Assignment 2

Ashish Anand, Akshay Parekh

CS595: Data Visualization

Due Date: February 9, 2020

January 29, 2020

1 Outline

Often a tabular data has implicit grouping, for example a data giving 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 each state in a single plot, without creating another data, we have to use the concept of **grouping**.

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

Objective of the assignment is the following:

- Introduce two concepts while summarizing and visualizing a tabular data: Grouping and Faceting.
- Extend your understanding of ggplot2 (R library for creating visualizations) features

Reference: **Data Visualization: A Practical Introduction. Kieran Healy**

2 Datasets

In this exercise, we will be working on the following datasets.

- gapminder
  - Install:: install.packages(“gapminder”)
  - Load :: library(gapminder)
- gss_sm (part of library socviz)
  - Install:: install.packages(“socviz”)
  - Load :: library(socviz)

3 Questions

Question 1. [20 points] Load the data gapminder and analyze different columns of data. Plot life expectancy over time for each country. Do you think the plot is meaningful? Justify your answer in markdown.

Hint: Explore **aes()** along with **geom_line()**.

Question 2. [20 points] As discussed in section1, *faceting* the data using extra variable means making small multiple plots for the same data.

I. Can you improve the plot from the first question by faceting the data? Which is the most appropriate variable (column) to facet the data? Plot and justify.
Hint: Explore **facet_wrap()**

II. We can facet the data based on more than one variable, Use *gss_sm* data to plot a *smoothed scatter plot* showing the relationship between the age of the respondent and the number of children they have. Facet this relation based on race and degree. Also, in markdown, describe your observation from the plot in brief.
Hint: Explore **facet_grid(), geom_smooth()**

Question 3. [10 points] *gss_sm* data contains the political view *(polviews)* variable. Plot a bar graph, with the bar in the chart represented by different political views.

Question 4. [10 points] Again using *gss_sm* data, visualize the *frequency plot*, with the bars representing different political views and each bar in the graph, is further categorized by a different religion. Also, visualize frequency plot faceted by variable *bigregion*.

Question 5. [40 points] Plot following from the *gss_sm* data.

I. Histogram. Showcasing *ages* into bins.

II. Modify the previous histogram, dividing the observation based on variable *race*. Each *race* should be represented in a different color.

III. 2 Stacked Density plots, dividing the observation based on the variable *race*:

- In x-axis, choose variable *age*
- In x-axis, choose variable *agegrp*.

Discuss which variable among the two *age* and *agegrp*, is more appropriate for plots like a *histogram* and *density plots* and why?

IV. Density plot

- x-axis: *age*.
- Division based on variable *degree*.
- Faceting based on variable income16.

Hint: Explore **geom_histogram(), geom_density()**