Assignment 3: Part I

Ashish Anand, Akshay Parekh
CS595: Data Visualization

Due Date: March 9, 2020

February 23, 2020

Outline

The third assignment will continue to explore the role of data visualization in exploratory data analysis. While the first two assignments focus on visualization of raw numbers of data, their distributions and how to visualize implicit grouping of data. This assignment primarily focus on summarizing/transforming the data through statistical models.

We often start data exploration by trying to understand implicit trend or relation among variables. Smoothing or curve-fitting are first few things to try to achieve that. We divide this assignment into two parts. This the first part and the second part will be given after mid-sem.

Objectives of the assignment are:

- How to use various smoothing functions? If you are using ggplot, then various smoothing functions can be called from the function `geom_smooth`. The function `geom_smooth` can call a range of regression models including *OLS*, *robust regression (rlm)*, *LOESS* etc to produce fit. [**Part I**]
- How to produce multiple fits in one plot with different colors and appropriate legends? [**Part I**]
- How to use visualization as a tool in model checking and validation? [**Part II**]

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

Datasets

In this exercise, we will again use *gapminder* dataset.

Questions

Question 1. In the previous exercise, we have used `geom_smooth()` function for smoothing the curve. In that function we can pass various smoothing methods.

- Using the *gapminder* dataset, plot a scatter plot with *log(gdpPercap)* on x-axis and *lifeExp* on y-axis. Generate a graph with following different smoothing methods (**Add legend** for clear distinction of each smooth curve) [15 points]

  1. use *lm* to fit three degree polynomial curve
  2. use *glm* [glm: generalized linear model]
  3. use *gam* with $y = \log(x)$ for smoothening [gam: generalized additive model]
  4. use *rlm* with $y = \text{poly}(x, 2)$ for smoothening [rlm: robust linear regression model]
  5. use *loess*

- In your report (markdown format), explain the following: [15 points]

  1. smoothing
  2. above mentioned different smoothing methods
  3. Which of these methods cannot be used for a large dataset and why?