

Homework 3

Due Tuesday, September 15, 2026 at 11:59 p.m.

Your Name Here

Table of contents

Problem 1: Final Project Prep (40pts)	1
Problem 2: Olympic Regression (25pts)	2
Problem 3: Contour Plots and Heat Maps (35 points)	3

Problem 1: Final Project Prep (40pts)

The goal of this problem is to help you get started on Professor McGovern's Global Data & AI Workforce Compensation Analytics project. Use the supplied salary archive and the executive-search client described in the project brief.

- a. (10pts) To help structure your report, come up with three potential client questions for the compensation archive. Each should help the executive-search steering committee make a recruiting decision rather than merely request a descriptive chart. List the questions here:

[PUT YOUR CLIENT QUESTION 1 HERE]

[PUT YOUR CLIENT QUESTION 2 HERE]

[PUT YOUR CLIENT QUESTION 3 HERE]

- b. (20pts) Describe a plot you can make for each client question and write a few sentences justifying why the plot is appropriate. Name the variables, subset, units, and one important limitation or sample-size check. To make you think more creatively, **you must describe a different type of plot for each question** (e.g., you cannot propose a histogram for all three questions).

[PUT YOUR PLOT DESCRIPTION FOR CLIENT QUESTION 1 HERE]

[PUT YOUR PLOT DESCRIPTION FOR CLIENT QUESTION 2 HERE]

[PUT YOUR PLOT DESCRIPTION FOR CLIENT QUESTION 3 HERE]

- c. (10pts) Create one of the plots described in part b. You do not need to write a full interpretation yet. **Make sure it is labeled appropriately and does not imply that record counts measure employer demand.**

PUT YOUR CODE AND PLOT HERE

Problem 2: Olympic Regression (25pts)

For this problem you'll work with a dataset curated via the #TidyTuesday project containing information about Olympic athletes. The following code reads in the dataset and filters it to just include USA athletes, selecting a subset of the columns, and tidies up medal columns you will use in this problem:

```
library(tidyverse)
usa_olympics <- read_csv('https://raw.githubusercontent.com/rfordatascience/
tidytuesday/master/data/2021/2021-07-27/olympics.csv') %>%
  filter(team == "United States") %>%
  dplyr::select(-team, -noc) %>%
  mutate(won_medal = as.numeric(!is.na(medal)),
         medal = ifelse(is.na(medal), "None", medal))
```

The dataset contains information on the performance of USA Olympic athletes. Each row corresponds to a unique athlete-event combination for a particular year (there are 17,847 in total). This means that athletes are repeated throughout the dataset but we will ignore that detail for this problem. We have information about the athlete such as their height, weight, age, and sex. We also have info about the Olympics event they competed in, including the year, season, and the event outcome for the athlete in terms of their medal placement (if any). You will focus on the following variables for this problem:

- **year**: The year that the Olympics occurred.
 - **season**: Type of Olympics, Summer or Winter.
 - **height**: Athlete height in cm.
 - **sex**: Athlete's reported sex.
- a. (10pts) For this part, using the `usa_olympics` dataset, make a graph that does the following:
- Displays **year** on the x-axis and **height** on the y-axis. (Make sure your plot avoids overplotting...)
 - Displays the *linear trend* between **year** and **height**. Display a 99% confidence interval when doing this.
 - Be sure that your graph has appropriate labels.

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

Then answer the following the questions:

- Describe the relationship you observe between the **height** and **year** based on the displayed regression line.

[PUT YOUR ANSWER HERE]

- Do any of the assumptions we make for linear regression appear to be violated based on your plot? (You do NOT need to turn in a residual versus fit plot to answer this question.)

[PUT YOUR ANSWER HERE]

b. (5pts) For this part, make a graph that does the following:

- Displays `year` on the x-axis, `height` on the y-axis, and color by `sex`. (Again, avoid overplotting...)
- Displays the *linear trend* between `year` and `height` with separate lines by `sex` mapped via color. Display a 99% confidence interval when doing this.
- Be sure that your graph has appropriate labels.

For this part, you only have to turn in the graph.

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

- c. (5pts) Do you observe a similar or different relationship between `height` and `year` for your lines in part (b) compared to part (a)? State if they are similar or different, and explain in 1-2 sentences.

[PUT YOUR ANSWER HERE]

- d. (5pts): Are the intercepts meaningful in this context? Would you feel comfortable using this regression model to predict USA Olympic athlete height in the year 3005? Explain in a few sentences.

[PUT YOUR ANSWER HERE]

Problem 3: Contour Plots and Heat Maps (35 points)

In this problem, we will use a dataset on students' academic performance, found here:

```
student_data <- read_csv("https://raw.githubusercontent.com/ryurko/DataViz-36613-Fall23/main/data/students.csv")
```

Details about the dataset are found here. However, the main things you need to know about this dataset are:

- Students' `Grade` is classified as Low (L), Medium (M), or High (H).
- Covariates: There are 15 variables on student characteristics and behaviors, 4 of which are quantitative.

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

- Create a scatterplot of `RaisedHands` and `VisitedResources` (make sure to set `alpha < 1`) with contour lines added using `geom_density2d()`.

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

- We discussed how contour lines use two bandwidths; `geom_density2d()` estimates these bandwidths by default. Now, copy-and-paste your above code, but make the bandwidth smaller by setting `h = c(10, 10)` within `geom_density2d()`.

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

- Compare and contrast the two plots in 1-3 sentences.

[PUT YOUR ANSWER HERE]

- b. (15pts) Similar to Part A, again make a scatterplot of `RaisedHands` and `VisitedResources` with contour lines, but with the following changes:
- Make the bandwidth of the contour lines larger by setting `h = c(80, 80)` within `geom_density2d()`
 - Set the color of the points according to `Grade` and the shape of the points according to `Gender`.

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

After you've made your plot, answer the following two questions:

- How many modes are there in the scatterplot? In your answer, also characterize/describe each mode in terms of `RaisedHands` and `VisitedResources`.

[PUT YOUR ANSWER HERE]

- In 1-3 sentences, characterize/describe each mode in terms of `Grade` and `Gender`.

[PUT YOUR ANSWER HERE]

- c. (10pts) For this part, you'll have to make two different heat maps (and all you'll need to do is turn in the two graphs). Please do the following:
- Make a heat map of `RaisedHands` and `VisitedResources` with points added but no contour lines (using the default bandwidth) with `stat_density2d`. Furthermore, change the default colors using `scale_fill_gradient()` and setting the `low` and `high` arguments in that function. Be sure that you use `geom_point()` *after* you use `stat_density2d` (otherwise, you won't be able to see the points).

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

- Make a hexagonal heatmap of `RaisedHands` and `VisitedResources`, but this time use `scale_fill_gradient2()`. Within `scale_fill_gradient2()`, specify a "medium count" color using the `mid` argument (similar to the `low` and `high` arguments). Within `scale_fill_gradient2()`, there is an argument called `midpoint` that specifies what a "medium density" is. The default is 0, which doesn't make sense for densities, because 0 is the lowest

possible value for densities. So, experiment and set `midpoint` equal to a non-zero number that you think makes sense given the range of counts observed in your hexagonal bins.

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

Hint: For the `midpoint` argument, your graph should be a gradient of three different colors that you've specified. If this isn't the case, you may have specified `midpoint` poorly.

You should end up with a hexagonal heat map that has a mix of three colors throughout it. **Honestly, you should be careful when making visualizations based on three colors in the display.** For the purpose of a density heat map, you should only focus on low to high changes in color. But now you know how to modify the color gradient to include a middle point cutoff for future reference.