

Homework 4

Due Tuesday, September 22, 2026 at 11:59 p.m.

[Your Name Here]

Table of contents

Problem 1: EDA for Final Project (10pts)	1
Problem 2: The Principal Components of Coffee (55 points)	1

Problem 1: EDA for Final Project (10pts)

For this problem, turn in one plot using the supplied workforce-compensation archive for Professor McGovern’s final project. The plot can be any visual of your choice, but it must meet these requirements:

- At a minimum your graph must display three variables. (It can even display more, if you prefer.)
- Be sure that your plot is appropriately titled and labeled such that it is clear what you are plotting.
- Your graph should answer a question about your dataset.
- The title, caption, or surrounding text must identify the years, subset, units, and one important evidence limitation.

```
# PUT YOUR CODE AND PLOT HERE
```

After you’ve made your plot, interpret it in 2-4 sentences and discuss why it is informative for the executive-search client’s recruiting decision. Do not interpret record frequency as employer demand.

[WRITE YOUR ANSWER HERE]

The best way to complete this question is to take a few days to think about your dataset and make a few plots that you think are interesting, and then only submit your best plot for this question. You **should not** just try to make an arbitrary multivariate graph as quickly as possible to complete this problem, because if we think your graph/interpretation is not well-motivated given the goals of your project, we may deduct up to 7.5pts for this problem.

Problem 2: The Principal Components of Coffee (55 points)

For this problem you’ll work with a dataset curated via the #TidyTuesday project containing information from the Coffee Quality Database. The following code reads in the dataset and performs some initial pre-processing:

```
library(tidyverse)
coffee_ratings <- read_csv('https://raw.githubusercontent.com/rfordatascience/
tidytuesday/master/data/2020/2020-07-07/coffee_ratings.csv') %>%
  filter(total_cup_points > 0)
```

a. (10pts) For this part, perform the following tasks:

- Create a subset of the `coffee_ratings` dataset that only contains the 11 quantitative grade columns (starting with `aroma` and ending with `moisture`, see the README above for info about the columns in this dataset). Call this `coffee_quant`.

```
# PUT YOUR CODE HERE
```

- Run principal components analysis (PCA) on `coffee_quant`. Be sure that the columns of `coffee` are *centered* and *standardized* when you run PCA.

```
# PUT YOUR CODE HERE
```

Then, answer the following questions: What proportion of the total variation in the data is accounted by the first principal component specifically? By the second principal component specifically? By both the first and second principal components together?

[PUT YOUR ANSWER HERE]

- b. (10pts) We discussed using a scree plot (or *elbow plot*) to determine how many principal components should be used for visualization. For this part, create an elbow plot for these data, where the component numbers (1 through 11) are on the x-axis, and the proportion of total variation is on the y-axis.

```
# library(factoextra)
# PUT YOUR CODE AND PLOT HERE
```

After you've made your plot: Based on the plot, how many principal components do you think should be used for visualizations and analyses? Provide a 1-2 sentence explanation.

[PUT YOUR ANSWER HERE]

- c. (15pts) Use `ggplot()` to make the following scatterplots:

- A scatterplot with the first principal component on the x-axis and the second principal component on the y-axis, with points colored by `total_cup_points`. **Choose an appropriate color gradient scale for `total_cup_points` that is different from the default.**

```
# PUT YOUR CODE AND PLOT HERE
```

- A scatterplot with the first principal component on the x-axis and the third principal component on the y-axis, with points colored by `total_cup_points`. **Again, choose an appropriate color gradient scale for `total_cup_points` that is different from the default.**

```
# PUT YOUR CODE AND PLOT HERE
```

Then, for your each plot, discuss your main observations in 1-3 sentences. In particular: does `total_cup_points` appear to be associated with the any of the principal components you viewed in your plots?

[PUT YOUR ANSWER HERE]

- d. (20pts) As discussed, principal components are not interpretable by themselves, thereby limiting the observations you can make about the data. In this part we'll explore in what ways the **first and third principal components** are related to the original variables in the data.

For this part, complete the following two tasks:

- 1) Use `fviz_pca_biplot()` to make a biplot of the first and third principal components (change the `axes` input to be `c(1, 3)`). When you make this plot, unfortunately, it may be difficult to see some of the variable names on the arrows; if that is the case for you, put `repel = TRUE` within `fviz_pca_biplot()`, such that you can see all of the variable names. Additionally, color the points by the `total_cup_points` variable using the `col.ind` argument. You can use the the commented out code as a template.

After you've made the biplot, discuss in 1-4 sentences in what ways `total_cup_points` is associated with the original variables in the data.

```
# fviz_pca_biplot(PUT YOUR CODE HERE,
#                 # Plot PC1 and PC3
#                 axes = ,
#                 # Change the alpha for the observations -
#                 # which is represented by ind
#                 alpha.ind = ,
#                 # Modify the alpha for the variables (var):
#                 alpha.var = ,
#                 repel = ,
#                 # Set the color of the points to decades variable:
#                 col.ind = coffee_ratings$total_cup_points,
#                 # Modify the color of the variables
#                 col.var = ) +
#   scale_color_gradient(low = "darkblue", high = "darkorange") +
#   labs(color = "Total cup points") +
#   theme(legend.position = "bottom")
```

[PUT YOUR ANSWER HERE]

- 2) In Part C you should have used the `prcomp()` to create the principal components. After you've done that, you should be able to edit and uncomment the following line of code:

```
# coffee_pca$rotation
```

where `coffee_pca` is the object appropriately defined by `prcomp()`. You should be able to see an 11-by-11 matrix, where the rows correspond to the original variables and the columns correspond to the principal components. Each column represents the *linear combination* of the variables that each principal component represents. In particular: Each number tells you how the variable is associated with the principal component. If a number is positive, that means coffees with a higher value of that principal component also tend to have a higher value of that variable. Similarly, if a number is negative, that means coffees with a *lower* value of that principal component tend to have a higher value of that variable.

Now look back at the first plot you made in Part C, which visualizes the first and third principal components (colored by `total_cup_points`). Given these details about the rotation matrix, discuss in 1-4 sentences in what ways `total_cup_points` is associated with the original variables in the data. Be sure to explain how you used the rotation matrix output to arrive at your answer.

```
# PUT OPTIONAL CODE TO ANSWER QUESTION HERE
```

[PUT YOUR ANSWER HERE]