Assignment 3: Part II

Ashish Anand, Akshay Parekh
CS595: Data Visualization
Due Date: March 20, 2020

March 12, 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 using different methods. Train a model using those methods following, [20 points]

- Divide the *gapminder* dataset based on column `Continent`. Training data will incorporate data from `1952` to `1992` for each continent and for testing/prediction use data from `1993` to `2007`.

  1. Train the linear model, where column `gdpPercap` is input data and predicted value is Column `lifeExp`. In R Markdown, print the summary of the learned model. Based on the data, we propose a hypothesis, *"lifeExp is dependent on gdpPerCap"*. Analyze the summary of model and explain, Is Hypothesis correct?
  2. Use the data from summary and plot a histogram to show how well does your model fits the data.
  3. Plot a graph showing the prediction on test data points, linear regression line fitting the data. Also draw the area covered by prediction intervals. Facet the data based on continents.

HINT: Explore `lm()`, `predict()`, `geom_ribbon()`, `library(broom)`:: `filter()`