Coding Challlenge 3

Public Health 460

Due: Febuary 18th 8:00 pm, 2022

This assignment is all about visualization using ggplot. Please turn in an HTML file and your .Rmd file

- Download the data called `gapminder.csv` from Moodle (Unit3 > Unit 3 readings and files folder).
- Use `ggplot2` and its functions for every question about plotting. Do not use the `qplot()` function.
- Include your code in your output.

The first set of questions revolve around a dataset from Gapminder.org. It includes data on life expectancy (in years) and GDP per capita of different countries.

Answer the questions below for credit

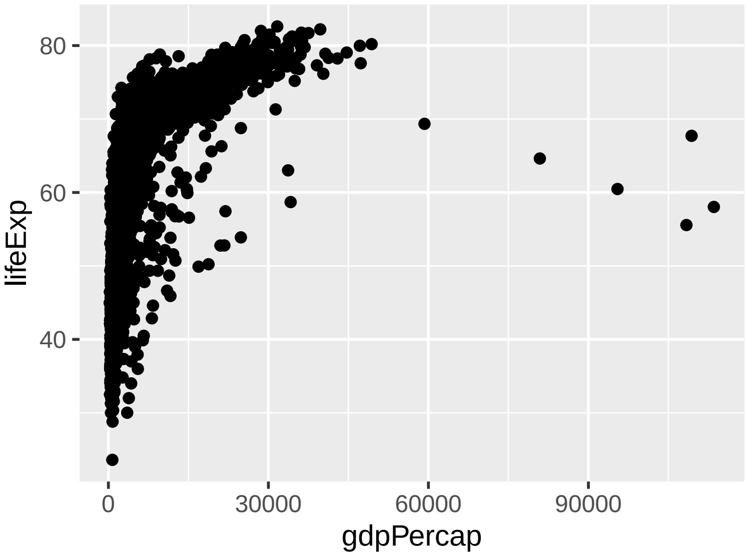
Part I: Gapminder dataset (10 points)

1) Read in the `gapminder.csv` from a relative file path. Load the `tidyverse`, `DT`, and `plotly` packages. (If you have not installed `DT` or `plotly` yet, run the code to install them once on your machine, but do not include that code to install them in your assignment.)

2) Use `View()`, `print()` or just click on the dataset in your "Environment" pane in RStudio to inspect the dataset. What is the unit of observation of this dataset? How are the samples distributed in time?

3) Use the `DT::datatable()` function to present an interactive table in your HTML file. [Note that we use the `::` notation here to show that you should use the `datatable()` function from the `DT` package. You can use this syntax to call a single function directly without having run `library(DT)` (as long as you have the `DT` package installed on your machine). This syntax can also be helpful when two loaded packages have functions with the same name, as you can specify exactly which function you want to use. For the purposes of the assignment you can use either method to use the `datatable()` function.]

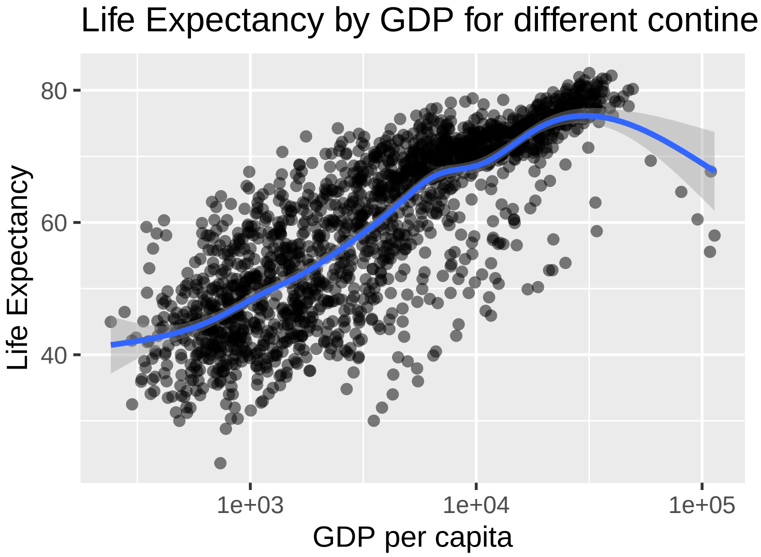
4) Create a simple scatter plot of life expectancy as the dependent variable by GDP per capita. Do not worry about titles or labels beyond what is given as the default. The plot should look like the following figure.



5) Using the same data as the graph above make the following adjustments to the plot

- Label the y-axis as "Life expectancy" and the x-axis as "GDP per capita".
- Add a title to your graph.
- Transform the variable GDP per capita on the x-axis, using a log-transform while still displaying the unlogged GDP values as axis labels.
- Add a smooth regression line using `geom_smooth()`.
- Make all the points slightly transparent so you can see individual points more clearly. Make sure that the transparency does not apply to the smooth line layer of the graphic.

It should look something like the graph shown below



6) Pick a new `ggplot` theme and use `theme_set()` to set a new theme for the remaining figures for this assignment. Re-plot the figure from question 5 with the new theme applied.

7) Build off of the previous graph, to examine each continent separately. First, facet the graph by continent as was shown in the SAT lecture. This way each continent has its own smaller graph next to the other continents. Second, remove the "error bars" from the smooth lines.

8) Pick 5 countries that you are interested in comparing. Create a new data frame that has only observations from those 5 countries. Build off of the figure created in question 5 (i.e. do not use facets). In this plot, color your points by country. Also, connect the points showing the trajectory of each country across time. (Hint: to connect the points use the `group` aesthetic and `geom_line()`.)

9) Use the `plotly::ggplotly()` function to create a quick interactive version of the figure created in the previous question.

Part II: Visualization taxonomy (4 points)

10) Chapter 2 of Modern Data Science with R discusses a taxonomy that can be used to describe data graphics. Choose a data graphic from https://flowingdata.com/ or another data journalism website. Copy the graphic or take a screenshot of the figure and insert the figure in your RMarkdown file. Answer the following questions:
    a) Identify the visual cues, coordinate system , and scale(s).
    b) Describe the variables visualized by the graphic. What are they? How do they relate to each visual cue you identified in the previous question?

Part II: Religious income dataset (6 points)

11) Read in the `relig_income` data set that is contained in the `tidyverse` package. This dataset provides partial results from the Religious Landscape Study conducted by the Pew Research Center. As provided, the dataset provides a cross-tabulation of the number of respondents who identified as being a member of each religious group with their reported income range.
    a) How many individuals' responses are included in this dataset? (Use code to calculate the answer. There is no specific "right" way to do this that we are looking for.)
    b) The `pivot_longer()` function allows a data analyst to make data "tidy" by taking a dataset where the values of one variable are encoded in column names. That is the situation here because the income variable is not a single value. It is called `pivot_longer()` because it increases the number of rows and decreases the number of columns. (There is a corrsponding `pivot_wider()` function as well.) Use `pivot_longer()` to create a single column for the variable `income`, rather than having a column for each level of income. Remove the "Don't know/refused" category for income. Select the Atheist, Buddhist and Muslim religious groups to plot.
    c) Plot income on the x-axis and the count of individuals on the y-axis using points and lines and a different color for each religious group. You will need to reorder the factor levels of income so that they are are in increasing order on the x-axis.
    d) Someone looks over your shoulder at the resulting plot and says: "Wow, atheists make more money than Buddhists, huh?" What is one problem with the statistical logic that they used to draw that conclusion?