Assignment 3 - Data Visualization. Due October 3, 11:59pm 2021

EPIB607 - Inferential Statistics[a]

[a]Fall 2021, McGill University

This version was compiled on September 19, 2021

All questions are to be answered in an R Markdown document using the provided template and compiled to a pdf document. You are free to choose any function from any package to complete the assignment. Concise answers will be rewarded. Be brief and to the point. Each question is worth 25 points. Label your graphs appropriately with proper titles and axis labels. Justify your answers. You may compile your reoport to pdf or to HTML. If you compile to HTML, then you must print the resulting HTML to pdf. Please submit the compiled pdf report to Crowdmark. You must also submit your code to Crowdmark. If you use the template, the code from your assignment will automatically appear at the end. Upload this code to Q5 in Crowdmark. You can upload a single pdf to Crowdmark, and then select the pages for a given question. See https://crowdmark.com/help/ for details.

Template

Please use the `.Rmd` template for Assignment 3 which is available on myCourses.

1. (25 points) Mask mandates in Kansas

This question is based on the article *Kansas began requiring masks, then virus cases dropped. Weeks later, 'the data is solid'*. In a short press conference, Dr. Lee Norman presents a figure which is reproduced here in Figure 1.

a) (5 points) What does this graph make it appear is happening? What would your conclusion be about the importance of a mask mandate?

b) (5 points) Provide the column names of the tidy dataset that would have been used to create Figure 1. For each column, name the aesthetic its being mapped onto.

c) (10 points) Using the data from Figure 1, create an alternative visualization and interpret the graph. Be sure to label your axes and provide a descriptive title.

d) (5 points) Comment on the differences between Figure 1 and the graph you created in part b). Does the conclusion you described from part a) still hold ?

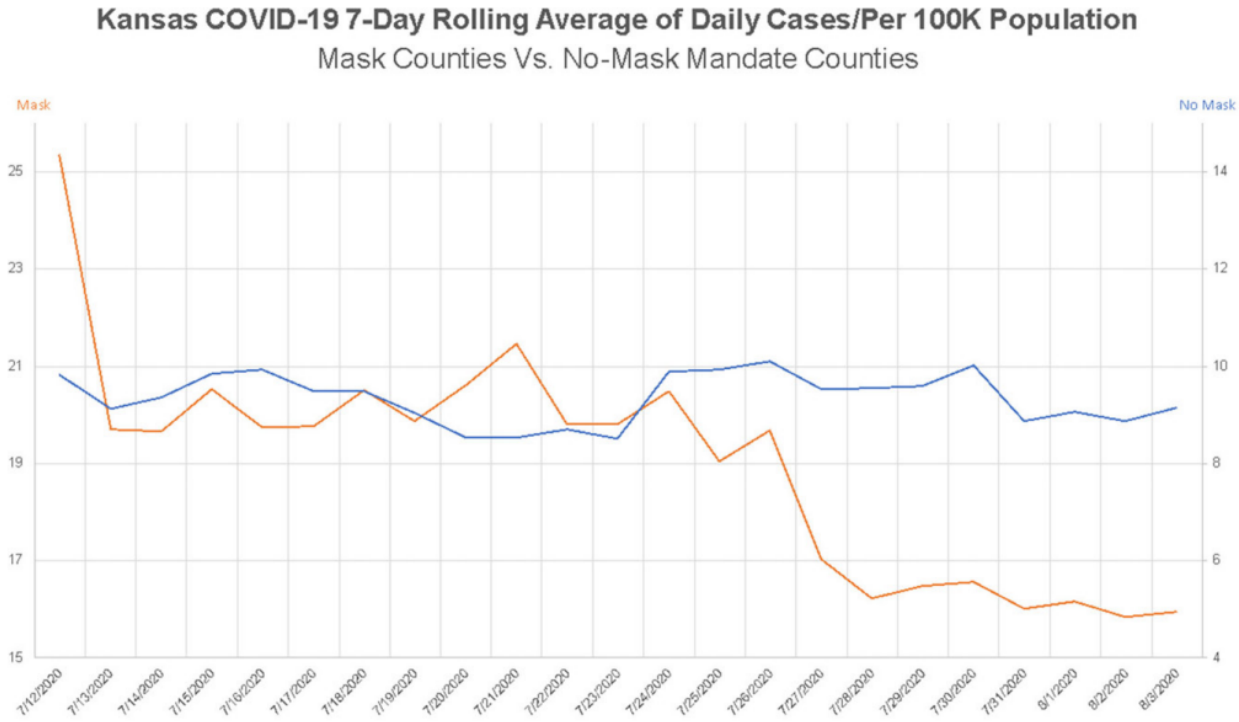


Fig. 1. Kansas COVID-19 7-Day rolling average of cases per 100k.

2. (25 points) Are Covid cases decreasing over time in Georgia?

This question is based on the data presented by the Georgia Department of Public Health and shown in Figure 2:

a) (5 points) List the mappings of variables onto visual cues. What does this graph make it appear is happening? Is this a good visualization - why or why not?

b) (10 points) The data in the file `georgiaCounties.csv` contains daily COVID19 data for the 5 counties of interest in Georgia. In addition to `date`, `state`, `county`, `population`, `daily_cases` and `daily_deaths` you are given `cases` which represents the cumulative number of cases, `deaths` represents the cumulative number of deaths, and `fips` which is the Federal Information Processing System codes for counties. The data can be read into `R` using the code shown below. Using this data, create an alternative version of Figure 2 and interpret the graph. Be sure to label your axes and provide a descriptive title.

```
georgia <- readr::read_csv(here::here("georgiaCounties.csv"),
                           col_types = c("Dfffddddd"))
```

c) (5 points) Comment on the differences between Figure 2 and the graph you created in part b). Does the conclusion you described from part a) still hold ?

d) (5 points) Plot the data for a longer time horizon and comment on any patterns you see. Contrast this with Figure 2 and the figure you created in part b.

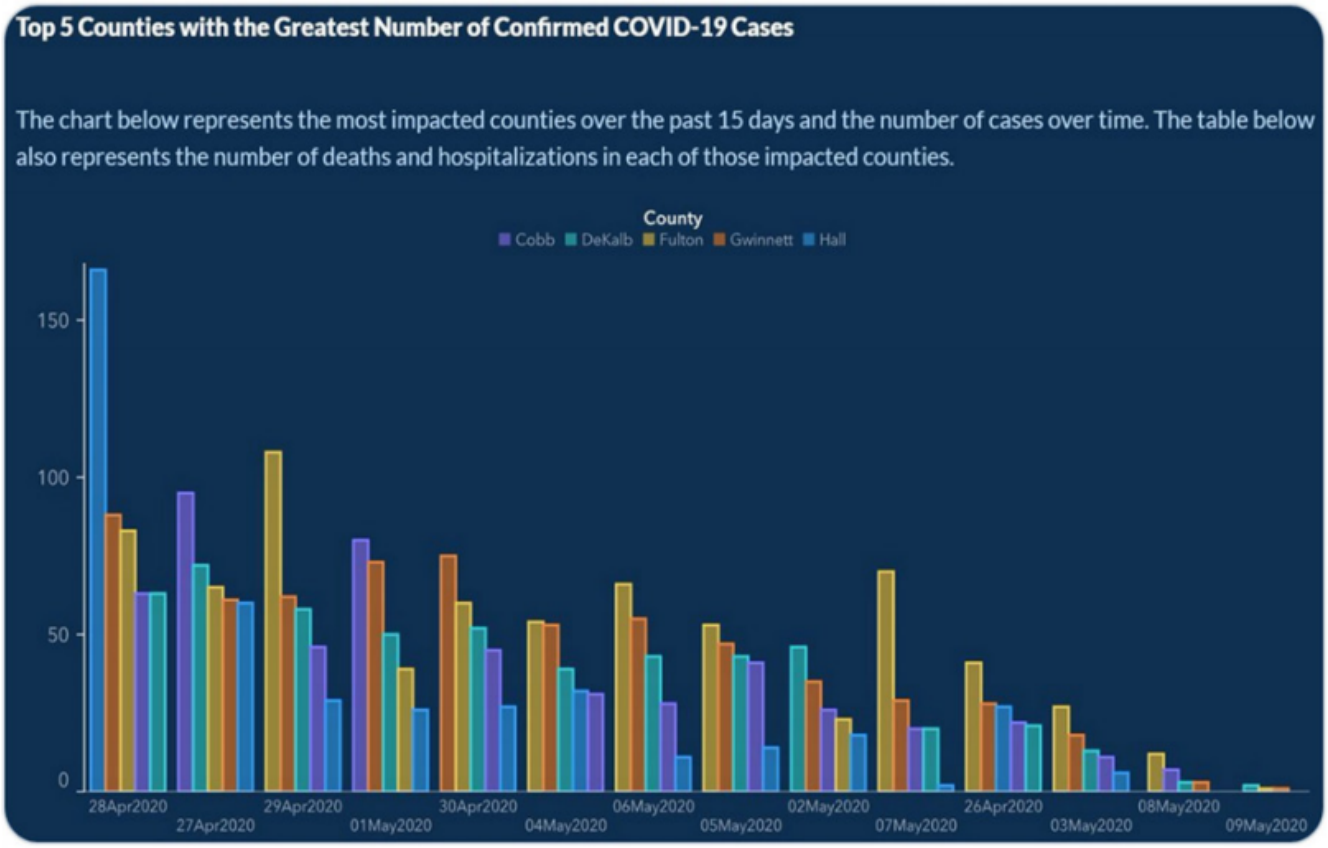


Fig. 2. Georgia Department of Public Health Figure.

3. (25 points) Food in America

Vox published a list of Charts that explain food in America. There are 40 maps, charts, and graphs that show where our food and drink comes from and how we eat it.

a) (20 points, 10 for best and 10 for least favorite) Pick your best and least favorite graphic, and briefly explain why using the taxonomy we learned in class (Sections 2, 3 and 5) (e.g. visual cues being used, one-to-one mapping, appropriate use of coordinate systems and color scales). Provide a link to each figure. For example, this link: Figure 18 was created using the following code: `[Figure 18](https://www.vox.com/a/explain-food-america#list-21)`

b) (5 points) For either of the two graphs you chose in part a), **describe** an alternative version and explain its advantages over the original. You do not need to actually create the figure.

4. (25 points) Geometries in ggplot2

a) (5 points) In Figure 4.6 of the course website notes, we plotted life expectancy vs. GDP per capita from the gapminder data with point and smooth geometries. What happens when you put the `geom_smooth()` function before `geom_point()` instead of after it? What does this tell you about how the plot is drawn? Think about how this might be useful when drawing plots.

b) (2.5 points) What happens if you map `color` to `year` instead of `continent`? Explain this behavior.

c) (2.5 points) Instead of mapping `color = year`, what happens if you try `color = factor(year)`? Explain this behavior.

d) (5 points) The `Oxboys` dataset, from the `nlme` package records the heights (`height`) and centered ages (`age`) of 26 boys (`Subject`), measured on nine occasions (`Occasion`). Figure 3 is a line plot of height vs. age. Is this an appropriate visualization of the data? If yes, interpret the figure. If no, explain what is wrong. Create an alternative visualization by modifying the code below and interpret this graph.

```
data(Oxboys, package = "nlme")
p <- ggplot(data = Oxboys, aes(x = age, y = height))
p + geom_line()
```

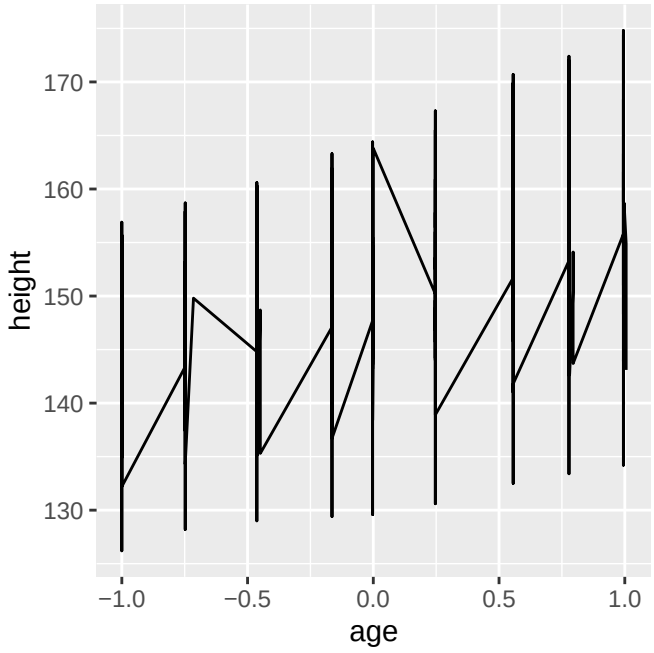
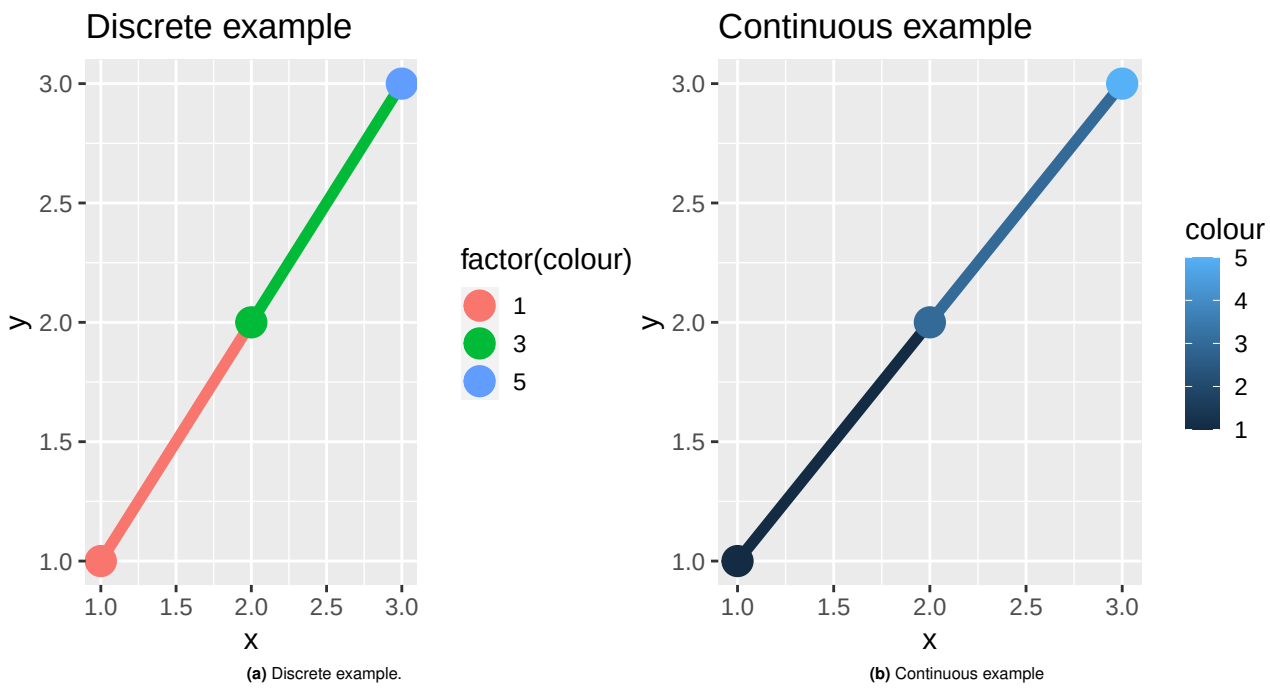


Fig. 3. Height vs age for the Oxboys data.

e) (10 points, 5 each) Figure 4 illustrates the difference between mapping continuous and discrete colours to a line and code is given below to reproduce these plots.

i) Explain the purpose of the `aes(group = 1)` code.

ii) What's the difference between `aes(group = 1)` and `aes(group = 2)`? Why?



(a) Discrete example.

(b) Continuous example

Fig. 4. Difference between mapping continuous and discrete colours to a line

```
df <- data.frame(x = 1:3, y = 1:3, colour = c(1,3,5))

ggplot(df, aes(x, y, colour = factor(colour))) +
  geom_line(aes(group = 1), size = 2) +
  geom_point(size = 5) + labs(title = "Discrete example")

ggplot(df, aes(x, y, colour = colour)) +
  geom_line(aes(group = 1), size = 2) +
  geom_point(size = 5) + labs(title = "Continuous example")
```