MECH476: Engineering Data Analysis in R

Chapter 7 Homework: Multivariate Exploratory Data Analysis

Ethan Rutledge

08 December, 2023

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] 0
```

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).

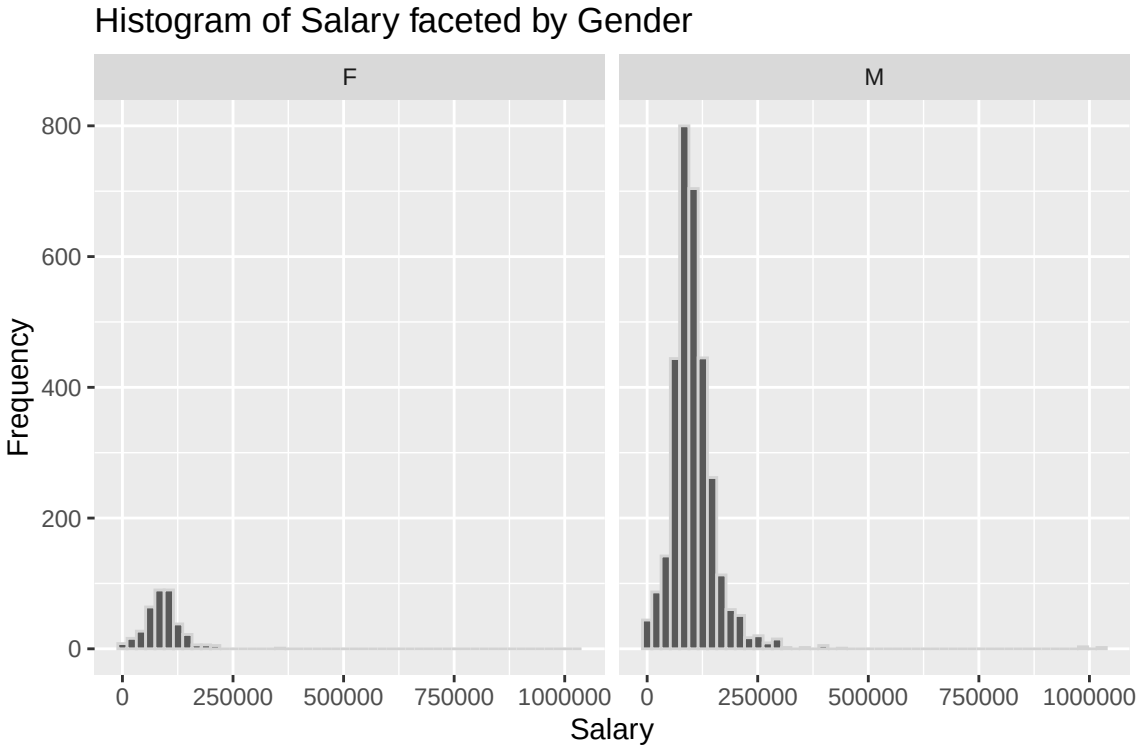
```
## [1] "F" "M"
```

Question 2: Univariate EDA

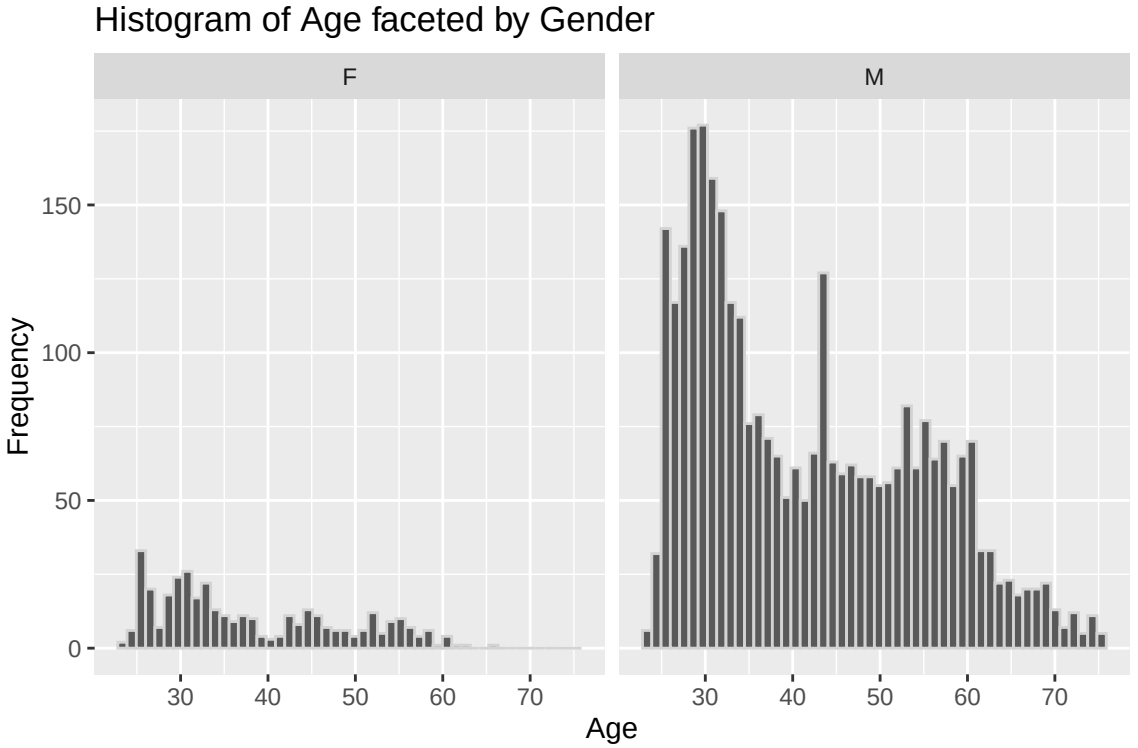
Using what you learned in Chapter 5, generate basic plots and/or descriptive statistics to explore `age`, `gender`, and `salary`. List whether each variable is continuous or categorical, and explain how and why you adjusted your EDA approach accordingly.

Question 3: Multivariate histograms

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

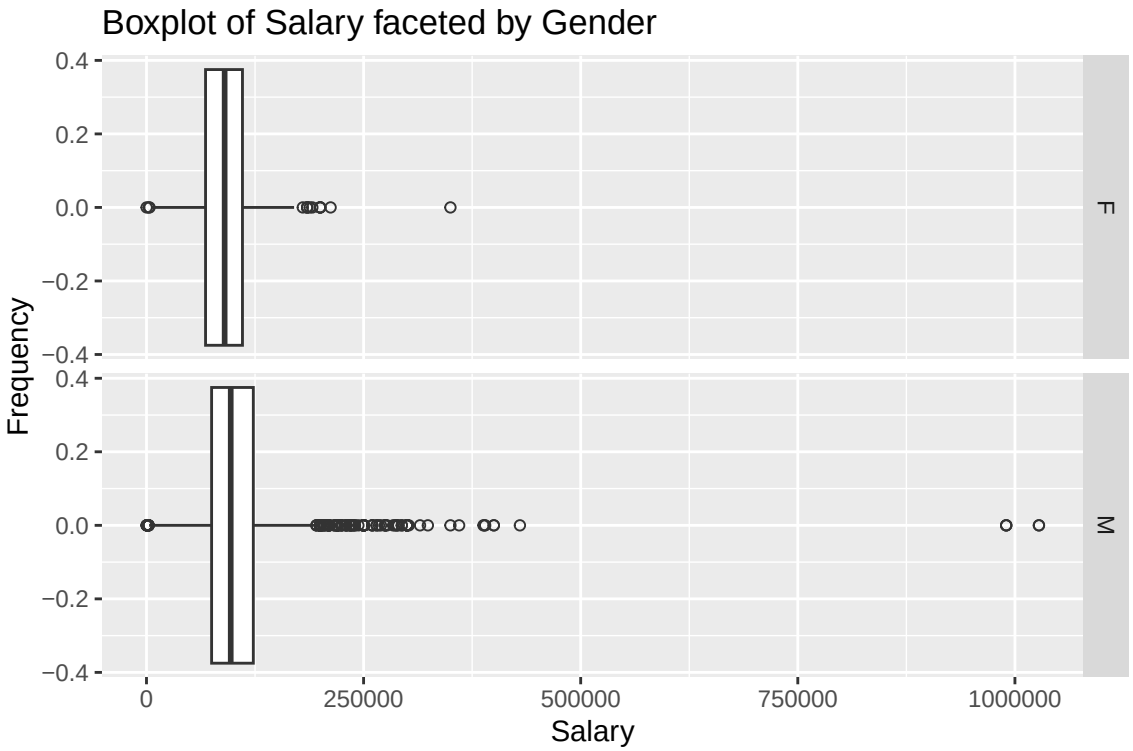


Create a histogram of `age`, faceted by `gender`. Add `bins = 50` and `color = "lightgrey"`.

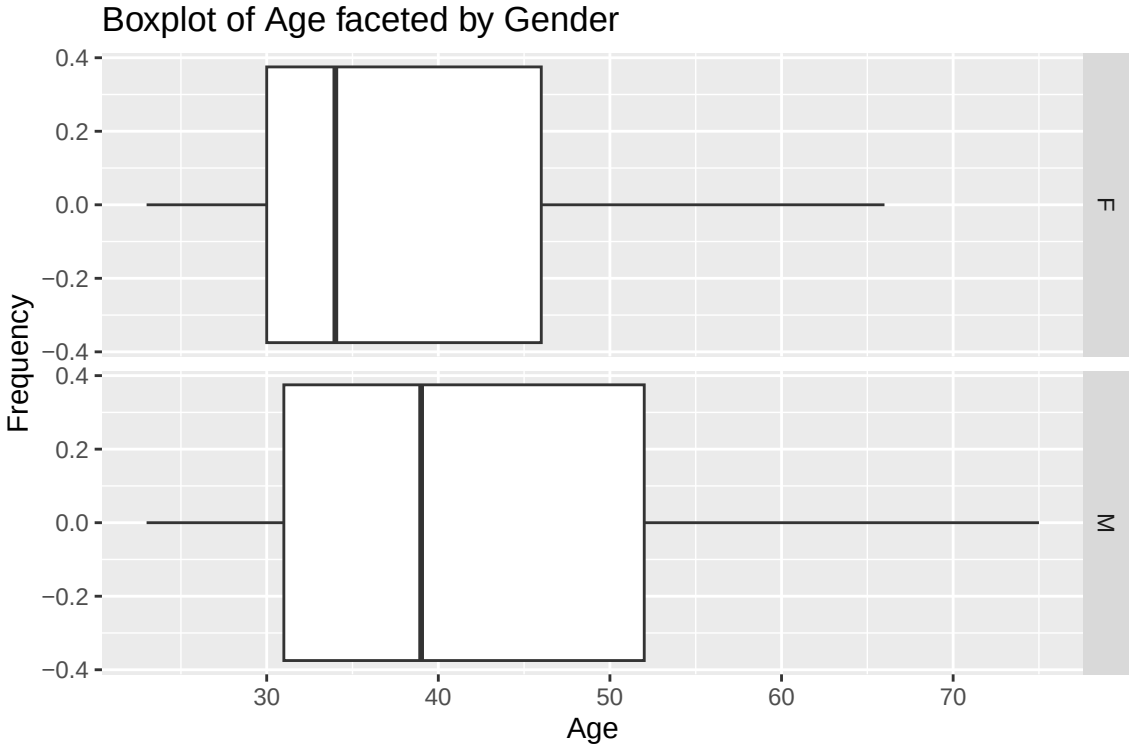


Question 4: Multivariate boxplots

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

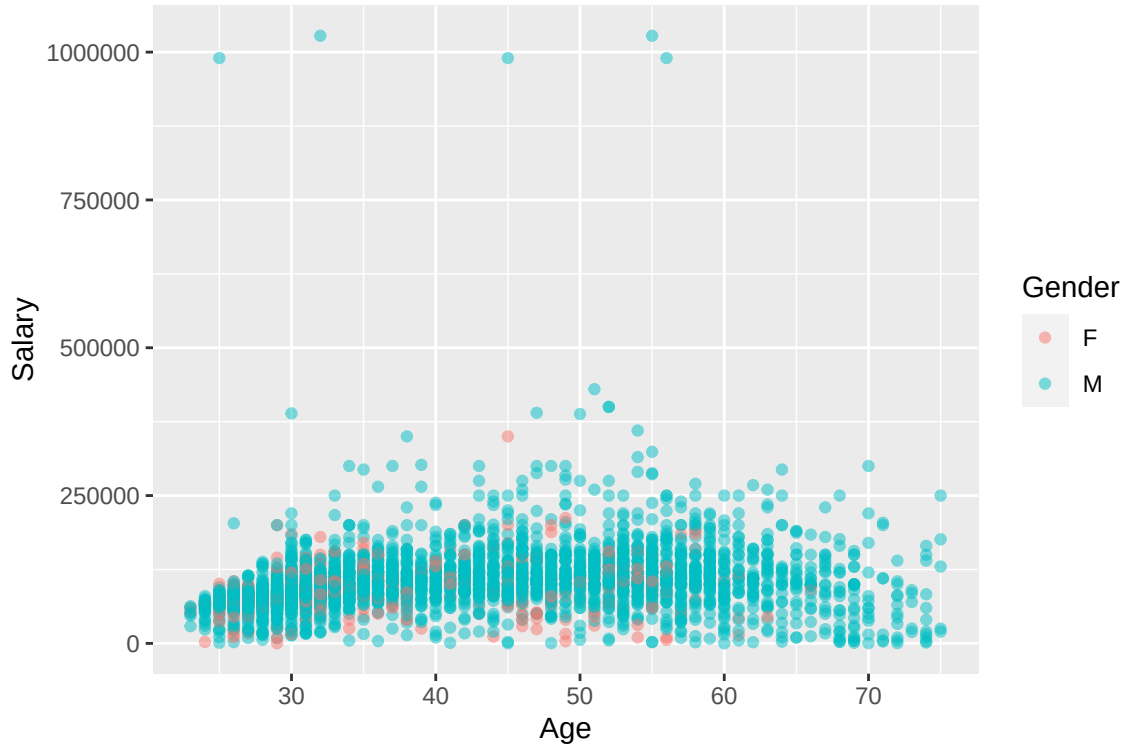


Create a boxplot of `age`, faceted by `gender`.



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.

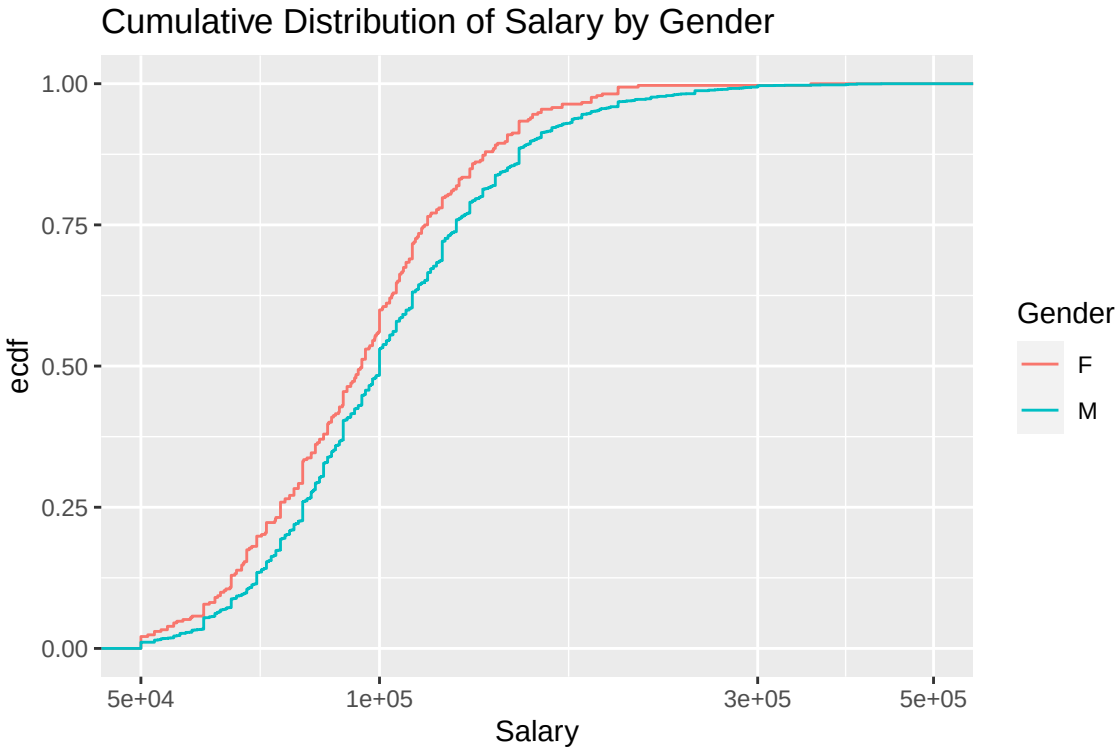
```
## [1] 0.2077451
```

Answer: The correlation is 0.208 which is very low

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.

Answer: The female salary clearly lags behind the male salary, especially at the high range.



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   min    Q1 median     Q3     max
##   <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 – $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?

#Answer: Yes, based off the data presented so far I believe there is a salary difference by gender across the pay scale. Other valuable information could be sample size, regions included, experience in the field, if maternity leave is considered or not.

Question 8: Hypothetical analysis

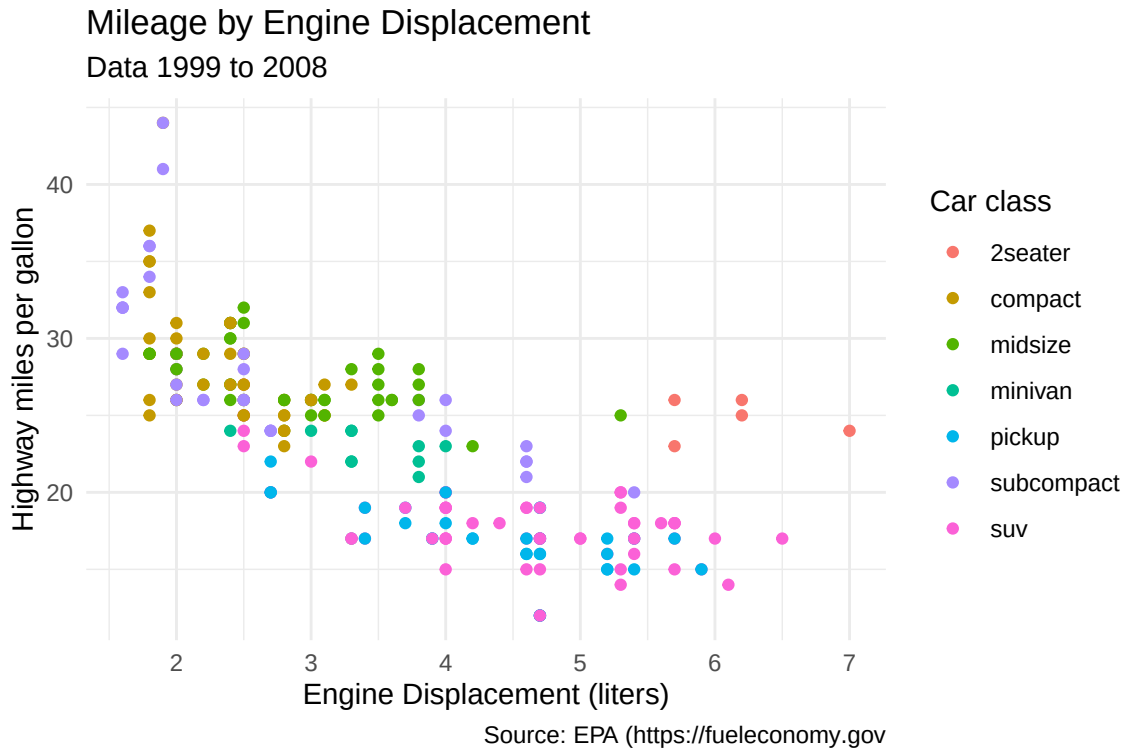
Think about what other variables you would like to include in an hypothetical analysis. From your perspective, what are the most important individual, family, and workforce factors related to salary—beyond gender and age?

#Answer: Some other factors that would be important are level of education, geographical location (cost of living), and field of engineering.

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)
library(janitor)
library(ggplot2)
# import and tidy salary data
salaries <- read_csv("../data/ME_salaries.csv") %>%
  clean_names() %>%
  mutate(gender = as.factor(gender)) %>%
  filter(salary != 0)

# number of observations with salary as 0
sum(salaries$salary == 0)

# number of factor levels
levels(salaries$gender)

# univariate eda
ggplot(salaries) + geom_histogram(aes(x=age))+labs(x="Age",y="Frequency", title = "Histogram of Ages")

ggplot(salaries) + geom_bar(aes(x=gender))+labs(x="Gender",y="Frequency", title = "Bar Graph of Gender

ggplot(salaries) + geom_histogram(aes(x=salary))+labs(x="Salary",y="Frequency", title = "Histogram of Sa
# histogram of salaries split by gender
ggplot(salaries) + geom_histogram(aes(x=salary), bins = 50, color = "lightgrey") + facet_wrap(vars(gende

# histogram of ages split by gender
ggplot(salaries) + geom_histogram(aes(x=age), bins = 50, color = "lightgrey") + facet_wrap(vars(gender)

# boxplots of salary data by gender
ggplot(salaries) + geom_boxplot(aes(x=salary), outlier.shape = 1) + facet_grid(vars(gender)) + labs(x =

# boxplots of age data by gender
ggplot(salaries) + geom_boxplot(aes(x=age), outlier.shape = 1) + facet_grid(vars(gender)) + labs(x = "Ag

# scatterplot of salary across age by gender
ggplot(salaries) + geom_point(aes(x=age, y = salary, color = gender), alpha = 0.5) + labs(x = "Age", y =

# correlation test
cor(salaries$age, salaries$salary)

# plot cdf of salary by gender
ggplot(salaries)+geom_step(aes(x=salary, color=gender), stat = "ecdf")+labs(x="Salary", color = "Gender
# calculate quantiles of salary by gender
quantiles <- salaries %>%
  group_by(gender) %>%
  summarize(min = min(salary),Q1 = quantile(salary, .25),median = quantile(salary, .5), Q3 = quantile(sa
  ungroup()
```

```
quantiles

# call mpg pdf - you need to recreate it
 ggplot(mpg) + geom_point(aes(x=displ, y=hwy, color=class)) + labs(x= "Engine Displacement (liters)", y=
```