431 Lab 02

Due **2022-09-26** at 9 PM | Last Edited 2022-08-23 12:46:21

Table of contents

Deadline . . . 2
Getting Help . . . 2
Learning Objectives . . . 2

Getting Started with Lab 02 **2**
The R Markdown Template for Lab 02 . . . 3
Obtaining the Data for Lab 02 . . . 3
Getting the Data into RStudio . . . 4

Question 1 (10 points) **4**

Question 2 (10 points) **4**

Question 3 (10 points) **5**

Question 4 (15 points) **5**

Question 5 (10 points) **5**

Question 6 (15 points) **5**

Question 7 (10 points) **5**

Question 8 (20 points) **6**

Include the session information **6**

Submitting the Lab **7**

Grading **7**
Late Penalties for Lab Work . . . 8

Deadline

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

- To receive full credit on a Lab, it must be received on Canvas no later than 59 minutes after the posted deadline. (This allows for small issues with uploading to Canvas to occur without penalty.)

Getting Help

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.

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.

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 `2022-431` directory on your machine.

- Into that `lab02` directory, you will download the R Markdown Template for Lab 02, which is called `YOURNAME-Lab02.Rmd`, 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, rather than `YOURNAME`.
- You will then download the data file `lab02_counties.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.

The R Markdown Template for Lab 02

In this Lab, you will analyze some data, and prepare a report in the form of an HTML file, using R Markdown. We have provided you with a very useful R Markdown document template for this assignment called `YOURNAME-Lab02.Rmd` 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 `2021-431` directory.) Then copy the template into your directory for Lab 02, specifically, that you created earlier.

You will build your response to all eight questions as an R Markdown file using the `YOURNAME-lab02.Rmd` template provided. Use the Knit button in RStudio to compile your work and create the HTML output. You'll want to do this multiple times as you go, to identify potential problems quickly.

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 2021 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_counties.csv` in our class Data and Code repository. This data file contains 3,142 rows (each row is a county) and 5 variables including:

Variable Name	Description
`state`	Two-letter postal abbreviation of the state name
`county_name`	Name of the county
`metro`	Whether or not the county is in a metropolitan area
`some_college`	Percentage of county residents who have completed some college
`female_pct`	Percentage of county residents who are female

Note that the `some_college` estimates come from American Community Survey 5-year estimates from 2015-19, and the `female_pct` estimates come from Census Population Estimates published in 2019.

Getting the Data into RStudio

If you've stored the `lab02_counties.csv` file in your R Project directory for Lab 02, you can then read the data into R and create an object called `lab02_data` containing the information with the following command, which is also part of the `YOURNAME-lab02.Rmd` file.

```
lab02_data <- read_csv("lab02_counties.csv")
```

If you've instead stored the `lab02_counties.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:

```
lab02_data <- read_csv("data/lab02_counties.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.

```
lab02_data <- read_csv(
 "https://raw.githubusercontent.com/THOMASELOVE/431-data/main/lab02_counties.csv"
 )
```

Running any of these commands in R should lead to the appearance of a `lab02_data` object in your R session. Don't run more than one command - just the one is what you need.

Question 1 (10 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 `lab02_data` 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.**

Question 2 (10 points)

Write a piece of R code that counts the number of observations (counties) in our 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.

Question 3 (10 points)

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

Question 4 (15 points)

Use the tools we've been learning in the `ggplot2` package to build a histogram of the `some_college` 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.

Question 5 (10 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 in terms of `some_college`.

Question 6 (15 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 `some_college` 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.

Question 7 (10 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 `some_college` relative to the other counties within its `metro` category. Two sentences should be sufficient here.

Question 8 (20 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.

Include the session information

At the end of your R Markdown file, please include a new code chunk to provide the **session information**. The result will look something like this.

```
sessionInfo()
```

```
R version 4.2.1 (2022-06-23 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22000)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] compiler_4.2.1  magrittr_2.0.3  fastmap_1.1.0   cli_3.3.0
 [5] tools_4.2.1     htmltools_0.5.2 rstudioapi_0.13 yaml_2.3.5
 [9] stringi_1.7.8   rmarkdown_2.14  knitr_1.39      stringr_1.4.0
[13] xfun_0.31       digest_0.6.29   jsonlite_1.8.0  rlang_1.0.2
[17] evaluate_0.15
```

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. The `sessionInfo()` command shown above is part of the template for this Lab.

Submitting the Lab

As mentioned, you should build your entire response as an R Markdown file using the `YOURNAME-lab02.Rmd` 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.

Submit **both** your revised R Markdown 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 R Markdown 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.

Grading

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

- This Lab will be graded on a scale from 0-100.
- 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 on the day after the submission deadline, and a grading rubric will be provided when the grades are made available, approximately one week after the submission deadline.

Late Penalties for Lab Work

- Labs that are turned in 1-12 hours after the deadline will lose 10% of available points.
- Labs turned in more than 12 but less than 72 hours after the deadline will lose 25% of available points.
- No extensions to Lab deadlines will be permitted this semester. Labs turned in more than 72 hours after the deadline will receive no credit.
- Note that your lowest lab score (out of Labs 1-7) over the course of the semester will be dropped before we calculate your lab grade.