STA035B Homework 3, due: 2/15, 9pm

Spencer Frei

Instructions

Upload a PDF file, named with your UC Davis email ID and homework number (e.g., sfrei_hw3.pdf), to Gradescope (accessible through Canvas). You will give the commands to answer each question in its own code block, which will also produce output that will be automatically embedded in the output file. All code used to answer the question must be supplied, as well as written statements where appropriate.

All code used to produce your results must be shown in your PDF file (e.g., do not use `echo = FALSE` or `include = FALSE` as options anywhere). Rmd files do not need to be submitted, but may be requested by the TA and must be available when the assignment is submitted.

Students may choose to collaborate with each other on the homework, but must clearly indicate with whom they collaborated.

Gapminder

We will be using the `gapminder` dataset; you can load this by installing the package `gapminder` and then loading it as a library. Inspect the tibble by typing `?gapminder` in the console.

```r
library(gapminder)
```

In this homework we aim to plot the minimum, median, and maximum life expectancy, population, and gdp per capita per continent over time. We need to first do a bit of data cleaning and preparation, and then we can use the power of ggplot.

Part (a)

Write a function `summary_stats` which takes in a tibble and a variable name and returns a tibble with the following summary statistics for the variable:

- minimum
- maximum
- median

Allow for an argument, `na.rm=`, which will specify whether or not the computation of these summary statistics will remove NA's or return NA's for any NA's in the computation. The function must use the `summarize()` function with `.groups = "drop"`.

```r
summary_stats <- function(data, var, na.rm=TRUE) {
  data |> summarize(
    min = min({{ var }}, na.rm = na.rm),
    median = median({{ var }}, na.rm = na.rm),
    max = max({{ var }}, na.rm = na.rm),
    .groups = "drop"
  )
}
```

If your code is correct, running `summary_stats(flights, dep_time)` should return

min	median	max
1	1401	2400

Part (b)

Using `across()` and `summary_stats()` from above, compute the minimum, median, and maximum for each of the columns `lifeExp`, `pop`, and `gdpPercap` per year and per continent (i.e., min/median/max of each of these variables per year and per continent - you are doing these operations over different countries in every year-continent pair.). Save the resulting tibble as `gapminder_summary`, and print the first few rows of the tibble by writing `gapminder_summary`.

```r
(gapminder_summary <- gapminder %>%
  group_by(continent, year) %>%
  summarize(
    across(
      c(lifeExp, pop, gdpPercap),
      list(min = min, median = median, max = max)
    )
  )
)
```

```
`summarise()` has grouped output by 'continent'. You can override using the
`.groups` argument.
```

```
# A tibble: 60 x 11
# Groups:   continent [5]
   continent  year lifeExp_min lifeExp_median lifeExp_max pop_min pop_median
   <fct>     <int>       <dbl>          <dbl>       <dbl>   <int>      <dbl>
 1 Africa     1952        30             38.8        52.7   60011   2668124.
 2 Africa     1957        31.6           40.6        58.1   61325   2885790.
 3 Africa     1962        32.8           42.6        60.2   65345   3145210
 4 Africa     1967        34.1           44.7        61.6   70787   3473692.
 5 Africa     1972        35.4           47.0        64.3   76595   3945594.
 6 Africa     1977        36.8           49.3        67.1   86796   4522666
 7 Africa     1982        38.4           50.8        69.9   98593   5668228.
 8 Africa     1987        39.9           51.6        71.9  110812   6635612.
 9 Africa     1992        23.6           52.4        73.6  125911   7140388.
10 Africa     1997        36.1           52.8        74.8  145608   7805422.
# i 50 more rows
# i 4 more variables: pop_max <int>, gdpPercap_min <dbl>,
#   gdpPercap_median <dbl>, gdpPercap_max <dbl>
```

Part (c)

Make the gapminder_summary table in "long" format so that we have variable names continent, year, lifeExp, pop, and gdpPercap, and measurement, where measurement is either "min", "max", or "median". Call the tibble gapminder_summary_long, and print the first few rows of the tibble.

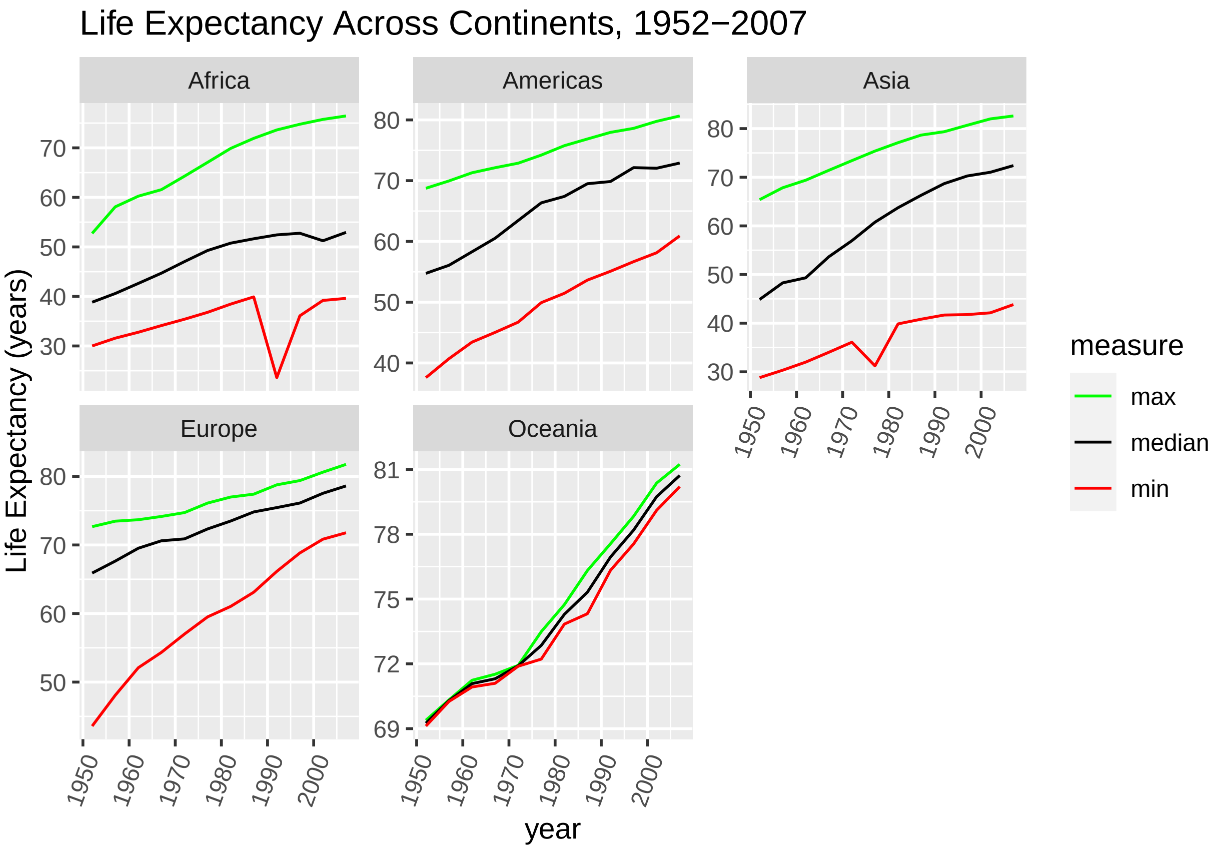
```
gapminder_summary_long <- (gapminder_summary %>%
   pivot_longer(cols = lifeExp_min:gdpPercap_max,
                names_to = c(".value", "measure"),
                names_sep = "_"
   )
)
```

Part (d)

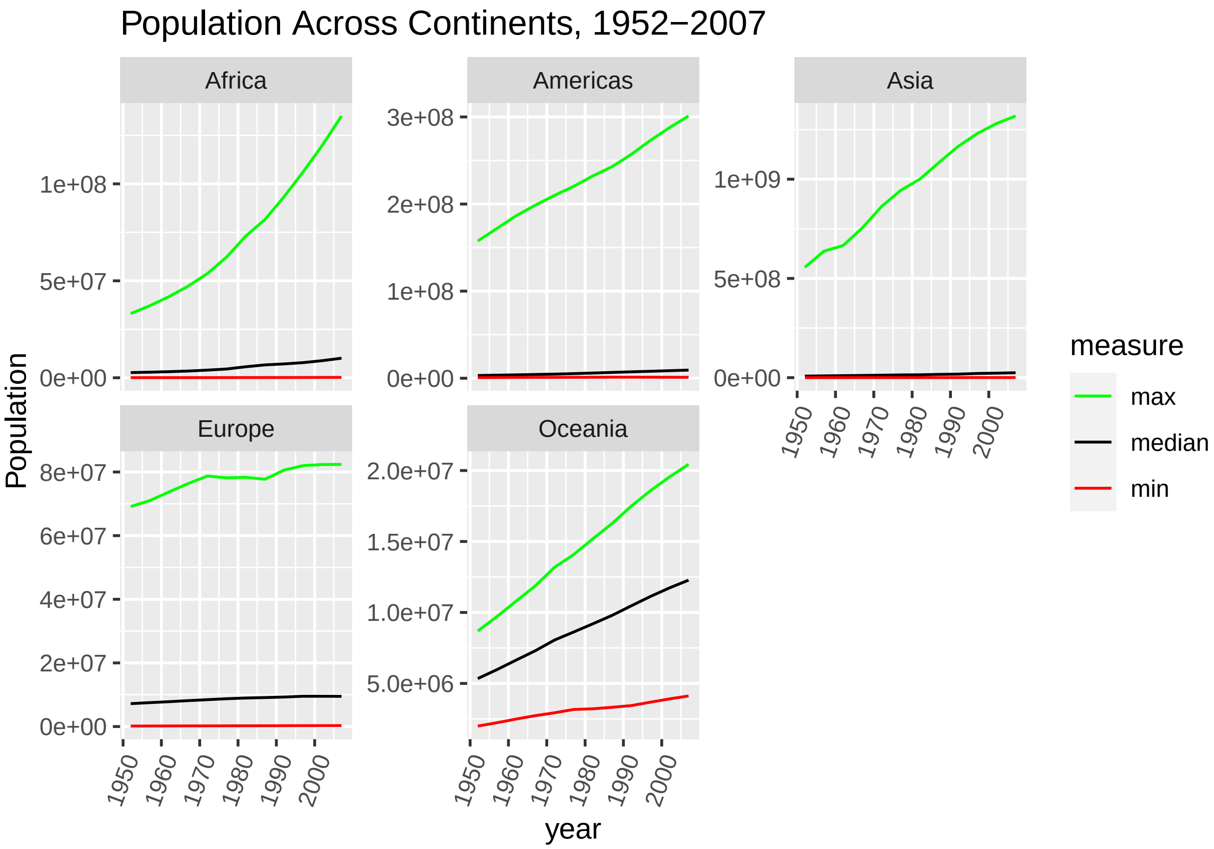
Create three ggplots, one for each of life expetancy, population, and gdp per capita. In each plot, we want to have 5 sub plots (using facet wrap or facet grid), one for each of the continents. In each subplot, we want three lines: the minimum of the variable (either lifeExp / pop / gdpPercap), the median of the variable, and the maximum of the variable. Do this by creating a function `plot_min_med_max()`, which takes as its input the variable name, y-axis label, and a plot title and returns a ggplot which has the 5 subplots in it. Your final 3 plots can be created by calling this function three times with inputs `lifeExp`, `pop`, and `gdpPercap` (no quotes!). Ensure that the plots have the following properties:

- The x-axis should be the same across subplots in a given plot, but the y-axis should scale separately for each sub-plot so that it is easy to visualize what is happening within each continent
- The y-axis should be a human readable form: not "lifeExp", "pop", or "gdpPercap", but something like "Life Expetancy (in years)", "Population", etc.
- There is a title to the plot which describes what the subplots describe at a high-level
- The minimum line should be in red, the median in black, and the maximum in green.
- The function `plot_min_med_max()` takes 3 arguments: variable name (not a string, just a sequence of characters), y axis label (a string), and a title (a string).
- All x-axis and y-axis labels are human-readable; you may need to adjust font sizes or the orientation of the labels (look up the `theme()` function, and inspect the `axis.text.x` argument and use the function `element_text()`) to do this.

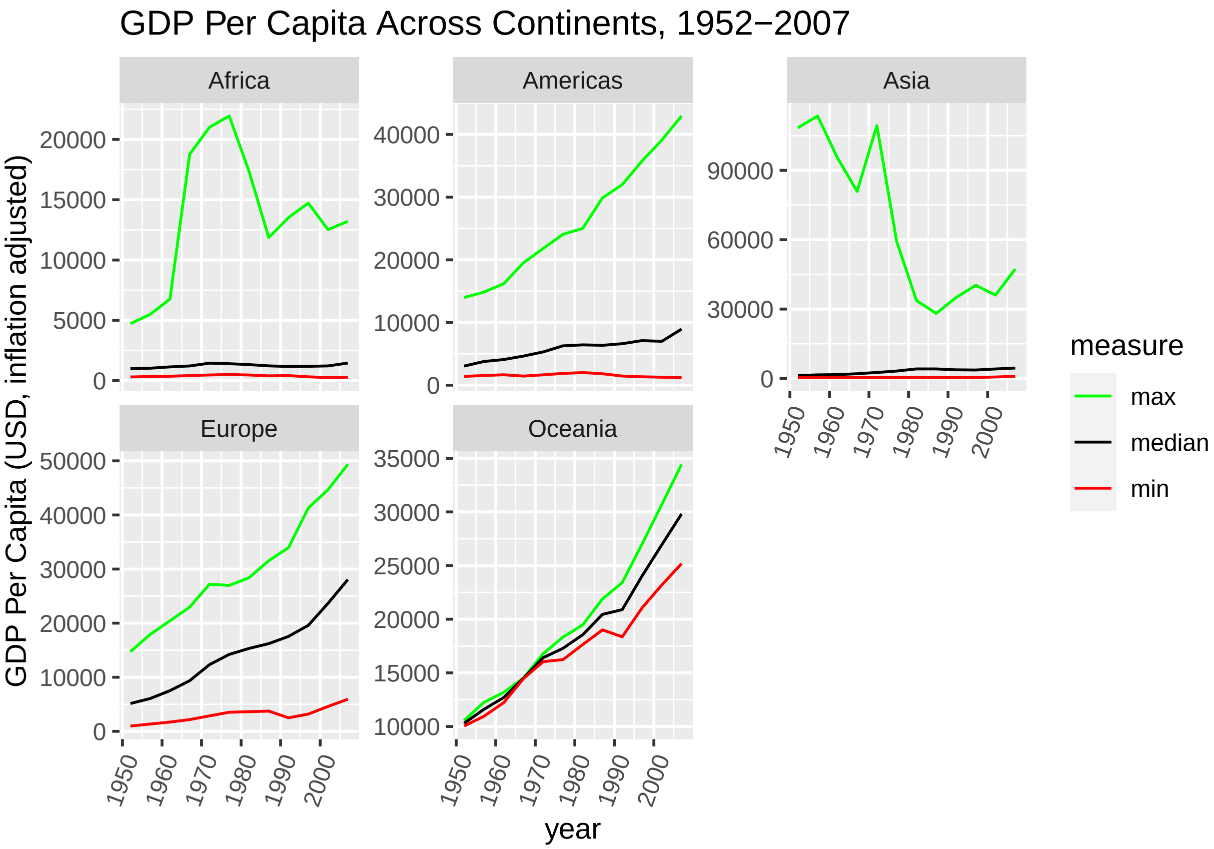
```r
plot_min_med_max <- function(variable, title, ylab) {
  gapminder_summary_long %>%
    ggplot(aes(x = year, y = {{variable}}, color = measure)) +
    geom_line() +
    facet_wrap(~continent, scales = "free_y") +
    scale_color_manual(values = c("min" = "red", "median" = "black", "max" = "green")) +
    labs(y = ylab,
         title = title) +
    theme(axis.text.x = element_text(angle = 70, hjust = 1))
}
plot_min_med_max(lifeExp, "Life Expectancy Across Continents, 1952-2007", "Life Expectancy (years)")
```



```r
plot_min_med_max(pop, "Population Across Continents, 1952-2007", "Population")
```



```
plot_min_med_max(gdpPercap, "GDP Per Capita Across Continents, 1952-2007", "GDP Per Capita (USD, inflati
```


GDP Per Capita Across Continents, 1952−2007
Africa
Americas
Asia
Europe
Oceania
GDP Per Capita (USD, inflation adjusted)
0
5000
10000
15000
20000
0
10000
20000
30000
40000
0
30000
60000
90000
0
10000
20000
30000
40000
50000
10000
15000
20000
25000
30000
35000
1950
1960
1970
1980
1990
2000
year
measure
max
median
min

Part (e)

Comment on any unusual trends that you notice in the life expectancy across continents. Which (country, year) pairs are responsible for any apparent outliers? Do you have any hypotheses for what might have caused them? Provide any R code which you use to do this.

```
gapminder %>%
  filter(continent == "Africa",
         between(year, 1990, 1995),
         lifeExp < 30
         )
```

```
# A tibble: 1 x 6
  country continent  year lifeExp     pop gdpPercap
  <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
1 Rwanda  Africa     1992    23.6 7290203      737.
```

```
gapminder %>%
  filter(continent == "Asia",
         between(year, 1975, 1980),
         lifeExp < 35
         )
```

```
# A tibble: 1 x 6
  country  continent  year lifeExp     pop gdpPercap
  <fct>    <fct>     <int>   <dbl>   <int>     <dbl>
1 Cambodia Asia       1977    31.2 6978607      525.
```

Rwanda in 1992 and Cambodia in 1977. These were years where genocides occurred in these countries.