MECH476: Engineering Data Analysis in R

Chapter 7 Homework: Multivariate Exploratory Data Analysis

Brad Portouw

20 October, 2022

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.

What are the levels in `gender`? (Ignore the fact that the values for this variable refer to “biological sex”, not “gender”, although the categories often overlap among people).

```
## [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`.

Create a histogram of `age`, faceted by `gender`. Add `bins = 50`.

Question 4: Multivariate boxplots

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

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.

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

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

At the top percentile (maximum)?

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

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

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?

(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
ME_salaries <- read.csv("../../Data/ME_salaries.csv")%>%
  dplyr::rename(salary = SALARY, age=AGE, gender = GENDER) %>%
  dplyr::filter(salary != 0 & !is.na(salary))

ME_salaries$gender <- as.factor(ME_salaries$gender)
#check to see how many zero salaries were removed:

# 15 zero salary rows were removed.

# number of observations with salary as 0
ME_sal_zeroes <- read.csv("../../Data/ME_salaries.csv")%>%
  dplyr::filter(SALARY == 0)%>%
  nrow()
# there were 15 rows removed for having 0 salaries.
# number of factor levels
levels(ME_salaries$gender)
#the levels associated with gender here are either Female (F) or Male(M), not considering other possibl
# univariate eda

# age of Mechanical Engineers in the data frame. Excluding data with no salary listed (0).
ggplot(data = ME_salaries, aes(x = age))+
  geom_bar( fill= "darkorchid")+
  ggtitle("Age of Mechanical Engineers")+
  xlab("Age")+
  ylab("Count")
# including a bar chart for age for more information:
ggplot(data = ME_salaries, aes(y = age))+
  geom_boxplot()+
  ylab("Age")+
  ylim(20,80)

# histogram of salaries split by gender

# histogram of ages split by gender

# boxplots of salary data by gender

# boxplots of age data by gender

# scatterplot of salary across age by gender

# correlation test

# plot cdf of salary by gender
# calculate quantiles of salary by gender

# call mpg pdf - you need to recreate it
#knitr::include_graphics("../figs/mpg-ch7-plot.pdf")
```