**Unit 6 Reading  
Sarkar, T. & Roychowdhury, S. (2019) Data Wrangling with Python. 1st ed. Packt.**

**Chapter 9 Application of Data Wrangling in Real Life**

Data is often inadequate for final analysis requirements, particularly if from one source. Have to distinguish which is relevant and non-relevant from different sources. Primary role of data wrangler is to pull data from various sources, format and clean it and combine ready for further analysis.

A data wrangling exercise to prepare the data to answer the following question is provided: **In India, did the enrolment in primary/secondary/tertiary education increase with the improvement of per capita GDP in the past 15 years?** Activity requires combing data from the UN with data from the world bank in a CSV, both have some missing and unnecessary data. Data will then be saved as a SQL database file.

Process involves importing data into a Pandas dataframe from two sources. Checking and removing redundant data and useful data. Filtering the dataframes based on these. Plotting to check for gaps. Filling those gaps (data cleaning) and finding solutions via google, a common practice when coding – borrow with pride!

The additional data is then imported, cleaned and filtered to find relevant data, it was also transposed as was displayed in columns for years. Data types also converted. Then the two datasets were merged into one dataframe.

This data frame was exported to a SQL database file using sqlite

**Reflection:**

This activity showed a full data wrangling flow and very practical application for pandas for preparing and making data fit for analysis. Simple reminders of checking data, removing redundancy, filling gaps, visualising and converting data types all shown. Writing back to a database file was a new method demonstrated to me.