CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 1: Introduction to Data Science

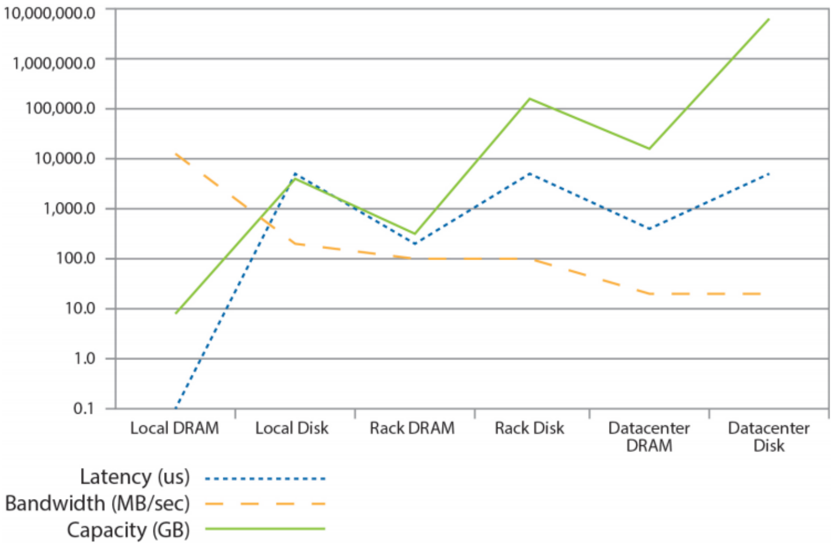
1. In the lecture, we have learnt the following plot on “The Cost of Moving Data Around Data Center”. Answer the following questions.

a) Why the latency of Rack DRAM is much higher than that of Local DRAM?

b) Why the latency of Local Disk is similar to that of Rack/Data Center Disk?

c) If we have an application with the *working set* smaller than Local DRAM, should we always put the working set into a single machine for efficiency? Justify your answer.

d) Since DRAM at all levels has a much better performance than Disk in terms of both latency and bandwidth, should we always put the data into DRAM? Justify your answer.



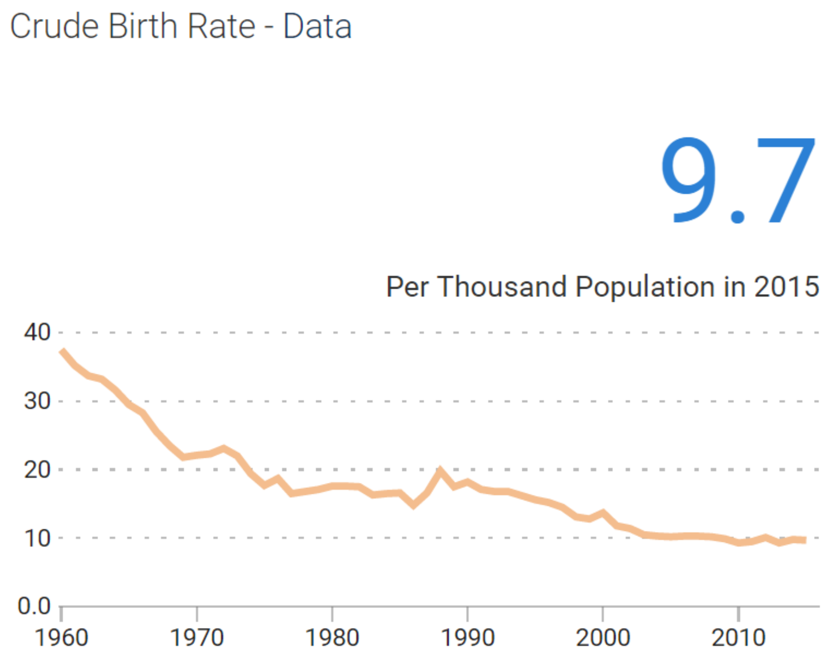
2. "A picture is worth a thousand words". An important aspect of big data is the presentation of the answers. In this question, we examine a number of plots from https://data.gov.sg/, which is an effort of public data analytics in Singapore. For each plot, write down your findings, share/discuss your findings with peer classmates and try to dig out more insights from discussion.

a) Crude birth rate

b) GDP at current market prices

c) Graduates from university first degree courses

d) Rainfall - monthly total

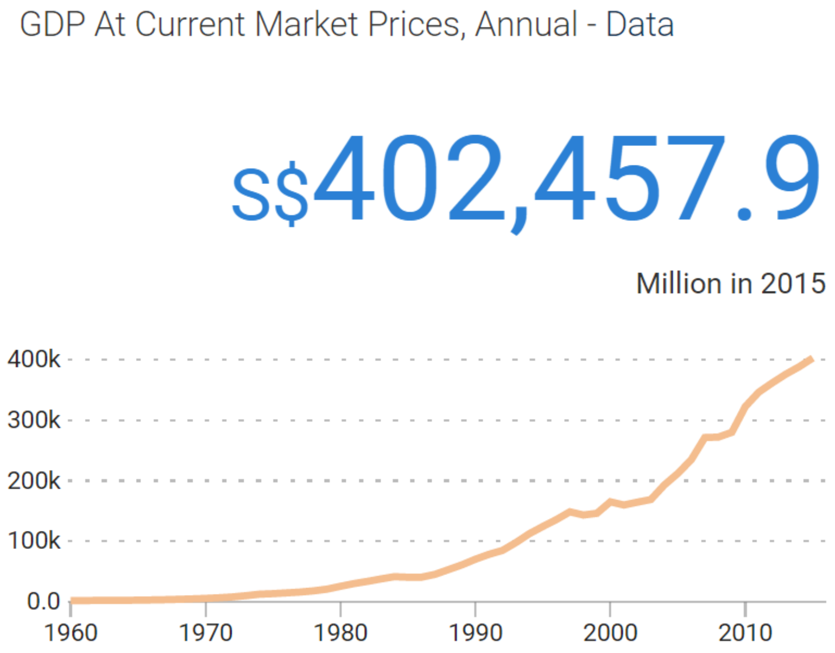


Crude Birth Rate - Data

9.7

Per Thousand Population in 2015

Figure (a)



GDP At Current Market Prices, Annual - Data

S$402,457.9

Million in 2015

Figure (b)

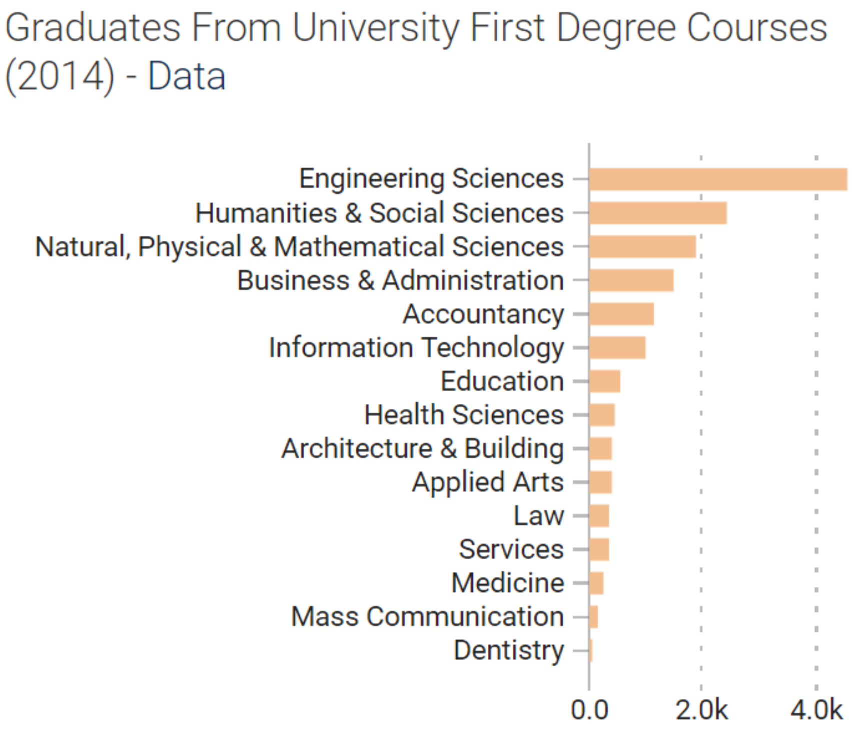


Figure (c)

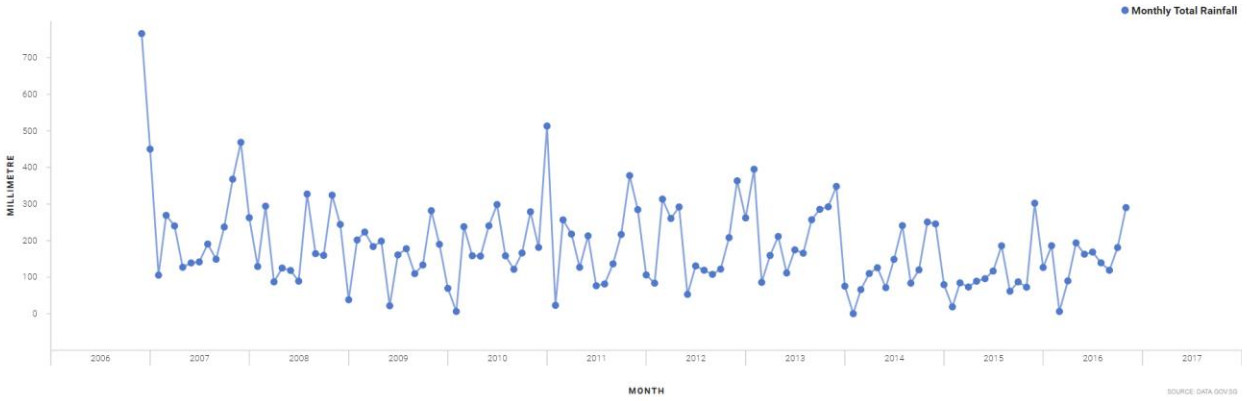


Figure (d)