PGDDS - Statistics and EDA - Gross Domestic Product Assignment

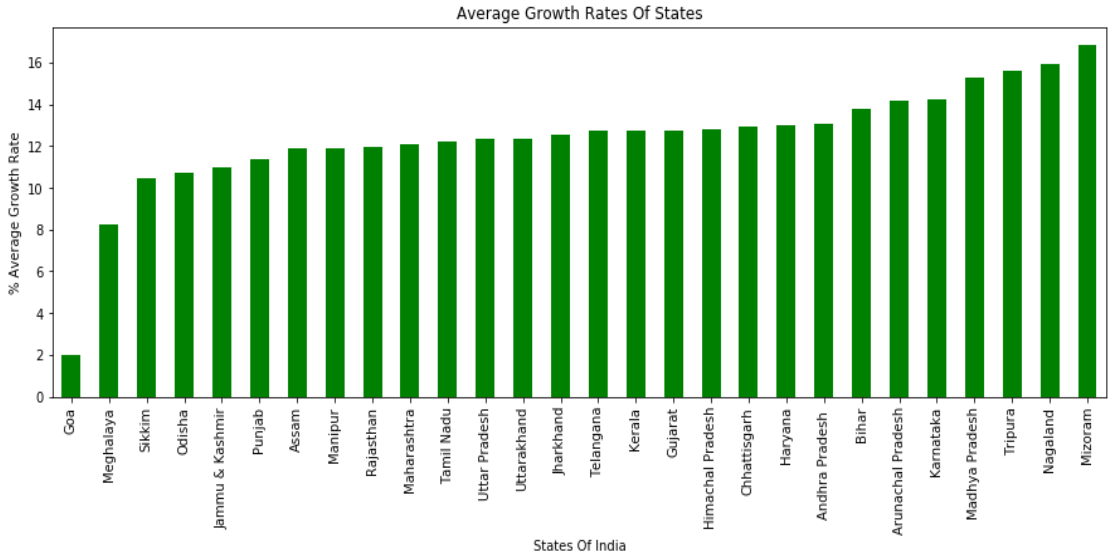
The overall goal of this project is to help the CMs focus on areas that will foster economic development for their respective states. Since the most common measure of economic development is the GDP, we need to analyse the GDP of the various states of India and suggest ways to improve it.

Part 1

Part 1a (Using Data-1A.csv)

1. Analyse and compare the GDPs of various Indian states
2. Divide the states into four categories based on the GDP per capita, and for each of these four categories, analyse the sectors that contribute the most to the GDP (such as agriculture, real estate, manufacturing, etc.)

Figure 1: Average Growth Rate of States



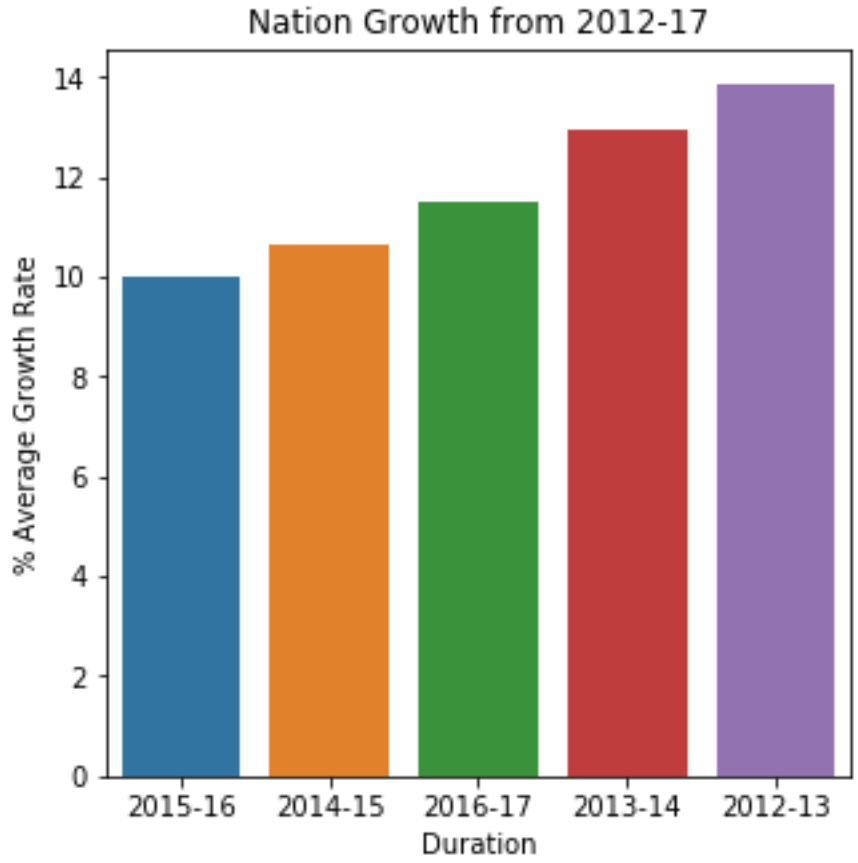
As seen in the above figure "Average Growth Rates Of States", we can compare the growth rate of states.

1. *The growth rate of Goa is the slowest whereas Growth rate of Mizoram is of the highest*
2. *The first three fastest growing states are*
    A. Mizoram
    B. Nagaland

C. Tripura

3. *The least three growing states are*
   A. Goa
   B. Meghalaya
   C. Sikkim

Below figure shows the national growth rate from 2012-2017



The growth rate of Nation is 11.8 and the growth rate of my home state Tamil Nadu is 12.25

My home state Tamil Nadu grows at 1.04 % to that of nation

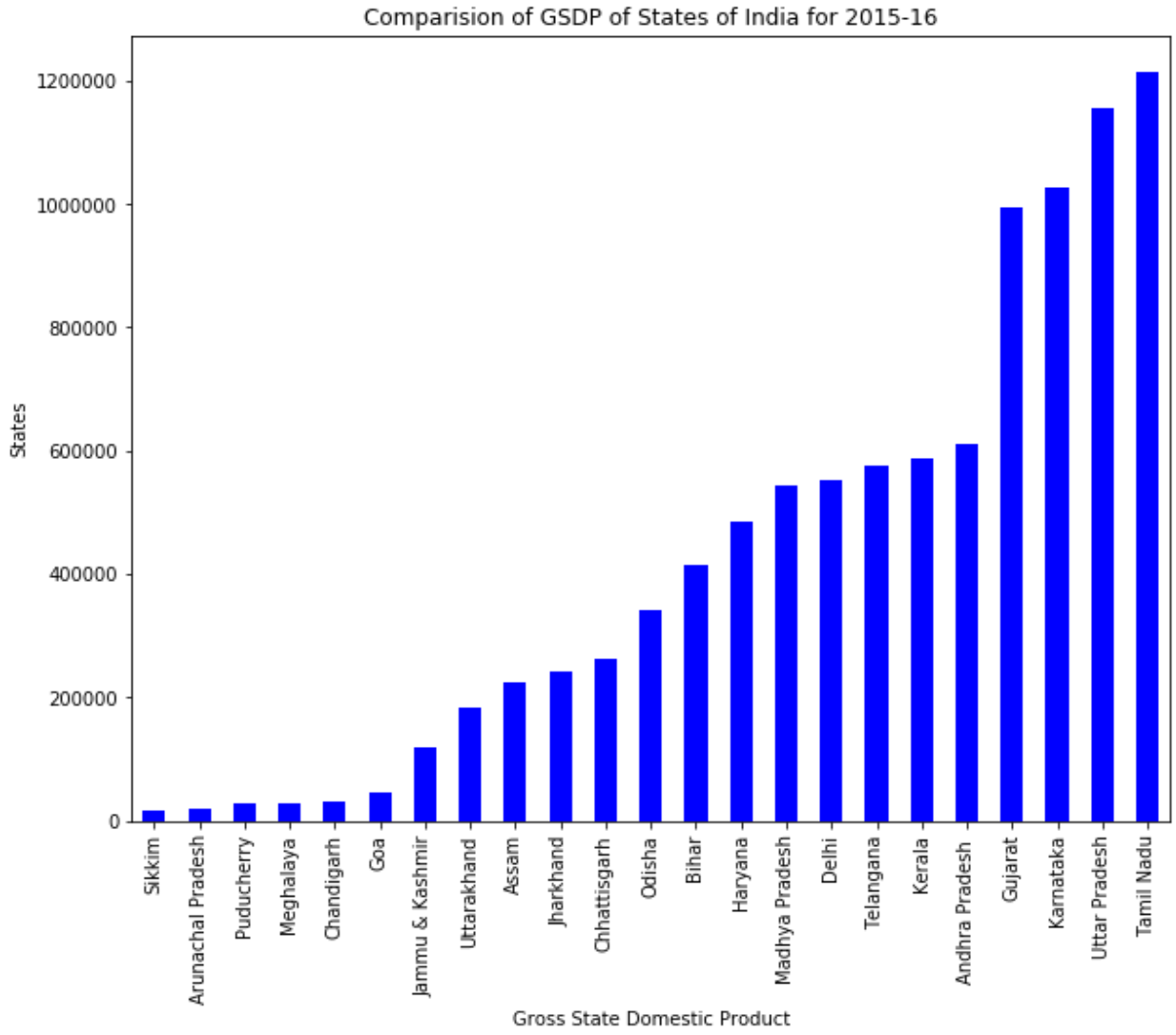
Plot the total GDP of the states for the year 2015-16

Steps Performed:

1. The *original data frame df is considered* as filtering the union territories is not expected for this part
2. iloc[:,2:-1] is used to *remove the first two columns* and consider the remaining columns
3. The result is *transposed* where in rows and columns are exchanged

4. The result is converted into a series by selecting the last column using: iloc[:,-1]
5. The output is then sorted in increasing order
6. Finally the null values are dropped using *dropna()*

Comparision of GSDP of States of India for 2015-16



Insights

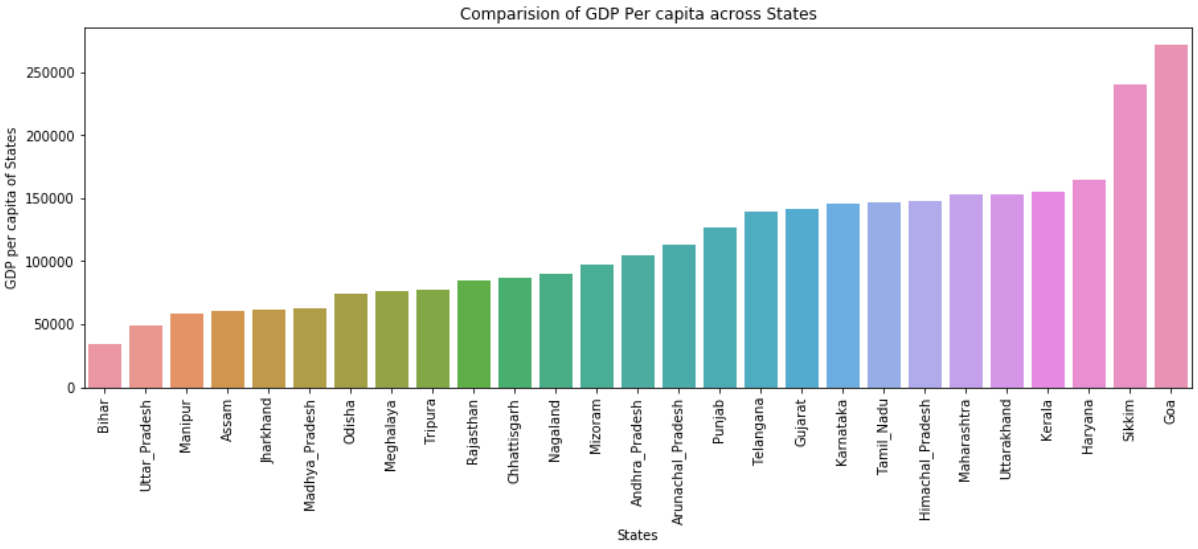
1. From the above image, we can conclude that ***Bar plot*** is best to compare data across states
2. The top 5 states based on total GSDP are
    A. Tamil Nadu
    B. Uttar Pradesh
    C. Karnataka
    D. Gujarat
    E. Andhra Pradesh

3. The bottom 5 states/UT based on total GSDP are
    A. Sikkim
    B. Arunachal Pradesh
    C. Puducherry
    D. Meghalaya
    E. Chandigarh

Part 1B

Plot the GDP per capita for all the states for the duration 2014-2015 excluding union territories.

1. *The first step is to read the *NAD files and merge them into a single data frame, for this we will use glob library package to read the paths of the files and read using pandas read_csv method*
2. *Then we need to consider only records for the duration 2014-2015*
3. *The files for union territories (Andaman, Chandigarh, Puducherry, Delhi) aren't considered/downloaded as per statement in Data Set page. Hence filtering them in data frame is not required*



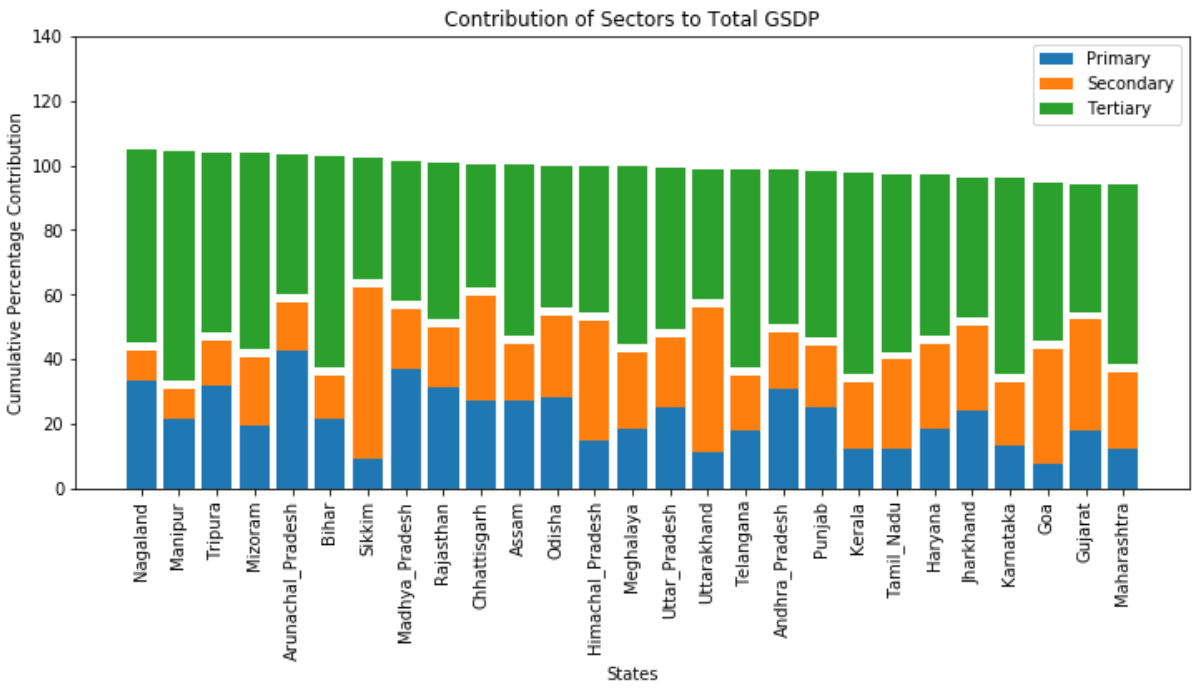
Insights

1. The top 5 states based on total Per capita GDP are
    A. Goa
    B. Sikkim
    C. Haryana
    D. Kerala
    E. Uttarakhand
2. The bottom 5 states based on total Per capita GDP are
    A. Bihar
    B. Uttar Pradesh

C. Manipur
D. Assam
E. Jharkhand

Finding State-Wise Primary, Secondary, Tertiary and Gross State Domestic Product

1. Creating 4 data frames (one for total GSDP, one for total primary sector earnings, one for total secondary sector earnings and one for total tertiary sector earnings
2. Renaming the same column names for easy differentiation
3. Merging all the data frame into a single data frame using State as join clause and dropping the Item column
4. The percentage contribution of primary, secondary and tertiary is computed by dividing against total GSDP Amount
5. The combined contribution percentage of primary, secondary, tertiary is computed by adding individual percentages of primary, secondary and tertiary
6. At last, the data frame is sorted by combined contribution for viewing in ascending order.
7. ***Stacked Bar Chart*** is best to represent the subset of primary/secondary/tertiary to total contribution.
8. Reference used for plotting stacked bar chart : https://matplotlib.org/3.1.1/gallery/lines_bars_and_markers/bar_stacked.html?highlight=stacked%20bar%20plot



Observations and Answers to Questions

1. *The sum of primary, secondary and tertiary is not equal to total GDP because there are other factors like* ***Taxes on Products (Positive impact on GSDP)*** *and* **Subsidies on products (negative impact on GSDP)**
2. Based on the above stacked bar plot, it is inferred that in all of the states except Arunachal Pradesh, Sikkim, Chhattisgarh, Uttarakhand tertiary products contributes to most of the growth
3. Sikkim earns the highest revenue from Secondary sector whereas Nagaland earns the least
4. Arunachal Pradesh earns the highest revenue from Primary Sector whereas Goa earns the least
5. Manipur earns the highest from Tertiary Sector whereas Sikkim earns the least

Categorization

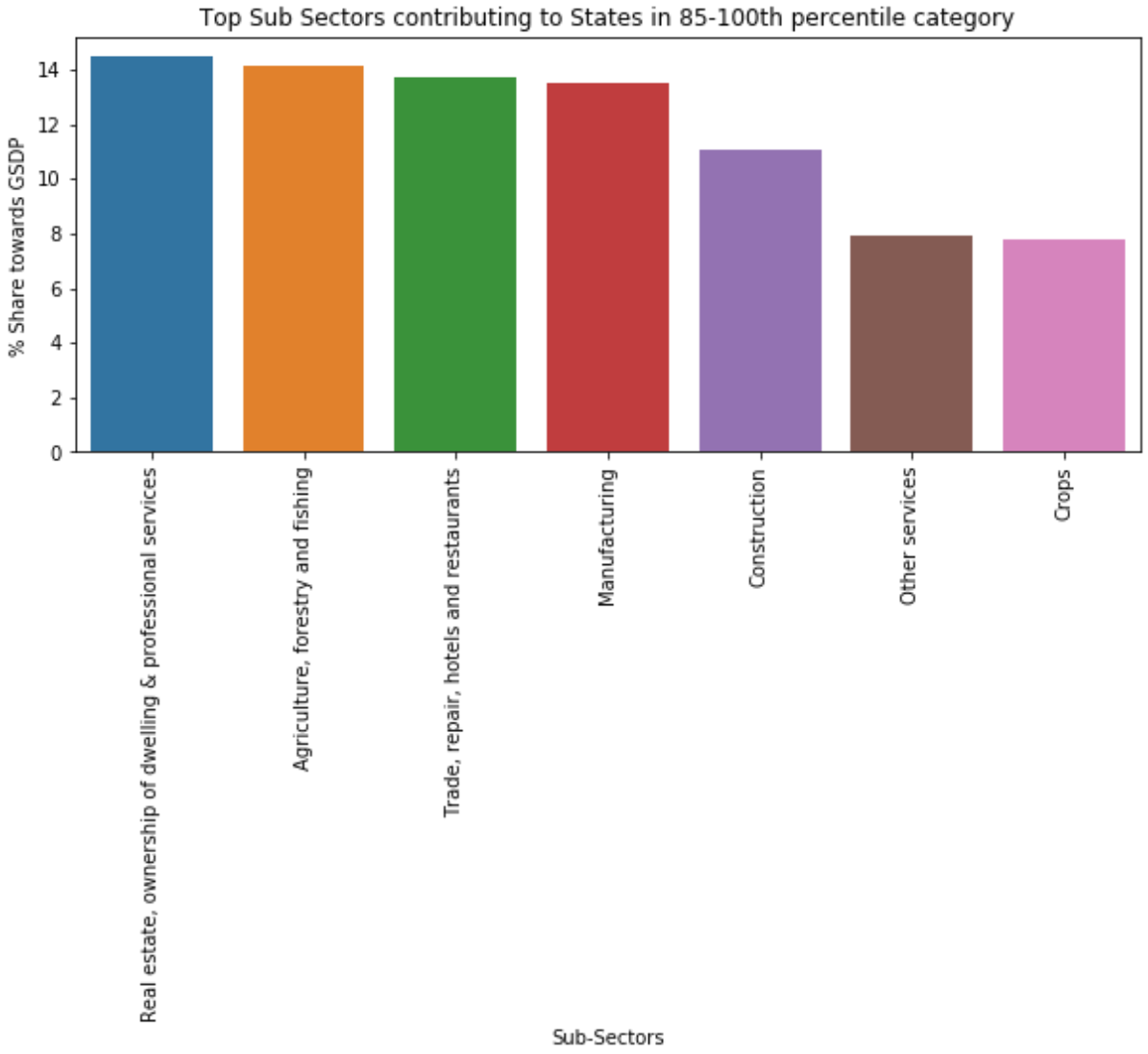
Categorise the states into four groups based on the GDP per capita (C1, C2, C3, C4, where C1 would have the highest per capita GDP and C4, the lowest). The quantile values are (0.20,0.5, 0.85, 1), i.e., the states lying between the 85th and the 100th percentile are in C1; those between the 50th and the 85th percentiles are in C2, and so on

Note:

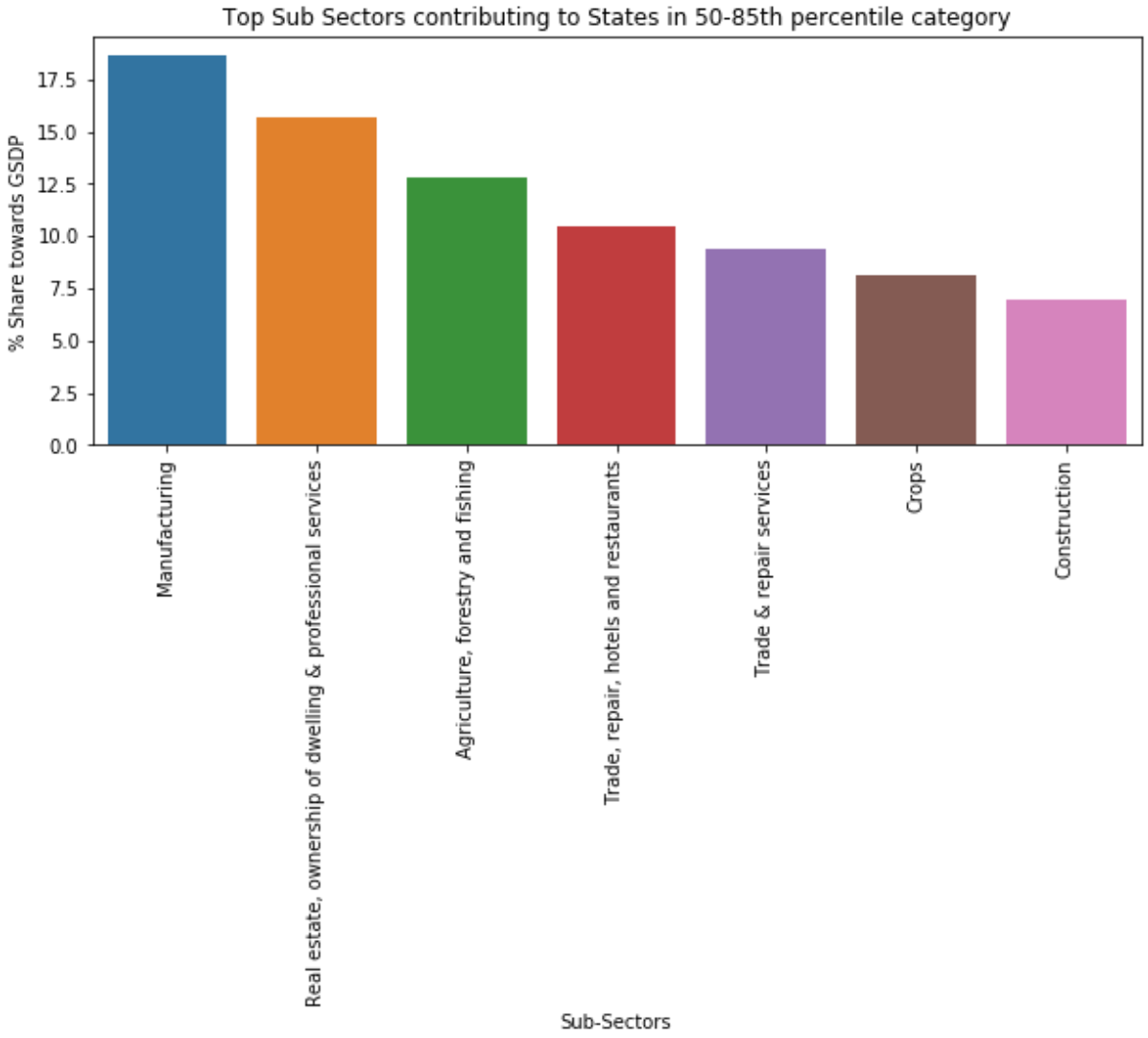
Examples for Sub-Sectors given in the Problem Description 2 is quite contradictory. For example, it indicates sub-sectors as agriculture, forestry and fishing, crops, manufacturing etc. On analysis of state wise data it is seen that, Crops is a sub-section under agriculture, forestry and fishing. Having said that, the subsequent analysis considers all the sub-sectors (i.e., both parent sub-sector and it's child sub-sector)

Some Examples:

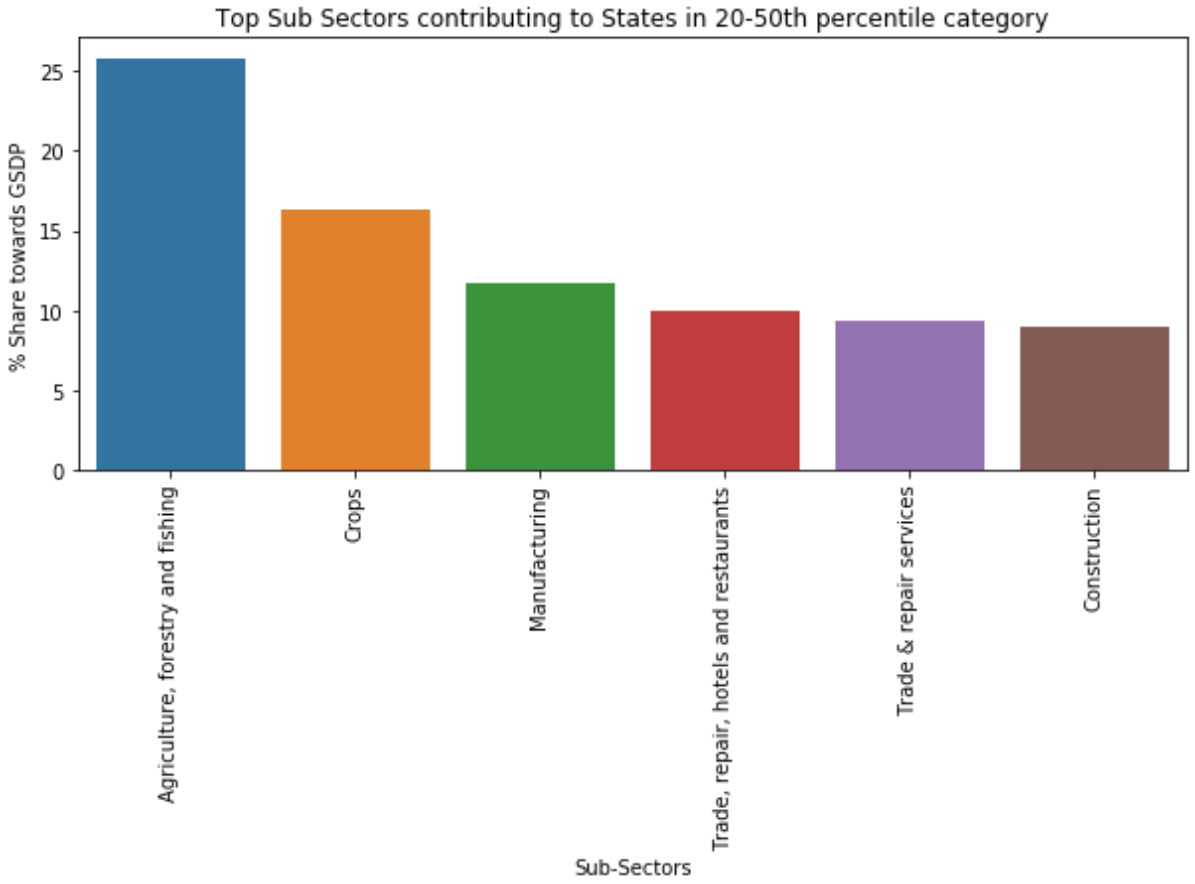
1. Agriculture, forestry and fishing **and** crops
2. Trade, repair, hotels and restaurants **and** Trade & repair services
3. Transport, storage, communication & services related to broadcasting **and** Railways



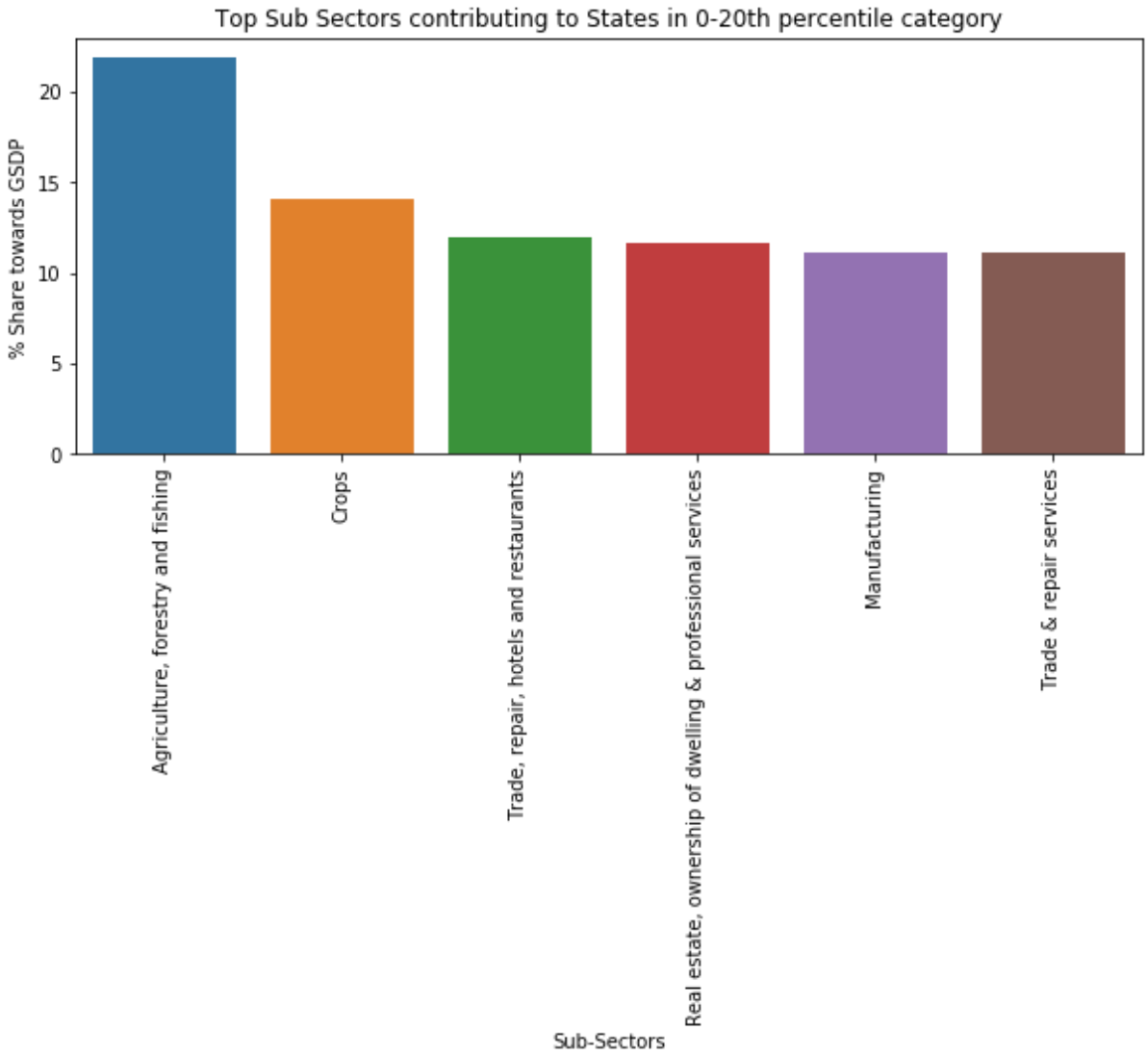
Top Sub-Sector(s) of C1 Category	%_of_GSDP
Real estate, ownership of dwelling & professional services	14.461049
Agriculture, forestry and fishing	14.119213
Trade, repair, hotels and restaurants	13.730076
Manufacturing	13.498187
Construction	11.05109
Other services	7.907258
Crops	7.811695



Top Sub-Sector(s) of C2 Category	**%_of_GSDP**
Manufacturing	18.62213
Real estate, ownership of dwelling & professional service	15.710184
Agriculture, forestry and fishing	12.825977
Trade, repair, hotels and restaurants	10.443537
Trade & repair services	9.422608
Crops	8.109086
Construction	6.932967



Top Sub-Sector(s) of C3 Category	%_of_GSDP
Agriculture, forestry and fishing	25.849557
Crops	16.312163
Manufacturing	11.676084
Trade, repair, hotels and restaurants	9.993973
Trade & repair services	9.288358
Construction	8.89223



Top Sub-Sector(s) of C4 Category	%_of_GSDP
Agriculture, forestry and fishing	21.88519
Crops	14.112128
Trade, repair, hotels and restaurants	11.9571
Real estate, ownership of dwelling & professional service	11.627645
Manufacturing	11.141726
Trade & repair services	11.092776

The Average GDP of C2 category is 2.56 times greater than C1
The Average GDP of C4 category is 1.45 times greater than C1
The Average GDP of C1 category almost matches with C3 category as: 0.99
Category C2 contributes most to the Overall GDP: 65153383.56

Note: Average is taken to compare against states/categories as count is different for each ca tegory

Sub-sectors are correlated to high GDP across categories

1. Agriculture, forestry and fishing
2. Manufacturing
3. Real estate, ownership of dwelling & professional services
4. Trade, repair, hotels and restaurants

We have already seen the sub-sectors which contribute the most by each category

Some Interesting Observations

1. *The states which are in C1 category are Kerala, Haryana, Sikkim and Goa. One common thing amongst them is they are relatively smaller in size when compared to other states of India. This could influence it's rate of GDP. There are certain outliers like Manipur which is opposite to this hypothesis*
2. *The people/state governments in C1 category earns more than other category people for the same man hours spent. This is assumed on the basis on per capita GDP of C1 category is of the highest*

The sectors which category C1 needs to focus on is shown below

Sub-Sector	Contribution to GSDP	%_of Contribution
Mining and quarrying	588961	0.577805
Railways	587419	0.576293
Services incidental to transport	281108	0.275784
Air transport	171388	0.168142
Water transport	44776	0.043928
Storage	18106	0.017763

The sectors which category C2 needs to focus on is shown below

Sub-Sector	Contribution to GSDP	%_of Contribution
Railways	2494518	0.425409
Services incidental to transport	1735162	0.295911
Air transport	659841	0.112528
Storage	368515	0.062846
Water transport	318192	0.054264
Road transport*	307147	0.05238

The sectors which category C3 needs to focus on is shown below

Sub-Sector	Contribution to GSDP	%_of Contribution
Services incidental to transport	987204	0.436498
Water transport	145480	0.064325
Storage	104405	0.046163
Services incidental to transport*	88721	0.039229
Air transport	85716	0.0379

The sectors which category C4 needs to focus on is shown below

Sub-Sector	Contribution to GSDP	%_of Contribution
Hotels & restaurants	1599466	0.864324
Services incidental to transport	238539	0.128902
Storage	192928	0.104255
Air transport	89408	0.048315
Water transport	7411	0.004005

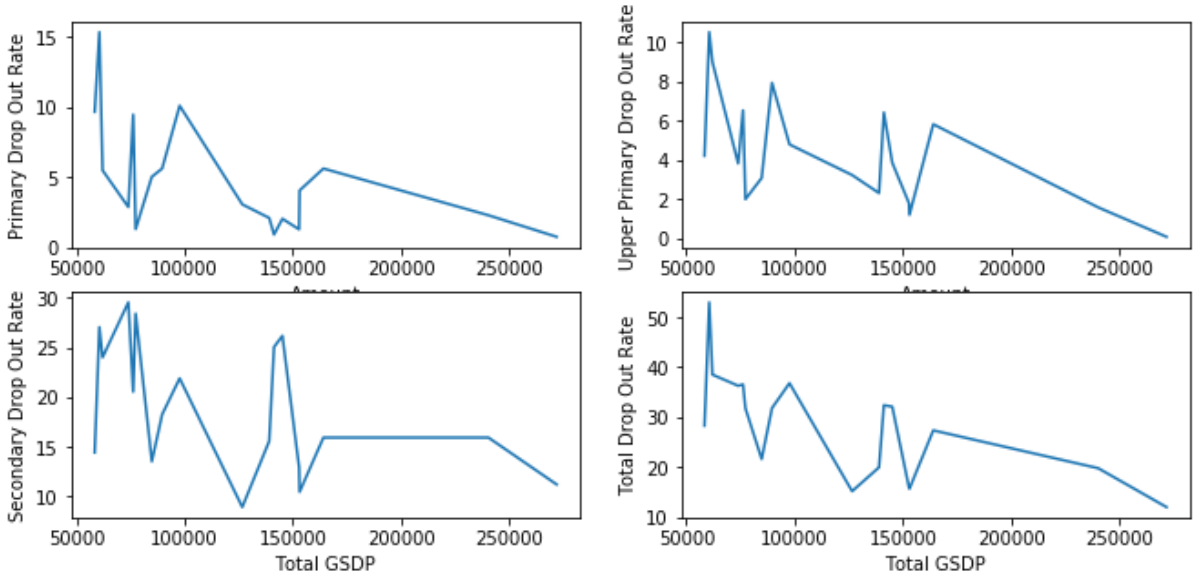
1. Across categories, potential of **Air Transport/Water Transport** needs to be enhanced by creating/improving infrastructures for these.
2. **Storage** is another un-tapped potential revenue which needs to be focussed
3. Tertiary Sector is contributing high towards GSDP. Hence, we need to maintain the same momentum on sub-sectors of tertiary sector to sustain the contribution of it towards GSDP and possibly improving it further

Recommendation:

1. For each category, concentrate more on the sub-sectors which are having very high potential (example TRANSPORT) and are contributing very less to GSDP. The government needs to reduce the losses occurring in these sub-sectors.
2. Focus intact and maintain the same momentum on sub-sectors of tertiary sector to sustain the contribution of it towards GSDP and possibly improving it further
3. As seen above, Average GSDP of C4 category is 1.45 times greater than C1 and the average GSDP of C1 category almost matches with C3 category as: 0.99. This is in sharp contrast to the per capita income of C4 and C3 category, as it is way less when compared to that of C1. Government needs to focus on restricting the population levels in C3 and C4 category to have a positive impact on GSDP for these categories
4. 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.

Part 2 GDP and Education Dropout Rates

Comparison of Total GSDP versus Drop Out Rates



Insights

1. Based on above sub-plots we can confirm that the ***GSDP increases with decrease in drop-out rate*** across Primary, Upper Primary and Secondary and Total of these for the year 2014-2015

Table of Population and Drop Out Rates

Population	State	Per capita GSDP	Primary	Upper Primary	Secondary	Total Drop Out
326780	Assam	60621	15.36	10.51	27.06	52.93
349660	Jharkhand	62091	5.48	8.99	24	38.47
11833	Mizoram	97687	10.1	4.78	21.88	36.76
32020	Meghalaya	76228	9.46	6.52	20.52	36.5
435220	Odisha	73979	2.86	3.81	29.56	36.23
633590	Gujarat	141263	0.89	6.41	25.04	32.34
635100	Karnataka	145141	2.02	3.85	26.18	32.05
20550	Nagaland	89607	5.61	7.92	18.23	31.76
38350	Tripura	77358	1.28	1.99	28.42	31.69
30873	Manipur	58442	9.66	4.2	14.38	28.24
266620	Haryana	164077	5.61	5.81	15.89	27.31
721610	Rajasthan	84837	5.02	3.07	13.48	21.57
367660	Telangana	139035	2.08	2.3	15.53	19.91
6330	Sikkim	240274	2.27	1.57	15.89	19.73
1172450	Maharashtra	152853	1.26	1.79	12.87	15.92
105820	Uttarakhand	153076	4.04	1.19	10.4	15.63
290673	Punjab	126606	3.05	3.22	8.86	15.13
14950	Goa	271793	0.73	0.07	11.15	11.95

Observations on Correlation between Population and Drop Out Rate

1. Based on the above table, we can conclude that there is no direct relationship between Population and Total Drop Out Rate (Cumulative value of Primary, Upper Primary and Secondary).
2. Highest population is at Uttar Pradesh whose drop out is 21.50% whereas the highest total drop-out rate (52.93%) is at Assam whose population is relatively in mid-range