MECH476: Engineering Data Analysis in R

Chapter 7 Homework: Multivariate Exploratory Data Analysis

Flynn Nyman

09 October, 2024

Load packages

Chapter 7 Homework

In Chapter 5, we briefly explored data on the salaries of engineering graduates from the National Science Foundation 2017 National Survey of College Graduates from a univariate perspective. Now, let's explore the relationships between multiple variables.

When a question asks you to make a plot, remember to set a theme, title, subtitle, labels, colors, etc. It is up to you how to personalize your plots, but put in some effort, and make the plotting approach consistent throughout the document. For example, you could use the same theme for all plots. I also like to use the subtitle as a place for the main summary for the viewer.

Question 1: Data wrangling

Within a pipeline, import the data from the .csv file, convert all column names to lowercase text (either "manually" with `dplyr::rename()`, or use `clean_names()` from the `janitor` package), convert `gender` from "numeric" to "factor", and drop any and all observations with `salary` recorded as 0. Assign this to a dataframe object with a meaningful name.

How many observations have a 0 (zero) value for salary? Note: The last question asked you to remove these observations from the resultant data frame.

```
## [1] 15
```

What are the levels in `gender`? (Ignore the fact that the observations refer to "biological sex", not "gender". *Gender* is now recognized as a fluid term with more than two options; *biological sex* - what was assigned at birth - is binary term).

```
## [1] M F
## Levels: F M
```

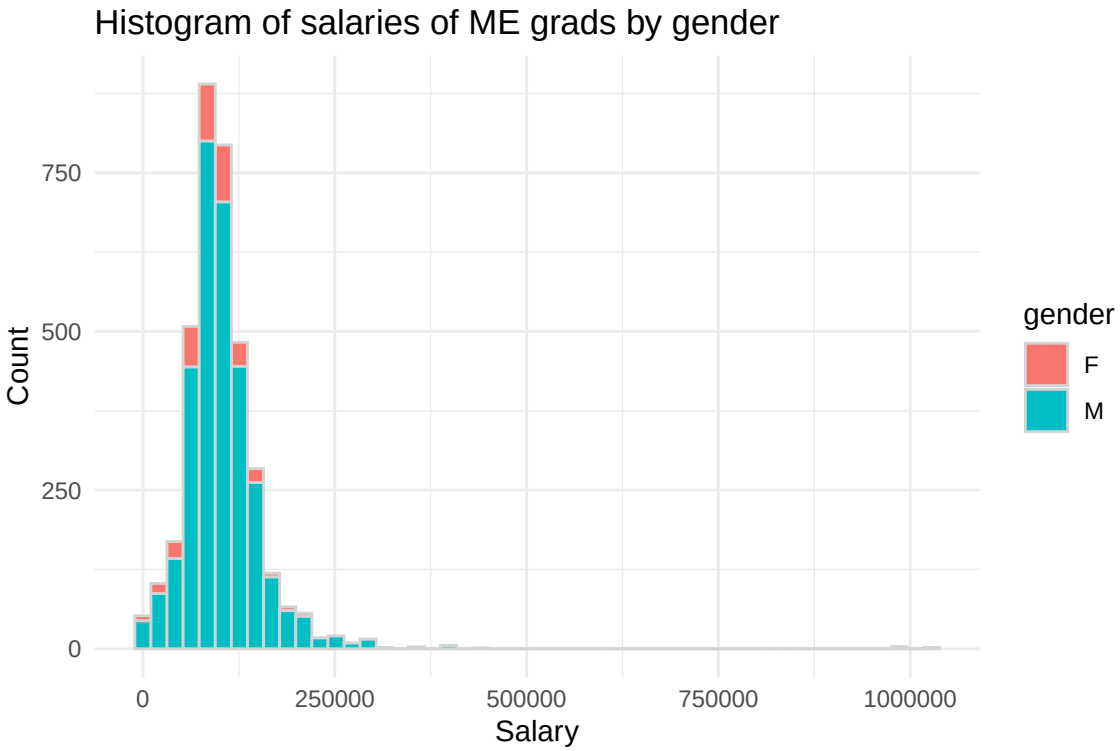
Question 2: Univariate EDA

Using what you learned in Chapter 5, generate basic plots and/or descriptive statistics to explore `age`, `gender`, and `salary`. List whether each variable is continuous or categorical, and explain how and why you adjusted your EDA approach accordingly.

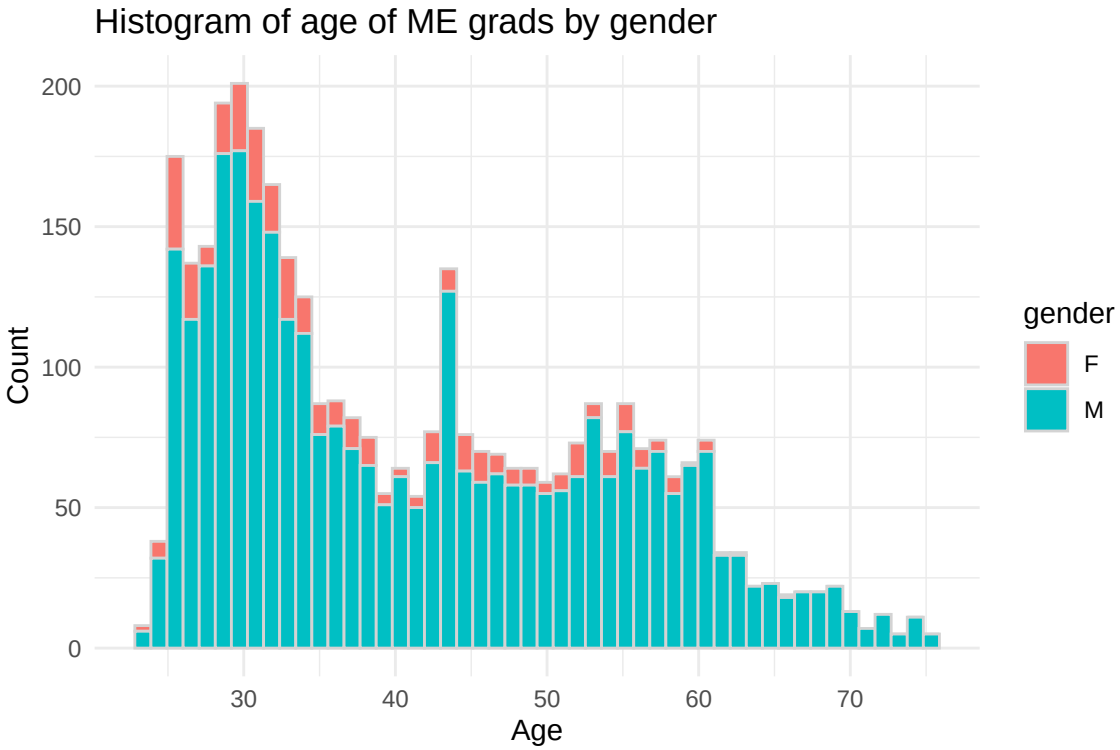
Age and salary are both continuous variables as there are many different observations in the data set and no limit for how many options they've can take on. Gender on the other hand is a caregorical variable as plot 2 shows that it only takes on two options in this dataset.

Question 3: Multivariate histograms

Create a histogram of `salary`, faceted by `gender`. Add `bins = 50` and `color = "lightgrey"`.

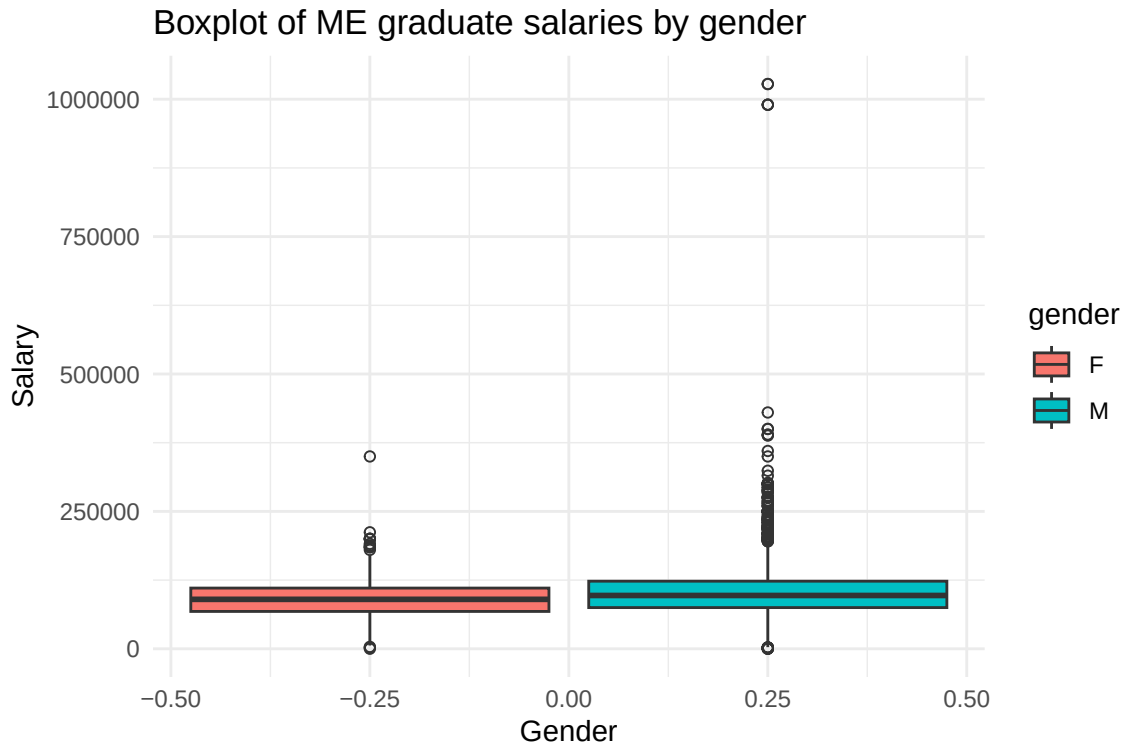


Create a histogram of `age`, faceted by `gender`. Add `bins = 50` and `color = "lightgrey"`.

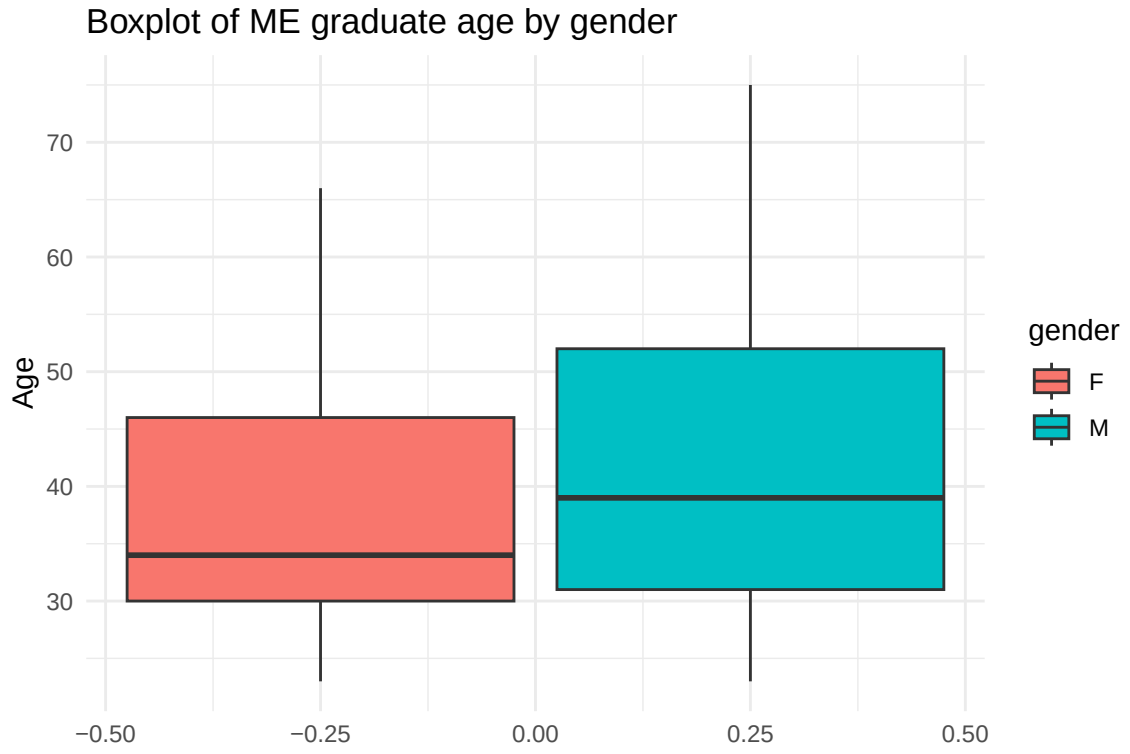


Question 4: Multivariate boxplots

Create a boxplot of `salary`, faceted by `gender`. Use `oulier.shope = 1` to better visualize the outliers.

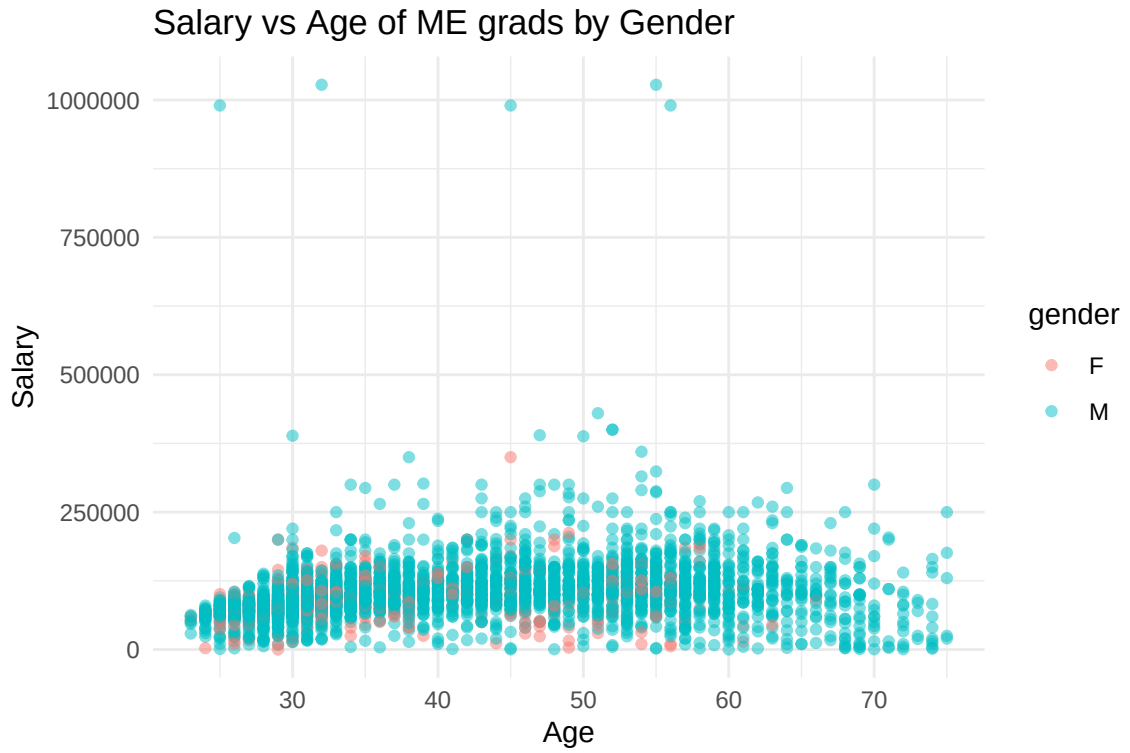


Create a boxplot of `age`, faceted by `gender`.



Question 5: Scatterplot and correlation

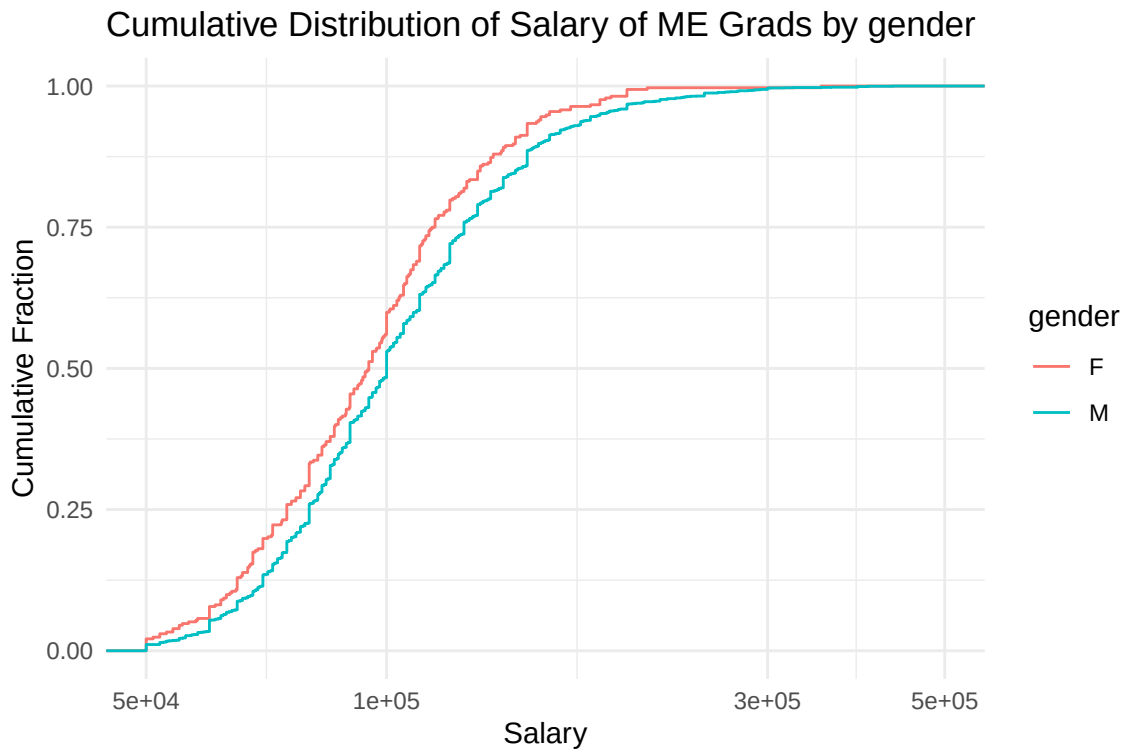
Create a scatterplot of `age` (x-axis) and `salary`, differentiating by `gender`.



Bonus point: Is there a correlation between an engineer's salary and age? What is the estimated Pearson correlation coefficient r? Run a formal test.

Question 6: Cumulative distribution function

Plot the cumulative distribution function of `salary` by `gender`. Adjust the x-axis with `scale_x_log10(limits = c(5e4, 5e5))` to zoom in a bit. What do you notice about the salaries for men and women? Hint: Remember there are greater differences the farther up you go on a log scale axis.



Question 7: Quantiles

Calculate the quantiles of `salary` by `gender`. You can either subset the data with `dplyr::filter()` and dataframe assignment, or you can group by, summarize by quantile, and ungroup.

Bonus point: Assign the output to a dataframe, and use inline code to call individual values when answering the following questions. Do not let R use scientific notation in the text output; check the knitted document.

```
##       0%      25%      50%      75%     100%
##      105    75000    97000   123000  1027653
```

```
##      0%     25%     50%     75%    100%
##     140   68000   90000  110513  350000
```

What is the difference in salary between men and women at the median?

- Median salary for women is $90,000
- Median salary for men is $97,000
- The difference at the median is $7,000

At the top percentile (maximum)?

- Maximum salary for women is $350,000
- Maximum salary for men is $1,027,653
- The difference at the maximum is $677,653

Do you think there is a salary difference by gender across the pay scale? What other information would you need to test your hypothesis?

Yes there is a salary difference by gender across the pay scale. To further test my hypothesis it would be helpful to have an equal amount of salary data on both female and male engineers.

Question 8: Hypothetical analysis

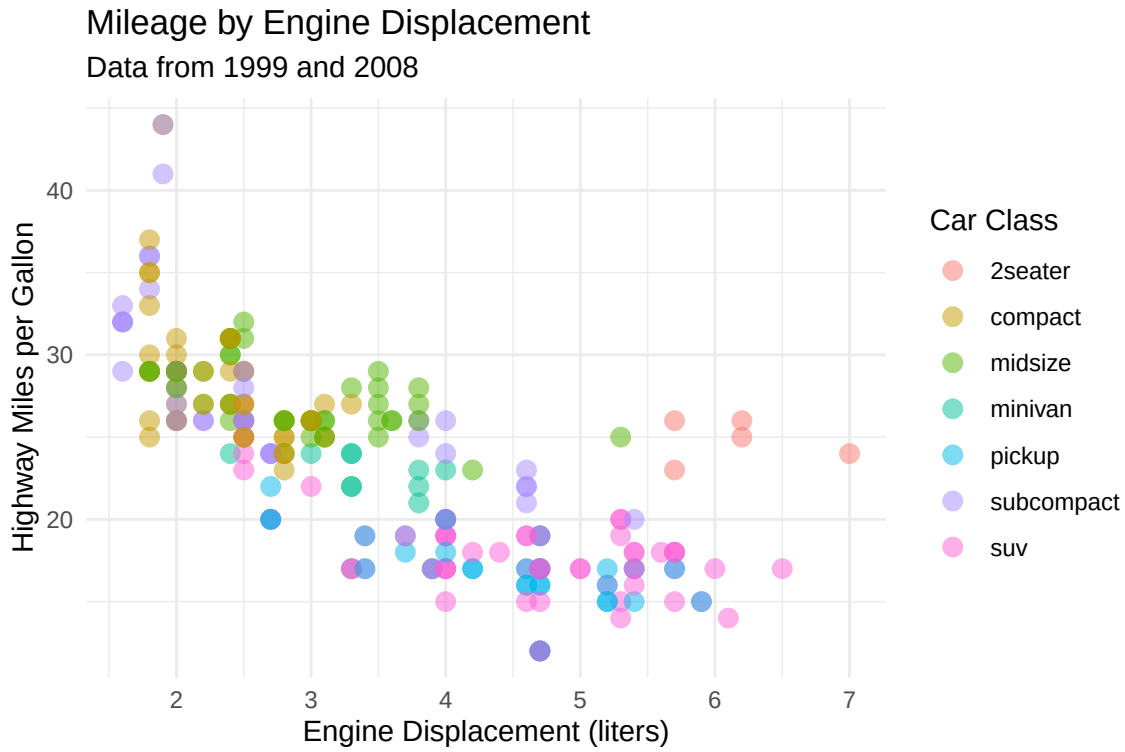
Think about what other variables you would like to include in an hypothetical analysis. From your perspective, what are the most important individual, family, and workforce factors related to salary—beyond gender and age?

Other factors that are important are: time at company, time as an engineer, whether they have any advanced degrees, and which engineers have certificates.

Question 9: Recreate plot

Recreate this plot with the `mpg` dataset. Remember to use `?mpg` for information on the dataset and the variables. How would you describe the correlation between the independent variable and dependent variable? Do you see any patterns when considering the third variable?

(View R Markdown PDF for image)



Appendix

```
# set global options for figures, code, warnings, and messages
knitr::opts_chunk$set(fig.width=6, fig.height=4, fig.path="../figs/",
                      echo=FALSE, warning=FALSE, message=FALSE)
# load packages for current session
library(tidyverse)
# import and tidy salary data
salaries <- read_csv("./ME_salaries.csv") %>%
  rename(salary = SALARY,
         age = AGE,
         gender = GENDER) %>%
  mutate(gender = factor(gender)) %>%
  subset(salary != 0)
# number of observations with salary as 0
nrow(read_csv("./ME_salaries.csv")) - nrow(salaries)
# number of factor levels
unique(salaries$gender)
# univariate eda
ggplot(salaries, aes(x = salary)) +
  geom_histogram(bins = 50, color = "black") +
  labs(x = "Salary", y = "Count", title = "Histogram of salaries of ME grads") +
  theme_minimal()

ggplot(salaries, aes(x = gender)) +
  geom_bar(fill = "darkgreen") +
  labs(x = "Gender", y = "Count", title = "Number of Engineers by Gender")
  theme_minimal()

ggplot(salaries, aes(x = age)) +
  geom_histogram(bins = 50, fill = "darkgreen", color = "black") +
  labs(x = "Age", y = "Count", title = "Age of ME Engineers") +
  theme_minimal()
# histogram of salaries split by gender
ggplot(salaries, aes(x = salary, fill = gender)) +
  geom_histogram(bins = 50, color = "lightgrey") +
  labs(x = "Salary", y = "Count", title = "Histogram of salaries of ME grads by gender") +
  theme_minimal()
# histogram of ages split by gender
ggplot(salaries, aes(x = age, fill = gender)) +
  geom_histogram(bins = 50, color = "lightgrey") +
  labs(x = "Age", y = "Count", title = "Histogram of age of ME grads by gender") +
  theme_minimal()
# boxplots of salary data by gender
ggplot(salaries, aes(y = salary, fill = gender)) +
  geom_boxplot(width = 1, outlier.shape = 1) +
  labs(y = "Salary", x = "Gender", title = "Boxplot of ME graduate salaries by gender") +
  theme_minimal()
# boxplots of age data by gender
ggplot(salaries, aes(y = age, fill = gender)) +
  geom_boxplot(width = 1, outlier.shape = 1) +
  labs(y = "Age", x = NULL, title = "Boxplot of ME graduate age by gender") +
  theme_minimal()
```

```r
# scatterplot of salary across age by gender
ggplot(salaries, aes(x = age, y = salary, color = gender)) +
  geom_point(alpha = 0.5) +
  labs(x = "Age", y = "Salary", title = "Salary vs Age of ME grads by Gender") +
  theme_minimal()
# correlation test

# plot cdf of salary by gender
ggplot(salaries, aes(x = salary, color = gender)) +
  stat_ecdf(geom = "step") +
  scale_x_log10(limits = c(5e4, 5e5)) +
  labs(x = "Salary", y = "Cumulative Fraction", title = "Cumulative Distribution of Salary of ME Grads
  theme_minimal()

# calculate quantiles of salary by gender
salaries_tibble <- as_tibble(salaries)
male_salaries <- filter(salaries_tibble, gender == "M")
male_quantiles <- quantile(male_salaries$salary, seq(0, 1, 0.25)) %>%
  print()

female_salaries <- filter(salaries_tibble, gender == "F")
female_quantiles <- quantile(female_salaries$salary, seq(0, 1, 0.25)) %>%
  print()
# call mpg pdf - you need to recreate it
#knitr::include_graphics("./mpg-ch7-plot.pdf")
mpg <- as.data.frame(mpg)
ggplot(mpg, aes(x = displ, y = hwy, color = class)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Engine Displacement (liters)", y = "Highway Miles per Gallon", title = "Mileage by Engine D
  theme_minimal()
```