Stat 220: Midterm I

January 30 2024

Name:

Total Points: 100

`Gapminder` data

Data includes health and income outcomes for 142 countries from 1952 to 2007 in increments of 5 years. The variables in the dataset are `country`, `continent`, `year`, `lifeExp`, `pop`, and `gdpPercap`. The descriptions for the variables are:

- `country` : name of the country, factor with 142 levels
- `continent`: name of the continent, factor with 5 levels
- `year` : ranges from 1952 to 2007 in increments of 5 years (12 distinct years)
- `lifeExp`: life expectancy at birth, in years
- `pop` : population
- `gdpPercap` : GDP per capita (US$, inflation-adjusted)

```
glimpse(gapminder)
Rows: 1,704
Columns: 6
$ country   <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", ~
$ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, ~
$ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, ~
$ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8~
$ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12~
$ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, ~
```

The distinct continents in the data are as follows:

```
gapminder %>% pull(continent) %>% unique()
[1] Asia     Europe   Africa   Americas Oceania
Levels: Africa Americas Asia Europe Oceania
```

Part 1: Data Wrangling *(10 points each)*

What do the following code chunks do? Provide a thorough and intuitive (2-3 sentences) description of the output from each of the following R chunks. The chunks produce a new data set. Please give the dimensions in addition to your description. Write your descriptions in regular English, without using variable names.

a.

```
gapminder %>%
  filter(year == 2007) %>%
  group_by(country) %>%
  summarize(median_lifeExp = median(lifeExp))
```

b.

```
gapminder %>%
  group_by(country) %>%
  arrange(year) %>%
  mutate(change_in_lifeExp = lifeExp - lag(lifeExp)) %>%
  summarize(average_change = mean(change_in_lifeExp, na.rm = TRUE))
```

c.

```
set.seed(143)
selected_countries <- gapminder %>%
  distinct(country) %>%
  slice_sample(n = 3) %>%
  pull(country)

gapminder %>%
  filter(country %in% selected_countries) %>%
  filter(year == 2007) %>%
  group_by(country) %>%
  summarize(median_lifeExp = median(lifeExp))
```

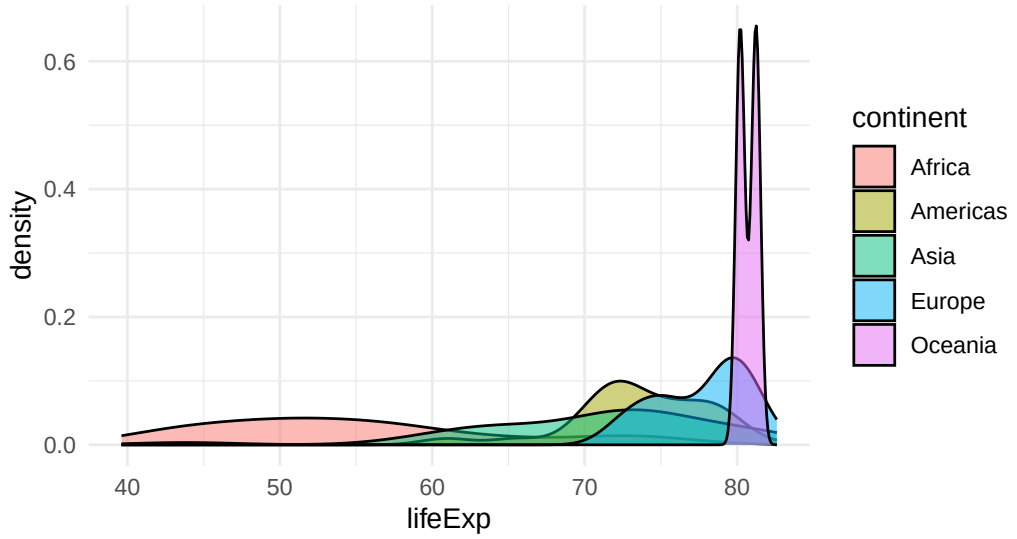
d.

```
set.seed(143)
gapminder %>%
  filter(continent == "Europe") %>%
  slice_sample(n = 3) %>%
  select(country) %>%
  inner_join(gapminder, by = "country") %>%
  group_by(country, year) %>%
  summarize(avg_gdpPercap = mean(gdpPercap)) %>%
  pivot_wider(names_from = year,
              values_from = avg_gdpPercap,
              names_prefix = "year_")
```

Part 2: Graphics *(5 points each)*

a. The density plot below visualizes the distribution of life expectancy across different continents in the year 2007. Suggest 2 ways to improve this plot's aesthetics and readability. Write just the code modifications below:

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = lifeExp, fill = continent)) +
  geom_density(alpha = 0.5)
```

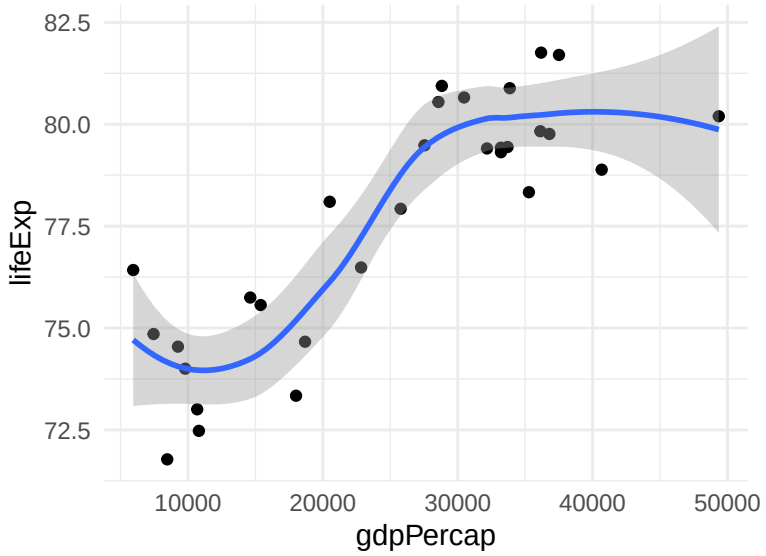


Answer:

b. Linear Model Plot Modification

Please suggest how you would modify the plot to color code the points by continent and make the smoother line purple. You don't need to provide a code from scratch. Just write the suggested changes.

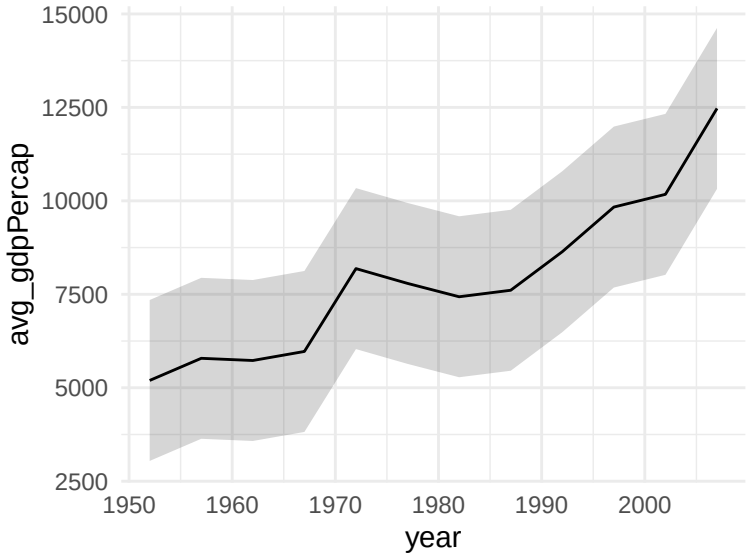
```
gapminder %>%
  filter(year == 2007, continent == "Europe") %>%
  ggplot(aes(x = gdpPercap, y = lifeExp)) +
  geom_point() +
  geom_smooth()
```



c. Ribbon Plot

The ribbon plot below is designed to display the range of one standard deviation above and below the mean for GDP per capita over time in Asia. To complete the code, fill in the necessary parts.

```
gapminder %>%
  filter(continent == "Asia") %>%
  ##### FILL IN i. #####
  summarize(avg_gdpPercap = mean(gdpPercap)) %>%
  ggplot(aes(x = year, y = avg_gdpPercap)) +
  geom_ribbon(aes(ymin = avg_gdpPercap - sd(avg_gdpPercap),
                  ymax = avg_gdpPercap + 2*sd(avg_gdpPercap)), alpha = 0.2) +
  geom_line()
```



d. Faceted Density Plot

Complete the following code to create a faceted density plot showing the distribution of population for each continent in 2007.

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = pop)) +
  geom_density() +
  ##### FILL IN i. #####
```

Part 3: Data Objects *(5 points each)*

```
x <- c(7, 5, 3, 9)
y <- c(FALSE, factor(c("cellar", NA)), 42)
z <- list(z1 = x,
          z2 = y,
          z3 = c("Pumpkin", "Spice"),
          z4 = matrix(10:18, nrow = 3),
          z5 = list(a = 20, b = "door"))
```

Consider the above objects to answer the following questions.

(a) What does the following evaluate to?

```
z[["z5"]][["b"]]
```

(b) What does the following evaluate to?

```
x * 2
```

(c) What does the following evaluate to?

```
x + y
```

(d) Write the code to extract the third column of the z4 matrix in z.

Part 4: What do the following codes/code chunks evaluate to?

(a)

```
library(lubridate)
time_length(interval(ymd("2000-01-01"), ymd("2005-01-01")), unit = "years")
```

(b)

```
library(lubridate)
ymd("2000-01-01") + weeks(1)
```

(c)

```
library(forcats)
f <- factor(c("low", "medium", "high"), levels = c("low", "medium", "high"))
new_order <- c("high", "low", "medium")
f <- fct_relevel(f, new_order)
levels(f)
```

(d)

```
library(forcats)
weather <- factor(c("sunny", "cloudy", "rainy", "snowy"))
weather <- fct_recode(weather, bright = "sunny", white = "snowy")
levels(weather)
```