Q1. Missing value handled by substituting the average of that particular feature in other states belonging to that particular region (class).

In case all the missing values in a particular group corresponding to a feature were missing, they were substituted by the overall mean across all states and union territories for that feature.

Q2. Top-5 states/union territories that are representative of India

i. Without normalising:

```
[1] "Maharashtra"
[1] "Tamil Nadu"
[1] "Uttar Pradesh"
[1] "Karnataka"
[1] "Gujarat"
```

ii. With normalizing:

```
[1] "Mizoram"
[1] "Chandigarh"
[1] "Tamil Nadu"
[1] "Lakshadweep"
[1] "Meghalaya"
```

Q3.

	nsdp_const	nsdp_cur	gdp_const	gdp_cur	growth_rate	sex_ratio	child_sex_ratio
nsdp_const	1.000	0.999	1.000	1.000	0.030	-0.061	-0.285
nsdp_cur	0.999	1.000	0.999	1.000	0.018	-0.052	-0.284
gdp_const	1.000	0.999	1.000	0.999	0.034	-0.064	-0.283
gdp_cur	1.000	1.000	0.999	1.000	0.022	-0.054	-0.283
growth_rate	0.030	0.018	0.034	0.022	1.000	-0.540	0.109
sex_ratio	-0.061	-0.052	-0.064	-0.054	-0.540	1.000	0.482
child_sex_ratio	-0.285	-0.284	-0.283	-0.283	0.109	0.482	1.000
dropout_rate	-0.355	-0.358	-0.355	-0.359	0.228	-0.038	0.430
enrolment_ratio	-0.349	-0.349	-0.349	-0.348	-0.238	0.205	0.087
enrolment_ratio_h	-0.193	-0.185	-0.194	-0.186	-0.228	0.197	0.019
literacy_rate_7	-0.069	-0.075	-0.067	-0.072	-0.183	0.110	0.081
toilet_boy	0.465	0.472	0.461	0.469	-0.108	0.219	-0.241
toilet_girl	0.466	0.472	0.462	0.469	-0.017	0.223	-0.160
drinking_water	0.464	0.467	0.461	0.465	0.007	0.156	-0.237
electricity	0.541	0.540	0.539	0.539	0.062	0.021	-0.348
computer	0.352	0.344	0.353	0.346	0.062	0.001	-0.243

	dropout_rate	enrolment_ratio	enrolment_ratio_highr	literacy_rate_7	toilet_boy	toilet_girl	drinking_water	electricity	computer
nsdp_const	-0.355	-0.349	-0.193	-0.069	0.465	0.466	0.464	0.541	0.352
nsdp_cur	-0.358	-0.349	-0.185	-0.075	0.472	0.472	0.467	0.540	0.344
gdp_const	-0.355	-0.349	-0.194	-0.067	0.461	0.462	0.461	0.539	0.353
gdp_cur	-0.359	-0.348	-0.186	-0.072	0.469	0.469	0.465	0.539	0.346
growth_rate	0.228	-0.238	-0.228	-0.183	-0.108	-0.017	0.007	0.062	0.062
sex_ratio	-0.038	0.205	0.197	0.110	0.219	0.223	0.156	0.021	0.001
child_sex_ratio	0.430	0.087	0.019	0.081	-0.241	-0.160	-0.237	-0.348	-0.243
dropout_rate	1.000	-0.259	-0.065	-0.335	-0.279	-0.194	-0.154	-0.534	-0.494
enrolment_ratio	-0.259	1.000	0.517	0.536	-0.278	-0.348	-0.434	-0.096	0.000
enrolment_ratio_h	-0.065	0.517	1.000	0.277	-0.139	-0.174	-0.293	0.058	0.071
literacy_rate_7	-0.335	0.536	0.277	1.000	-0.249	-0.356	-0.487	0.176	0.431
toilet_boy	-0.279	-0.278	-0.139	-0.249	1.000	0.951	0.849	0.658	0.453
toilet_girl	-0.194	-0.348	-0.174	-0.356	0.951	1.000	0.904	0.673	0.388
drinking_water	-0.154	-0.434	-0.293	-0.487	0.849	0.904	1.000	0.561	0.265
electricity	-0.534	-0.096	0.058	0.176	0.658	0.673	0.561	1.000	0.799
computer	-0.494	0.000	0.071	0.431	0.453	0.388	0.265	0.799	1.000

Conclusion:

Economy – All the four variables are extremely highly correlated (correlation > 0.999) among themselves. Therefore if we were to select feature naively based on just correlations without further analysis, it would be wise to select one of these four features viz. nsdp_const, nsdp_cur, gdp_const, gdp_cur, as they are highly representative of each other. Moreover, all other features have approximately the same correlation with these four features.

Demography – We see a moderately negative correlation between sex_ratio and growth_rate while a slight positive correlation between child_sex_ratio which might be indicative of female feticide in the recent past in places where the growth rate is high but a recent change in the conditions thereby improving the sex ratio in childs, maybe due to drives undertaken to curb the problem of female feticide thereby educating new younger parents.

Education – Features representing facilities in schools such as toilets, drinking water, electricity and computers are moderately positively correlated among themselves which should be quite obvious.

Q4. Using multiclass relief algorithm, we got the following results:

Economy:

nsdp_const	nsdp_cur	gdp_const	gdp_cur
-0.006396032	0.006762507	-0.008579350	0.005273463

Demography:

growth_rate	sex_ratio	child_sex_ratio
0.05349169	0.04747105	0.16283239

Education:

dropout_rate	enrolment_ratio	enrolment_ratio_h
-0.16429372	0.07727160	-0.04787824
literacy_rate_7	toilet_boy	toilet_girl
-0.44605781	-0.13081380	-0.22768003
drinking_water	electricity	computer
-0.30363478	-0.08183553	0.10566383

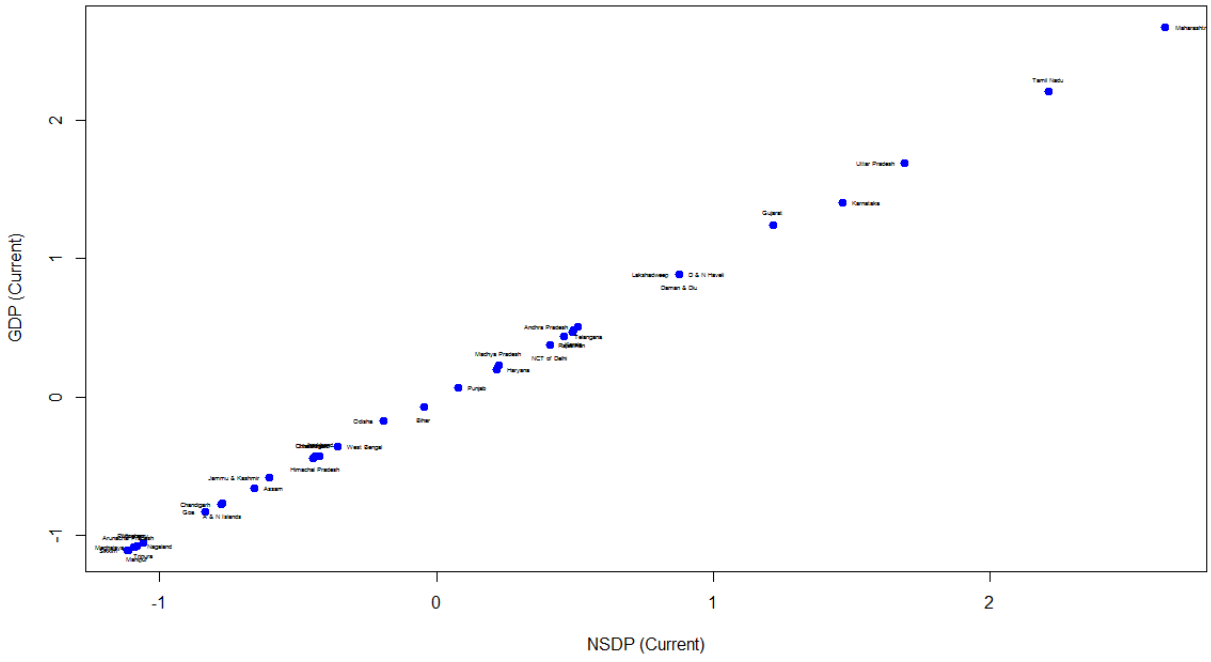
Across categories:

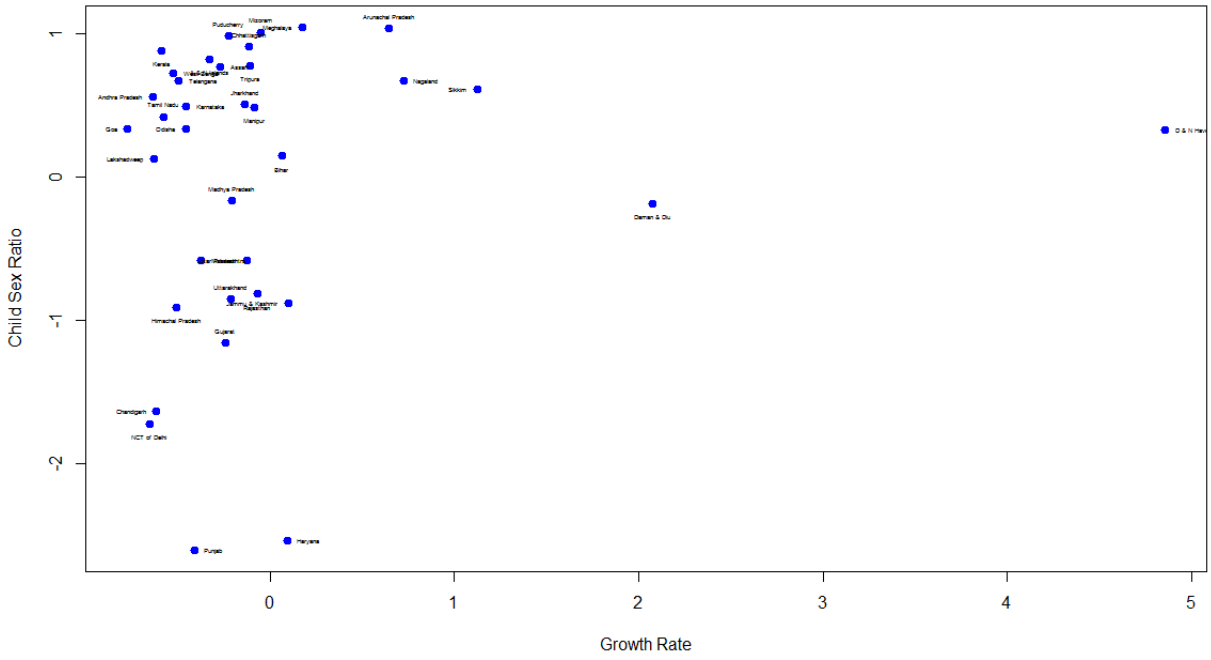
nsdp_const	nsdp_cur	gdp_const
0.22657712	0.23464358	0.22373391
gdp_cur	growth_rate	sex_ratio
0.23285845	0.08475644	0.04331105
child_sex_ratio	dropout_rate	enrolment_ratio
0.17498036	0.03714870	0.01392531
enrolment_ratio_h	literacy_rate_7	toilet_boy
0.10687115	-0.02315977	0.01443851
toilet_girl	drinking_water	electricity
-0.00863802	-0.01882878	0.05312635
computer		
0.03300127		

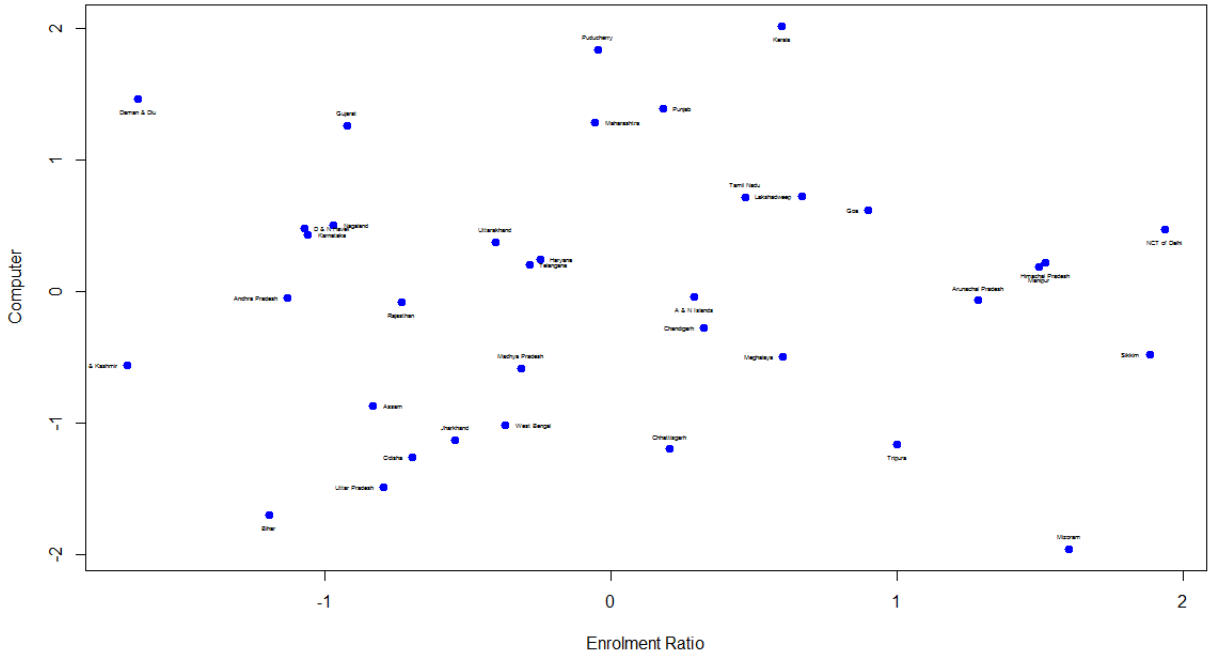
Across categories (after removing highly correlated variables)

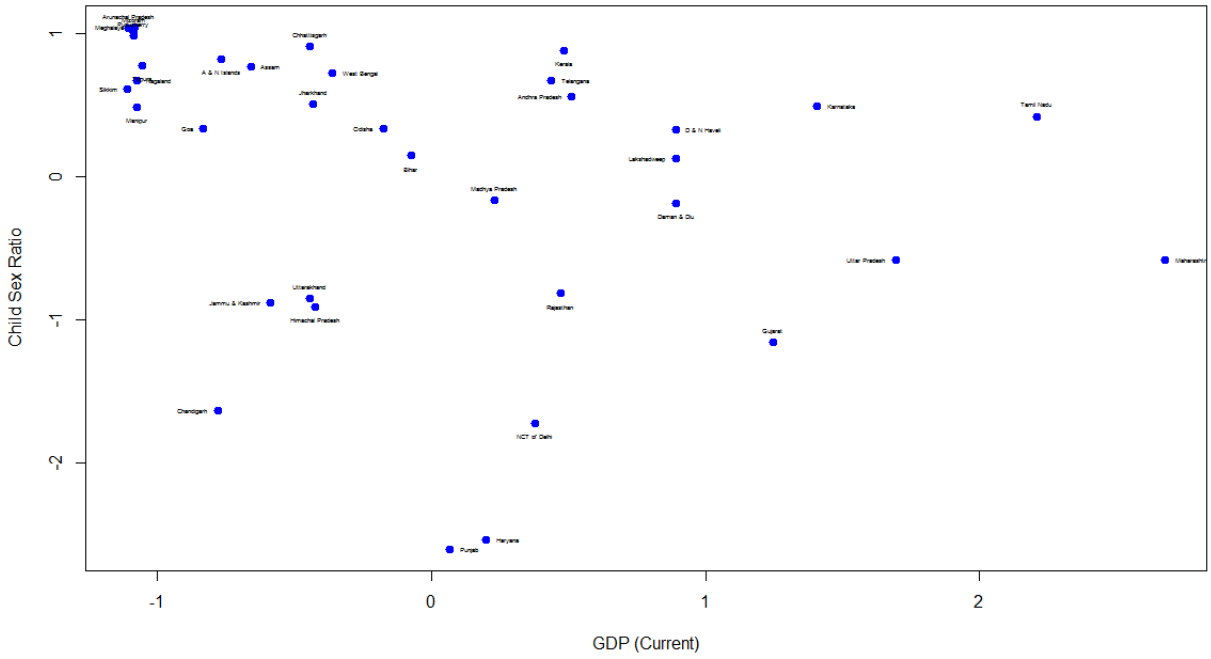
gdp_cur	growth_rate	child_sex_ratio
0.264311035	0.067028271	0.165895347
dropout_rate	enrolment_ratio	electricity
0.038764105	-0.002362554	0.057691158

Q5. Scatter Plots

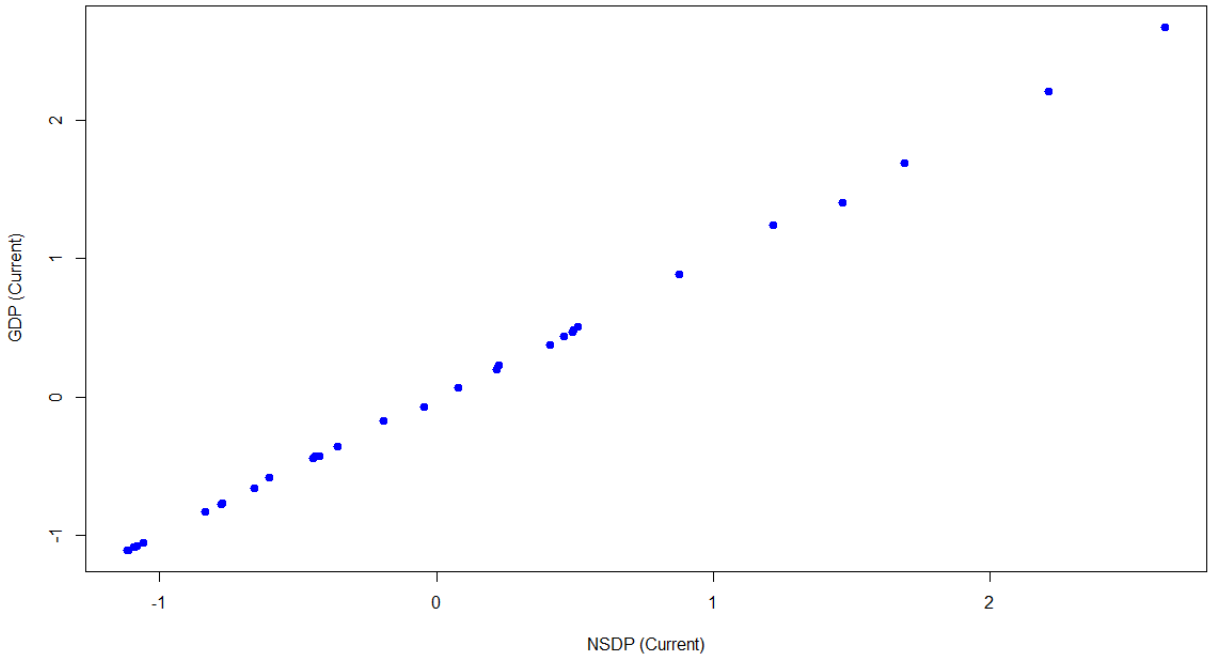


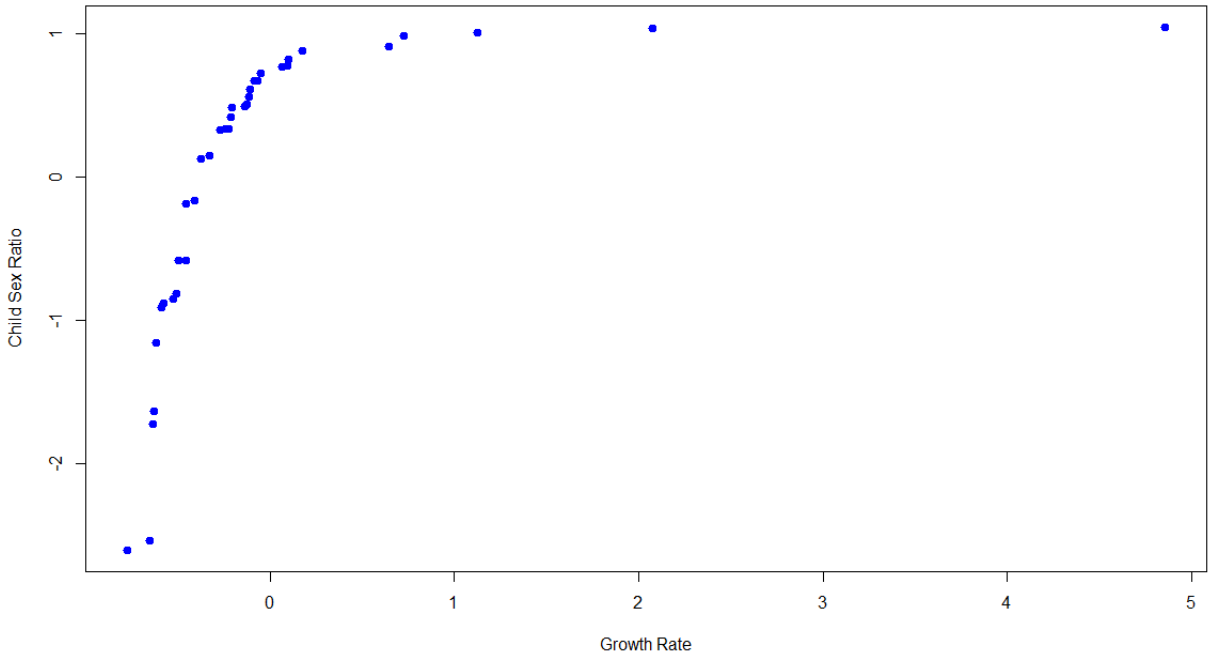


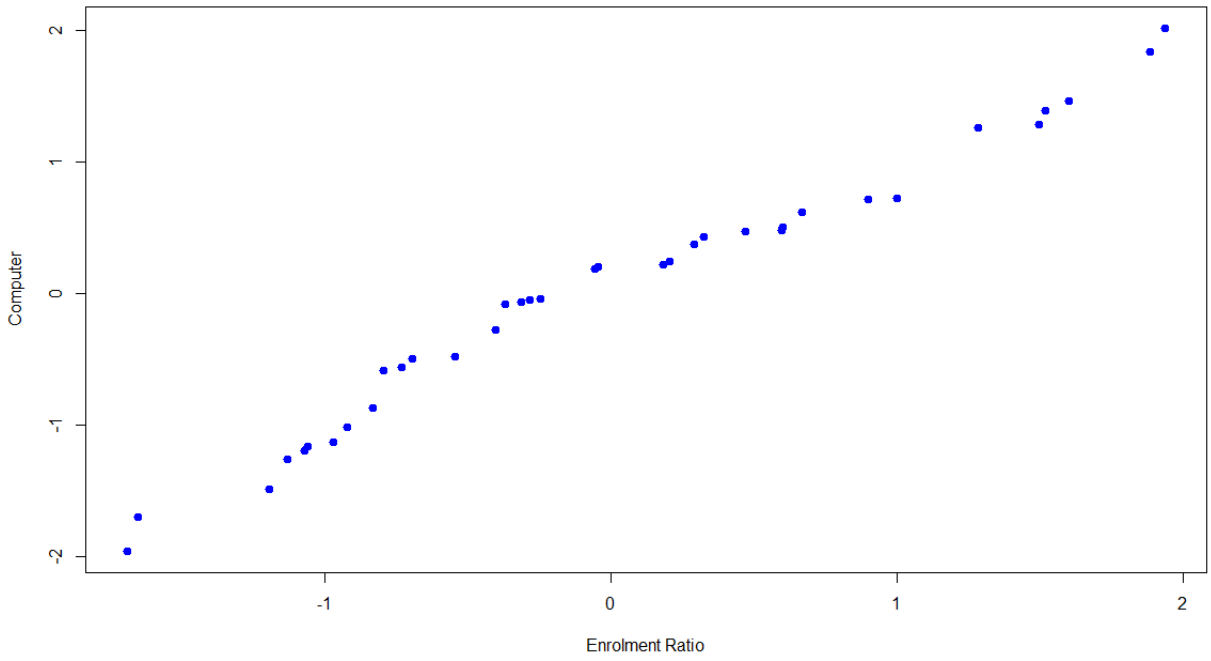

Computer
Enrolment Ratio
2
1
0
-1
-2
-1
0
1
2
Puducherry
Kerala
Daman & Diu
Gujarat
Punjab
Maharashtra
Tamil Nadu
Lakshadweep
Goa
NCT of Delhi
Uttarakhand
Haryana
Telangana
Himachal Pradesh
Manipur
Arunachal Pradesh
Andhra Pradesh
Rajasthan
A & N Islands
Chandigarh
Meghalaya
Sikkim
Madhya Pradesh
Assam
Jharkhand
West Bengal
Chhattisgarh
Odisha
Tripura
Uttar Pradesh
Bihar
Mizoram

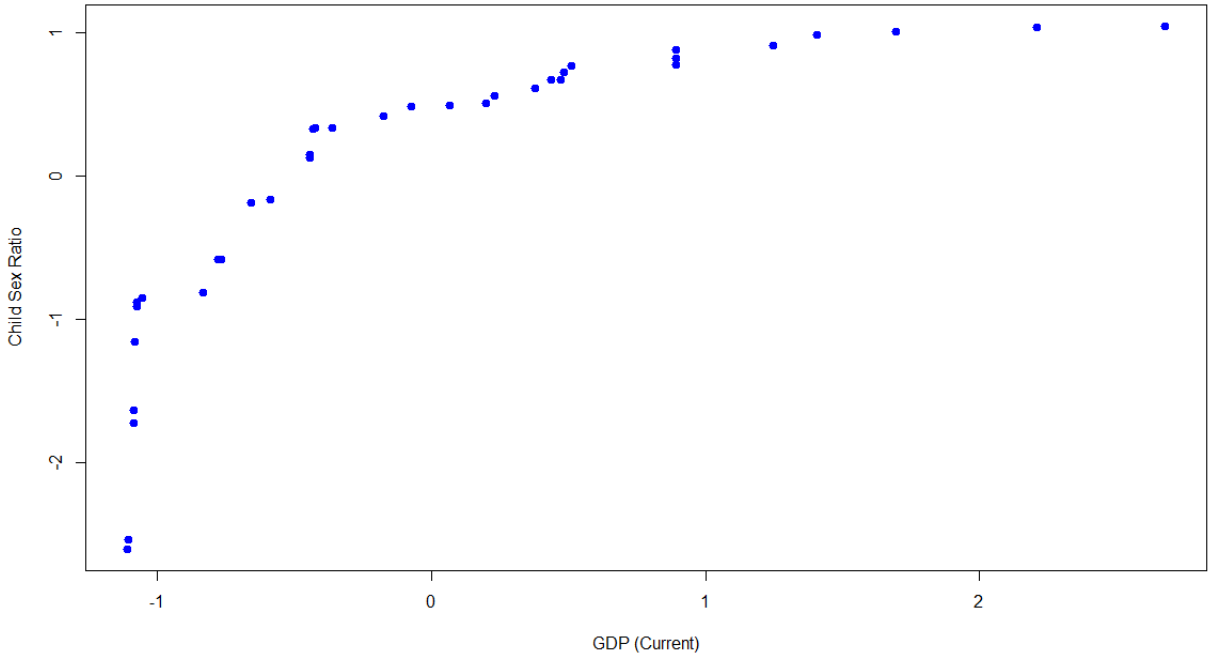

Child Sex Ratio
GDP (Current)
1
0
-1
-2
-1
0
1
2
Chhattisgarh
Kerala
A & N Islands
Assam
West Bengal
Telangana
Sikkim
Jharkhand
Andhra Pradesh
Karnataka
Tamil Nadu
Manipur
Goa
Odisha
D & N Haveli
Bihar
Lakshadweep
Madhya Pradesh
Daman & Diu
Uttar Pradesh
Maharashtra
Uttarakhand
Jammu & Kashmir
Himachal Pradesh
Rajasthan
Gujarat
Chandigarh
NCT of Delhi
Haryana
Punjab

Quantile-quantile plots






Computer
Enrolment Ratio
2
1
0
-1
-2
-1
0
1
2


Child Sex Ratio
GDP (Current)
1
0
-1
-2
-1
0
1
2

Q6. Intuitive Partitioning

gdp_cur:

5th percentile: 19405 = low

95th percentile: 818216 = high

Minimum: 14523

Maximum: 1040211

Rounding yields low' = 0, high' = 840000

1st level (4) partitions: (1, 210000), (210001, 420000), (420001, 630000), (630001, 840000)

Since max > high', new interval: (840001, 1050000)

2nd level partitions: (1, 70000), (70001, 140000), … (770001, 840000) and (840001, 1050000)

child_sex_ratio

5th percentile: 849 = low

95th percentile: 967 = high

Minimum: 822

Maximum: 969

Rounding yields low' = 800, high' = 1000

1st level (4) partitions: (800, 850), (851, 900), (901, 950), (951, 1000)

1st interval adjusted to (822, 850)

2nd level partitions: (822, 840), (841, 850), (851, 860) … (961, 970)