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 (3-5 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 %in% c(1952, 2007)) %>%
```

```
  group_by(continent, year) %>%
  summarize(median_gdpPercap = median(gdpPercap),
            median_lifeExp = median(lifeExp),
            median_pop = median(pop))
```

b.

```
gapminder %>%
  mutate(gdp_total = gdpPercap * pop) %>%
  group_by(continent, year) %>%
  summarize(gdp_continent = sum(gdp_total)) %>%
  pivot_wider(names_from = year,
              values_from = gdp_continent,
              names_prefix = "year_")
```

c.

```
set.seed(143)
selected_countries <- gapminder %>%
  distinct(country, continent) %>%
  group_by(continent) %>%
  slice_sample(n = 1) %>%
```

```
  pull(country)

gapminder %>%
  filter(country %in% selected_countries) %>%
  filter(year %in% c(1952, 2007)) %>%
  group_by(country, year) %>%
  summarize(median_gdpPercap = median(gdpPercap),
            median_lifeExp = median(lifeExp),
            total_pop = sum(pop)) %>%
  pivot_longer(cols = -c(country, year),
               names_to = "stat",
               values_to = "value")
```

d.

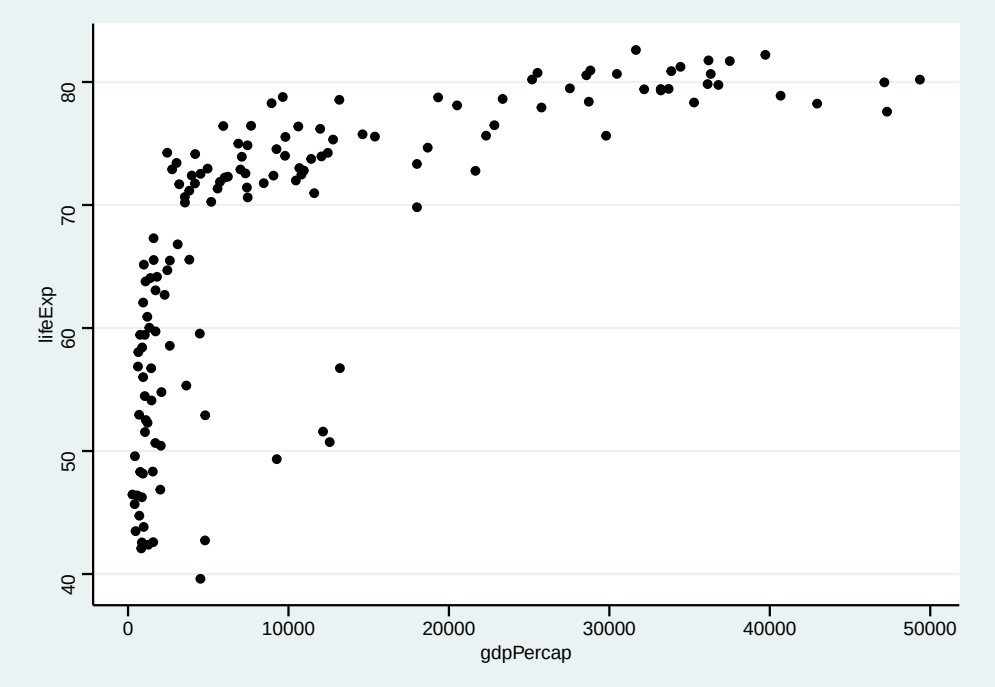
```
set.seed(143)
gapminder %>%
  filter(continent == "Asia") %>%
  distinct(country) %>%
  slice_sample(n = 5) %>%
  inner_join(gapminder, by = "country") %>%
  filter(year %in% c(1952, 2007)) %>%
  group_by(country, year) %>%
```

```r
  summarize(median_gdpPercap = median(gdpPercap),
            median_lifeExp = median(lifeExp),
            total_pop = sum(pop)) %>%
  pivot_longer(cols = -c(country, year),
               names_to = "stat",
               values_to = "value") %>%
  pivot_wider(names_from = year,
              values_from = value,
              names_prefix = "year_")
```

Part 2: Graphics *(15 points each)*

a. The scatter plot below visualizes the relationship between GDP per capita and life expectancy of countries in the year 2007. What are 5 ways you could improve the aesthetics and readability of this plot by following best data visualization practices? Also write 5 code modifications on the space provided below:

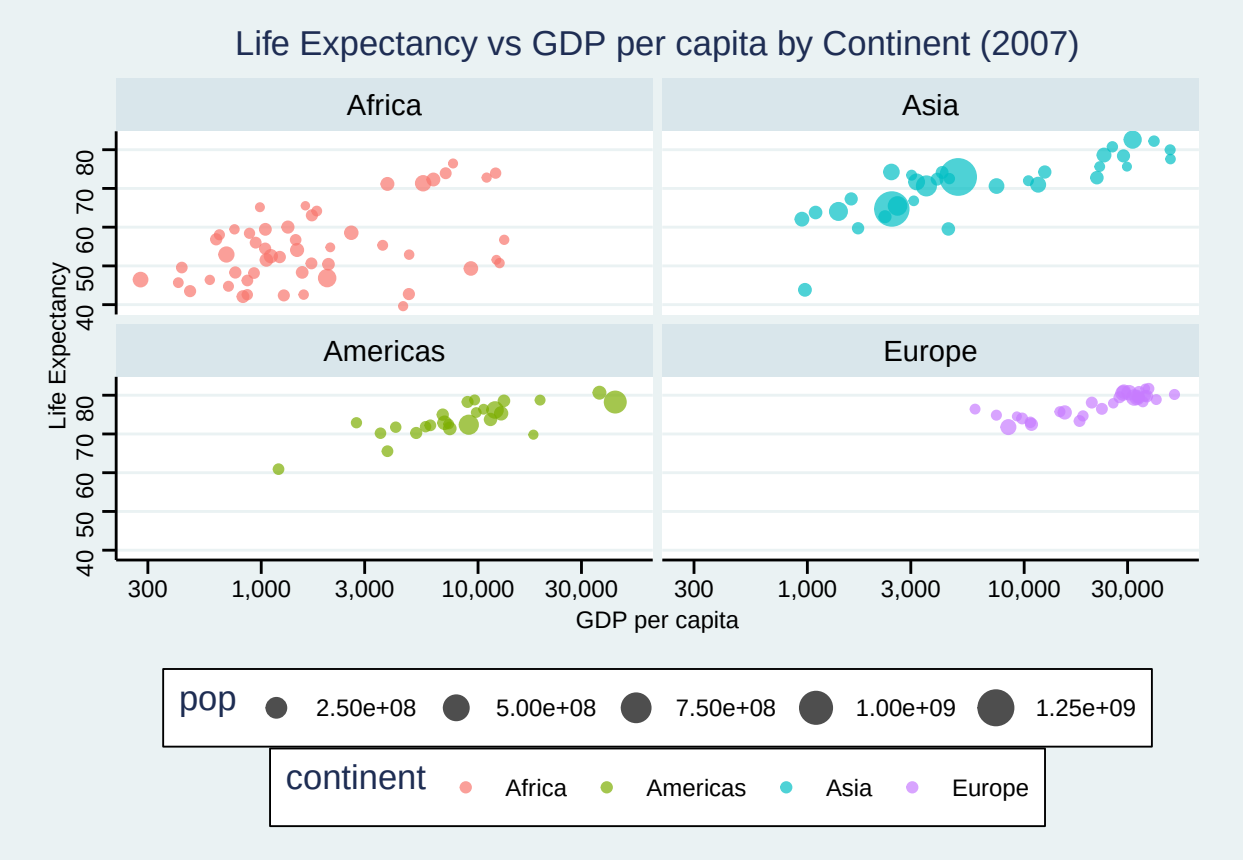
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp)) +
  geom_point()
```



Answer:

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp, color = continent)) +
  geom_point(aes(size = pop))
```

b. The partial code used to generate the plot below is given with placeholders for code snippet. Please provide the appropriate code snippet.




```
gapminder %>%
  ##### FILL IN i. ##### %>%
  ##### FILL IN ii. ##### %>%
  group_by(continent) %>%
  mutate(avg_lifeExp = mean(lifeExp)) %>%
  ungroup() %>%
  ##### FILL IN iii. ##### %>%
  ##### FILL IN iv. ##### +
  geom_point(alpha = 0.7) +
  labs(title = "Life Expectancy vs GDP per capita by Continent (2007)",
       x = "GDP per capita",
       y = "Life Expectancy") +
  scale_fill_brewer(palette = "Set1") +
  scale_x_log10(labels = scales::comma) +
  ##### FILL IN v. ##### +
  theme(legend.position = "bottom")
```

i. Create a new column called "year_date" that converts the "year" column into a date object

ii. Filter the data to only include rows where the year is 2007 and exclude continent "Oceania"

iii. Reorder the levels of the "continent" factor based on the "avg_lifeExp" column and store the result in a new column called "continent_reordered."

iv. Create a ggplot2 scatter plot with "gdpPercap" on the x-axis, "lifeExp" on the y-axis, point sizes representing the "pop" column, and points color-coded by the "continent" column.

v. Create a faceted plot based on the "continent_reordered" column, with 2 columns of panes.

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

```
x <- 4:1
y <- c(TRUE, factor(c(NA, "b")), 1)
z <- list(z1 = x, z2 = y, z3 = c("Carleton", "college"), z4 = matrix(1:9, nrow = 3))
```

Consider the above objects to tell what each of the following code chunks evaluate to? Briefly explain your answer.

(a)

```
y[3]
```

(b)

```
z[["z3"]][1]
```

(c)

```
z[x][[2]][[2]]
```

(d)

```
x - y
```

(e)

```
unlist(z)
```

(f)

```
typeof(unlist(z))
```