MECH476: Engineering Data Analysis in R

Chapter 7 Homework: Multivariate Exploratory Data Analysis

Michael Thill

11 December, 2025

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).

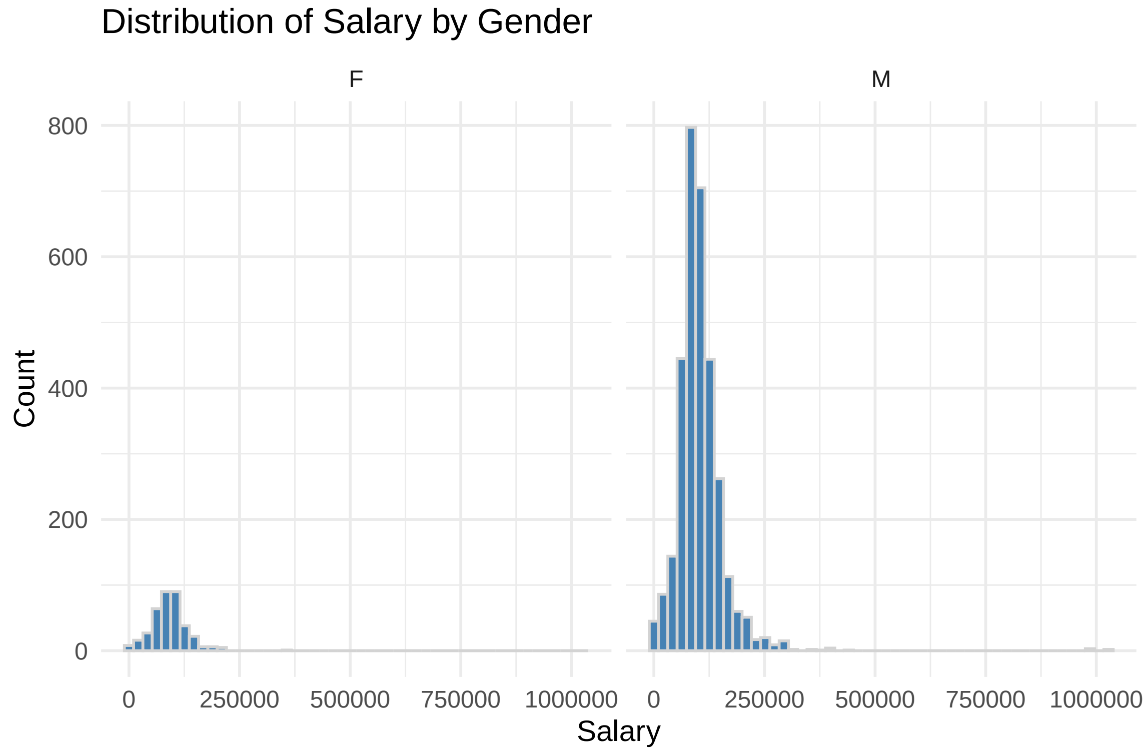
```
## [1] "F" "M"
```

Question 2: Univariate EDA

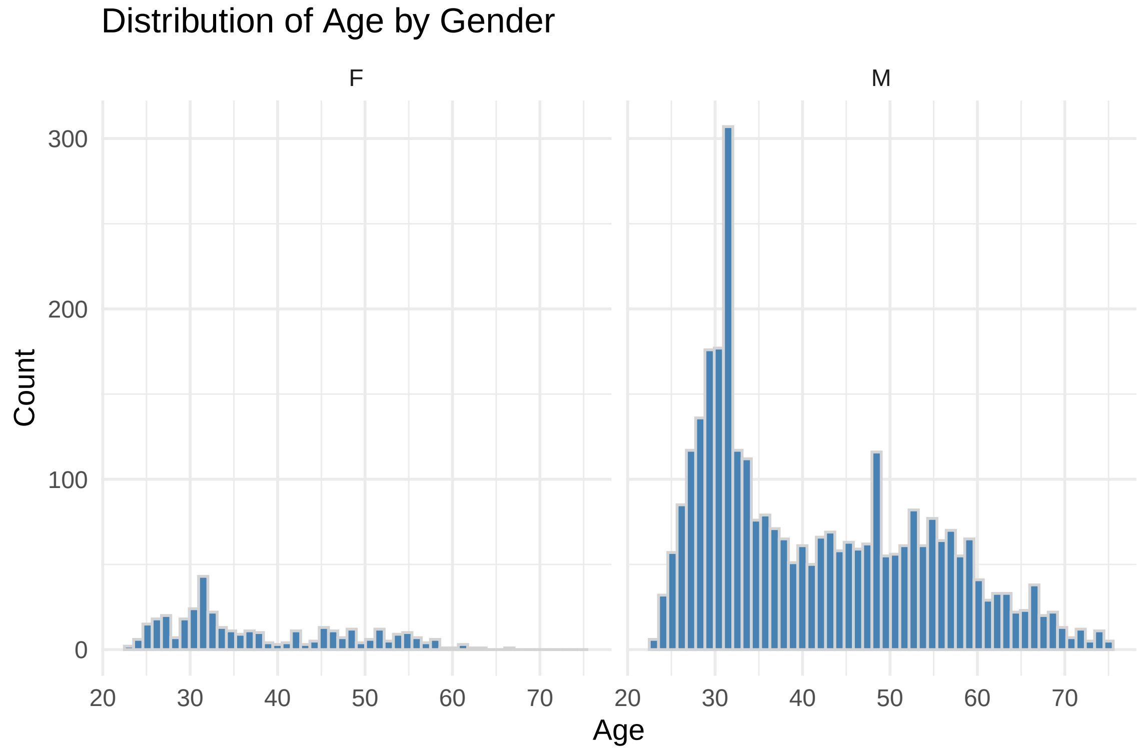
Using what you learned in Chapter 5, generate basic plots and/or descriptive statistics to explore `age`, `gender`, and `salary`. List whether each variable is continuous or categorical, and explain how and why you adjusted your EDA approach accordingly.

Question 3: Multivariate histograms

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

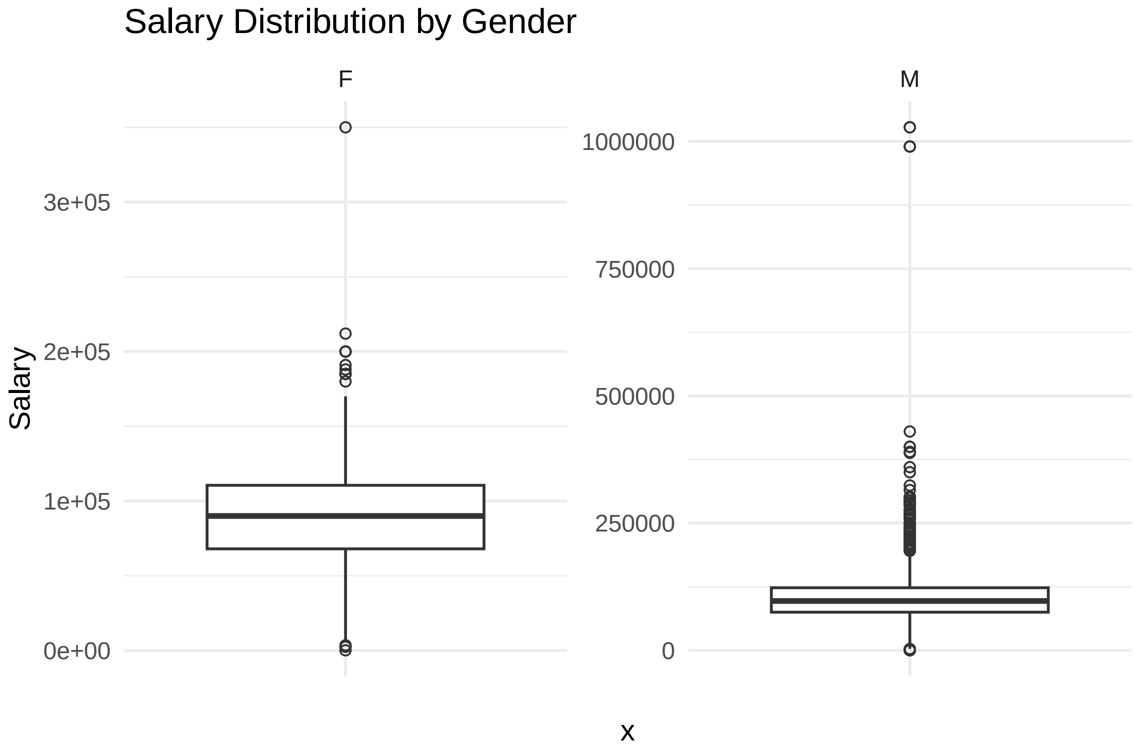


Create a histogram of `age`, faceted by `gender`. Add `bins = 50` and `color = "lightgrey"`.

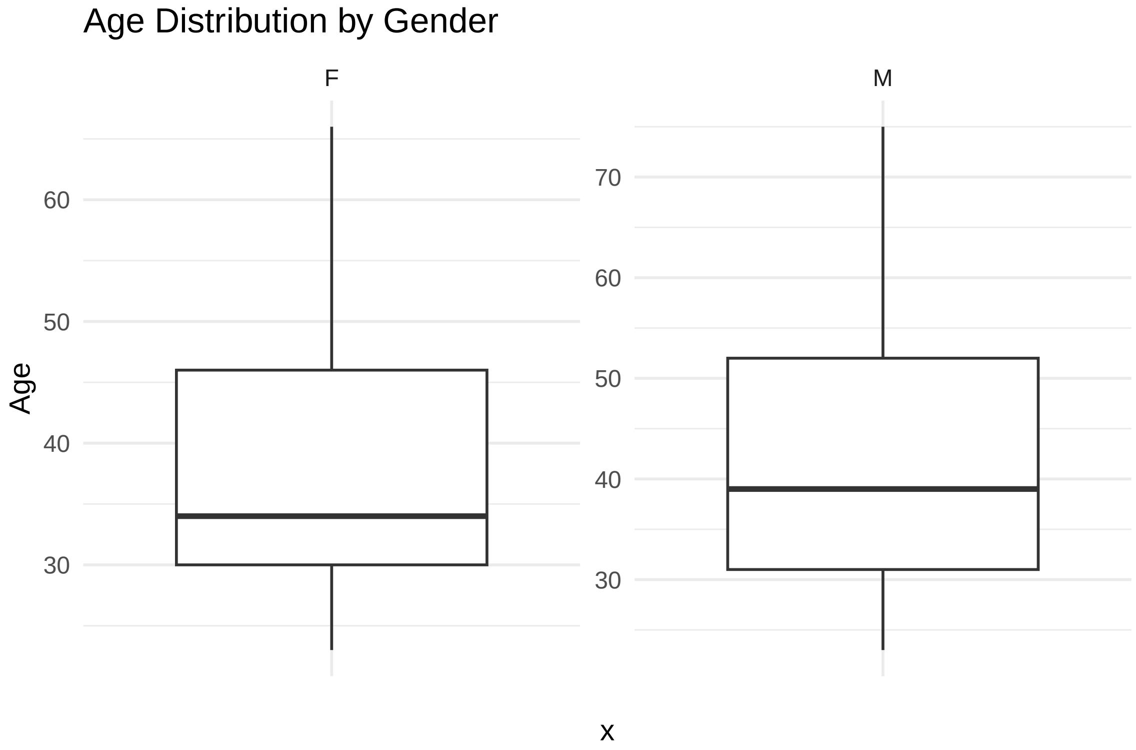


Question 4: Multivariate boxplots

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

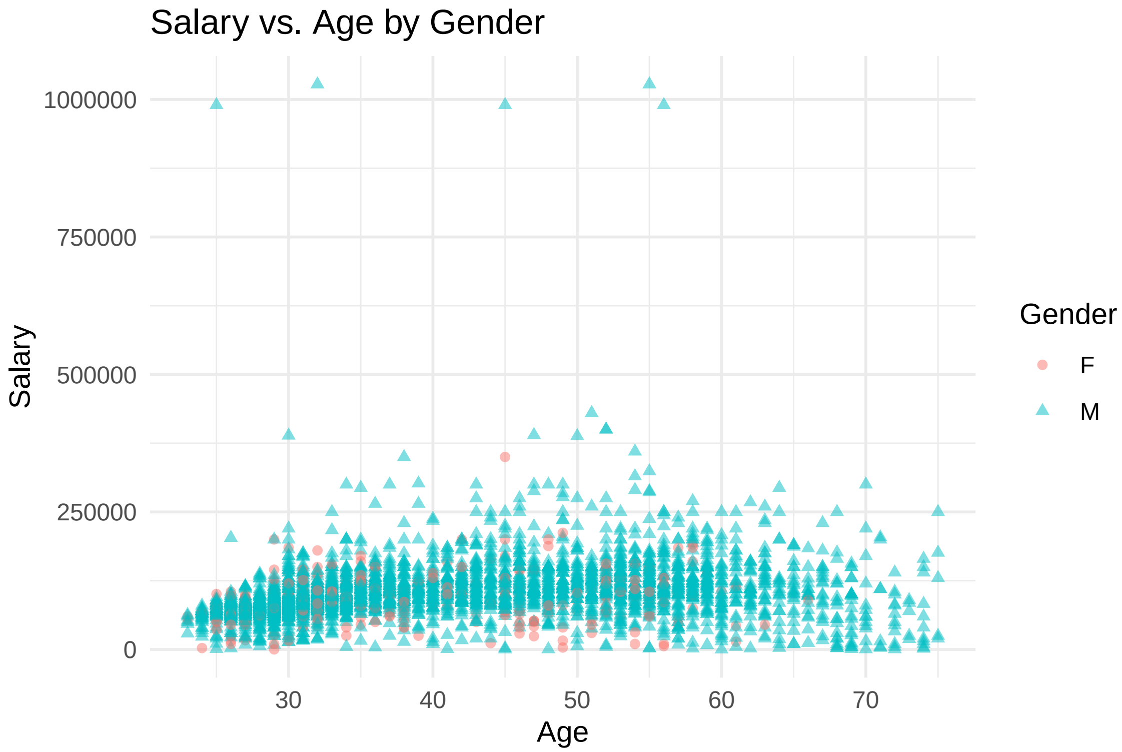


Create a boxplot of `age`, faceted by `gender`.



Question 5: Scatterplot and correlation

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



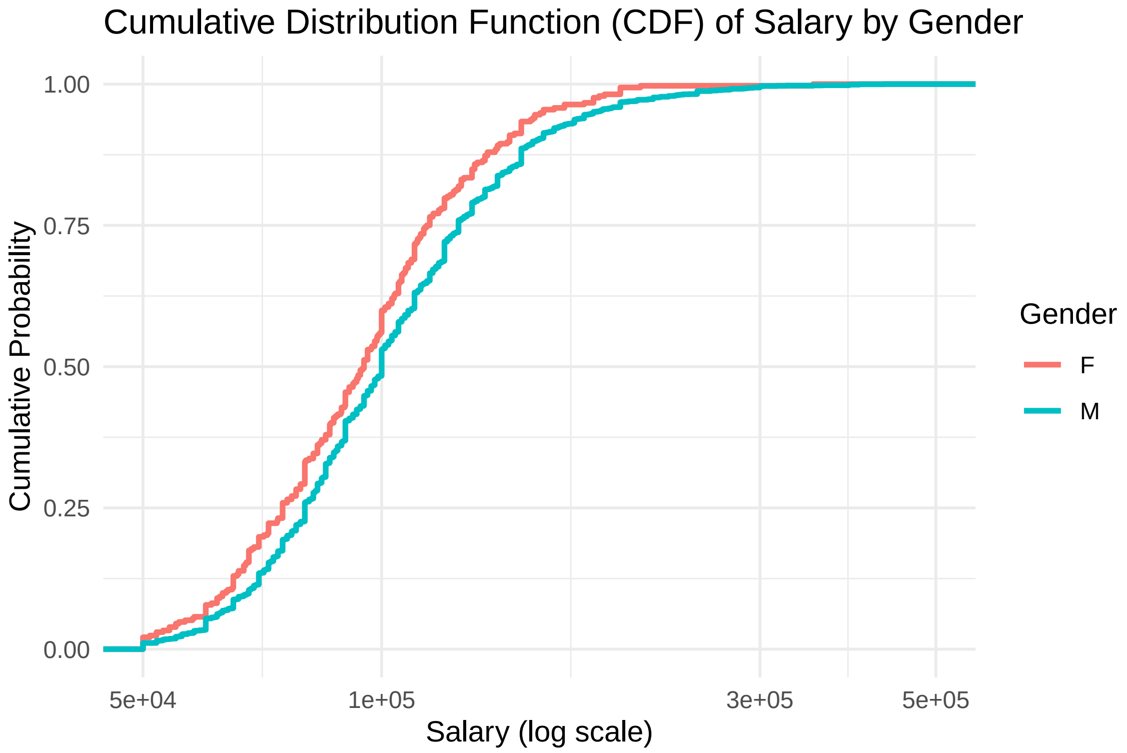
Bonus point: Is there a correlation between an engineer's salary and age? What is the estimated Pearson correlation coefficient r? Run a formal test.

```
## [1] 0.2077451
```

```
##
##  Pearson's product-moment correlation
##
## data:  salary_data_raw$age and salary_data_raw$salary
## t = 12.741, df = 3599, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1762777 0.2387883
## sample estimates:
##       cor
## 0.2077451
```

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.

```
## # A tibble: 2 x 6
##   gender    q0   q25   q50    q75    q100
##   <fct>  <dbl> <dbl> <dbl>  <dbl>   <dbl>
## 1 F        140 68000 90000 110513  350000
## 2 M        105 75000 97000 123000 1027653
```

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

- Median salary for women is 90000
- Median salary for men is 97000
- The difference at the median is 7000

At the top percentile (maximum)?

- Maximum salary for women is 350000
- Maximum salary for men is 1027653
- The difference at the maximum is 677653

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

Examining the 25th percentile, median, and 75th percentile indicates a consistent salary gap favoring males. To further validate this hypothesis, the cumulative distribution function (CDF) should be analyzed. The CDF shows that at every percentile, males earn more than females, confirming that the wage gap persists across the entire income distribution.

Question 8: Hypothetical analysis

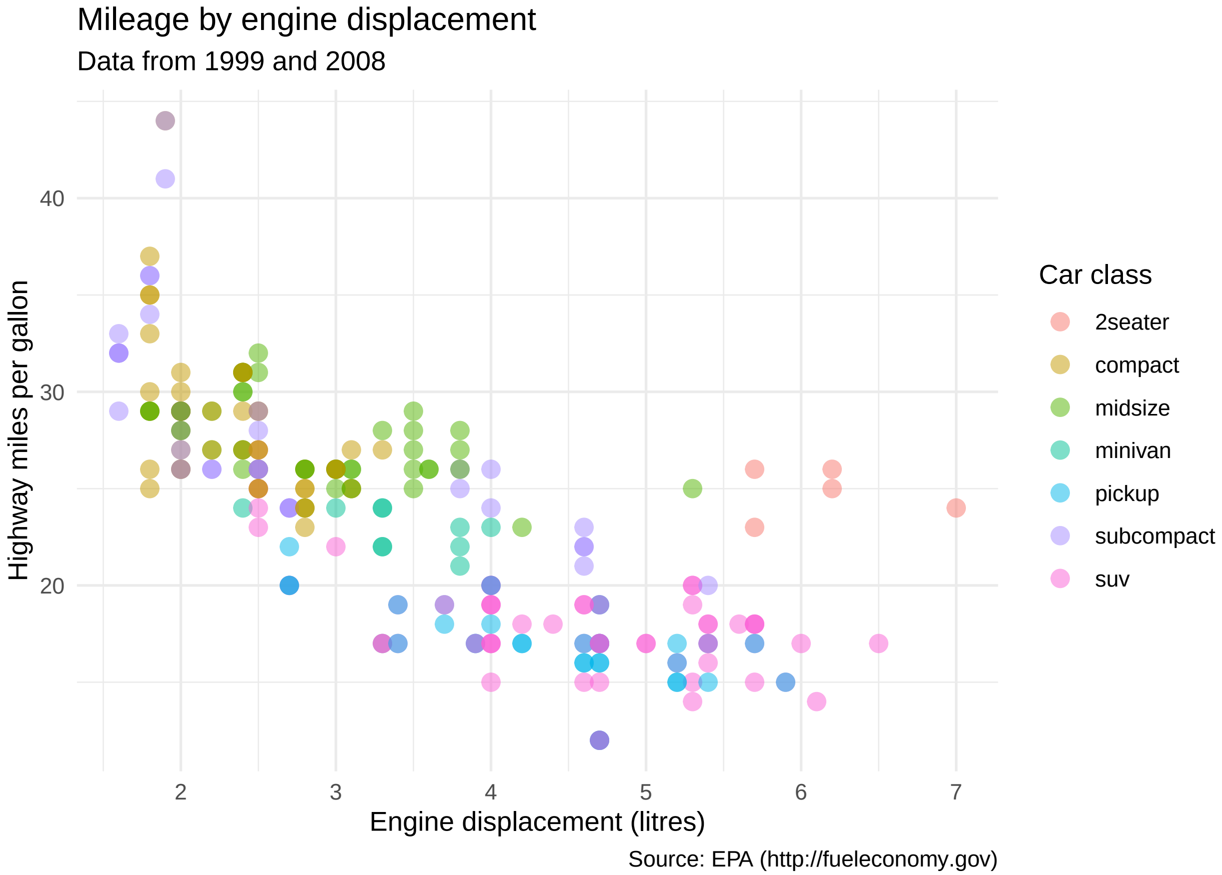
Think about what other variables you would like to include in an hypothetical analysis. From your perspective, what are the most important individual, family, and workforce factors related to salary—beyond gender and age?

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?

Beyond gender and age, education level plays a significant role in salary outcomes. In the engineering industry, individuals with a master's degree or PhD generally earn higher salaries than those with only a bachelor's degree. Years of experience is also a key determinant of income, as compensation typically increases with greater tenure and accumulated industry experience.

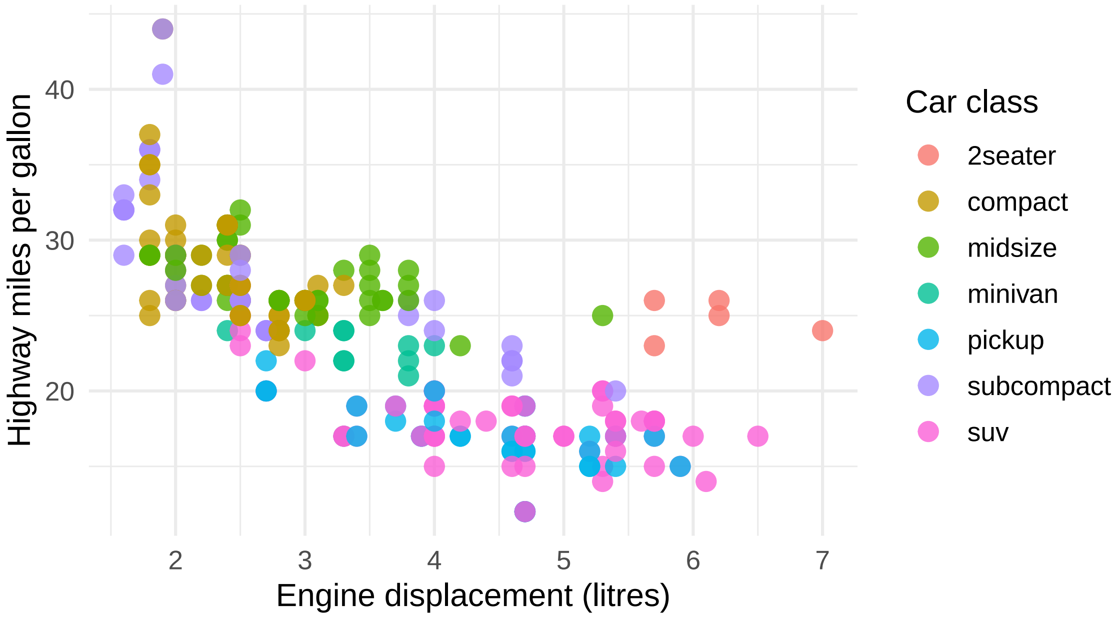
(View R Markdown PDF for image)



Mileage by engine displacement

Data from 1999 and 2008

Source: EPA (http://fueleconomy.gov)



Mileage by Engine Displacement

Data from 1999 and 2008

Source: EPA (http://fueleconomy.gov)

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)
library(readr)
library(dplyr)
library(stringr)
library(ggplot2)
library(knitr)
library(lubridate)
library(janitor)

# import and tidy salary data
salary_data_raw <- readr::read_csv(file = "../data/ME_salaries.csv") %>%
  clean_names() %>%
  mutate(gender = as.factor(gender)) %>%
  filter(salary != 0)

# number of observations with salary as 0
Salary_zero = readr::read_csv(file = "../data/ME_salaries.csv") %>%
  clean_names() %>%
  filter(salary == 0) %>%
  nrow()
Salary_zero

# number of factor levels
levels(salary_data_raw$gender)
# univariate eda and histogram of age
age_history <- ggplot2::ggplot(salary_data_raw, aes(x = age)) +
  geom_histogram(
    bins = 30,
    fill = "steelblue",
    color = "white",
    alpha = 0.7
  ) +
  labs(
    title = "Distribution of Employee Age",
    x = "Age",
    y = "Count"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )
age_history
salary_histo <- ggplot2::ggplot(salary_data_raw, aes(x = salary)) +
  geom_histogram(
    bins = 30,
    fill = "steelblue",
    color = "white",
    alpha = 0.7
  ) +
  labs(
    title = "Salary Distribution",
    x = "Salary",
    y = "Count"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )
salary_histo
gender_box <- ggplot(salary_data_raw, aes(x = gender, y = salary, fill = gender)) +
  geom_boxplot() +
  labs(
    title = "Salary by Gender",
    x = "Gender",
    y = "Salary"
  ) +
  scale_fill_manual(values = c("male" = "#0072B2", "female" = "#D55E00")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "none"
  )
gender_box

# histogram of salaries split by gender
ggplot(salary_data_raw, aes(x = salary)) +
  geom_histogram(bins = 50, color = "lightgrey", fill = "steelblue") +
  facet_wrap(~ gender) +
  labs(
    title = "Distribution of Salary by Gender",
    x = "Salary",
    y = "Count"
  ) +
  theme_minimal()
# histogram of ages split by gender
ggplot(salary_data_raw, aes(x = age)) +
  geom_histogram(bins = 50, color = "lightgrey", fill = "steelblue") +
  facet_wrap(~ gender) +
  labs(
    title = "Distribution of Age by Gender",
    x = "Age",
    y = "Count"
  ) +
  theme_minimal()
```

```r
# boxplots of salary data by gender
ggplot(salary_data_raw, aes(x = "", y = salary)) +
  geom_boxplot(outlier.shape = 1) +
  facet_wrap(~ gender, scales = "free_y") +
  labs(title = "Salary Distribution by Gender",
       y = "Salary") +
  theme_minimal()
# boxplots of age data by gender
ggplot(salary_data_raw, aes(x = "", y = age)) +
  geom_boxplot(outlier.shape = 1) +
  facet_wrap(~ gender, scales = "free_y") +
  labs(title = "Age Distribution by Gender",
       y = "Age") +
  theme_minimal()

# scatterplot of salary across age by gender
ggplot(salary_data_raw, aes(x = age, y = salary, color = gender, shape = gender)) +
  geom_point(alpha = 0.5) +
  labs(
    title = "Salary vs. Age by Gender",
    x = "Age",
    y = "Salary",
    color = "Gender",
    shape = "Gender"
  ) +
  theme_minimal()
# correlation test
correlation_r <- cor(salary_data_raw$age, salary_data_raw$salary, method = "pearson")
correlation_test <- cor.test(salary_data_raw$age, salary_data_raw$salary, method = "pearson")
correlation_r
correlation_test
# plot cdf of salary by gender
ggplot(salary_data_raw, aes(x = salary, color = gender)) +
  stat_ecdf(size = 1) +
  scale_x_log10(limits = c(5e4, 5e5)) +
  labs(
    title = "Cumulative Distribution Function (CDF) of Salary by Gender",
    x = "Salary (log scale)",
    y = "Cumulative Probability",
    color = "Gender"
  ) +
  theme_minimal()
# calculate quantiles of salary by gender
salary_quantiles <- salary_data_raw %>%
  group_by(gender) %>%
  summarize(
    q0 = quantile(salary, 0),
    q25 = quantile(salary, 0.25),
    q50 = quantile(salary, 0.5),
    q75 = quantile(salary, 0.75),
    q100 = quantile(salary, 1)
  ) %>%
  ungroup()
```

```
salary_quantiles
# call mpg pdf - you need to recreate it
knitr::include_graphics("./mpg-ch7-plot.pdf")
ggplot(mpg, aes(x = displ, y = hwy, color = class)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(
    title = "Mileage by Engine Displacement",
    subtitle = "Data from 1999 and 2008",
    x = "Engine displacement (litres)",
    y = "Highway miles per gallon",
    color = "Car class",
    caption = "Source: EPA (http://fueleconomy.gov)"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold"),
    plot.subtitle = element_text(size = 10),
    plot.caption = element_text(hjust = 0, size = 8),
    legend.position = "right"
  )
```