# Constructing a socio-economic index for districts of Odisha

## Data Wrangling:

Data for the calculation of index was collected from the following data sources. All of the sources are Government organisations working in some related field.

**Data Sources:**

- Socio-Economic Caste Census 2011, http://www.secc.gov.in/
- Odisha Economic Survey 2014-2015, Planning and Coordination Department, Government of Odisha. http://www.odisha.gov.in/pc/Download/Economic_Survey_2014-15.pdf
- Annual Health Survey 2012-2013 Odisha, Census http://www.censusindia.gov.in/vital_statistics/AHSBulletins/AHS_Factsheets_2012-13/FACTSHEET-Odisha.pdf
- District Information System for Education, http://www.dise.in/

The data was consolidated and manipulated using different sources to get a high quality and informative dataset.

**Indicators Selected For Index Calculation:**

1. Net District Domestic Product
2. Net Enrolment ratio in Upper Primary
3. Percentage of Pucca households to Total households
4. Percentage of Rural Income greater than 10000
5. Percentage of Women who received complete ANC
6. Crude Birth Rate
7. Literacy Rate
8. Sex Ratio at Birth
9. Under 5 mortality Rate
10. Percentage of Households with Toilet Access within Premises
11. Railway Access (Kilometres of Railway Line in each District)

A number of other indicators were also considered but only one indicator among similar indicators was selected for further procedure. For example: Among Infant mortality rate and Under 5 mortality Rate , only the latter was selected as both are highly correlated and the latter indicator provided more information and was expected to be a better indicator for the socio-economic status.

## Calculating Index:

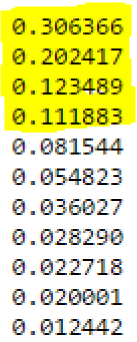
The following steps were involved for calculating the index from the dataset. Python's library Scikit-Learn was used to implement the following steps.

**Standardizing Data:**

StandardScaler from the scikit-learn library was used to scale all the variables in the dataset. This function standardizes the variables by removing the mean and scaling to unit variance. Centering and scaling happen independently on each variable by computing the relevant statistics.

**Applying Principal Component Analysis:**

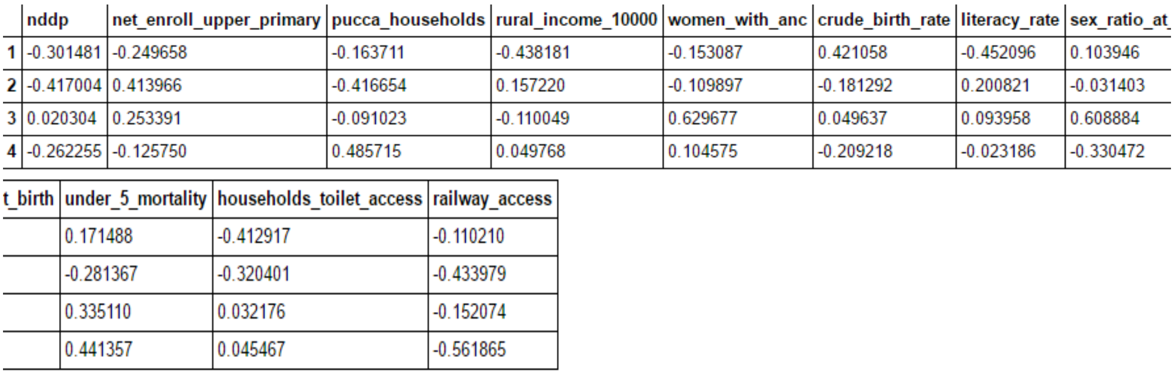
Following are the results of the principal component analysis describing the variance being explained by each principal component. This was done using PCA function from the scikit-learn library.



```
0.306366
0.202417
0.123489
0.111883
0.081544
0.054823
0.036027
0.028290
0.022718
0.020001
0.012442
```

The first four components were used to compute the index as they explain about 74 % of the variation in the data.

Following are the loading scores of each the 4 components for all the 11 variables. (The table is broken into two rows to make it legible)



| | nddp | net_enroll_upper_primary | pucca_households | rural_income_10000 | women_with_anc | crude_birth_rate | literacy_rate | sex_ratio_at_birth | under_5_mortality | households_toilet_access | railway_access |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.301481 | -0.249658 | -0.163711 | -0.438181 | -0.153087 | 0.421058 | -0.452096 | 0.103946 | 0.171488 | -0.412917 | -0.110210 |
| 2 | -0.417004 | 0.413966 | -0.416654 | 0.157220 | -0.109897 | -0.181292 | 0.200821 | -0.031403 | -0.281367 | -0.320401 | -0.433979 |
| 3 | 0.020304 | 0.253391 | -0.091023 | -0.110049 | 0.629677 | 0.049637 | 0.093958 | 0.608884 | 0.335110 | 0.032176 | -0.152074 |
| 4 | -0.262255 | -0.125750 | 0.485715 | 0.049768 | 0.104575 | -0.209218 | -0.023186 | -0.330472 | 0.441357 | 0.045467 | -0.561865 |

A variable with a positive loading indicates a negative association to the component. A negative loading simply means that the results need to be interpreted in the opposite direction from the way it is worded. Higher value of NDDP in the original data indicate better socioeconomic circumstances, hence the negative sign on this variable means a higher economic situation. Another example is that Lower value of Under 5 mortality rate means a better socioeconomic situation; therefore a positive value indicates a better socioeconomic situation.

**Calculating the socio-economic index**:

Now the first step in the computation of a single index is calculating the component scores. Since I decided to use four components, I will calculate component scores for 4 components. Component scores are the scores of each district, on each component.

To compute the component scores for a given district for a given component, the district's standardized score on each variable is multiplied by the corresponding loading score of the variable for the given component, and summed these products.

The four factors explained a total of 74.42% of the variation with the first, second, third and fourth components explaining 30.64 per cent, 20.24 per cent, 12.35 per cent and 11.19 per cent.

Therefore, the importance of the components in measuring overall socioeconomic condition is not the same. Using the proportion of these percentages as weights on the component score, a Non- standardized Index (NSI) was developed for each district, using the formula:

$$NSI = \left(\frac{30.64}{74.42}\right) * ComponentScore1 + \frac{20.24}{74.42} * ComponentScore2 + \frac{12.35}{74.42} * ComponentScore3 + \frac{11.19}{74.42} * ComponentScore4$$

This index measures the socioeconomic status of one district relative to the other on a linear scale. The value of the index can be positive or negative, making it difficult to interpret. Therefore, a Standardized Index (SI) was developed, the value of which can range from 0 to 100, using the formula:

$$SI = \frac{NSI\ of\ District - Min(NSI)}{Max(NSI) - Min(NSI)} * 100$$

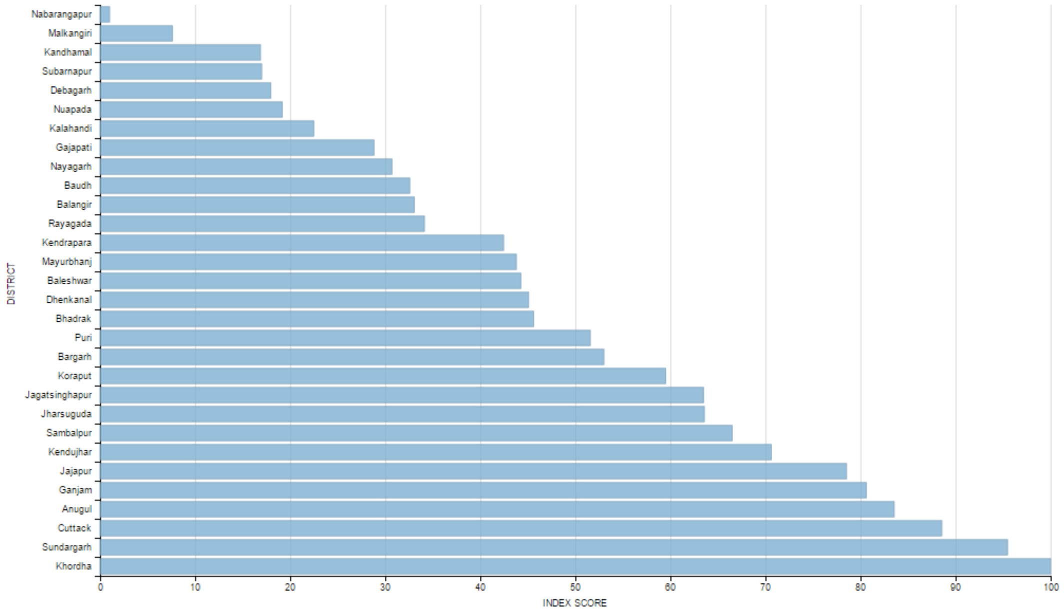
## Result:

The districts are ranked on the basis of the Standardized Index as calculated above. The ranking can be seen from the following table.

| **DISTRICT** | **INDEX** | **RANK** |
|---|---|---|
| Khordha | 100 | 1 |
| Sundargarh | 95.46713 | 2 |
| Cuttack | 88.55253 | 3 |
| Anugul | 83.54409 | 4 |
| Ganjam | 80.60963 | 5 |
| Jajapur | 78.53203 | 6 |
| Kendujhar | 70.61109 | 7 |
| Sambalpur | 66.49382 | 8 |
| Jharsuguda | 63.56995 | 9 |
| Jagatsinghapur | 63.48013 | 10 |
| Koraput | 59.49175 | 11 |
| Bargarh | 53.00308 | 12 |
| Puri | 51.5664 | 13 |
| Bhadrak | 45.60316 | 14 |
| Dhenkanal | 45.07823 | 15 |
| Baleshwar | 44.25831 | 16 |
| Mayurbhanj | 43.78864 | 17 |
| Kendrapara | 42.45157 | 18 |
| Rayagada | 34.118 | 19 |
| Balangir | 33.06272 | 20 |
| Baudh | 32.57699 | 21 |
| Nayagarh | 30.7014 | 22 |
| Gajapati | 28.8136 | 23 |
| Kalahandi | 22.47478 | 24 |
| Nuapada | 19.15856 | 25 |
| Debagarh | 17.96048 | 26 |
| Subarnapur | 17.00335 | 27 |
| Kandhamal | 16.87693 | 28 |
| Malkangiri | 7.608468 | 29 |
| Nabarangapur | 1.0 | 30 |

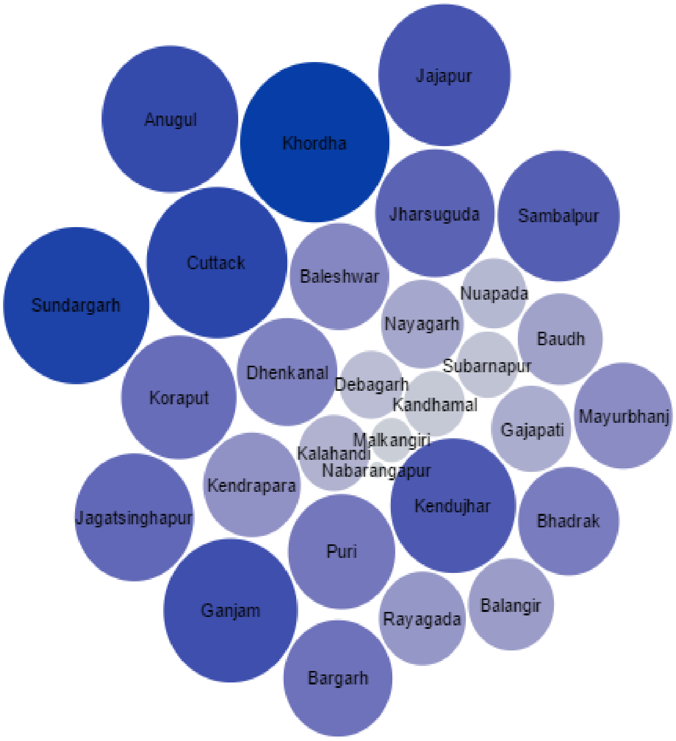
The following bar graph compares the indices of the districts.

DISTRICTS OF ODISHA BY SOCIO-ECONOMIC INDEX

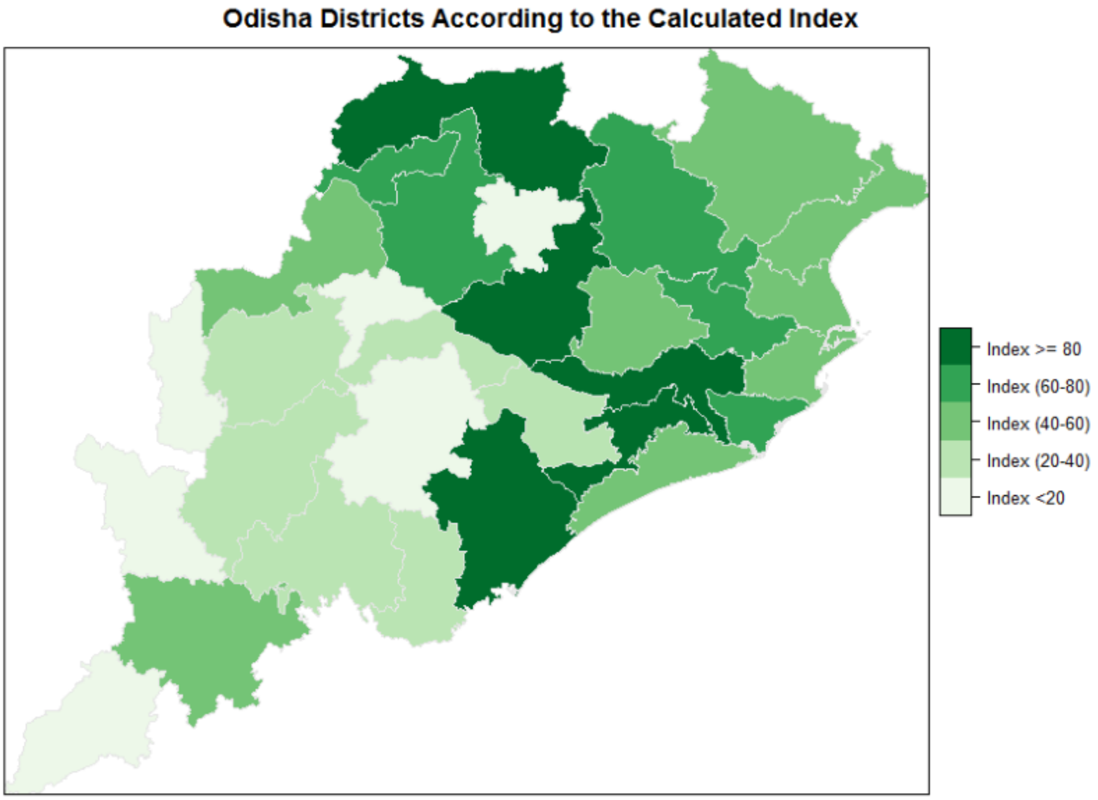


Here is the link to this visualization: http://tinyurl.com/DistrictRanked

To get a quick sense of the relative rankings the following visualization can be helpful. The darker the colour hue the higher the socio-economic index, hence more developed the district. Also greater the size of the bubble greater the index



The following visualization shows Map of Odisha with districts coloured according to the index to get a better idea of the socio-economic development geographically.



Odisha Districts According to the Calculated Index