A3: Insert Assignment Name Here

EPIB 607 - FALL 2021

your name and McGill ID

compiled on October 07, 2021

1 Mask mandates in Kansas

(a)

Figure 1. appears to show that the 7-day rolling average of daily new cases of COVID-19 in the counties where mask is mandatory has a **decreasing trend**; while the number in the no-mask mandate counties **remains in a similar level**. After July 24th, it seems that the number in mask mandate counties **dropped below** the number in no-mask mandate counties. Based on the first impression of this figure, people might draw conclusion that the mask mandate policy is **effectively reducing** the new cases and the mask mandate counties become "safer' ' (less new cases) than the no-mask mandate counties.

(b)

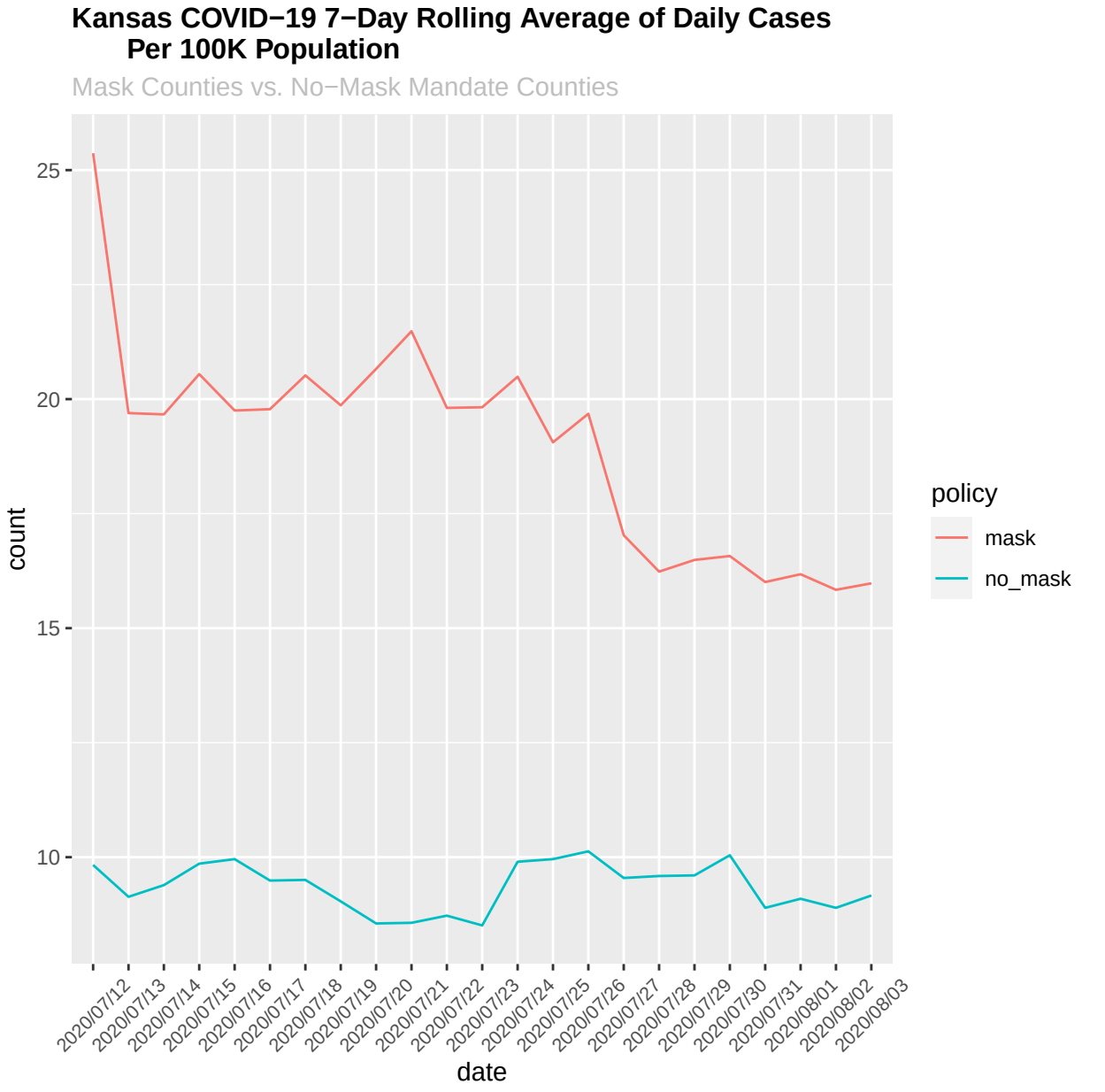
The column names should be `date`, `count`, and `policy`. `count` refers to the 7-day rolling average of daily new COVID-19 cases, and policy is binary, referring to whether the number is for mask mandate or no-mask mandate counties.

The variable `date` should be mapped to the x-axis and `count` should be mapped to the y-axis, and the line color is specified by `policy`.

(c)

```
p<- ggplot(data=q1.dat,
       aes(x=date, y=count, group = policy, colour=policy))
p + geom_line() +
  labs(title = "Kansas COVID-19 7-Day Rolling Average of Daily Cases
       Per 100K Population",
```

```
    subtitle = "Mask Counties vs. No-Mask Mandate Counties")+
theme(axis.text.x = element_text(angle = 45, size=8, vjust=0.5)) +
theme(plot.title = element_text(color = "black", size = 12, face = "bold"),
plot.subtitle = element_text(color = "grey"))
```



Interpretation: The new figure shows two trajectories of the 7-day rolling average of daily cases per 100k population in mask-mandate counties and in no mask mandate counties from Kansas. The number in mask mandate counties starts high around 25 and drop to 20 in a day. It remains between 20-22 for 12 days and decreases quickly to around 16. The number in no-mask mandate counties are stable between 8-10 the whole time.

(d)

Figure 1 is misleading with the double y-axes and the time frame. With the new figure, we realize that:

1. The number for mask-mandate counties on the first day (July 12th) is much higher than the rest data points and it adds a lot to the "feeling of decreasing". Excluding the first day, the new cases for mask-mandate countries fluctuates from around 20 (July 13) to 16 (Aug 2nd), while the numbers for no-mask mandate counties are much lower, and stable between 8-10.

2. The intersection in Figure 1 is also deceiving as the numbers in mask mandate counties are consistently higher than that in no mask mandate counties.(It could be explained partially by decision-making strategy – counties with more severe pandemics may first adopt stricter public health regulations.)

3. The x-axis shows a short period of time from July 13th to Aug 2nd, a little over 2 weeks. With a 7-day rolling average, it is harder to tell the number decreasing or increasing as the change is thinned by the previous 6 days' readings.

I would conclude that, in such a short period, there is no strong evidence to show that the mandatory mask policy is making a huge impact on the daily new cases. But based on the decreasing trend, the mask mandate policy could be effective in reducing the new cases in the long run.

(You still get full mark to (d) with other conclusions and proper justification)

2 Are Covid cases decreasing over time in Georgia?

(a)

Counties mapped on colour, date mapped on x-axis, number of cases mapped on y-axis(2 points)

The graph makes it appear that cases are decreasing over time in the 5 counties (1 point)

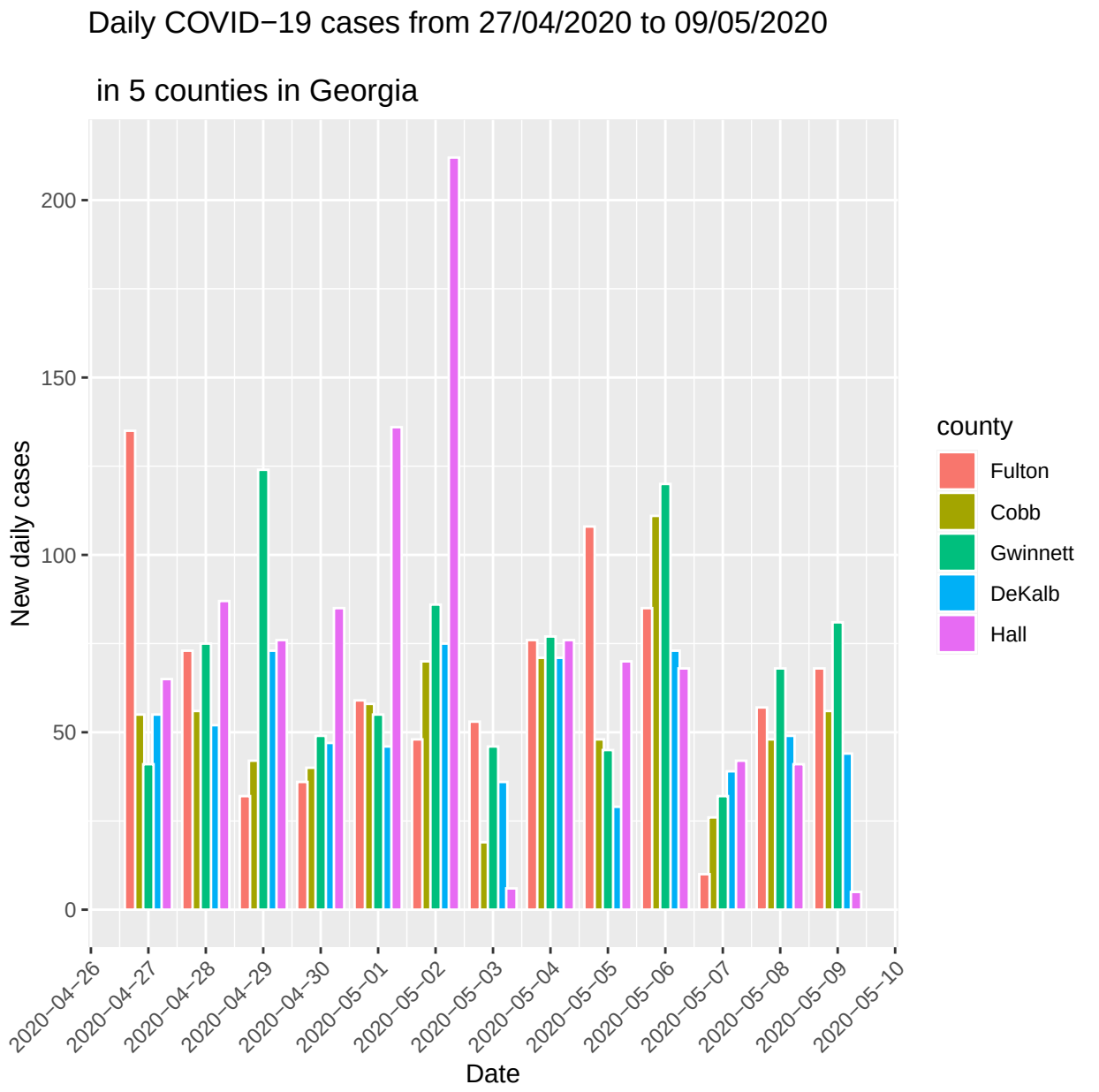
Problems with this visualization (2 points):
- The x-axis values aren't mapped in chronological order (-1 if not mentioned), they seem to be ordered in a descending order of number of cases, which is deceiving.
- The counties (bars) weren't ordered in the same way for all dates on the x-axis, which makes it harder to observe the trend for each county (-0.5 if not mentioned)

Other problems:
- Cases aren't standardized so we can't compare counties to each other
- Graph isn't well-labelled

Even though that graph labels didn't specifically state that the it was daily new cases, we can deduce that it is not cumulative cases. Cumulative cases can't be decreasing, they only plateau (when there are no new cases) and increase with the addition of new cases.

(b)

Graph (7 points: 3 points for labels (x-axis, y-axis, title), 4 points overall)



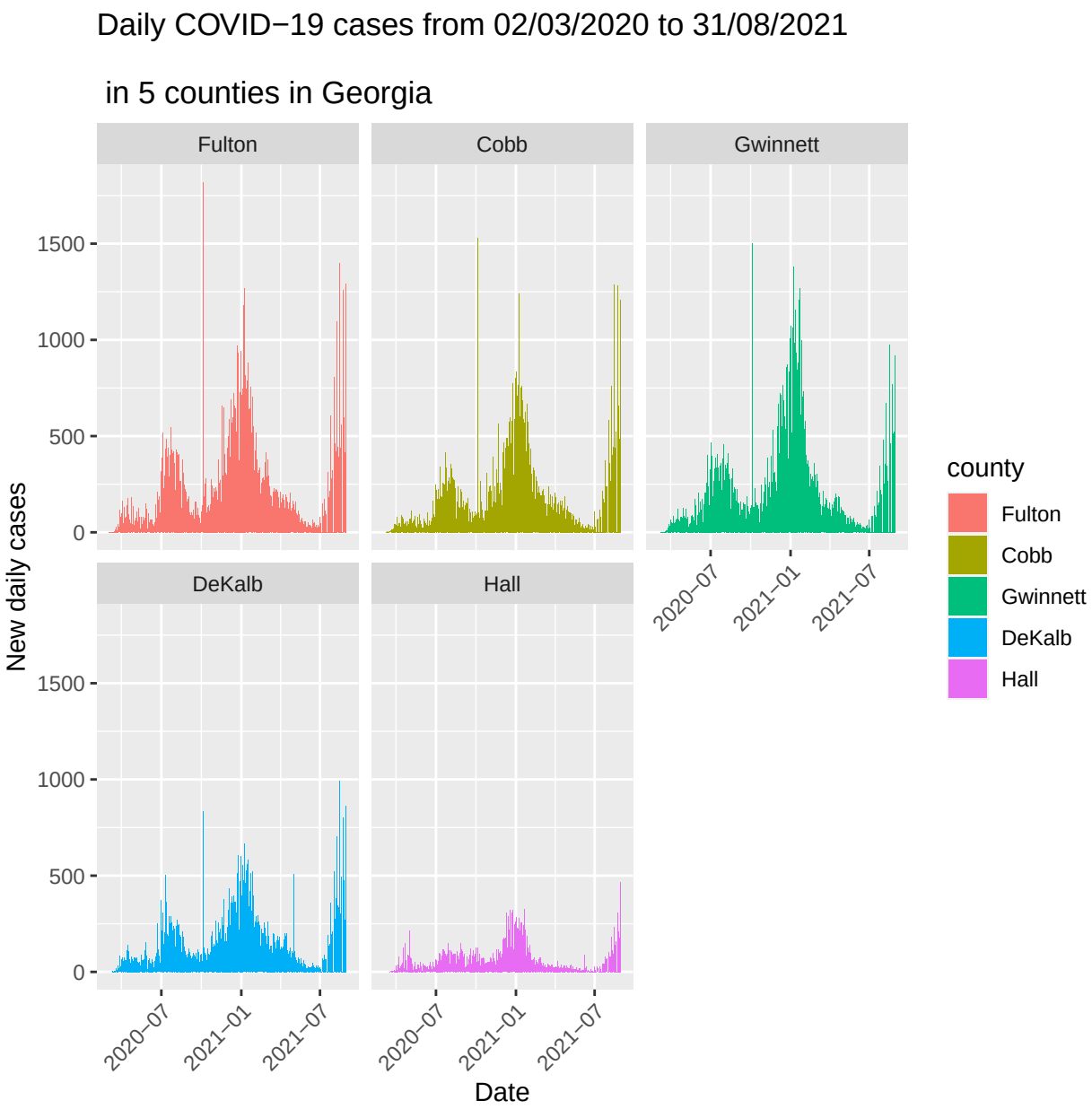
The graph shows that cases are fluctuating over time. Most of the trends of the different counties are similar except for Hall, cases increased significantly on May 2nd then started to decrease notably over the following 8 days. (3 points for interpretation)

(c)

Differences: x-axis in chronological order and bars (counties) are in the same order for all dates (2 points) Conclusion doesn't hold, cases are not really decreasing over time across the counties, maybe only in Hall new cases are decreasing but the trend across the counties is mainly fluctuating over the 15-day period. (3 points)

(d)

Graph (2.5 points)



Figures 2 and 2b are mainly snapshots of figure 2c. They represent a period where cases were really low and relatively stable. Figure 2c shows that there have been three main waves of COVID-19 cases in Georgia (summer 2020, winter 2021 and summer 2021). It kind of shows that cases are fluctuating in the sense that they increase over specific periods of time significantly, then decrease again. However there is an overall trend since specific periods of significant increase can be observed in Fig2c while in Fig 2b the fluctuation seemed to be random. (2.5 points)

3 Food in America

(a)

You should be able to identify the **variables** used in the figure, the **variable type** (categorical/continuous/etc.), the **aesthetics** that each variable is mapped to, other **visual cues** used in the figure, whether there is **one-to-one mapping** of aesthetics/scales, and **justify whether the choice of aesthetics/scales/visual cues are appropriate and effective**.

Sample Answer:

Favourite figure: Figure 33

This figure displays information about public drinking laws by state. Public drinking laws is qualitative ordinal variable while state is qualitative nominal variable. The figure uses position and colour as visual cues. The status of public drinking laws is mapped one-to-one to colour, and the state is mapped one-to-one to position (geographical location within the United States). The colour scale clearly distinguishes between different laws, but it is also sequential with colder colours representing no statewide bans or laws allowing public drinking in designated areas, and warmer colours representing no statewide bans or prohibition of public drinking in most municipalities. Colour is also used as a tool to highlight towns that allow public drinking. The position scale is easy to follow as it simply represents geographical location, and one can visually assess whether public drinking laws are clustered geographically.

Least favourite figure: Figure 23

This figure displays information about fruit and vegetable consumption (percentage of adults consuming fruit 2+ times per day and vegetables 3+ times per day) by state. Fruit and vegatable consumption are each continuous variables that have been discretized into bins. State is a qualitative nominal variable. The figure uses position, colour, and pattern as visual cues. Fruit and vegetable consumption are mapped to the same colour scale, which also includes patterns (diagonal lines vs. solid colour). State is mapped one-to-one to position (geographical location within the united states). The mapping of fruit/vegetable consumption to both colour and pattern is confusing, particularly as the diagonal lines seem to be a visual cue to highlight certain states when in fact they are just part of the continuous scale. Further, the use of the same colour/pattern scales for both fruit and vegetable consumption but on different position scales (i.e. maps) make it difficult to assess the two together.

(b)

You should be able to describe the **variables** you would include in your figure, which **aesthetics** they would be mapped to, and **justify their advantages over the original figure**.

Sample Answer:

For Figure 23, I would plot fruit and vegetable consumption on the same figure/position scale. I would map fruit consumption one-to-one to a sequential colour scale (no pattern) and vegetable consumption one-to-one to a pattern scale, and overlay the two onto the map of the US. This would allow me to visually assess fruit and vegetable consumption by state together, and would be a one-to-one mapping for each variable.

4 Geometries in ggplot2

Link

(a)

Since the layer of the plot is additive (2 point), when geom_smooth() function is put before geom_point(), the output generated by "geom_point()" is on top of that generated by "geom_smooth()".(1 point) Thus, some part of the smoothed line is not visible in our plot. (1 point) This reminds us to pay attention to the order of commands when doing the visualization. (1 point)

(b)

By switching the color from continent to year, we try to visualize the trend of GDP vs LifeExp as changed by year. (1 point) We can see that generally around 2000, the GDP and LifeExp is higher that that around 1960. Compared to the categorical variable "continent", "year" is treated as continuous. Thus, instead of distinct colors and multiple smoothed trend, it gives us a gradient color scale with one trend. (state "continuous" and "color scale" difference, 1.5 points)

(c)

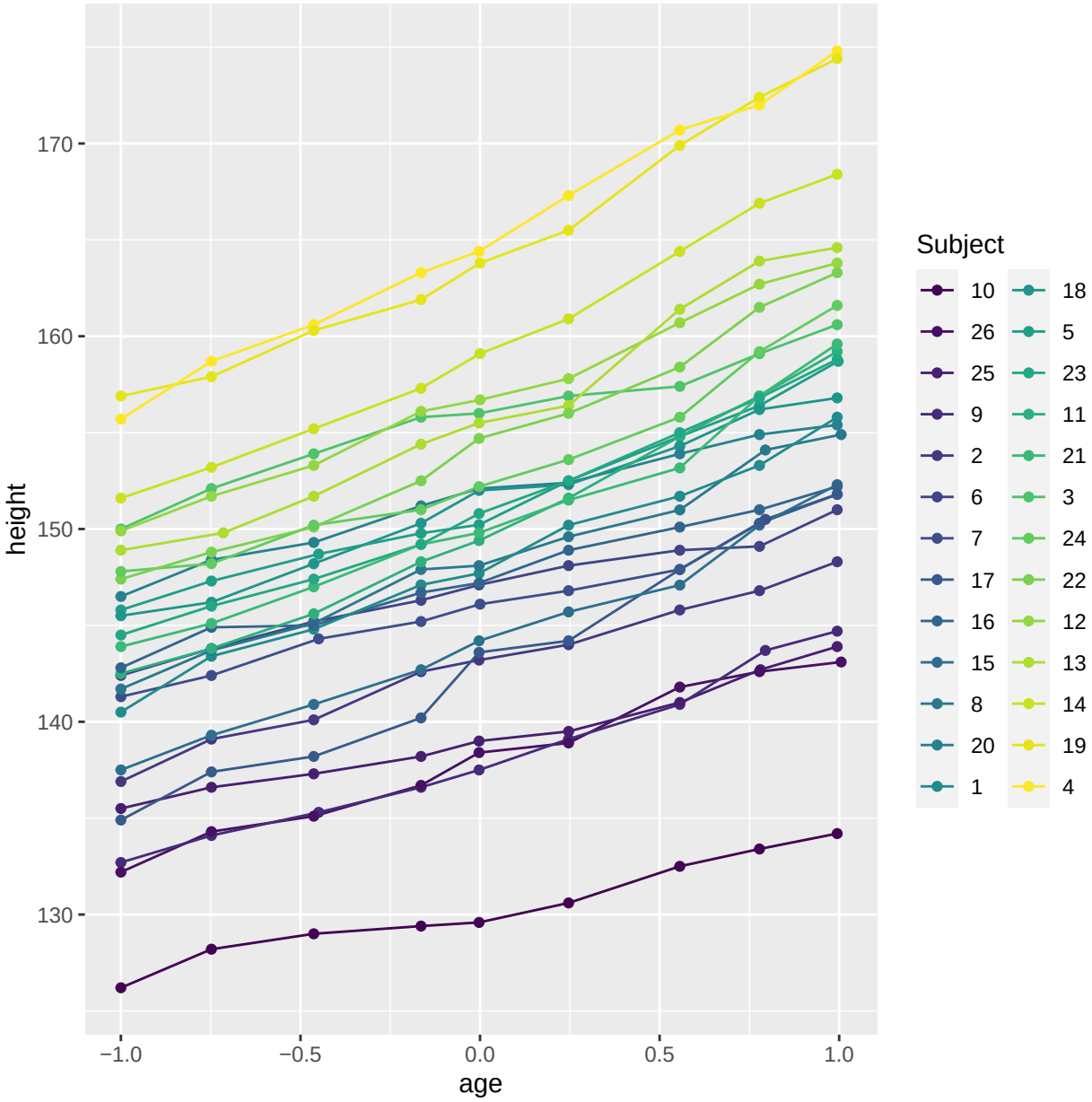
When we factorize the year, we change it from a continuous variable to a discrete/categorical variable. (1 point) The color becomes more vibrant and we can see in which year exactly these data were recorded. We also have multiple smoothed line now. (1.5 points)

(d)

No, it's not a good visualization.We should plot the height growth with respect to each boy, so we need to group and color the data by "Subject" in the dataset. To better show the measured data, I add the function "geom_point()". (As long as your visualization is based on each subject, I'll give you full mark)

```
library(ggplot2)

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



(e)

i) For the discrete case, since we only have one observation within each group (color), we need to specify "group=1" to override the default value. The value here is simply a dummay variable/placeholder. Therefore, we force the line between two different color to be drawn. We don't need to specify that for continuous variable. (As long as you say it's a placeholder/dummy variable, you should be fine.)

ii) Nothing happens when we change "group=1" to "group=2". (3 points) Actually, if we simply want to override the default setting, we can put whatever number here. (the trait of place holder. 2 points)

Code

```
## ---- Setup ------------------------------------------------------------
# set default chunk options here
knitr::opts_chunk$set(
  echo = FALSE,           # don't show code
  warning = FALSE,        # don't show warnings
  message = FALSE,        # don't show messages (less serious warnings)
  cache = FALSE,          # set to TRUE to save results from last compilation
  fig.align = "center",    # center figures
  fig.asp = 1           # fig.aspect ratio
)
## ---- Question-1 -------------------------------------------------------

library(tidyverse)
library(ggplot2)
library(reshape2)

mask<-read.csv("a1d1.csv", header=F)
no_mask<-read.csv("a1d2.csv", header=F)

# creating long format tidy data
colnames(mask)<-c("date", "mask")
colnames(no_mask)<-c("date", "no_mask")
temp<-merge(mask,no_mask, by="date")
q1.dat <- melt(temp, id="date")
colnames(q1.dat)<- c( "date", "policy", "count")
p<- ggplot(data=q1.dat,
       aes(x=date, y=count, group = policy, colour=policy))
p + geom_line() +
  labs(title = "Kansas COVID-19 7-Day Rolling Average of Daily Cases
       Per 100K Population",
      subtitle = "Mask Counties vs. No-Mask Mandate Counties")+
  theme(axis.text.x = element_text(angle = 45, size=8, vjust=0.5)) +
  theme(plot.title = element_text(color = "black", size = 12, face = "bold"),
  plot.subtitle = element_text(color = "grey"))
## ---- Question-2 -------------------------------------------------------
library(tidyverse)
#reading data
georgia <- readr::read_csv(here::here("georgiaCounties.csv"),
                           col_types = c("Dfffddddd"))

# summary(georgia)
```

```r
#can't have negative daily_cases, nor daily_deaths,
#so will consider those typos and will remove - sign
georgia$daily_cases <- abs(georgia$daily_cases)
georgia$daily_deaths <- abs(georgia$daily_deaths)

#check
#summary(georgia)

# subsetting the entries for the period of interest
# (15 days from 2020-04-26 to 2020-05-10)
q2_dat <- georgia %>%
  filter(date > "2020-04-26" &
           date < "2020-05-10")

ggplot(data = q2_dat,
       aes(y = daily_cases,
           x = date,
           fill = county)) +
  geom_bar(stat = "identity",
           color = 'white',
           position = position_dodge(0.8)) +
  xlab("Date") +
  ylab("New daily cases") +
  ggtitle("Daily COVID-19 cases from 27/04/2020 to 09/05/2020
          \n in 5 counties in Georgia") +
  scale_x_date(date_breaks = "1 day") +
  theme(axis.text.x = element_text(angle = 45,
                                   hjust=1))
ggplot(data = georgia,
       aes(y = daily_cases,
           x = date,
           fill = county)) +
  geom_bar(stat = "identity") +
  xlab("Date") +
  ylab("New daily cases") +
  ggtitle("Daily COVID-19 cases from 02/03/2020 to 31/08/2021
          \n in 5 counties in Georgia") +
  facet_wrap(~county) +
  xlim(as.Date(c('2020-03-02', '2021-08-31'),
               formate = "%d/%m/%Y")) +
  theme(axis.text.x = element_text(angle = 45,
                                   hjust=1))
## ---- Question-3 ------------------------------------------------------
```

```r
## ---- Question-4 ----------------------------------------------------------

library(ggplot2)

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