Data

The data is sourced from https://data.gov.in/, an Open Government Data (OGD) platform of India. The download instructions are provided in the next segment. The data for GDP analysis of the Indian states is divided into two parts:

- Data I-A: This dataset consists of the GSDP (Gross State Domestic Product) data for the states and union territories.

- Data I-B: This dataset contains the distribution of GSDP among three sectors: the primary sector (agriculture), the secondary sector (industry) and the tertiary sector (services) along with taxes and subsidies. There is separate dataset for each of the states. You are expected to read the dataset for the available states and join these (in Python) if needed.

There are two parts to this project. In the first part, you will analyse and compare the GDPs of various Indian states (both total and per capita). The GDP of a state is referred to as the GSDP (Gross State Domestic Product). Then, you will divide the states into four categories based on the GDP per capita, and for each of these four categories, you will analyse the sectors that contribute the most to the GDP (such as agriculture, real estate, manufacturing, etc.).

In the second part, you will analyse whether GDP per capita is related to dropout rates in schools and colleges.

Part-I: GDP Analysis of the Indian States

For each of the following steps of analysis, choose an appropriate type of plot for comparing the data. Also, ensure that the plots are in increasing or decreasing order for better comparison. For example, if you make a bar plot to compare the GDPs of the states, ensure that the bars are in either increasing or decreasing order of GDP.

Part I-A:

- For the analysis below, use the Data I-A.

- First, you need to load the data in Python properly and then clean it. This also involves the treatment of missing values, you can choose to drop the row or column as well. Remember this will affect your next analysis and results drastically.

- Plot a graph for rows " % Growth over previous year" for all the states (not union territories) whose data is available, use as much data as possible for this exercise. Use the best fit line to represent the growth for each state. Draw a similar line graph for the nation as well.
    - How will you compare the growth rates of any two states?
    -
    - Which states have been growing consistently fast, and which ones have been struggling? Rank top 3 fastest and 3 slowest-growing states.
    -
    - What is the Nation's growth rate?
    - What has been the growth rate of your home state, and how does it compare to the national growth rate?
- Plot the total GDP of the states for the year 2015-16:
    - Which Plot will you use for this? Why? (Remeber to plot the graph in a way such as it is easier to read and compare)
    - Identify the top 5 and the bottom 5 states based on total GDP.
    - What insights can you draw from this graph? What states are performing poorly? (Remember: this will not be solely based on total GDP)

Part I-B:

- For the analysis below, use Data I-B. You can also use Data I-B along with Data I-A if required. Also, perform the analysis only for the duration 2014-15.
- Filter out the union territories (Delhi, Chandigarh, Andaman and Nicobar Islands, etc.) for further analysis, as they are governed directly by the central, not state governments.
- Plot the GDP per capita for all the states.
    - Identify the top 5 and the bottom 5 states based on the GDP per capita.
    - Find the ratio of the highest per capita GDP to the lowest per capita GDP.
- Plot the percentage contribution of the primary, secondary and tertiary sectors as a percentage of the total GDP for all the states.
    - Which plot will you use here? Why?
    - Why is (Primary + Secondary + Tertiary) not equal to total GDP?
    - Can you draw any insight from this? Find correlation of percentile of the state (% of states with lower per capita GDP) and %contribution of Primary sector to total GDP.

- Categorise the states into four groups based on the GDP per capita (C1, C2, C3, C4, where C1 would have the highest per capita GDP and C4, the lowest). The quantile values are (0.20,0.5, 0.85, 1), i.e., the states lying between the 85th and the 100th percentile are in C1; those between the 50th and the 85th percentiles are in C2, and so on.
    - Note: Categorisation into four groups will simplify the subsequent analysis, as otherwise, comparing the data of all the states would become quite exhaustive.
- For each category (C1, C2, C3, C4):
    - Find the top 3/4/5 sub-sectors (such as agriculture, forestry and fishing, crops, manufacturing etc., not primary, secondary and tertiary) that contribute to approximately 80% of the GSDP of each category.

        - Note-I: The nomenclature for this project is as follows: primary, secondary and tertiary are named 'sectors', while agriculture, manufacturing etc. are named 'sub-sectors'.

        - Note-II: If the top 3 sub-sectors contribute to, say, 79% of the GDP of some category, you can report "These top 3 sub-sectors contribute to approximately 80% of the GDP". This is to simplify the analysis and make the results consumable. (Remember, the CEO has to present the report to the CMs, and CMs have limited time; so, the analysis needs to be sharp and concise.)

    - Plot the contribution of the sub-sectors as a percentage of the GSDP of each category.

Now that you have summarised the data in the form of plots, tables, etc., try to draw non-obvious insights from it. Think about questions such as:

- How does the GDP distribution of the top states (C1) differ from the others?
- Which sub-sectors seem to be correlated with high GDP?
- Which sub-sectors do the various categories need to focus on?
-

Ask other such relevant questions, which you think are important, and note your insights for category separately. More insights are welcome and will be awarded accordingly.

- Finally, provide at least two recommendations for each category to improve the per capita GDP.

Part-II: GDP and Education Dropout Rates

In Part-I, you would have noticed that (one) way to increase per capita GDP is by shifting the distribution of GDP towards the secondary and tertiary sectors, i.e., the manufacturing and services industries. But these industries can thrive only when there is an availability of educated and skilled labour.

In this part of the analysis, you will investigate whether there is any relationship between per capita GDP with dropout rates in education.

Data

Data II: This section will require the dropout rate dataset apart from the dataset that you used in Part-1 of the case study. Download instructions are provided in the next segment.

Part-II: GDP and Education

- Analyse if there is any correlation of GDP per capita with dropout rates in education (primary, upper primary and secondary) for the year 2014-2015 for each state. Choose an appropriate plot to conduct this analysis.
    - Is there any correlation between dropout rate and %contribution of each sector (Primary, Secondary and Tertiary) to the total GDP?
- You have the total population of each state from the data in part I. Is there any correlation between dropout rates and population? What is the expected trend and what is the observation?
- Write down the key insights you draw from this data:
    - Form at least one reasonable hypothesis for the observations from the data

Important note: All your code has to be submitted in one Jupyter Notebook. For every checkpoint, keep writing the code in one well-commented Jupyter Notebook, which you can submit at the end.