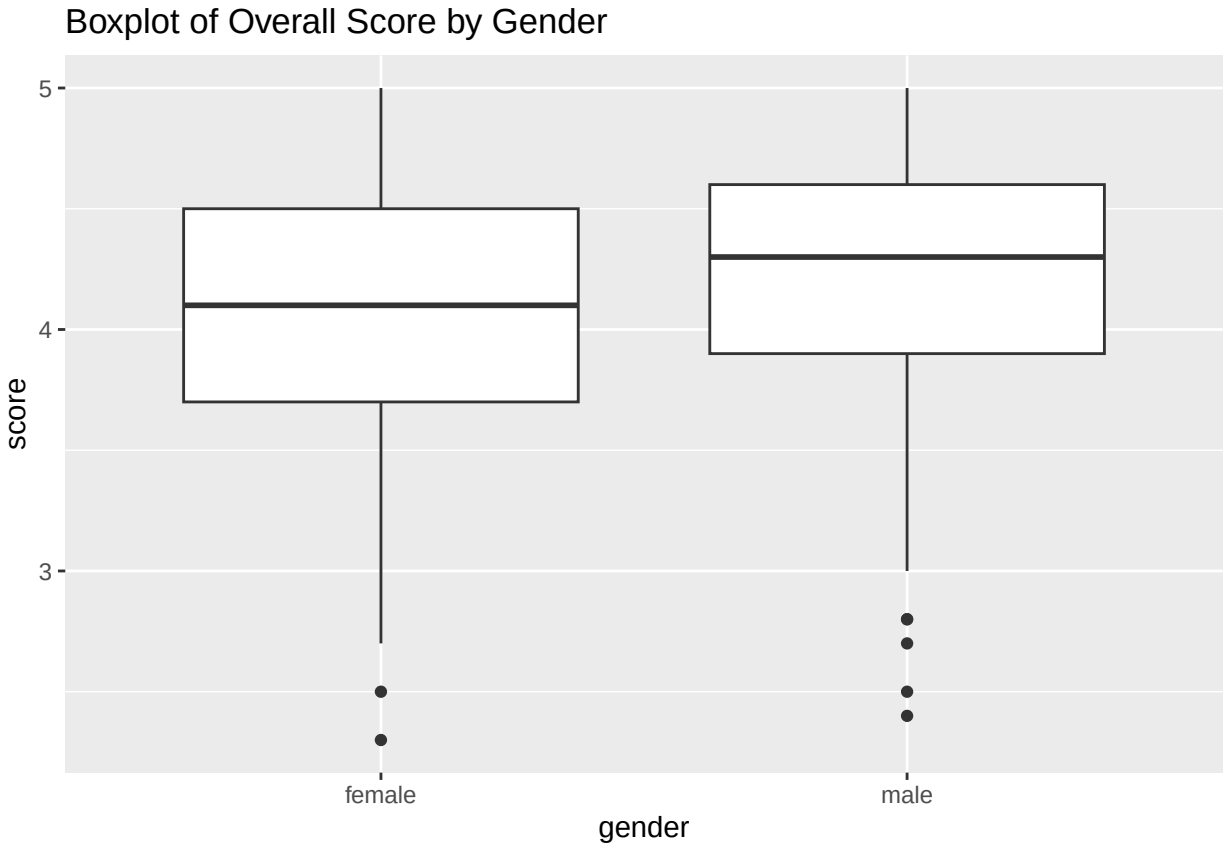



```
# Scatter plot between overall score and class percentage evaluation
data %>%
  ggplot(aes(x = cls_perc_eval, y = score)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatter Plot of Overall Score vs. Class Percentage Evaluation")
```

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




```r
# Side-by-side boxplots of overall score by gender
data %>%
  ggplot(aes(x = gender, y = score)) +
  geom_boxplot() +
  labs(title = "Boxplot of Overall Score by Gender")
```


Boxplot of Overall Score by Gender
5
4
3
score
female
male
gender