Assignment 3

Prof. Ashish Anand
CS595: Data Visualization
Due Date: March 14, 2023
Posted on: February 21, 2023

Outline

Often tabular data has implicit grouping, for example, data having several financial statistics of the last 5 years for all states in India. Now the respective two columns are state names and years. If we want to visualize one particular attribute over the five years for all states together or a single state in a single plot, without creating another tabular data, we have to use the concept of **grouping**. Basically, you are utilizing categorical variable `state` to group all rows with the same label.

Similarly, if we want to visualize one particular attribute for each state separately but keep those plots next to each other or in a grid fashion, then we have to use the concept of **faceting**.

In this assignment, we will continue to focus on exploratory data analysis through various plots. Some of the relevant plots come from the family of visualizing density or distribution of continuous variables. **Histograms**, **density**, **scaled density** are the examples of such plots.

The objectives of the assignment are the following:

- Introduce the two concepts while summarizing and visualizing tabular data: Grouping and Faceting.
- Extend your understanding of features of the chosen library (ggplot2 or seaborn or any other)
- Use of grammar of graphics concepts as much as possible to generate your plots.
- Learn the importance of **Data Cleaning** and **Data Normalization**. There may be some noisy or missing data. To keep only meaningful data, you have to do Data Cleaning. Similarly, for certain plots, you have to do appropriate normalization.

  Do make sure you mention all the steps of Data Cleaning and/or Data Normalization if they are performed.

References:

1. Chapter 4. Data Visualization: A Practical Introduction. Kieran Healy

2. Notes on Grammar of Graphics shared in MS Teams group
3. Section II: Data Visualization, Introduction to Data Science
4. A Grammar of Graphics for Python

Dataset

Please download the data country_profile.csv from the Assignment 3 directory in the Files section of MS Teams group.

The dataset contains information on several economic, social, and environmental factors for all countries along with their region information.

Questions

Question 1. [10 points] We want to compare the GDP of countries. The relevant variable is named as **GDP: Gross domestic product (million current US$)**. Use **box-plot** or equivalent appropriate plot to compare GDP distribution *region-wise*. **Region** is a categorical variable in the data.

Make sure to discuss Data Cleaning and/or Data Normalization steps if they were required to perform.

Discuss, whether keeping the Region variable on *x-axis* or *y-axis* makes any difference aesthetically.

Question 2. [10 points] Include an extra categorical variable indicating **continents**. Repeat question 1 by comparing the GDP distribution continent-wise.
Note the difference between continents and regions

Question 3. [10 points] Repeat question 1, i.e., comparison of the GDP distribution. However, this time, apply **Faceting** to generate plots, where each plot compares the GDP distribution of regions within a continent. In other words, **faceting** is done on the **continent** variable.

Question 4. [10 points] Compare the different continents based on employment in the different sectors.

The relevant variables are **Employment: Agriculture (% of employed)**, **Employment: Industry (% of employed)**, and **Employment: Services (% of employed)**.

Question 5. [10 points] We want to understand the **expenditure on Health (% of GDP)** in different regions of each continent. Discuss that through an appropriate visualization.