CAPSTONE REPORT

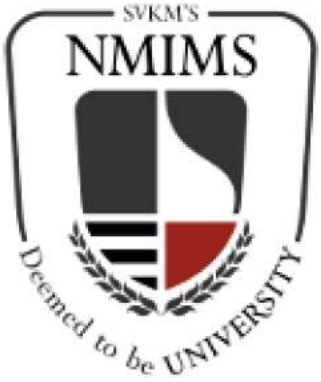
YASH DALAL

BTECH DATA SCIENCE

SEMESTER VIII

SAP ID – 70091017009

MENTOR: DR. ARINDAM CHAUDHURI



DEPARTMENT OF DATA SCIENCE

MUKESH PATEL SCHOOL OF TECHNOLOGY, MANAGEMENT AND ENGINEERING

NMIMS, MUMBAI

APRIL 2021

ACKNOWLEDGEMENT

I express my sincere gratitude to my mentor Dr. Arindam Chaudhuri. It was a great learning experience to work on this project. This project has brought out the data mining, data pre - processing and data visualization skills in me.

Dr. Arindam Chaudhuri has been a strong support throughout this semester. His guidance and experience have helped me during the course of this project.

My sincere thanks to Prof. Sarada Samantaray (HOD, Data Science) for his support and inspiration to the batch of BTech Data Science 2017 – 2021. I would also extend my gratitude to all the faculty members of BTech Data Science for their constant motivation and help.

Yash Dalal.

INDEX

STATISTICAL ANALYSIS OF SEMI URBAN DISTRICTS OF INDIA

AIM –

To find the factors and the correlation between them that can lead to further development of the semi – urban districts of India.

INTRODUCTION –

In India, there are total 718 districts in 28 states and 8 union territories. These 718 districts can be broadly classified into three types – urban, semi – urban and rural. According to Collins dictionary, semi - urban is an area which is neither fully urban nor fully rural. It has a mixture of rural and urban characteristics.[1] Semi - urban districts are those which have an urban population between 35% - 60%. Semi – urban districts have an economy which is based on primary, secondary as well as tertiary sectors. This gives it a rural as well as an urban aspect. The economies of urban areas are mostly confined to tertiary sector while rural areas have more focus upon primary sector. There are a total 75 semi - urban districts in India.

The main aim of this project is to find relationships between the variables and to know the extent of influence on each other. Most of the insights of this project are based upon unsupervised machine learning techniques. When the data points are segregated into clusters it becomes easy to find factors which can be improved upon.

'Statistical analysis of semi – urban districts of India' is a data analysis and clustering project. The most important task of this project is data mining and data pre - processing. It is often said that 80% of the time in any data analytics project is consumed by data mining and remaining for data pre – processing, exploratory data analysis and algorithm training. Extensive Exploratory Data Analysis is performed to get an in – depth knowledge of the data. Data Visualization which is rightly called Data Storytelling, reveals a lot of information in itself. Moreover, hierarchical clustering method is used widely in this project to identify the cluster of districts facing similar kind of issues. The results from the clustering methods will help to identify the key problems faced by the districts in the cluster.

There has not been much focus of the researchers on the semi – urban districts of India. Most of the research available has been performed on a sample of districts selected or belonging to a particular area or state. This project is unique in all aspects as it not only focuses on demographics of these districts but also other factors which are crucial like economy, health, land use and ground water quality.

PROJECT TIMELINE –

The project timeline consists of a schedule according to which tasks are completed. This capstone project was started in the month of January 2021 and completed in April 2021. As the deadline for submitting the project was known well in advance, it was easy to schedule the data mining, data transformation, exploratory data analysis, clustering and data visualization tasks. It also includes the time to create a documentation in the form of detailed report and also a research paper on the topic. The figure on the next page gives a detailed summary of the tasks performed during each of the months – January, February, March and April of 2021. All the tasks were completed as per the schedule decided.

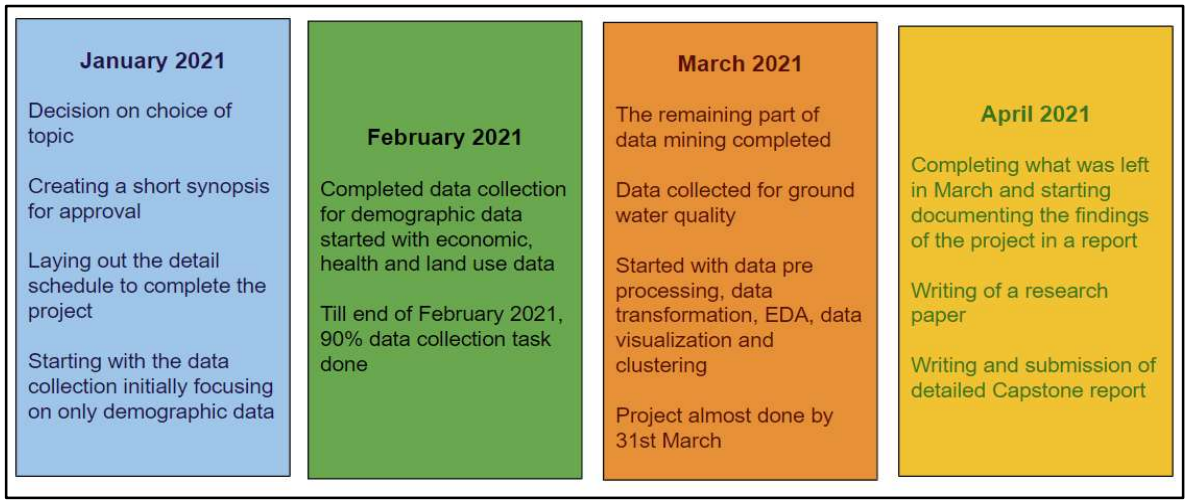


Fig1. Project Timeline

DATA MINING –

The data for this project has been gathered from multiple sources such as Central and State government websites, third party websites, district websites, websites of the ministries and various newspaper articles. The data has been segregated into demographic, health, economic, ground water quality and land use.

Semi – urban districts considered –

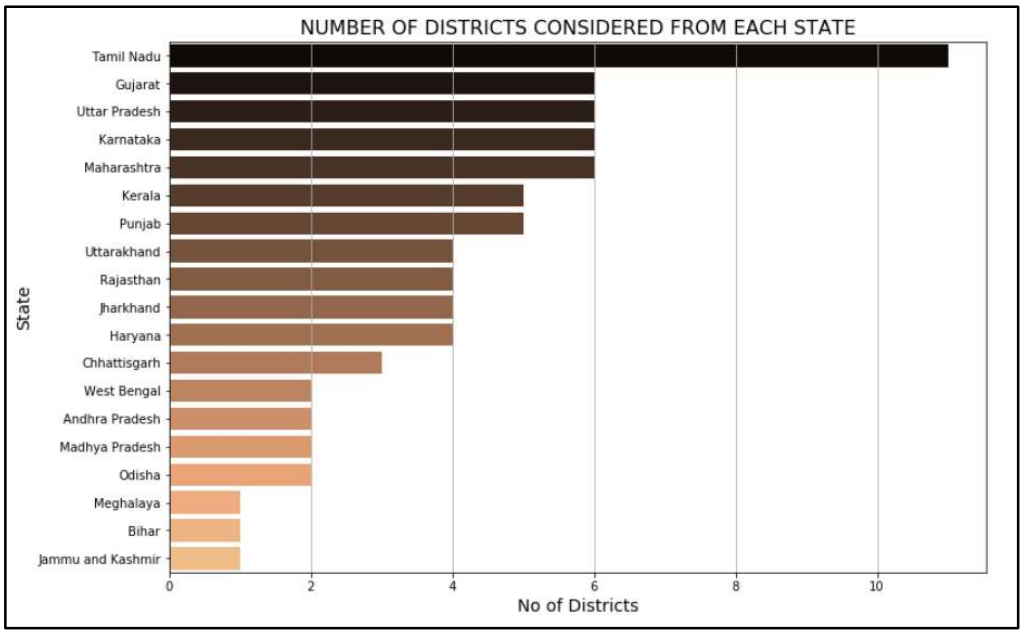


Fig2. Number of districts considered from each state

The above figure depicts the number of districts considered from each state in India. Tamil Nadu has maximum 11 semi – urban districts. It is followed by Gujarat, Uttar Pradesh, Maharashtra and Karnataka having 6 semi – urban districts each. Kerala and Punjab have 5 semi – urban districts each. Uttarakhand, Rajasthan, Jharkhand and Haryana have 4 districts included in this project. Chhattisgarh has 3 while West Bengal, Andhra Pradesh, Madhya Pradesh and Odisha have 2 each. Jammu and Kashmir, Bihar and Meghalaya have only one district under this category. Overall trend that can be inferred from this graph is

that states with high population and industrialization have greater amount of semi – urban districts than those having less population and less industrialization.

The 75 semi – urban districts from 19 states considered for this project are as follows –

State	No of Districts	Districts considered
Kerala	5	Kasaragod, Malappuram, Alappuzha, Kollam and Thiruvananthapuram.
Tamil Nadu	11	Karur, Theni, Madurai, Tirunelveli, Salem, Tiruchirappalli, Thoothukkudi, Erode, Namakkal, Virudhunagar and Vellore.
Karnataka	6	Dharwad, Bellary, Shimoga, Dakshina Kannada (Mangalore), Mysore and Gadag.
Andhra Pradesh	2	Visakhapatnam and Krishna (Vijayawada).
Maharashtra	6	Akola, Amravati, Chandrapur, Aurangabad, Nashik and Raigad.
Gujarat	6	Gandhinagar, Rajkot, Jamnagar, Bhavnagar, Valsad and Vadodara.
Madhya Pradesh	2	Ujjain and Jabalpur.
Chhattisgarh	3	Durg, Korba and Raipur.
Odisha	2	Sundergarh and Khorda.
West Bengal	2	Darjeeling and Hooghly.
Jharkhand	4	Bokaro, Dhanbad, Ranchi and East Singbhum (Jamshedpur).
Bihar	1	Patna
Uttar Pradesh	6	Bareilly, Agra, Varanasi, Meerut, Jhansi and Gautam Buddha Nagar (Noida).
Haryana	4	Panipat, Ambala, Rohtak and Yamunanagar.
Rajasthan	4	Jaipur, Ajmer, Jodhpur and Kota.
Punjab	5	Ludhiana, SAS Nagar (Mohali), Patiala, Jalandhar and Amritsar.
Uttarakhand	4	Dehradun, Haridwar, Udham Singh Nagar and Nainital.
Jammu and Kashmir	1	Jammu
Meghalaya	1	East Khasi Hills (Shillong). [2]
19 States	75	Total

Note – There are 2 districts of Telangana (RangaReddy and SangaReddy) that fall under this category, but as they were carved out from larger districts when Telangana was formed in 2014, historical data cannot be accurately extracted. Bardhaman (Burdwan) district of West Bengal was separated into 2 new districts in 2017, therefore data for the same cannot be extracted accurately. [3][4][5][6]

Variables used for the project -

The most important aspect of this project is the amount of data mined from the sources. It can be broadly categorised as Economic, Demographic, Health, Land Use and Ground Water Quality.

Main columns are those columns which have been directly extracted from the web sources. Derived columns are those which are created by performing mathematical computations on the main columns. The formulae for the computation of derived columns are given in Mathematical model part of this report.

Demographics data – [2]

Sr No	Main columns		Sr No	Derived Columns
1	District		1	Percentage of urban population (in %)
2	State		2	Total population growth (in %)
3	Area (in sq.km.)		3	Male population growth (in %)
4	Total population 2011		4	Female population growth (in %)
5	Male population 2011		5	Total literacy growth (in %)
6	Female population 2011		6	Male literacy growth (in %)
7	Urban population 2011		7	Female literacy growth (in %)
8	Rural population 2011		8	Child population proportion (in %)
9	Total literacy 2011		9	Sex Ratio growth (in %)
10	Male literacy 2011			
11	Female literacy 2011			
12	Total population 2001			
13	Male population 2001			
14	Female population 2001			
15	Total literacy 2001			
16	Male literacy 2001			
17	Female literacy 2001			
18	Sex Ratio 2011			
19	Sex Ratio 2001			
20	Density of population (per sq. km.)			
21	Working class population (in %)			
22	Child population under 6 years 2011			

Health data – [7]

Sr No	Main Columns		Sr No	Derived Columns
1	Infant Mortality Rate 2011 (per thousand)		1	Growth in Infant Mortality Rate (in %)
2	Infant Mortality Rate 2001 (per thousand)		2	Growth in Under 5 Mortality Rate (in %)
3	Under 5 Mortality Rate 2011 (per thousand)		3	Population per Government Hospital
4	Under 5 Mortality Rate 2001 (per thousand)			
5	Number of Government Hospitals			

Economic data – [8]

Sr No	Main Columns
1	Below poverty line population (in %)
2	Per capita income (in Rs)

Land Use data – [9]

Sr No	Main Columns		Sr No	Derived Columns
1	Net sown area (sq.km.)		1	Percentage of area sown (in %)
2	Forest area (sq.km.)		2	Percentage of forest land (in %)
3	Net irrigated area (sq.km.)		3	Percentage of sown area irrigated (in %)
4	Rainfall (in mm.)			

Ground Water Quality data – [10]

Sr No	Main Columns
1	Salinity (> 3000 µS/cm)
2	Chloride (> 1000 mg/litre)
3	Fluoride (> 1.5 mg/litre)
4	Iron (> 1 mg/litre)
5	Arsenic (> 0.05 mg/litre)
6	Nitrate (> 45 mg/litre)
7	Total ground water quality

Problems faced during Data Mining step –

1. Some states do not have district wise information about certain variables.
2. Lack of uniformity in units used (eg. Some states use agricultural land in hectares while some in sq. km.)
3. Data not available for more than 30% of the districts. (eg. Secondary schools present in the district). Such variables cannot be considered for analysis.

MATHEMATICAL MODEL –

The formulae used to compute the derived columns are given as below:

1. Urban population (%) = $\frac{(urban\ population * 100)}{total\ population\ 2011}$

   This gives the percentage of population in that particular district which live in urban areas.

2. Population growth (%) = $\frac{(total\ population\ 2011 - total\ population\ 2001)}{total\ population\ 2011} * 100$

   The population growth from 2001 to 2011 is given by this formula. The percentage rise of total population seen by a district can be calculated.

3. Male Population growth (%) = $\frac{(male\ population\ 2011 - male\ population\ 2001)}{male\ population\ 2011} * 100$

   Similar to the total population growth formula, this only takes into account the rise in male population between 2001 and 2011.

4. Female Population growth (%) = $\frac{(female\ population\ 2011 - female\ population\ 2001)}{female\ population\ 2011} * 100$

   The formula is the same as Male population growth, just the difference here is that it calculates the growth in number of females during a given time.

5. Literacy growth (%) = $\frac{(total\ literacy\ 2011 - total\ literacy\ 2001)}{total\ literacy\ 2011} * 100$

   The growth in literacy rate from 2001 to 2011 is given by this formula. The growth is calculated in terms of percentage.

6. Male Literacy growth (%) = $\frac{(male\ literacy\ 2011 - male\ literacy\ 2001)}{male\ literacy\ 2011} * 100$

   The growth in male literacy rate is computed using this formula.

7. Female Literacy growth (%) = $\frac{(female\ literacy\ 2011 - female\ literacy\ 2001)}{female\ literacy\ 2011} * 100$

   Similar to the previous two formulae (5 and 6), this also calculates growth rate but for female literacy levels between 2001 and 2011.

8. Child population proportion (%) = $\frac{Child\ population * 100}{total\ population}$

The above formula gives the proportion of children below 6 years of age with respect to the total population.

9. Sex ratio growth (%) = $\frac{(sex\ ratio\ 2011 - sex\ ratio\ 2001)}{sex\ ratio\ 2011} * 100$

Sex Ratio Growth formula is similar to population and literacy growth formulae. It calculates the percentage growth in sex ratio levels between the given time frame.

10. Growth in IMR (%) = $\frac{(IMR\ 2011 - IMR\ 2001)}{IMR\ 2011} * 100$

IMR is the acronym for Infant mortality rate. The above formula gives the percentage increase or decrease in the Infant Mortality rate between a given time frame.

11. Growth in U5MR (%) = $\frac{(U5MR\ 2011 - U5MR\ 2001)}{U5MR\ 2011} * 100$

U5MR is the acronym for Under 5 Mortality Rate. The formula similar to IMR, computes the increase or decrease in U5MR levels.

12. Population per Government hospital = $\frac{Total\ population\ 2011}{Number\ of\ Government\ Hospital}$

Population per Government Hospital gives the number of people covered by one government hospital.

13. Percentage of area sown (%) = $\frac{(Net\ sown\ area * 100)}{Total\ area}$

The percentage of area utilized for agricultural purposes out of the total land area of that district.

14. Percentage of area under forests (%) = $\frac{(Forest\ area * 100)}{Total\ area}$

The percentage of area under forest out of total area of that district.

15. Percentage of sown area under irrigation (%) = $\frac{(Net\ irrigated\ area * 100)}{Net\ sown\ area}$

Proportion of area under irrigation out of total area under agriculture.

All the derived columns have been computed in Python. These derived columns are very useful to find the change in specific factors during the given period. Derived columns are pivotal in analysis and inferences.

The project mostly concentrates on Exploratory Data Analysis (EDA), data visualization and clustering algorithms. The following points will throw light on the basic understanding of the above-mentioned concepts.

- Exploratory Data Analysis (EDA) –

  The definition of Exploratory Data Analysis states that it is the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. EDA helps to gather information about the data before actually applying algorithms on it.[11]

- Data Visualization –

  In simple terms, data visualization is graphical representation of data. These visualizations are useful to find insights, trends and outliers. There are various types of visualization graphs in which data can be represented, such as scatter plots, bar plots, box plots, histograms, so on and so forth. They usually reveal much more information than the numbers can. [12]

- Clustering –

Clustering is a method to find similar groups in a dataset. There are many types of clustering algorithms. Firstly, the connectivity models which consists of hierarchical clustering, where one single cluster is further partitioned into many, based on the distances between them. Usually, these models are very easy for interpretation. Secondly, the centroid models are based upon the distance of that particular point from the centroid of the cluster. K-means is an example of this technique.

→ Hierarchical clustering –

Hierarchical clustering starts with the formation of one distinct cluster for one individual point. Then these points are merged into same clusters after computing the distance between them. The decision on the number of clusters to be chosen can be inferred from a dendrogram. The best choice of the number of clusters is based upon the vertical line that traverses maximum distance without being further divided. The biggest advantage of hierarchical based clustering algorithm is that the results are reproducible (they do not change even after multiple runs through the code).[12]

→ Metrics –

The two commonly used metrics for distance measurement are Euclidean and Manhattan distances. The formulae for both of them are as given below.

1. Euclidean distance = $(\Sigma (ai - bi))^{½}$
2. Manhattan distance = $\Sigma |ai - bi|$

Apart from these two, there are also some more metrics used in computing distance such as the Mahalanobis distance, Squared Euclidean distance and Maximum distance. [13]

→ Agglomerative clustering methods –

Agglomerative clustering method is a part of hierarchical clustering. In this method, all the data points are considered as separate clusters initially. These are then combined in subsequent iterations based on the distance between them until they merge to form one single cluster. This can be represented in a diagram known as dendrogram as mentioned earlier. Agglomerative clustering can further be divided into four distinct types – [14]

1. Complete linkage clustering – In this technique all pairwise dissimilarities are computed and then the largest values between the clusters are considered as the threshold for dividing the data points. Complete linkage method tends to produce more compact clusters. [15]

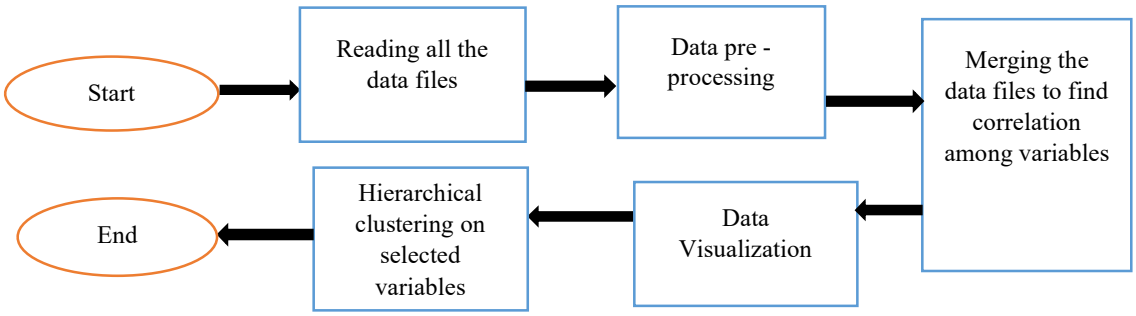
2. Single linkage clustering – In this type of agglomerative clustering technique, the pairwise distances for the two closest points are computed and then it is combined to form one cluster. It usually creates long and loose clusters. The drawback of this method is that points which are very farther from each other may also be considered in the same cluster. [16]

3. Average linkage clustering – Average linkage clustering calculates all the pairwise dissimilarities between the points and then takes average of them.

4. Ward's minimum clustering – This method minimizes the within cluster variance. At each step the clusters with minimum between-cluster distances are merged. [14]

FLOW CHART –

Algorithmic flowchart conveys the steps involved in making of a project. This project consists of five main steps and they are as follows:



Step 1: The data gathered has been segregated in five different Excel (comma separated value) files. The files are-

a. Demographics data
b. Economic data
c. Health data
d. Land Use data
e. Ground Water Quality data

Step 2: Data pre - processing forms an important step in any data science project. This step includes renaming the columns, finding out the null values if any, identifying the unique column from all the five files.

Step 3: On the basis of the unique primary key column (the name of the district is the primary key column in this project as it is unique), the csv files are merged. Computing the summary statistics, correlation heatmaps and extensive exploratory data analysis are the tasks performed. EDA helps to understand the data and what relationships can be established.

Step 4: Data Visualization is also called Data Storytelling. Visuals or graph tell us much more information than numbers can reveal. Most of the inferences can be established with this method. This step is the backbone of this project. Graph such as scatter plots, bar plots, box plots and heatmaps are used for this purpose.

Step 5: Hierarchical Clustering algorithms are applied on the variables to find the districts which fall under one cluster where a specific problem is prevalent. Moreover, what variables can be improved upon to bring these districts on par with those who are performing better.

The algorithm for hierarchical clustering is as follows –

1. Selecting the desired columns to form clusters.
2. Deciding the appropriate linkage for the clustering analysis and formation of dendrogram.
3. Selection of k (number of clusters to be formed) on the basis of dendrogram.
4. Fitting the dataset on the model.
5. Visualizing the clusters with the help of graph.

The project has been completed in Python language. The platform used for data cleaning, merging, Exploratory Data Analysis, Data Visualization and Clustering is Jupyter Notebooks.

The libraries used for this project are –

- NumPy (Numerical processing tasks)
- Pandas (Data pre-processing and transformation)
- Matplotlib (Visualization library)
- Seaborn (Visualization library)
- Scikit learn (For Clustering Analysis)
- SciPy – Scientific Python library (Used for Agglomerative Hierarchical Clustering)

RESULTS AND INFERENCES –

Result is the final outcome of this project. The result is mainly based upon data visualization and hierarchical clustering analysis.

1. Demographic Studies

→ Population Studies –

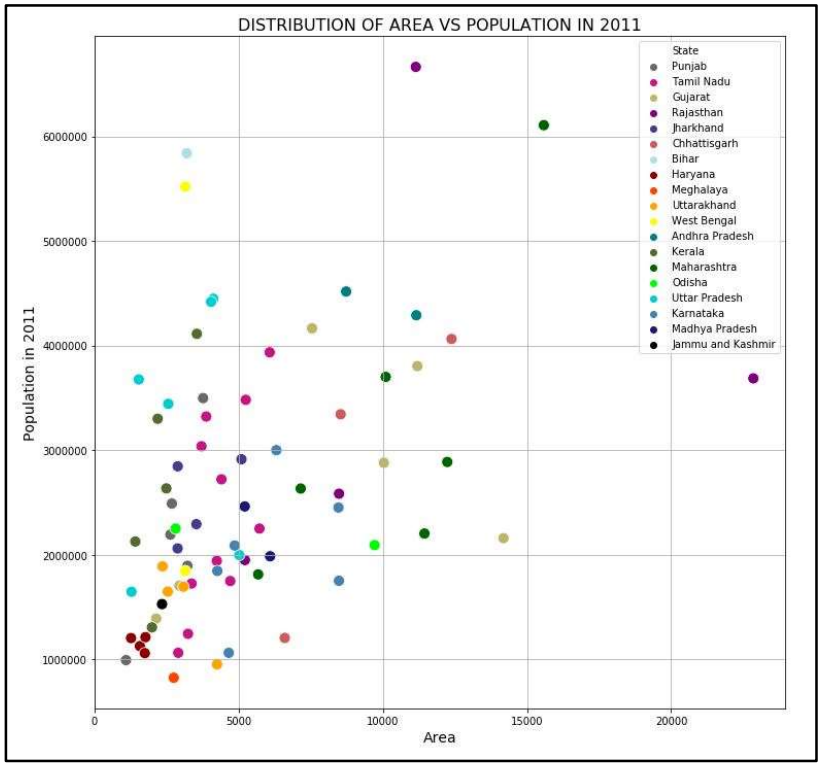


Fig3. Distribution of area and population in 2011

Overall, there is a linear trend between Area and Population. There are some outliers in this scatter plot. The point on the extreme right is Jodhpur district of Rajasthan. It is the semi – urban district with largest area of 22850 sq.km. It is followed by Nashik district of Maharashtra having an area of 15582 at second place and Jamnagar district of Gujarat at third place having an area of 14184 sq.km. On the other hand, SAS Nagar (Mohali) district in Punjab is the smallest semi – urban district by area. Its area is 1094 sq.km. Panipat in Haryana having 1268 sq.km. and Gautam Buddha Nagar (Noida) of Uttar Pradesh having 1282 sq.km. occupy second and third place respectively in terms of smallest areas.

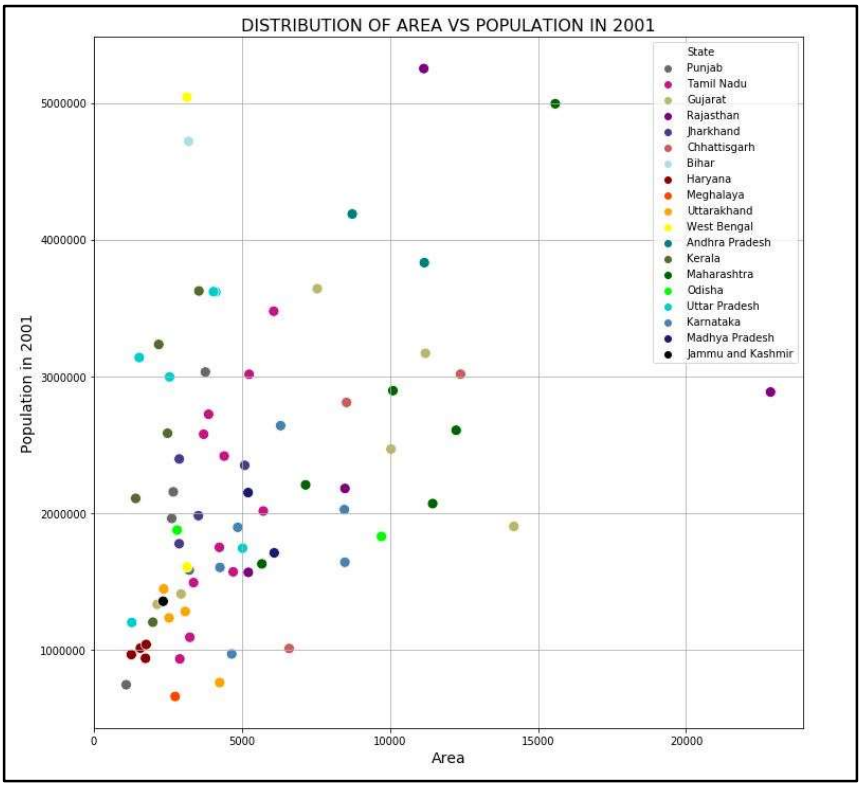


Fig4. Distribution of area and population in 2001

The table below depicts the most populous semi – urban districts in 2001 and 2011.

Rank	District	Population 2011	Rank	District	Population 2001
1	Jaipur	6663971	1	Jaipur	5251071
2	Nashik	6107187	2	Hooghly	5041976
3	Patna	5838465	3	Nashik	4993796
4	Hooghly	5519145	4	Patna	4718592

From the above table, it is evident that Jaipur in Rajasthan is the most populous semi – urban district in 2001 as well as 2011. Hooghly district of West Bengal has seen slow population growth in the decade. Nashik (Maharashtra) and Patna (Bihar) which were at third and fourth position respectively in 2001. In 2011, they surpassed Hooghly to stand on second and third position.

The table below gives information about the least populated semi – urban districts in 2001 and 2011.

Rank	District	Population 2011	Rank	District	Population 2001
1	East Khasi Hills	825992	1	East Khasi Hills	660923
2	Nainital	954605	2	SAS Nagar	746987
3	SAS Nagar	994628	3	Nainital	762909

East Khasi Hills district of Meghalaya has been the smallest semi – urban district in terms of population for 2001 as well as 2011. SAS Nagar (Mohali) of Punjab has seen faster population growth in the decade from 2001 to 2011. The Nainital district of Uttarakhand occupies the second place in terms of low population in 2011.

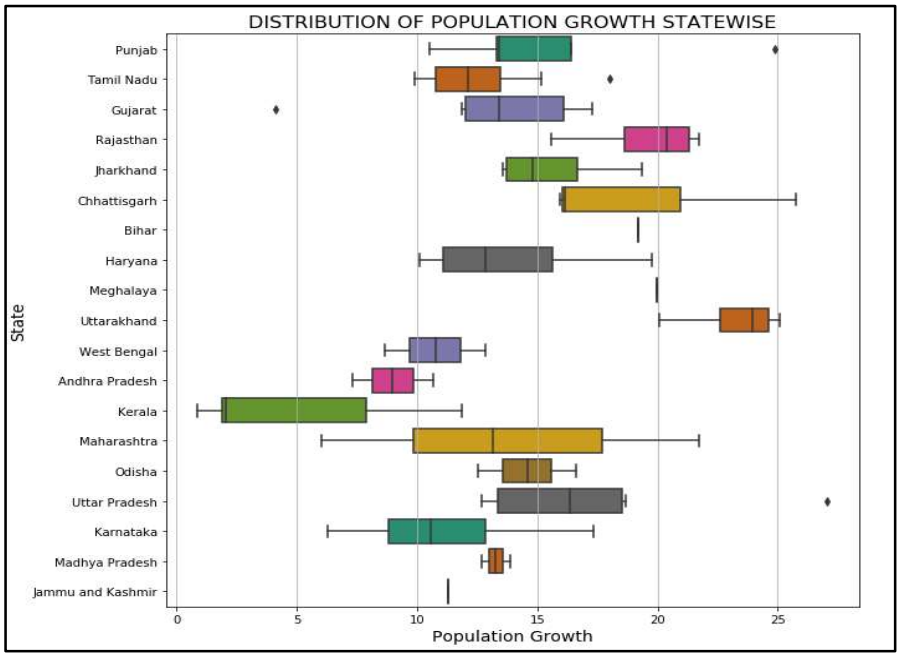
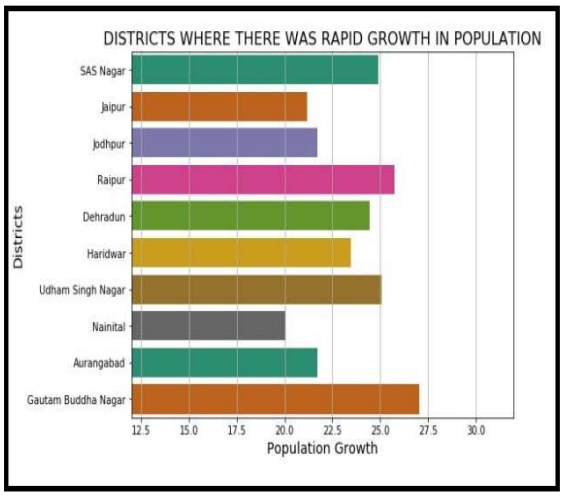


Fig5. Population growth state – wise

The above boxplot shows the overall population growth from 2001 to 2011 state - wise. States like Uttarakhand, Chhattisgarh, Rajasthan with some districts of Punjab, Maharashtra and Uttar Pradesh have witnessed high population growth (more than 20%). Meanwhile, Kerala state has seen slowest growth in population during the same period.

The next graphs will give information on the districts which have witnessed rapid and slow growth in population.



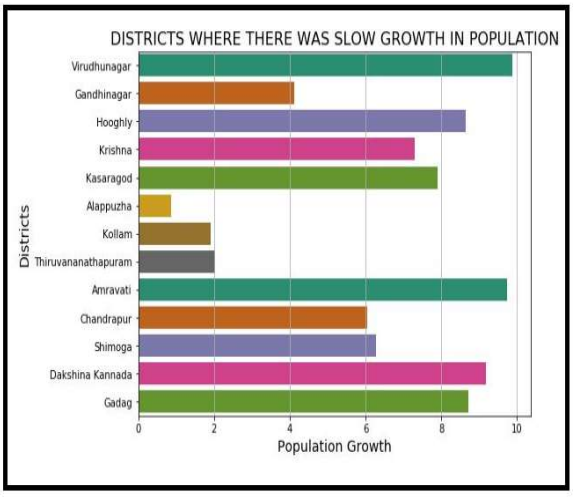


Fig6. Districts with rapid growth of population

Fig7. Districts with slow growth in population

Gautam Buddha Nagar (Noida) of Uttar Pradesh has seen the fastest growth in population at nearly 27% from 2001 to 2011. It is followed by Raipur the capital of Chhattisgarh at 26%. This is followed by Udham Singh Nagar district of Uttarakhand and SAS Nagar district of Punjab with nearly 25% growth in population. Most of the districts in this list come from Northern part of India.

The graph on the right depicts the districts which have witnessed slow growth in population. Alappuzha district of Kerala has seen the slowest population growth followed by Kollam and Thiruvananthapuram of Kerala. All the three districts have a growth rate of less than 2%. Most of the districts witnessing slow population growth are from Southern part of India.

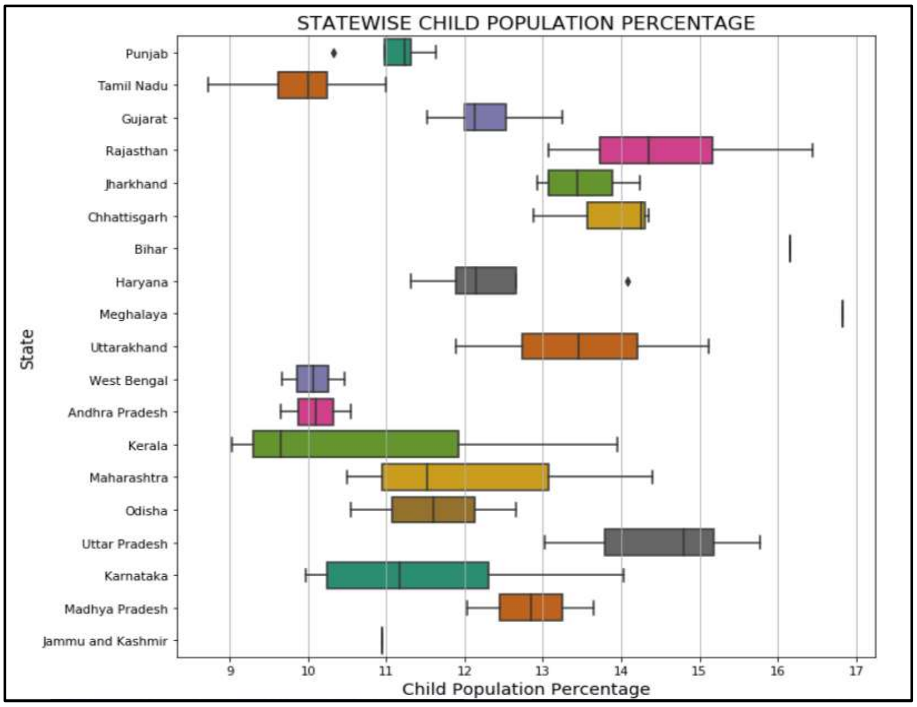
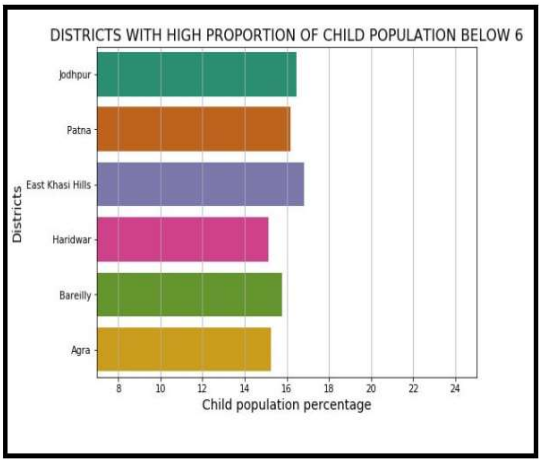


Fig8. State wise percentage of child population below 6 years

The above boxplot depicts the percentage of state - wise child population below 6 years. The districts in Meghalaya, Bihar, Rajasthan and Uttar Pradesh have more than 15% of their population in 2011 below 6 years of age. Some districts of Kerala, Tamil Nadu, Andhra Pradesh and West Bengal have child population below 10%. Average child population below 6 years of age is between 12 – 14% in semi – urban districts of India.



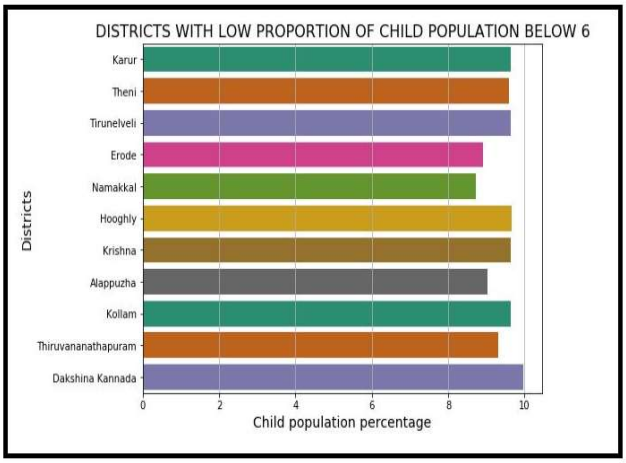


Fig9. Districts with high child population

Fig10. Districts with low child population

The figure on the left depicts the districts with high proportion of children below the age of 6. East Khasi Hills (Shillong) district of Meghalaya has approximately 17% of population below the age of 6. It is followed

by Jodhpur (Rajasthan) and Patna (Bihar) at just above 16%. The other three districts are Bareilly (UP), Haridwar (Uttarakhand) and Agra (UP). All the districts having high proportion of child population are from North India.

The figure on the right shows the districts having low proportion of child population. Lowest child population among semi – urban districts can be seen in Namakkal district of Tamil Nadu at 8.7%. The second place is taken by Erode district of Tamil Nadu having 8.9% child population. Alappuzha and Thiruvananthapuram districts of Kerala occupy 3rd and 4th position respectively with 9% and 9.3%. Most of the districts having low child population are from South India.

Clustering analysis on Population growth –

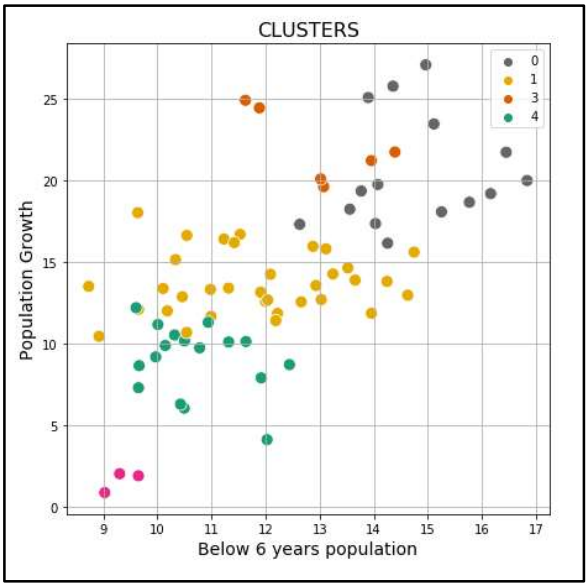


Fig11. Clusters containing Population growth and child population

There is a linear trend between Child population and Population Growth. For the above graph, hierarchical clustering method is used. The number of clusters considered is 5 with Manhattan distance metric. Linkage method used for analysis is 'complete'. This helps us to infer that as the child population increases, the overall population also increases. However, there are some outliers in this model. The districts in orange have low child population proportion but the population has increased rapidly. These districts are – SAS Nagar from Punjab, Jaipur and Kota from Rajasthan, Dehradun and Nainital from Uttarakhand and Aurangabad from Maharashtra.

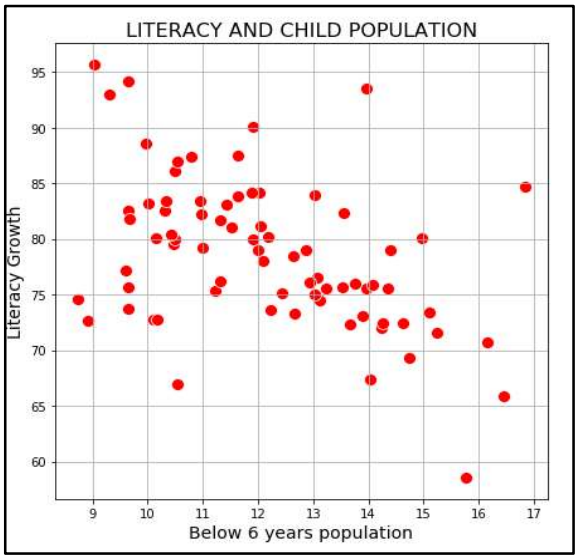


Fig12. Relationship between literacy and child population percentage

Literacy and child population percentage has a negative linear relationship. As the level of literacy increases, the proportion of child population decreases.

→ Urbanization Studies –

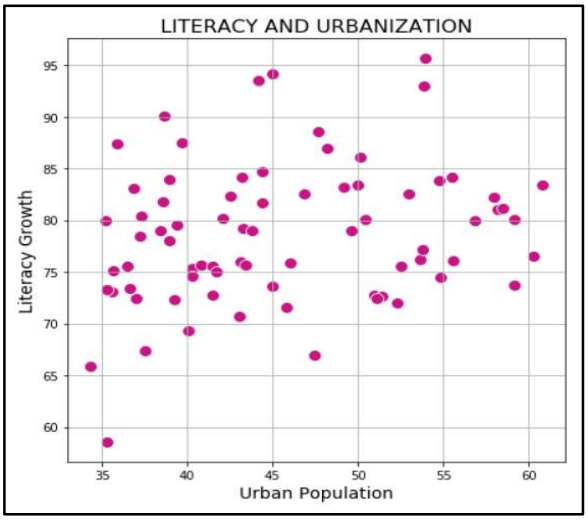


Fig13. Relationship between Literacy and Urbanization

From the above graph it is evident that, as literacy increases the proportion of population living in urban areas increases. This shows that rural areas do not have much job opportunities for literate people and hence they have to move to urban areas.

→ Literacy Studies -

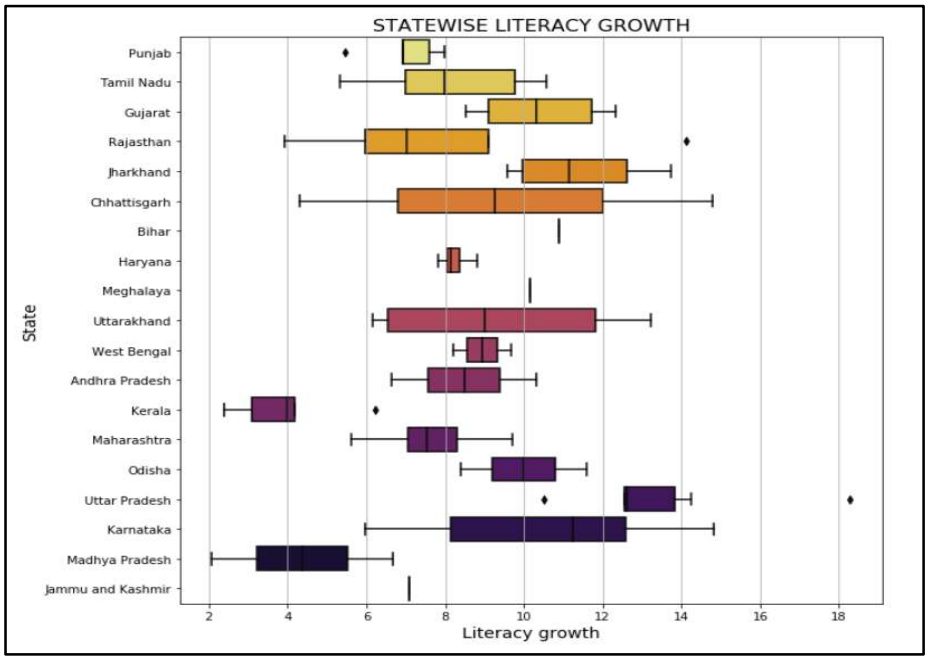


Fig14. State-wise Literacy Growth

The above figure depicts the state – wise literacy growth from 2001 to 2011. Kerala, Madhya Pradesh and certain districts of Rajasthan and Chhattisgarh have seen low rise in literacy levels. All the districts of Kerala have a literacy rate of more than 90%. On an average, the semi – urban districts have seen their literacy rate grow by 8 – 10%. For some districts of Uttar Pradesh this rate is much higher (around 15% increase).

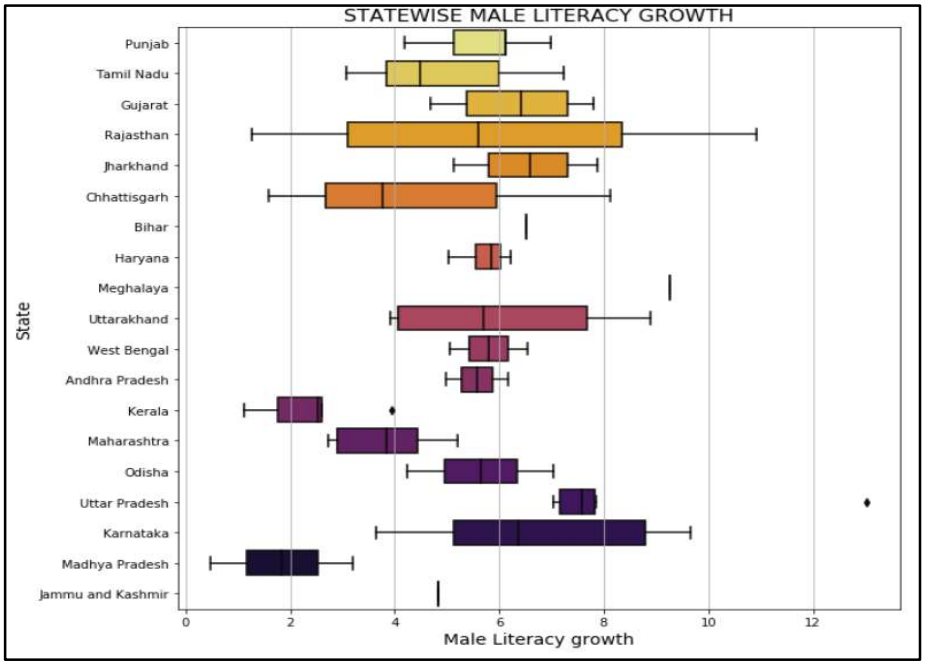


Fig15. State – wise male literacy growth

Unlike overall literacy, male literacy has not seen much increase in the decade from 2001 to 2011. On an average, male literacy levels rose by 4 – 6%. Madhya Pradesh has seen very low rise in male literacy levels

with Kerala and some districts of Rajasthan and Chhattisgarh. States like Rajasthan, Chhattisgarh, Uttarakhand and Karnataka have a lot of disparity in terms of male literacy growth.

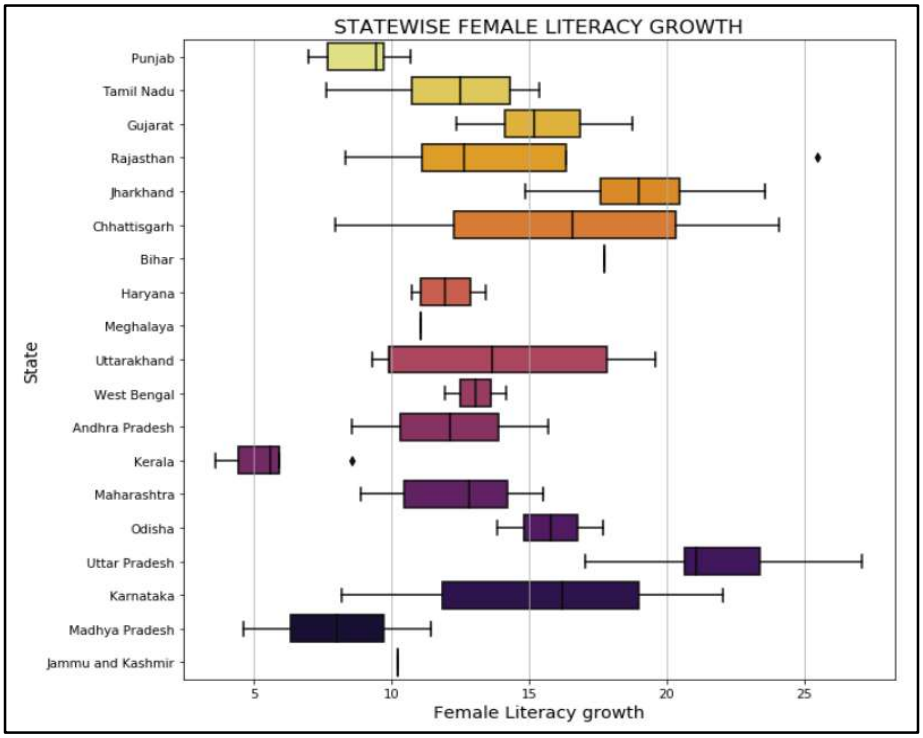


Fig16. Female literacy growth state – wise

The female literacy has grown at a faster pace than male literacy levels. Districts from Uttar Pradesh, Jharkhand, Rajasthan and Karnataka have seen a rapid growth in female literacy. Average growth of female literacy from 2001 to 2011 is between 10 – 16%.

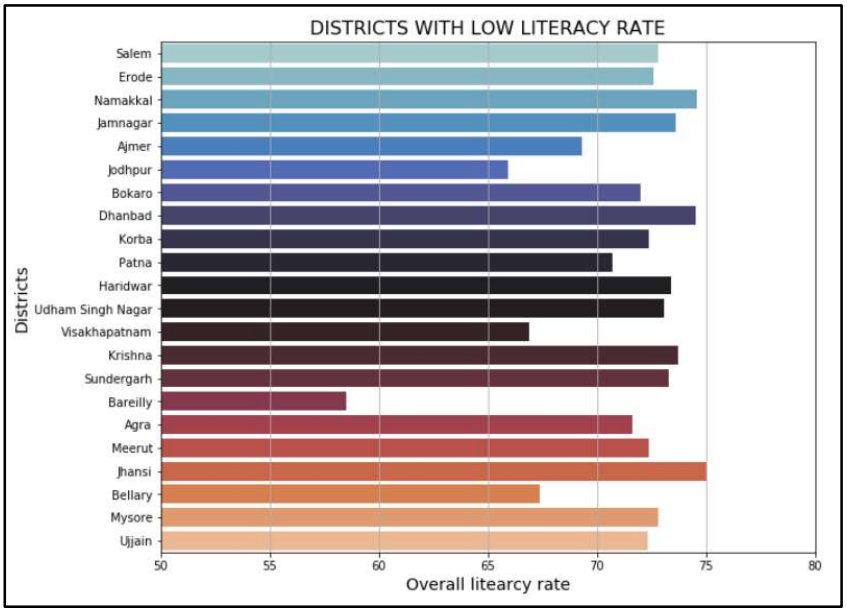


Fig17. Districts having low literacy rate

The below table gives the information about the districts with lowest literacy rate:

Rank	District	Literacy 2011	Rank	District	Literacy 2001
1	Bareilly	58.5%	1	Bareilly	47.8%
2	Jodhpur	65.9%	2	Jodhpur	56.6%
3	Visakhapatnam	66.9%	3	Bellary	57.4%
4	Bellary	67.4%	4	Visakhapatnam	60%
5	Ajmer	69.3%	5	Korba	61.7%

From the table, it can be inferred that 4 of the districts present in 2001 with low literacy rates are also present in 2011. Bareilly district of Uttar Pradesh has the lowest literacy rates of all semi – urban districts in India.

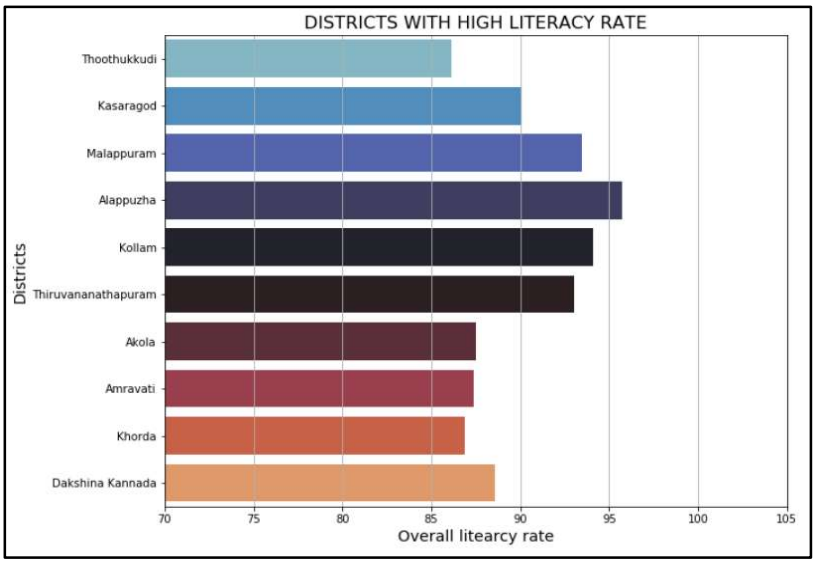


Fig18. Districts with high literacy rate

Rank	District	Literacy 2011	Rank	District	Literacy 2001
1	Alappuzha	95.7%	1	Alappuzha	93.4%
2	Kollam	94.1%	2	Kollam	91.2%
3	Malappuram	93.5%	3	Malappuram	89.6%
4	Thiruvananthapuram	93%	4	Thiruvananthapuram	89.3%
5	Kasaragod	90.1%	5	Kasaragod	84.5%

Kerala has clean swept all the districts in terms of literacy rate. Moreover, all the semi - urban districts of Kerala have maintained their ranks even though they have seen a rise in literacy. Apart from the districts of Kerala, Dakshina Kannada (Mangalore) of Karnataka, Akola and Amravati districts of Maharashtra, Thoothukkudi district of Tamil Nadu and Khorda district of Odisha have high literacy rate.

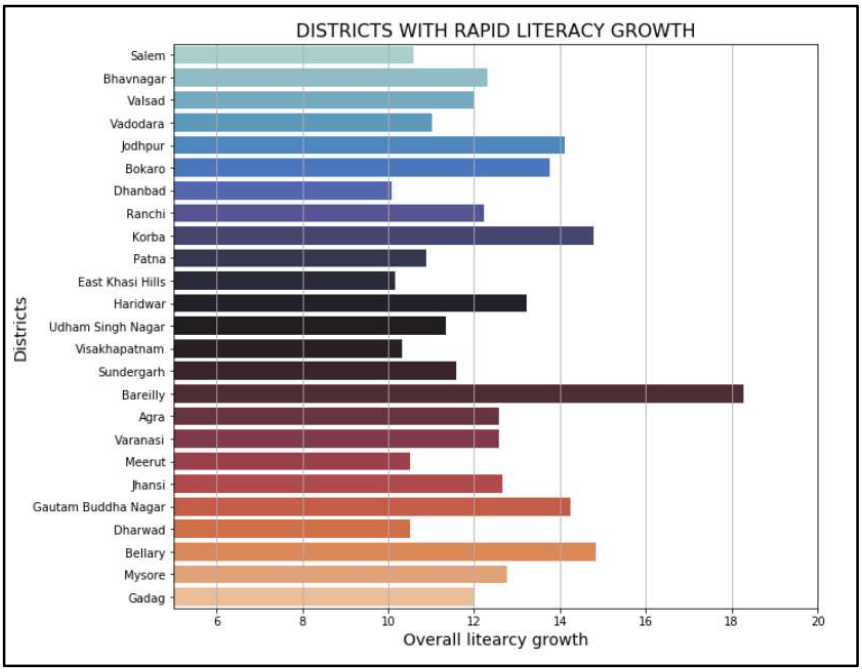


Fig19. District witnessing rapid rise in literacy levels

The above plot shows the districts which have seen a rapid rise in literacy levels. Bareilly district of Uttar Pradesh though having a low literacy rate, has the highest rise in literacy levels. Other districts which have witnessed an increase of 14% in literacy rate are – Jodhpur (Rajasthan), Korba (Chhattisgarh), Gautam Buddha Nagar (Noida) (UP) and Bellary (Karnataka).

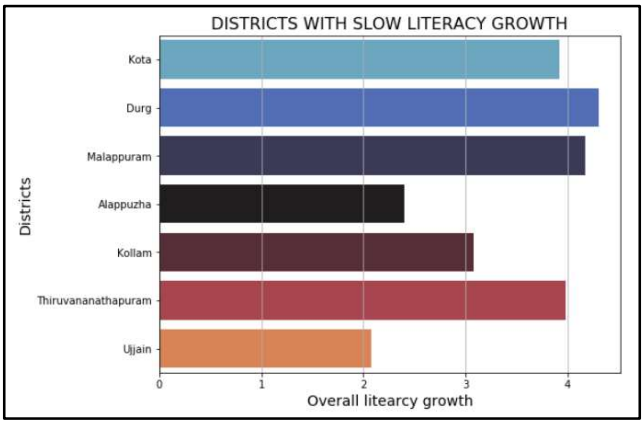


Fig20. Districts witnessing slow rise in literacy levels

Ujjain district of Madhya Pradesh has seen only 2% rise in literacy level in a decade. The other non – Kerala districts are Kota (Rajasthan) and Durg (Chhattisgarh) with around 4% increase in overall literacy levels.

→ Working Class Studies

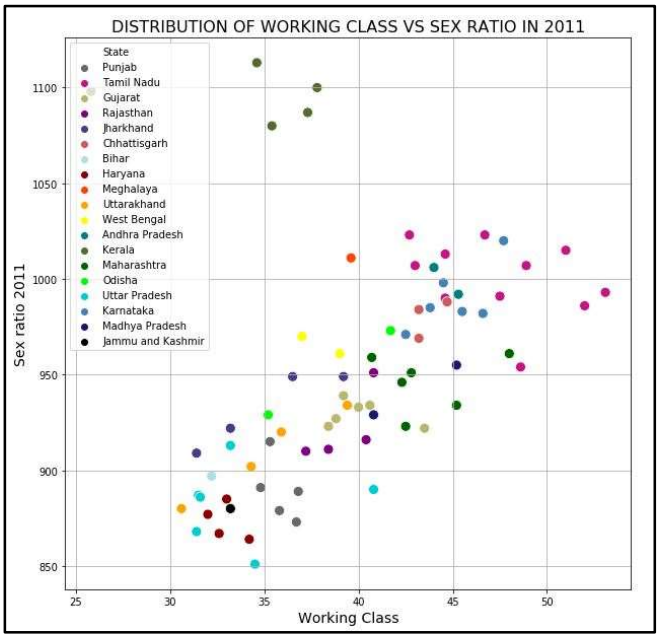


Fig21. Distribution of Working Class with respect to Sex Ratio 2011

Sex Ratio and Working-Class proportion of total population exhibit an increasing linear relationship. Except for the districts of Kerala located on the top portion of the graph all can be seen following this trend. Towards the right side of the graph, districts of Tamil Nadu and Karnataka can be seen. They have high sex ratio and also have nearly half of the population categorised as working class. On the other hand, near the bottom left corner of the graph are the districts of Haryana, Uttar Pradesh and Punjab having low sex ratio as well as low working-class population. This graph shows one of the most important highlights of the project. Females are equally important in the economy of a particular region. Better sex ratio also means higher proportion of working - class group. Better the sex ratio, better are the opportunities for women in economic activities of the region.

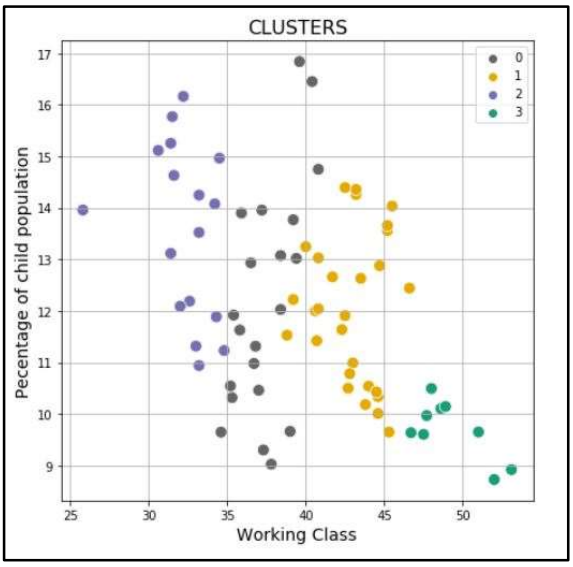


Fig22. Clustering analysis of Working Class with child population

The above scatter plot depicts the Working - class population with respect to percentage of Child population. There is a decreasing linear relationship between the two. More the working - class population, lower is the child population. The cluster of districts formed in the right bottom corner depict the low percentage of child population and high percentage of working class. The districts in this cluster are – Karur, Theni, Tirunelveli, Salem, Erode, Namakkal, Virudhunagar from Tamil Nadu, Chandrapur from Maharashtra and Dakshina Kannada (Mangalore). All of these districts are located in Southern part of India.

2.Economic Studies

Studying the economic factors of semi – urban districts is one of the most important part of this project.

→ Below Poverty Line –

The definition of Below poverty line in India :- A person who earns less than Rs 32 a day is said to be below poverty line.

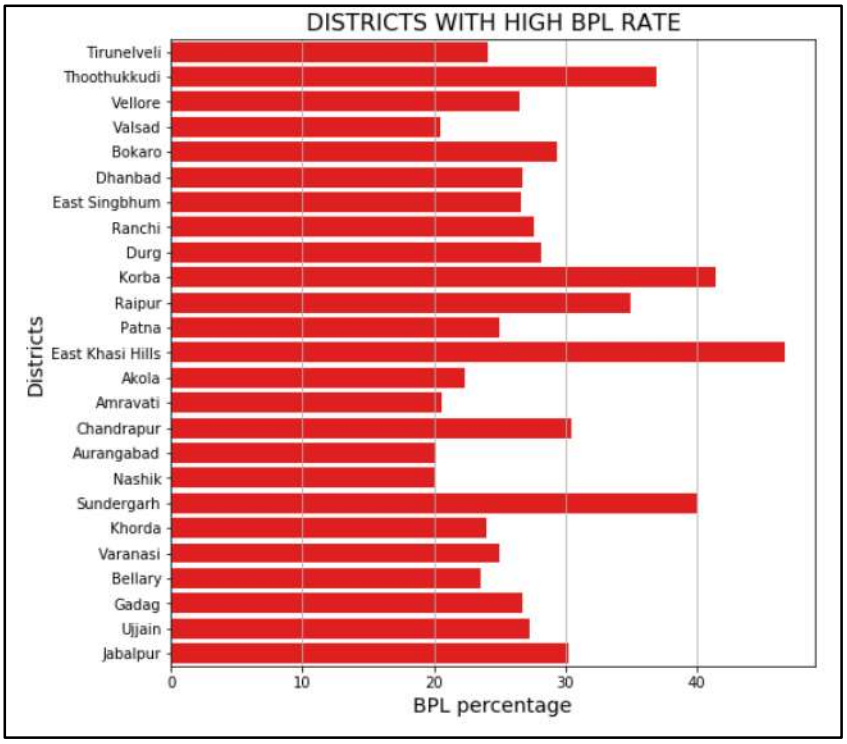


Fig23. Districts with high BPL population

East Khasi Hills (Shillong) district of Meghalaya has highest proportion of BPL population among semi – urban districts. It has 46% people below poverty line. Second place is occupied by Korba district of Chhattisgarh at 41%. Sundergarh district (Odisha) comes third at 40% BPL population followed by Thoothukkudi (Tamil Nadu) 37% and Raipur (Chhattisgarh) with 35%.

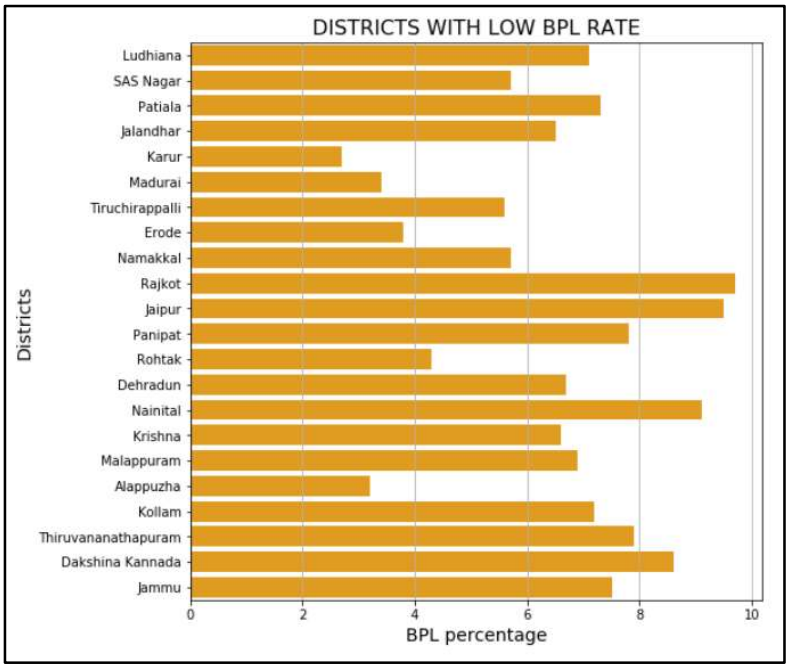


Fig24. Districts with low BPL population

Karur district of Tamil Nadu has the lowest people below poverty line at 2.5%. It is followed by Alappuzha (Kerala) at 3%, Madurai at 3.5%, and Erode at 3.8% of Tamil Nadu and Rohtak (Haryana) at 4.2%.

Clustering Analysis on BPL –

Agglomerative hierarchical clustering was applied. The linkage method used is 'complete' and affinity (metric) for distance is Manhattan.

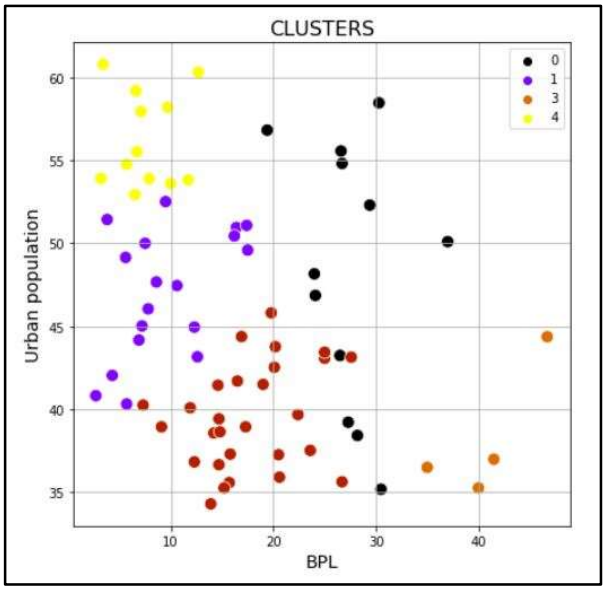


Fig25. Clustering analysis on BPL percentage with respect to Urban population

The yellow cluster on the top left corner of the graph has low BPL rate and high percentage of population living in urban areas. These districts are – Ludhiana, SAS Nagar, Jalandhar, Amritsar all from Punjab, Theni and Madurai from Tamil Nadu, Rajkot (Gujarat), Kota (Rajasthan), Dehradun (Uttarakhand), Krishna (Andhra Pradesh), Alappuzha and Thiruvananthapuram from Kerala.

On the right bottom corner, a cluster has been formed of districts in orange. The main characteristic of these districts is they have low urban population and very high BPL population. These districts are – Korba and Raipur from Chhattisgarh, East Khasi Hills (Meghalaya) and Sundergarh (Odisha).

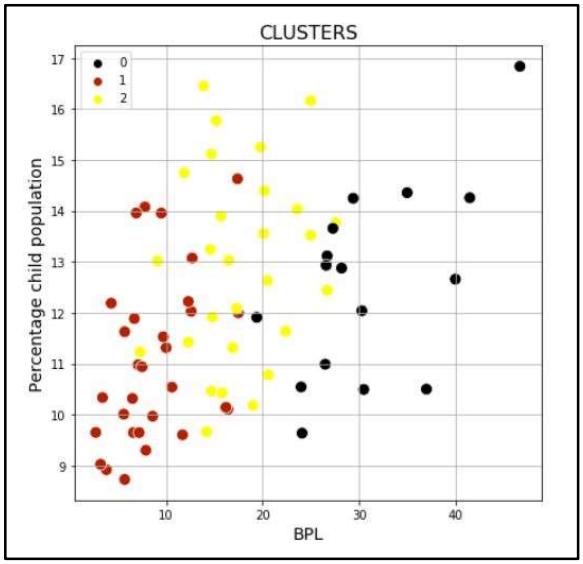


Fig26. Clustering analysis for BPL rate with respect to child population percentage

From the above scatterplot it is evident that, as the child population proportion increases, the population living under Below Poverty Line increases too. The red cluster consists of those districts having low child population and low BPL rate. Yellow cluster is the largest cluster amongst the three. This group consists of districts having moderate Below Poverty Line population. The districts in black needs attention as they have very high BPL population. Moreover, majority of them have child population under 6 years consisting of more than 14% of entire population.

→ Per Capita Income –

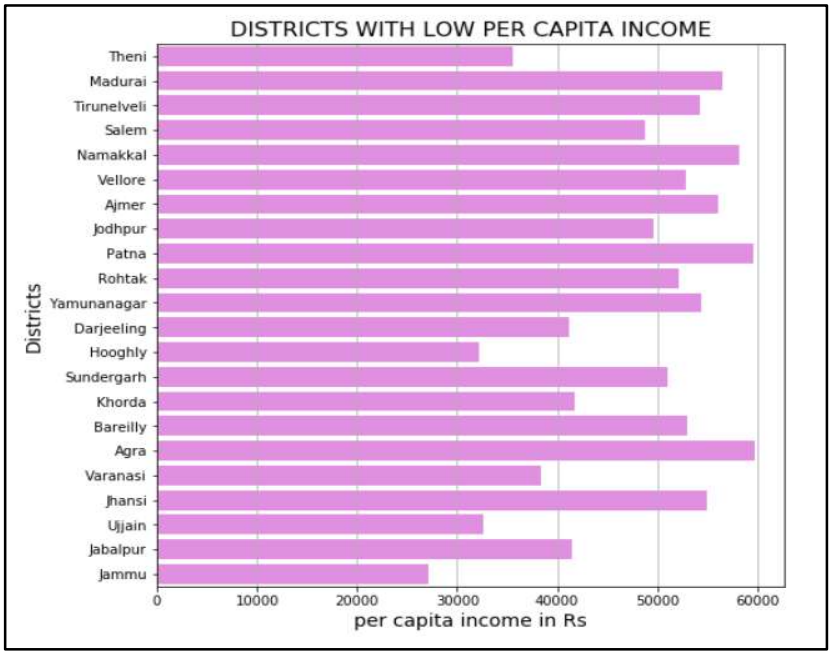


Fig27. Districts with low per capita income

The above graph represents the districts with low annual per capita income. The below table gives information about the five districts with lowest per capita income among all semi – urban districts. From the table, it can be inferred that low per capita income has no relation with below poverty line population.

District	Average Per capita income (in Rs)	Below Poverty Line
Jammu (Jammu and Kashmir)	27095	7.5%
Hooghly (West Bengal)	32223	14.2%
Ujjain (Madhya Pradesh)	32567	27.3%
Theni (Tamil Nadu)	35539	11.7%
Varanasi (Uttar Pradesh)	38407	25%

The districts having low below poverty line population despite having low per capita income signify the economic equality present in that particular district.

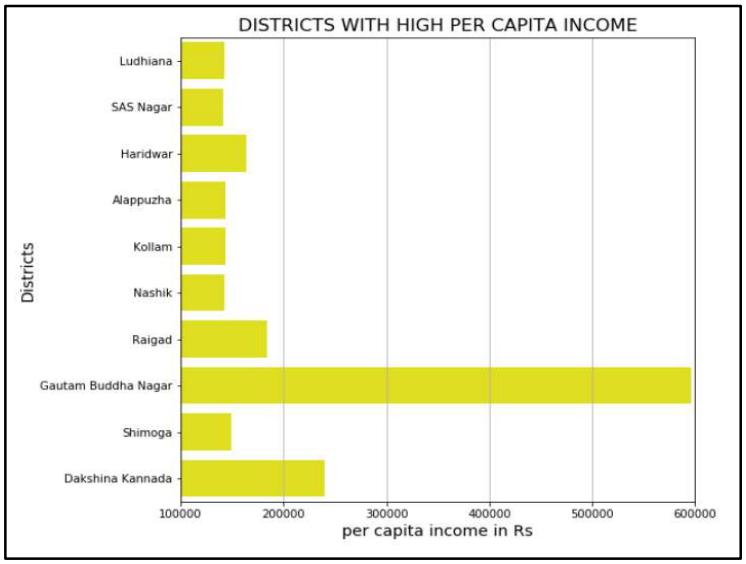


Fig28. Districts with high per capita income

The above plot depicts the semi – urban districts having highest per capita income. Gautam Buddha Nagar (Noida) district of Uttar Pradesh has extremely high per capita income. It is more than double of the second placed Dakshina Kannada district of Karnataka.

District	Average Per capita income (in Rs)	Below Poverty Line
Gautam Budhha Nagar (Noida)	595551	14.2%
Dakshina Kannada (Mangalore)	240448	8.6%
Raigad (Maharashtra)	184215	12.3%
Haridwar (Uttarakhand)	163869	14.7%
Shimoga (Karnataka)	148979	15.8%

Some districts despite having high per capita income have high below poverty line population. This may be possible because of the extreme economic inequality present.

Note - For further analysis Gautam Buddha Nagar district has been dropped. Moreover, per capita income district wise data is not available for the states of Gujarat, Jharkhand and Chhattisgarh. Hence, not considered for analysis on per capita income studies only.

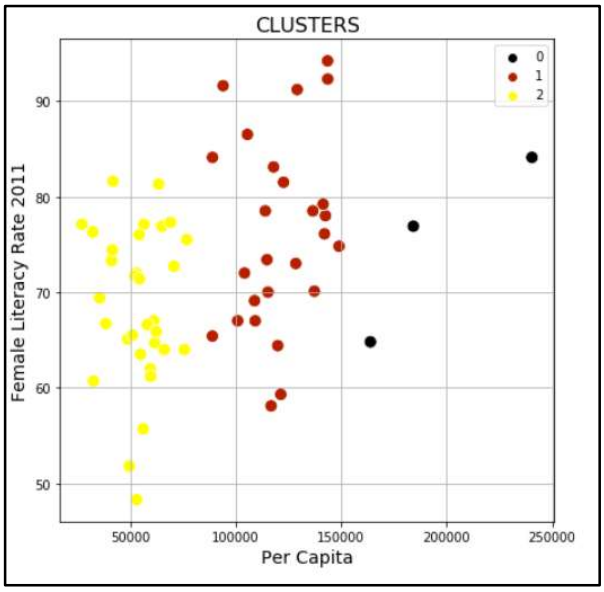
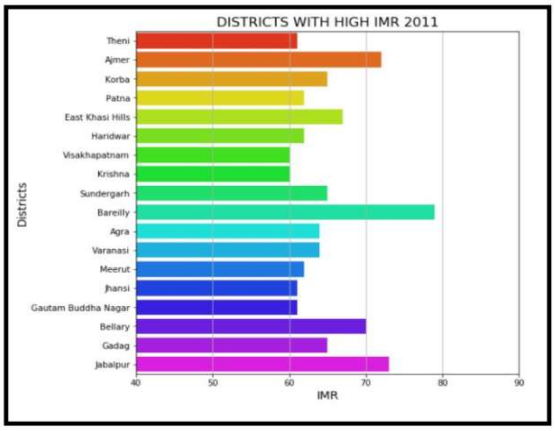


Fig29. Clustering analysis for per capita income and Female literacy 2011

Female literacy plays a very important role in driving per capita income for a region. The black dot on extreme right side depicts the Dakshina Kannada district (Karnataka). It has a high female literacy rate. The districts which are low in female literacy are also low in per capita income.

3.Health Studies

→ Infant Mortality Rate (IMR) -



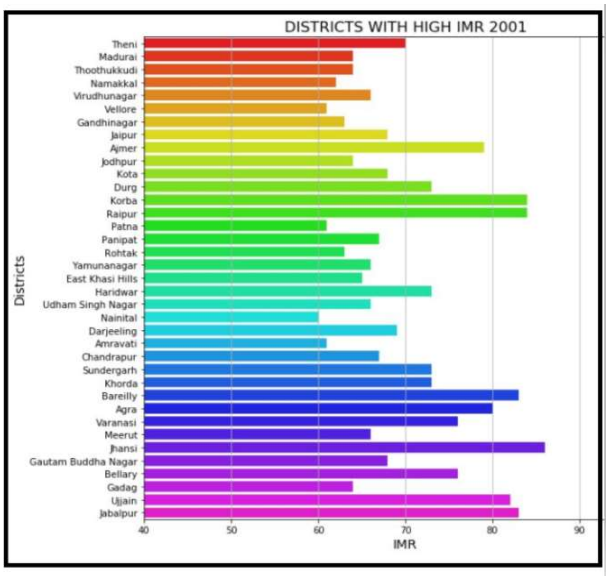


Fig30. Districts having high IMR 2011

Fig31. Districts with high IMR 2001

The above two graphs represent the Districts having high Infant Mortality Rate. In 2001, 37 districts had IMR more than 60 per thousand. This number significantly reduced to 18 districts in 2011. Also, in 2001, Korba (Chhattisgarh), Raipur (Chhattisgarh), Bareilly (UP), Jhansi (UP), Ujjain (MP) and Jabalpur (MP) had IMR more than 80 per thousand. In 2011, none of the districts had their IMR above 80 per thousand. The highest IMR is seen in Bareilly (UP) at 78 per thousand, followed by Jabalpur (MP) and Ajmer (Rajasthan) having 72 per thousand and 71 per thousand respectively.

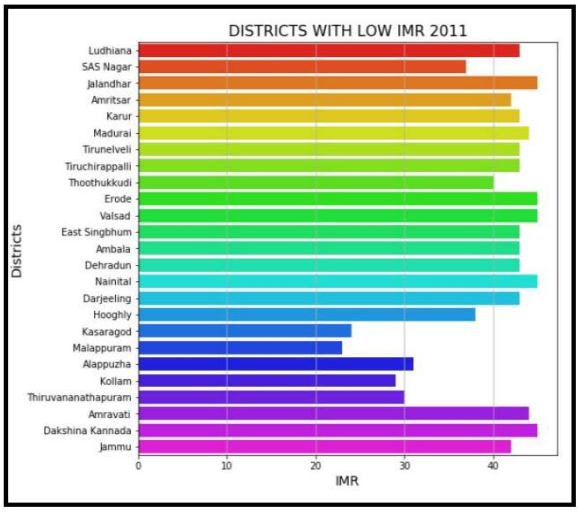


Fig32. Districts with low IMR 2011

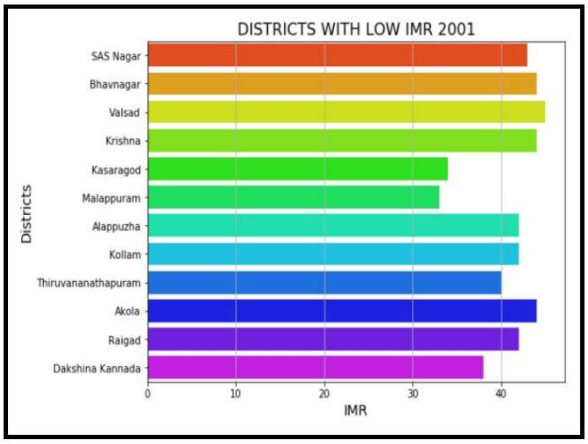


Fig33. Districts with low IMR 2001

The above graphs represent the districts which have low Infant Mortality Rate. In 2001,12 districts had IMR below 45 per thousand. This number rose to 25 districts having IMR less than 45 per thousand. Kerala has the lowest IMR in the country and that too by a very large margin.

→ Under Five Mortality Rate (U5MR)–

The five semi – urban districts of Kerala have the least Under Five Mortality Rate. Malappuram has the least U5MR at 29 per thousand children. It is followed by Kasaragod having 30 per thousand at second place. The third place is occupied by Kollam at 38 per thousand.

The districts having U5MR more than 90 per thousand in descending order are – Bareilly (114), Jabalpur (105), Ajmer (102), Bellary (98), East Khasi Hills (94), Gadag (92), Korba (91) and Sundergarh (90).

The Under 5 Mortality Rate has also decreased from 2001 to 2011.

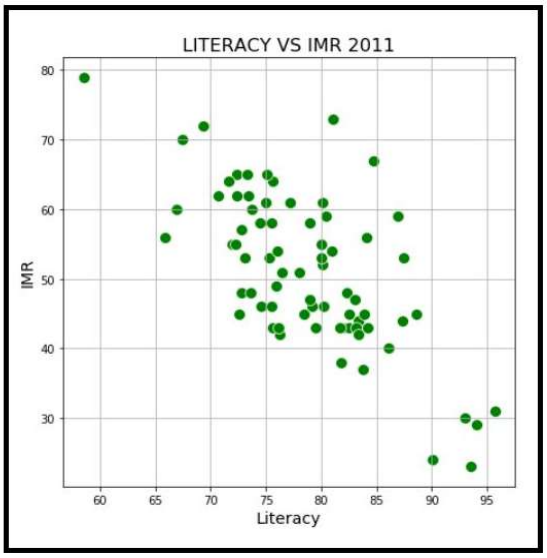


Fig34. Literacy rate and IMR 2011

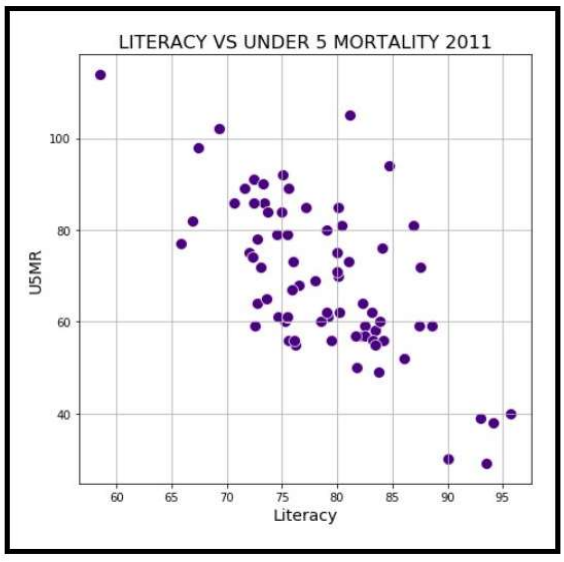


Fig35. Literacy rate and U5MR 2011

The graph on the left side, shows the relationship between Literacy and Infant Mortality Rate. The graph on the right depicts the relationship between Literacy and Under 5 Mortality Rate. Literacy is very

important factor related to child health. This means that literate parents can provide proper nourishment to the new born infants.

→ Population per Government Hospital –

Government Hospitals play an important role in the life of the people living in semi – urban areas. Most of the population in these areas is dependent on these Government hospitals for medical care.

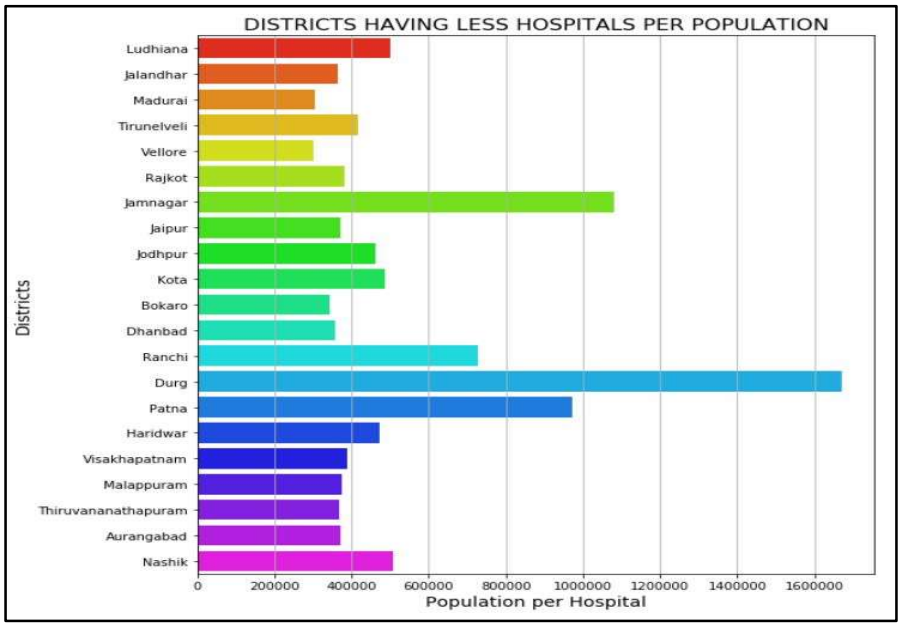


Fig36. Districts having more population per Government hospital

Durg district of Chhattisgarh has highest population per Hospital. One Government Hospital in Durg serves 16.5 Lakh people. Second place is occupied by Jamnagar district in Gujarat. 10.5 Lakh population is covered by 1 Government Hospital. Patna and Ranchi both have more than 6 Lakh people per Government Hospital.

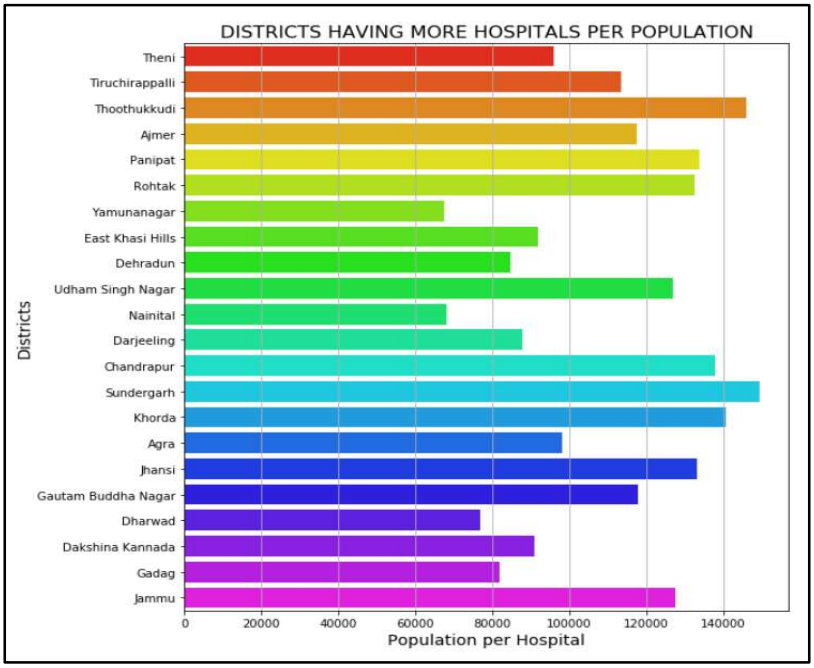


Fig37. Less population per Government Hospital

Nainital district of Uttarakhand has the least population per Government Hospital at 65000 per one Government Hospital. Yamunanagar district of Haryana comes on second place having 67000 and third place is occupied by Dharwad district in Karnataka having 76000 per Government Hospital.

4. Land Use

→ Rainfall -

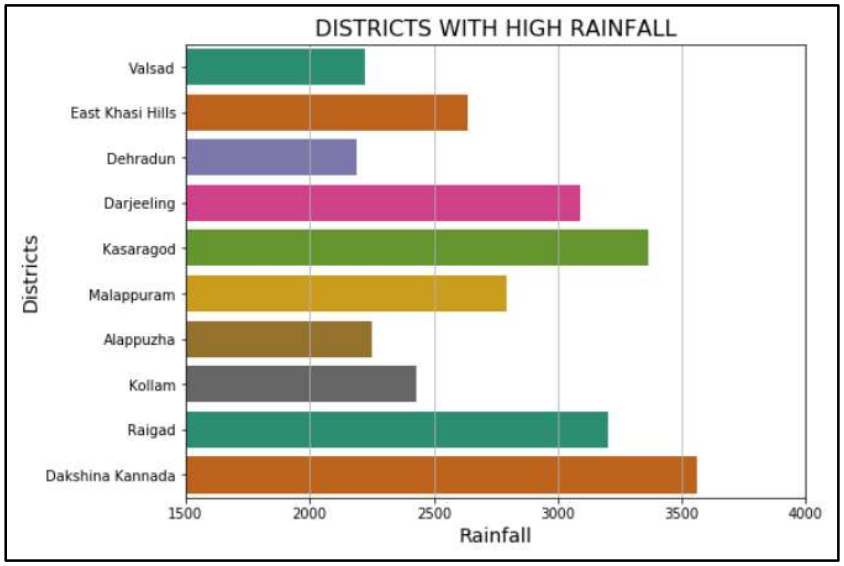


Fig38. Semi – urban Districts with high rainfall in mm

Dakshina Kannada (Mangalore) district of Karnataka has highest average annual rainfall of 3550 mm. It is followed by Kasaragod (Kerala) at 3550 mm, Raigad (Maharashtra) at 3250 mm and Darjeeling having 3100 mm. The other districts in this list are – Malappuram, East Khasi Hills, Kollam, Valsad, Alappuzha and Dehradun.

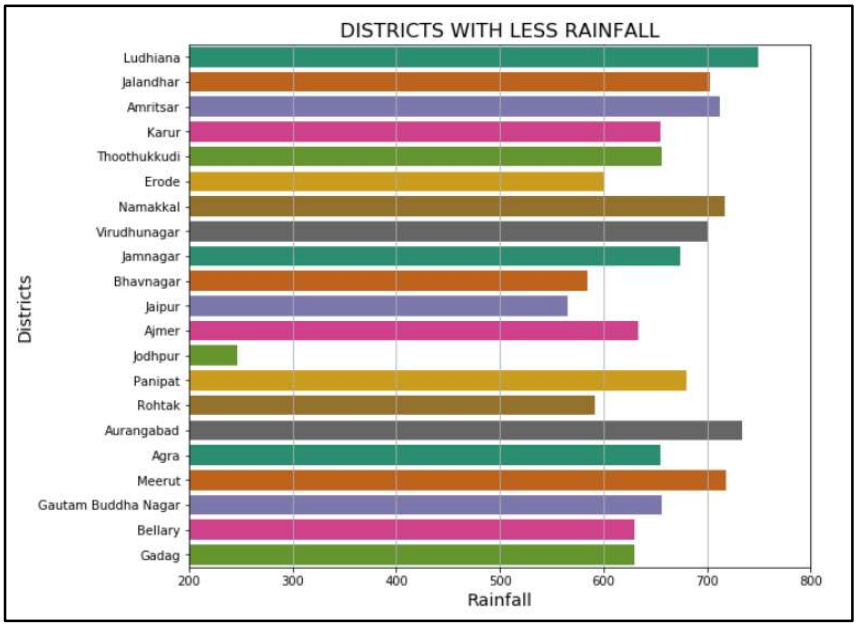


Fig39. Districts with less rainfall in mm

Jodhpur is the only district having rainfall of less than 500mm annually. Jodhpur on an average gets 250 mm of rainfall in a year.

→ Forest Area -

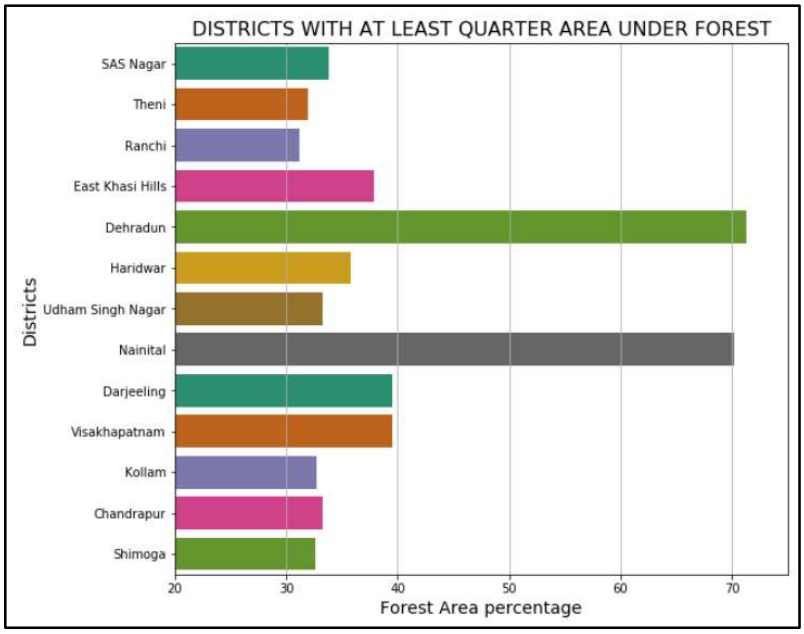


Fig40. Districts with at least quarter of their area under forests

Dehradun district of Uttarakhand has 72% area under forests. It is followed by Nainital (Uttarakhand) at 70%. Darjeeling (West Bengal) and Visakhapatnam (Andhra Pradesh) have 39% of their total area occupied by forests.

→ Irrigated Area -

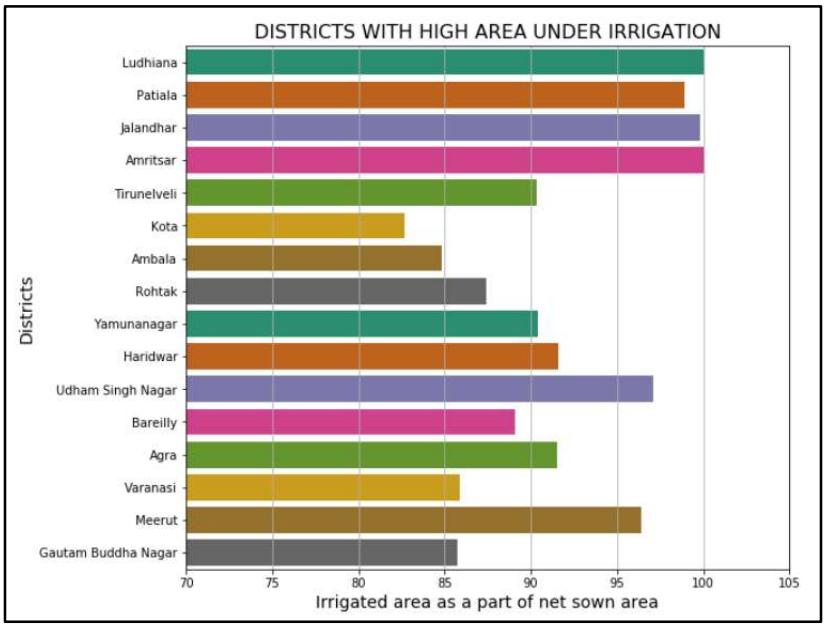


Fig41. Districts with irrigated area as percentage under net sown area

The districts of Punjab like Ludhiana, Patiala, Jalandhar and Amritsar have almost 100% of their net sown area irrigated. Meerut (UP) and Udham Singh Nagar (Uttarakhand) are the other 2 districts having more than 95% of their agricultural land irrigated. All of the districts in this list are located in Northern part of India.

5.Ground water quality data

Sr No	Main Columns
1	Salinity (> 3000 μS/cm)
2	Chloride (> 1000 mg/litre)
3	Fluoride (> 1.5 mg/litre)
4	Iron (> 1 mg/litre)
5	Arsenic (> 0.05 mg/litre)
6	Nitrate (> 45 mg/litre)
7	Total ground water quality

The columns in the Ground water quality data are as given in the table above. All the columns are binary in nature (1 for present, 0 for absent). Total ground water quality is the summation of these columns of binary numbers. This column has been computed taking summation of the 1's present in the individual columns such as Salinity, Chloride, Flouride, Iron, Arsenic and Nitrate for an individual district. For example, if a district has Salinity, Flouride ions and Iron in its ground water then the total ground water quality is 3. The higher the number, bad the quality of water. The table given below segregates the ground water quality based on the column of Total ground water quality.

Excellent (0)	Good (1)	Satisfactory (2)	Bad (3)	Critical (4)	Severe (5)
SAS Nagar	Ludhiana	Patiala	Tirunelveli	Karur	Salem
Gandhinagar	Jalandhar	Amritsar	Tiruchirappalli	Virudhunagar	Namakkal
Valsad	Madurai	Theni	Erode	Vellore	Bhavnagar
Dhanbad	Bokaro	Thoothukkudi	Jamnagar	Rajkot	Jaipur
Durg	Patna	East Singbhum	Kota	Vadodara	Rohtak
Udham Singh Nagar	East Khasi Hills	Korba	Ranchi	Ajmer	Krishna
Nainital	Dehradun	Raipur	Nashik	Jodhpur	Chandrapur
Darjeeling	Haridwar	Ambala	Mysore	Panipat	Bellary
Raigad	Hooghly	Yamunanagar	Ujjain	Visakhapatnam	
Varanasi	Kasaragod	Malappuram		Amravati	
	Thiruvananthapuram	Alappuzha		Agra	
	Aurangabad	Kollam		Gadag	
	Bareilly	Akola		Dharwad	
	Meerut	Sundergarh			
	Jhansi	Khorda			
	Gautam Buddha Nagar	Shimoga			
	Dakshina Kannada	Jabalpur			
Total: 10	Total: 17	Total: 17	Total: 9	Total: 13	Total: 8

CONCLUSION –

Education or literacy has been found out to be the biggest driving force for an area's development. Literate males and females can increase the working - class population of an area. This not only increases the overall per capita income of the family but also helps decrease poverty. Child health and literacy goes hand in hand. The relationship is equal for both male as well as female literacy. Literate parents can take better care of their children and thus bring down Infant Mortality Rate and Under 5 Mortality Rate. Many of the districts where there is significant amount of people below poverty line, high infant death rate are the ones having low literacy levels. High literacy can shape a path for an area to become economically, socially and environmentally strong.

A dream of making India a global superpower can only be achieved when not only the male child but also the female child is educated. Education can only bring about changes in the semi – urban districts lagging in developmental and social infrastructure. This same solution can be applicable to rural areas as well which are further behind the semi – urban districts in developmental hierarchy.

This project mainly focusses upon the demographic, economic, health, land use and ground water quality aspects. Further research into human development, industries, pollution, crime rate, so on and so forth can be done to find out more problems faced by these districts.

The biggest limitation in this project is the lack of data availability. There are many websites which are accessible only to the authorized personnel. If such a data is made available then better analysis can be performed, which will definitely help the Government Institutes for future planning and overall development.

References –

[1] https://www.collinsdictionary.com/dictionary/english/semiurban

[2] https://www.census2011.co.in [Individual districts data]

[3] https://en.wikipedia.org/wiki/Telangana

[4] https://sangareddy.telangana.gov.in/about-district/

[5] https://en.wikipedia.org/wiki/Ranga_Reddy_district

[6] https://en.wikipedia.org/wiki/Bardhaman_district

[7] Dr. Sandhya Ahuja, "Indirect Estimates of District wise IMR and Under 5 Mortality using Census 2011 data", National Health Systems Resource Centre (NHSRC), New Delhi.

[8] https://www.livemint.com/topic/poverty%20grid [Individual states' article] (2014)

[9] https://agricoop.nic.in/, Department of Agriculture, Cooperation and Farmers Welfare. [Individual district reports]

[10] Central Ground Water Board Ministry of Water Resources, Government of India, "Ground Water Quality in Shallow Aquifers of India", 2010.

[11] https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[12] https://www.tableau.com/learn/articles/data-visualization

[13] https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

[14] https://uc-r.github.io/hc_clustering

[15] https://en.wikipedia.org/wiki/Complete-linkage_clustering

[16] https://en.wikipedia.org/wiki/Single-linkage_clustering