431 Lab 02

Deadline: Tuesday 2023-09-19 at Noon

431 Staff

2023-08-09

Table of contents

0.1 Getting Help

Lab 02 has 8 questions, all of which you need to complete by the deadline specified on the Course Calendar.

You are welcome to discuss Lab 02 with Professor Love, the teaching assistants or your colleagues, but your answer must be prepared by you alone. Don't be afraid to ask questions, using any of the methods described on our Contact Us page.

0.2 Learning Objectives

1. Be comfortable using R to import and manage data.
2. Become familiar with the `tidyverse` packages and their functions
3. Be able to build and interpret a figure using R.
4. Use figures to contextualize specific data points of interest.

0.3 Getting Started with Lab 02

To start, create a directory on your computer for `lab02`. We suggest this be a directory you control, called something like `lab02`, and we recommend you create it as a subdirectory of a `2023-431` directory on your machine.

- Into that `lab02` directory, you will download the Quarto Template for Lab 02, which is called `431-lab02-template.qmd`, as described below. After you download the template file to your directory, you will want to rename it to substitute in your actual name in the file name, perhaps calling it `yourname-lab02.qmd`.
- You will then download the data file `lab02-data.csv`, also described below, into the same `lab02` directory, or perhaps into a subdirectory called `data` within your `lab02` directory.

After you've downloaded the relevant files, open RStudio, and use the **File … New Project … Existing Directory** menu to create an R Project in your `lab02` directory in which you will do all of your work for Lab 02.

0.4 The Quarto Template for Lab 02

In this Lab, you will analyze some data, and prepare a report in the form of an HTML file, using Quarto. We have provided you with a very useful Quarto document template for this assignment called `431-lab02-template.qmd` that you should use to complete your work.

- The template is part of the Data and Code repository for the course. Follow the instructions posted there to download all of the files you'll need in a ZIP file, including the template to an easy place to find them on your computer (we suggest a `431-data` subdirectory in your `2023-431` directory.) Then copy the template into your directory for Lab 02, specifically, that you created earlier.

Build your entire response as a Quarto file using the template provided. Then use the Knit button in RStudio to create the resulting HTML document. Be sure to remove all of the instructions included in the original template before submitting your work, and also be sure to review the HTML result to ensure that it looks clean and clear, that the labels on your plots and other output are easy to read, and that it doesn't retain any unnecessary warning messages or other material that distracts from your work. Be sure also that your name is listed as the author of the work.

0.5 Submitting Lab 02

Submit **both** your revised Quarto file **and** the HTML output file to the Lab 02 section in the Assignments folder in Canvas by the deadline specified in the Course Calendar. We will need both the Quarto and HTML file submitted before we can grade your work.

Again, we encourage you in the strongest possible terms to **ask questions**, using any of the approaches described on our Contact Us page.

0.6 Our Labs Page

Visit our 431-Labs page for more information, including:

- how to get help with the Lab,
- our late policy,
- details on answer sketches and grading rubrics
- how your lab will be graded, and
- how we handle requests to regrade a Lab.

0.7 Obtaining the Data for Lab 02

For this Lab, we have prepared a CSV (comma-separated version) file which contains a small subset of data from the 2022 County Health Rankings. The County Health Rankings data provide some useful information on how the health of US residents is affected by where they live, and we will use data from these Rankings several times this semester.

You can find the CSV file for Lab 02, called `lab02_data.csv`, in our class Data and Code repository. This file contains 2,564 rows (each row is a county) and 7 variables including:

Variable Name	Description
`state`	Two-letter postal abbreviation of the state name
`county`	Name of the county
`metro`	Whether or not the county is in a metropolitan area
`life_exp`	Average years a county resident can expect to live
`covid_mort`	county's COVID-19 age-adjusted mortality per 100K in 2020
`year`	2022 for all rows, represents CHR 2022 data
`fipscode`	Code describing the state and county (default sorting order)

0.7.1 A few notes

1. A county is listed here in the "metro" rather than "non-metro" category within the `metro` variable if fewer than 25% of county residents live in a rural area. These estimates are based on Census 2010 Population Estimates.
2. The `covid_mort` value for a county includes all deaths occurring between January 1, 2020 and December 31, 2020 due to COVID-19, per 100,000 population (age-adjusted). It was estimated using the 2020 National Center for Health Statistics - Mortality files.
3. The `life_exp` value for a county is the average number of years a county resident can expect to live. It was estimated using the 2018-2020 National Center for Health Statistics - Mortality files.
4. Data were not available in all counties measured by County Health Rankings, but our sample of 2,564 counties each have complete data on all variables.

0.8 Getting the Data into RStudio

If you've stored the `lab02_data.csv` file in your R Project directory for Lab 02, you can then read the data into R and create an object called `lab2data` containing the information with the following command, which is also part of the `431-lab02-template.qmd` file.

```
lab2data <- read_csv("lab02_data.csv")
```

If you've instead stored the `lab02_data.csv` file in a sub-directory called `data` within your R Project directory for Lab 02, you can accomplish the same task by modifying the command to read:

```
lab2data <- read_csv("data/lab02_data.csv")
```

Note that there is also an approach which pulls the raw data directly from Github, as demonstrated below, but we don't recommend this approach for this Lab.

```
lab2data <- read_csv(
 "https://github.com/THOMASELOVE/431-data/blob/main/data-and-code/lab02-data.csv"
 )
```

Running any of these commands in R should lead to the appearance of a `lab2data` object in your R session. Don't run more than one command - just the one is what you need.

1 Question 1 (5 points)

Write a piece of R code that filters the observations (counties) in the data set to only the following midwest states: Ohio (OH), Indiana (IN), Illinois (IL), Michigan (MI), and Wisconsin (WI). Specifically, take our `lab2data` and create `midwest_data` which contains counties in these states. Hint: the pipe `|>` and `filter` function should be a large part of your code. **The rest of the assignment will use this smaller set of counties.**

2 Question 2 (5 points)

Write a piece of R code that counts the number of observations (counties) in the midwestern states data that you created in Question 1, within each of the five states in which we are interested. Hint: The `count` function and the pipe `|>` should be a big part of your code.

3 Question 3 (5 points)

Use the `filter()` and `select()` functions in R to obtain a result which specifies the `life_exp`, `covid_mort` and `metro` status of Cuyahoga County in the state of Ohio.

4 Question 4 (8 points)

Use the tools we've been learning in the `ggplot2` package to build a histogram of the `life_exp` results across all of the midwest counties represented in the data subset you created in Question 1. Create appropriate (that is to say, meaningful) titles for each axis and for the graph as a whole (don't simply use the default choices.) We encourage you to use something you find more attractive than the default gray fill in the histogram.

5 Question 5 (5 points)

Based on your results in Questions 3 and 4, write a short description (2-3 sentences) of Cuyahoga County's position relative to the full distribution of counties shown in your Question 4 plot, in terms of `life_exp`.

6 Question 6 (7 points)

Use `ggplot2` to build a single plot (a pair of histograms after faceting would be one approach, or perhaps a comparison boxplot) which nicely compares the `covid_mort` distribution for counties within metropolitan areas to counties outside of metropolitan areas. Again, make an effort to build and incorporate useful titles and labels so that the resulting plot stands on its own, rather than just accepting all of the defaults that appear.

7 Question 7 (5 points)

Write a short description of where Cuyahoga County falls within the plot you built in Question 6. Specifically, comment on the position of Cuyahoga County in terms of `covid_mort` relative to the other counties within its `metro` category. Two sentences should be sufficient here.

8 Question 8 (10 points)

By now, we'd like you to have read through Chapter 3 of David Spiegelhalter's *The Art of Statistics*. In the above questions we, broadly, examined the relationship between county metropolitan status and the percent of residents who have completed some college. In our first step, we limited to just counties in 5 midwestern states. Reflecting on Chapter 3 of *The Art of Statistics*, please write a brief essay (100-150 words) that discusses the process of inductive inference and how that influences the conclusions we can draw from our work in this assignment. As always, use complete and clear English sentences in your essay.

9 Include the session information

Use the code below to include your session information.

```
sessioninfo::session_info()
```

This will produce considerable output, which we want to see, although I've hidden it in these instructions. Providing the session information helps with reproducibility. It lets us see what packages you have loaded on your machine, and some other information about your R session that can be helpful in understanding any problems you run into.

What Happens After Submission?

We will summarize some of the more interesting responses to Question 8 after the Lab has been graded.

- Note that the teaching assistants will review your responses to all Questions carefully to assess clarity of writing, attention to detail, and adherence to grammatical and syntax requirements. Spelling, grammar, syntax and the rest all matter for grading purposes in this and all other assignments this term.

A detailed answer sketch for this Lab will be provided 48 hours after the submission deadline, and a grading rubric will be provided when the grades are made available, approximately one week after the submission deadline.