**What story do the top 500 cities of India tell the world?**

**-:- Table of Content -:-**

|  |  |  |
| --- | --- | --- |
| **S.NO.** | **TOPIC** | **PAGE** |
| **1.** | **Introduction** | **6** |
| **2.** | **Scope of Analysis** | **8** |
| **3.** | **Existing System** | **9** |
| **4.** | **Source of dataset** | **10** |
| **5.** | **Objectives** | **11** |
| **6.** | **ETL Process** | **12** |
| **7.** | **My project ETL Workflow sheet** | **14** |
| **8.** | **Analysis on dataset** | **15** |
| **9.** | **List of Analysis with results** | **47** |
| **10.** | **Future Scope** | **48** |
| **11.** | **References** | **51** |
| **12.** | **Bibliography** | **52** |

**-:- Introduction -:-**

In India, as in other countries, the relationship between citizens and government is increasingly mediated by information systems, and e-governance is clearly seen as the way forward for efficient delivery of public services.

Today, more than half of the world’s population live in cities with more than six devices per person connected to the internet.

Census 2011 was the 15th Census of India and 7th Census after independence.

The population of the country as per the provisional figures of Census 2011 is 1210.19 million of which 623.7 million (51.54%) are males and 586.46 million (48.46%) are females.

The population of India has increased by more than 181 million during the decade 2001-2011.

According to the Census of 2011, the total population of India is 1,21,01,93,422. Uttar Pradesh (199.5 million) is the most populous state in the country followed by Maharashtra with 112 million.

Here are the five largest populous states of the country –

* Uttar Pradesh – 19,95,81,477
* Maharashtra – 11,23,72,972
* Bihar – 10,38,04,637
* West Bengal – 9,13,47,736
* Andhra Pradesh – 8,46,65,533

Five least populous state of the country

* Lakshadweep – 64,429
* Daman & Diu – 2,42,911
* D & N Haveli – 3,42,853
* A & N Island – 3,79,944
* Sikkim – 6,07,688

According to the census 2011, the literacy rate in India is 74% (Literacy of males – 82.14% and literacy of females – 65.46%).

In this report I have merged the data set of the census 2011 of Indian Cities with Population more than 1 Lac and City wise number of Graduates from the Census 2011. The main aim of this report is to create a visualization of where the future cities of India stands today.

From this project one can able to find out the various details like total population, male population, female population, literacy rate, sex ratio, male graduates, female graduates, population under 0-6 years and many more.

**-:- Scope of the Analysis -:-**

**There is following reasons which attract me to do this data analysis. The Scope of this Analysis is mentioned below.**

**-:- Existing System -:-**

I have collected the dataset from “[**https://www.kaggle.com/zed9941/top-500-indian-cities/home**](https://www.kaggle.com/zed9941/top-500-indian-cities/home)**.**” I have merged the dataset to the census 2011 report. The dataset I have collected from the site was the raw data only. After collecting the raw data, I have analysed and visualised the dataset and then made some conclusion. According to the dataset one can find out the following results-

* Female populations, Male populations and the Total population.
* One can also find 0\_6 years male and female populations, literacy rate of male and female.
* We are also able to find out the sex\_ratio, total\_graduates, total\_female\_graduates, total\_male\_graduates.

I have also used some of the mathematical, statistical, logical functions in this project. Some of the analysis are- population (0-6 years) 2001-2911 registered -3.08 percent growth with -2.42 for males and -3.80 for females (According to the census 2011).

The dataset I have collected have drawbacks or limitations too. The existing system i.e. dataset don’t have able to provide some results or conclusion or we can say that we are not able to find anything from the dataset because the dataset provided by **Arijit Mukherjee** was the raw data.

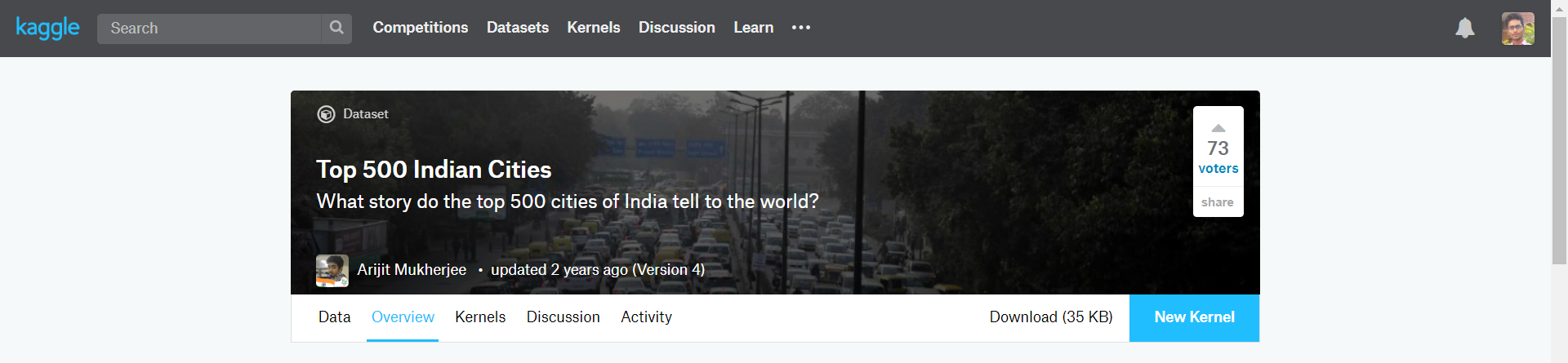
The dataset that was taken from the site have only columns and rows and these rows and columns are filled by some values like population\_male, population\_female, population\_total etc. By looking at the raw dataset one cannot able find out total literacy rate, population 0-6 years, total population, state name, district code because the data have too much rows and columns.

So last but not least, the limitations or drawbacks of the existing system are just because of raw dataset.

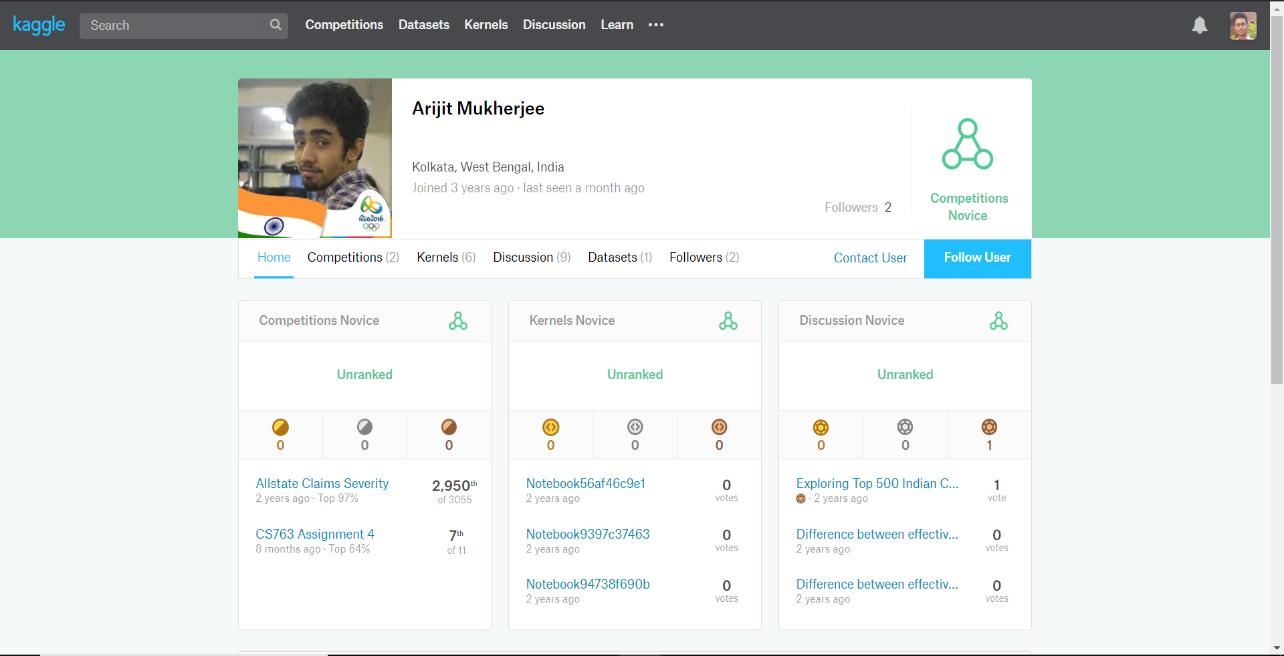
**-:- Source of dataset -:-**

I have used Kaggle website for my dataset. This dataset is 260kb in xlsx format having 30+ column and 500 rows.

**Source:** [**https://www.kaggle.com/zed9941/top-500-indian-cities/home**](https://www.kaggle.com/zed9941/top-500-indian-cities/home)



**Dataset Provider:** [**https://www.kaggle.com/zed9941**](https://www.kaggle.com/zed9941)

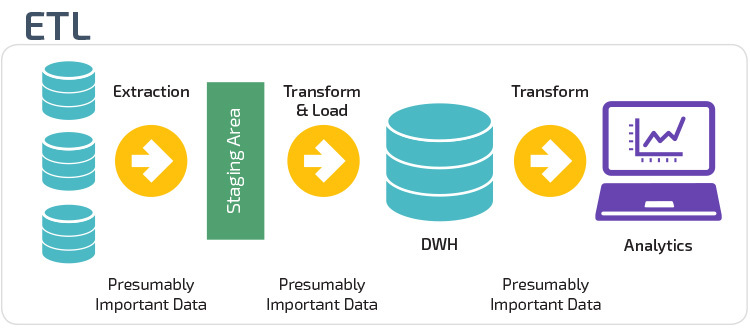


**:- Objectives:-**

Here are the objectives of this project-

1. Population analysis of city.
2. 0-6 years population analysis.
3. Literacy rate in all 500 cities.
4. Sex Ratio adult as well as 0-6 years old.
5. Education standard and effective education rate.

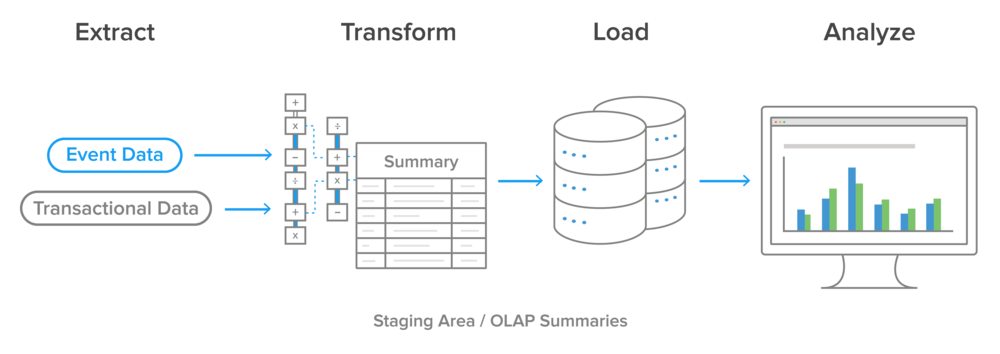
**-:- ETL Process -:-**



ETL process encompasses data extraction, transformation, and loading.

**ETL** is the process by which data is extracted from data sources that are not optimized for analytics, moved to a central lost, and optimized for analytics.

**Traditional ETL Process**

****

**Data** is extracted from online transaction processing (OLTP) databases, today more commonly known just as transactional databases, and other data sources. Data is then transformed in a staging area. These transformations cover both data cleansing and optimizing the data for analysis. The transformed data is then loaded into an online analytical processing (OLAP) database, today more commonly known as just an analytics database.

**Modern ETL Process**

Modern technology has changed most organizations approach to ETL, for several reasons. The biggest is the advent of powerful analytics warehouse like Google BigQuery. These newer cloud-based analytics databases have the horsepower to perform transformations in place rather than requiring a special staging area.

**What does the term Extract, Transform and Load means?**

* **Extract**is the process of reading data from a database. In this stage, the data is collected, often from multiple and different types of sources. Extraction of data is the most important step of ETL which involves accessing the data from all the Storage Systems. The storage systems can be the RDBMS, Excel files, XML files, flat files, ISAM (Indexed Sequential Access Method), hierarchical databases (IMS), visual information etc. Being the most vital step, it needs to be designed in such a way that it doesn’t affect the source systems negatively. Extraction process also makes sure that every item’s parameters are distinctively identified irrespective of its source system.
* **Transform** isthe process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data. Transformation is the next process in the pipeline. In this step, entire data is analysed and various functions are applied on it to transform that into the required format. Generally, processes used for the transformation of the data are conversion, filtering, sorting, standardizing, clearing the duplicates, translating and verifying the consistency of various data sources.
* **Load** is the process of writing the data into the target database. Loading is the final stage of the ETL process. In this step, the processed data, i.e. the extracted and transformed data, is then loaded to a target data repository which is usually the databases. While performing this step, it should be ensured that the load function is performed accurately, but by utilizing minimal resources. Also, while loading you have to maintain the referential integrity so that you don’t lose the consistency of the data. Once the data is loaded, you can pick up any chunk of data and compare it with other chunks easily.

**ETL Components**

* Auditing and logging
* Fault tolerance
* Notification support
* Handling of multiple source formats
* Accuracy
* Scalability

**-:- My project ETL Workflow sheet -:-**

**-:- Analysis on dataset -:-**

**1.** **Introduction:**

I have taken this dataset from Kaggle. This is one of the best compact and reasonable datasets which I came to know. Initial this dataset have 500 rows and 22 columns. But Forming ETL process. We created some additional columns also. Dataset was in Comma-separated values (CSV) format. We transform CSV file into Excel file.

**2.** **General Description:**

The dataset I have collected have drawbacks or limitations too. The existing system i.e. dataset don’t have able to provide some results or conclusion or we can say that we are not able to find anything from the dataset because the dataset provided by Arijit Mukherjee was the raw data.

**3.Specific Requirements, functions and formulas:**



The most basic and important requirement for doing the project is Ms Excel. Without this we can’t able to do this project.

One can also use SQL Server Management Studio and SQL Server Data Tools.

Also, there is a need of the dataset. After collecting the raw data, I visualized the data and then performed some functions on it.

We are also using “Power Bi” for our project. Microsoft offers one-month trial for students.so I learnt from net and used here.

Here are some of the functions and formulas –

**Mathematical Functions:**

* **SUM –** This function is used to adds all the values within a cell range.

**SUM(cell address : cell address)**

* **ROUND -** The round function is used to round a number to a specified number of digits.

**ROUND (number, number\_of\_digits)**

* **RAND ()-** This function is used to returns a random number greater than or equal to 0 and less than 1.

**RAND ()**

* **MOD-** This function is used to find the remainder after dividing a number by another number.

**MOD (number,divisor)**

**Statistical Functions:**

* **AVERAGE –** To calculate the average of a range of cells, use the average function.

**AVERAGE (number1, number2, ………)**

* **AVERAGEIF –** To average cells based on one criteria, use the averageif function.
* **MEDIAN –** To find the median use the median function.
* **MODE –** To find the most frequently occurring number, use the mode function.
* **MIN –** To find the minimum value, use the min function.
* **MAX –** To find the maximum value, use the max function.

**Information Functions:**

* **CELL function - :** Returns information about the formatting, location, or contents of a cell.
* **ISBLANK**: Return TRUE if value is BLANK.
* **ISTEXT:** Returns TRUE if value is text.
* **ISODD:** RETURN TRUE if value is Odd.
* **ISFORMULA:** Return TRUE if formula.

**Array Formulas –** Array formulas are powerful formulas that enable you to perform complex calculations that often can’t be done with standard worksheet functions.

They are also referred to as "Ctrl-Shift-Enter" or "CSE" formulas, because you need to press Ctrl+Shift+Enter to enter them.

**Logical Functions:**

A function that determines whether a condition is true or false is called a logical function.

Excel supports several logical functions such as **AND, FALSE, IF, NOT, OR** and **TRUE.**

* **IF** - IF function, which uses a logical test to determine whether an expression is true or false, and then returns one value if true or another value if false.
* **AND** - The **AND** Function returns **TRUE** if all conditions are true and returns **FALSE** if any of the conditions are false.
* **OR -** The **OR** function returns **TRUE** if any of the conditions are **TRUE** and returns **FALSE** if all conditions are false.
* **NOT-** The **NOT** function changes **TRUE** to **FALSE**, and **FALSE** to **TRUE**.

**Volatile Functions:**

A Volatile Function is one that causes recalculation of the formula in the cell where it resides **every time Excel recalculates**.

**Some of the volatile functions are as follows-**

**RAND (), NOW (), TODAY ()**

* **NOW () –** Get the current date and time

**Syntax =** NOW ()

* **TODAY () -** Get the current date

**Syntax =** TODAY ()

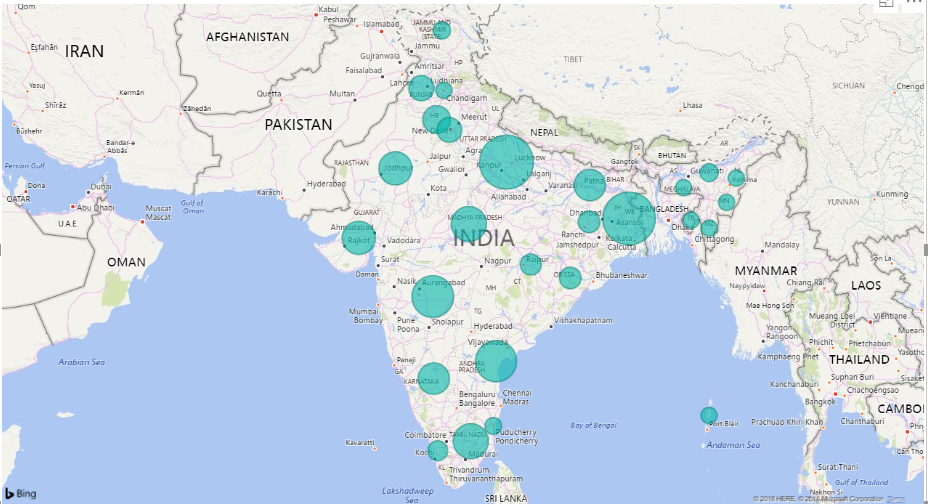
* **RAND () -** Get a random number between 0 and 1

**Syntax =** RAND ()

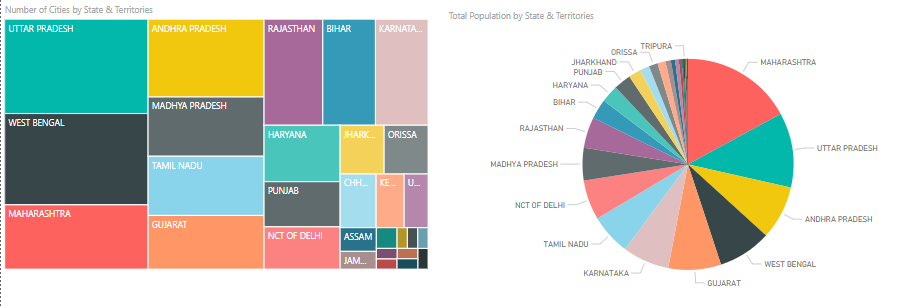
**4.Analysis results**

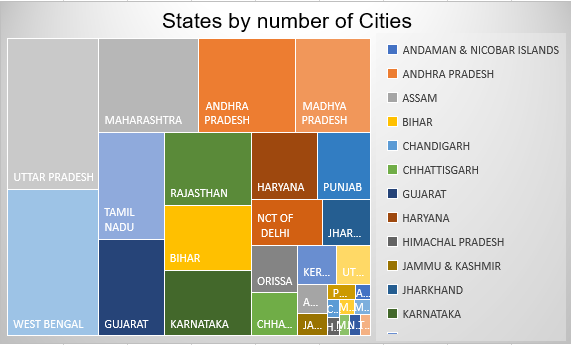
**5.Visualization:**

**1.States by Number of cities in Top 500(Using Power Bi)**

**xx**

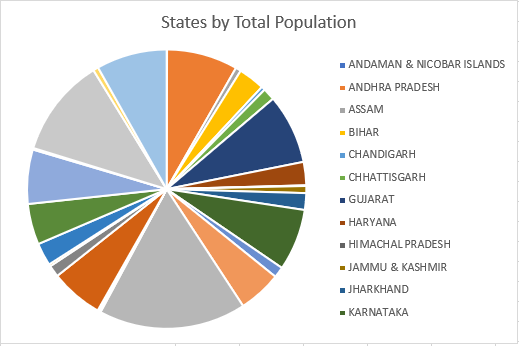
Bigger states of the country have high number of large cities.

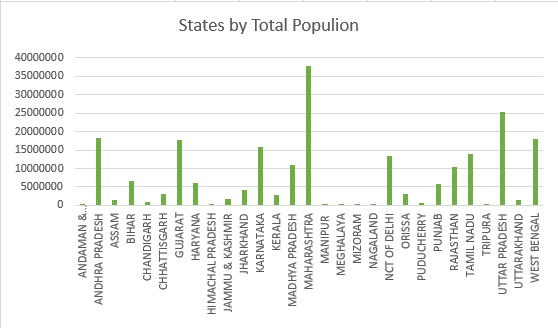




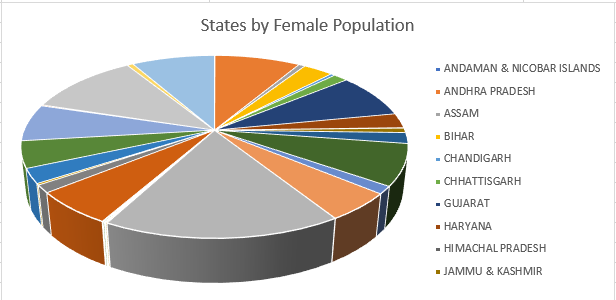
**2. State by Total Population in Top 500 cities**

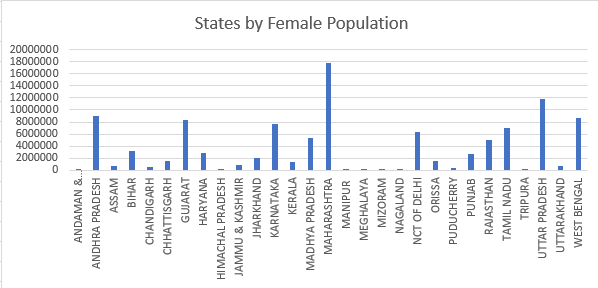
States which are economic power-houses of the country have more population in the cities.

****

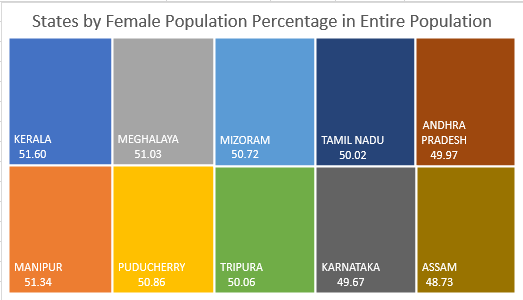
****

**2A****. State by Female Population**

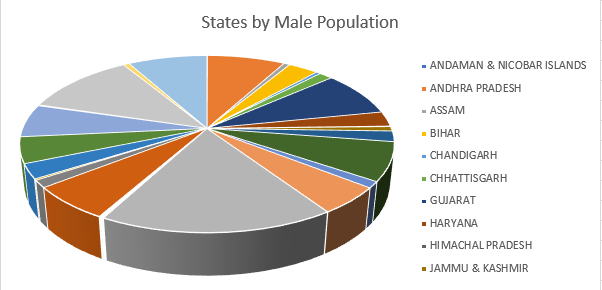
****

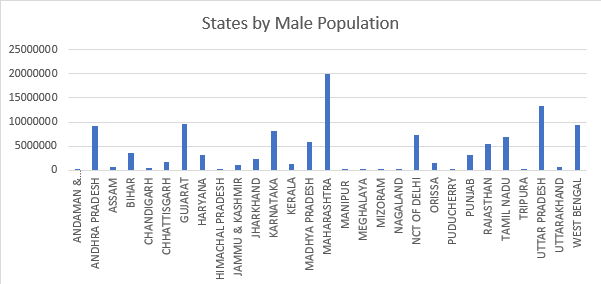
****

**Southern and North Eastern states have high female population rate**.

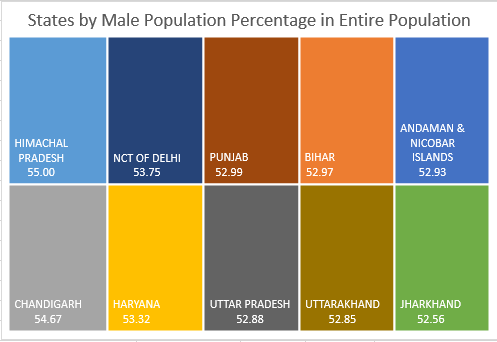


**2B. State by male Population**



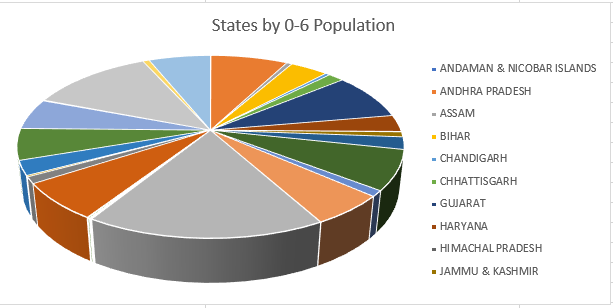


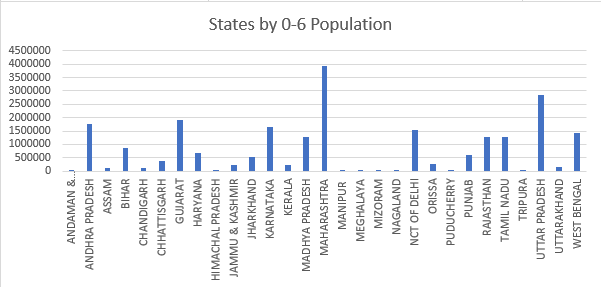
**Northern States of the country have high male population rate**

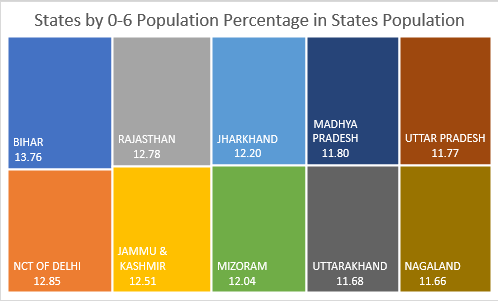


**3. State by Total Population 0-6year in Top 500 cities**

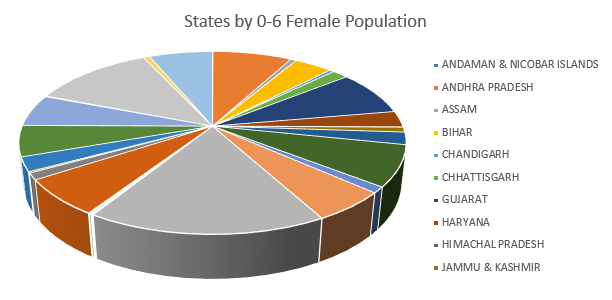
**Bigger states of the country have high number of large cities.**

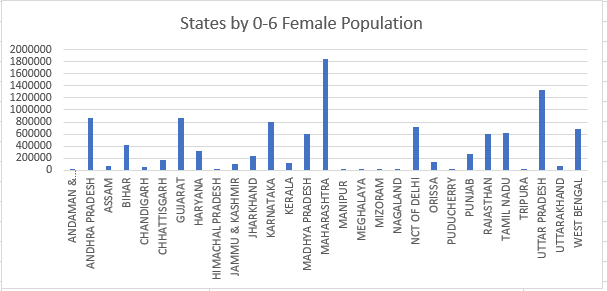




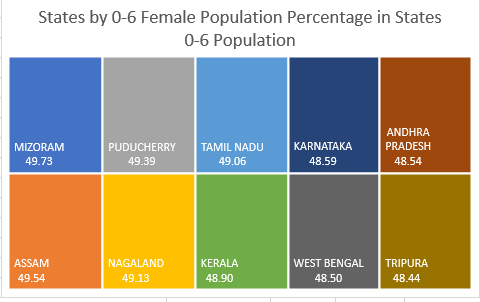


**3A. State by 0-6-year Female Population**

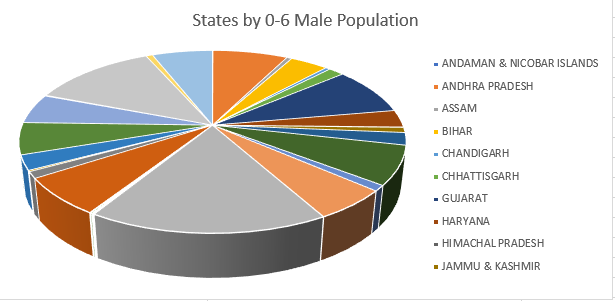


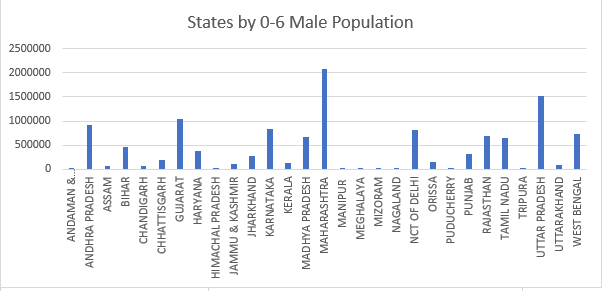


**Southern and North Eastern states have high 0-6 years female population rate**

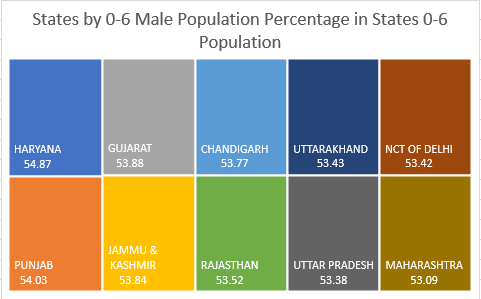


**3B. State by 0-6-year male Population**

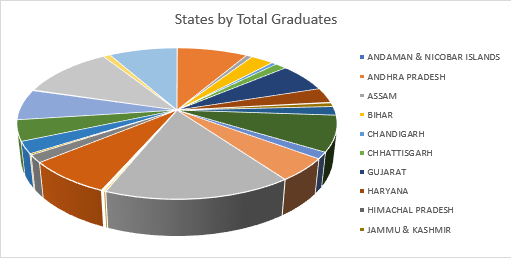


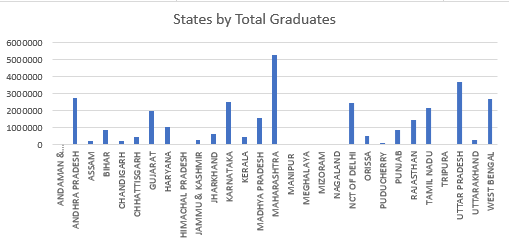


**States which are economic power-houses of the country have more population in the cities.**

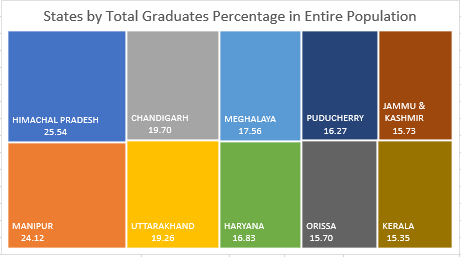


**4. State by total Graduates**

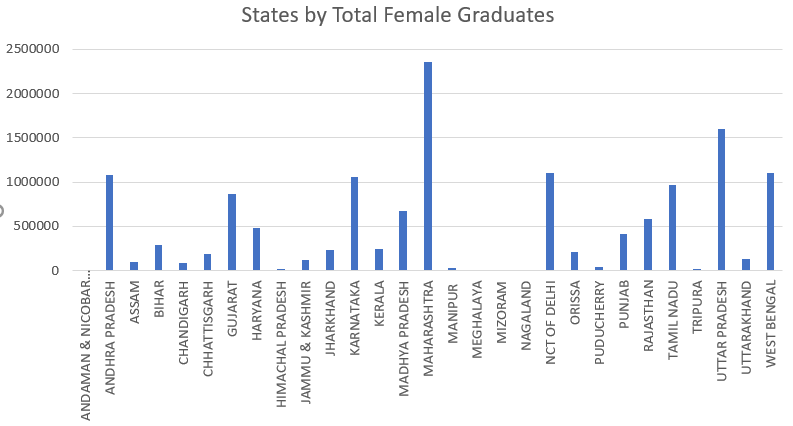
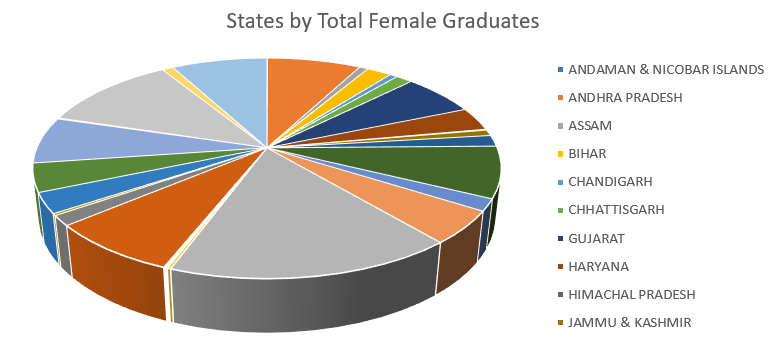


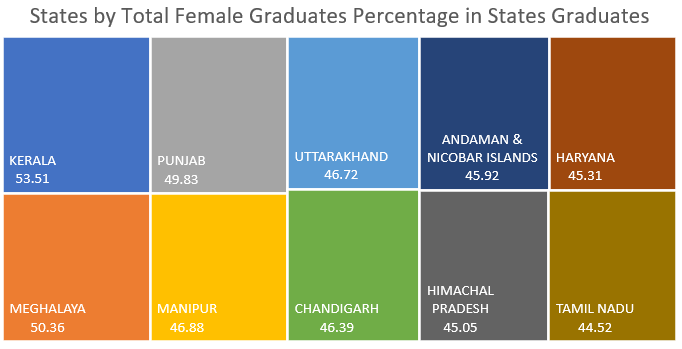


**North Eastern states along with few northern states have high number of graduates in the country**

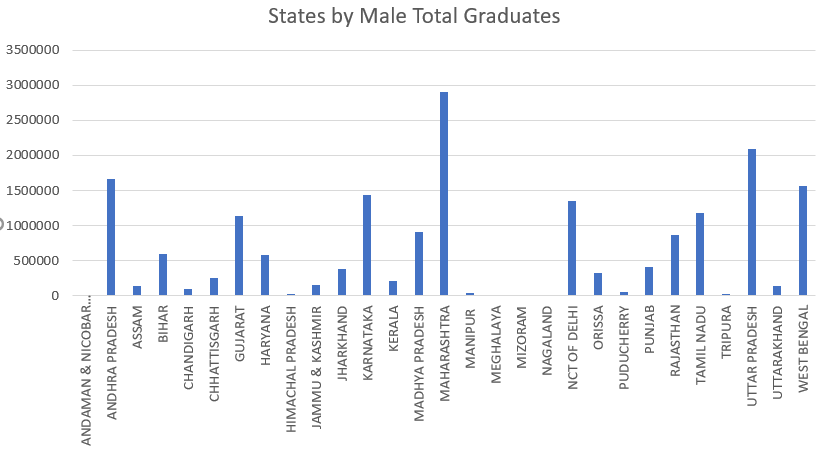
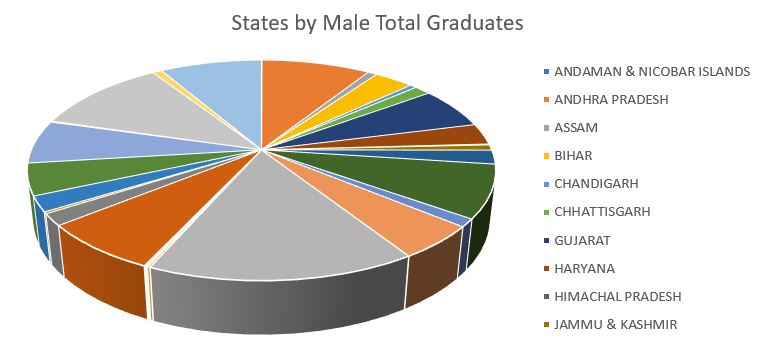


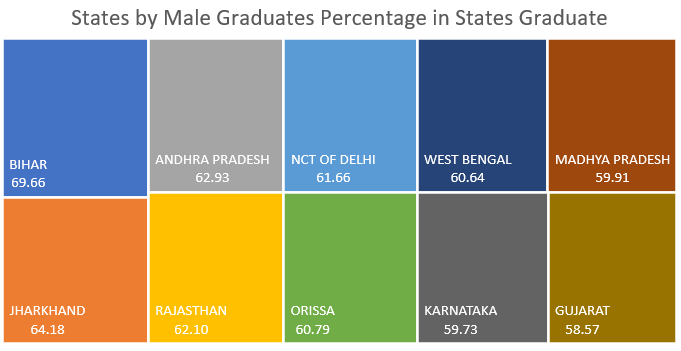
**4A. State by Female total Graduates**

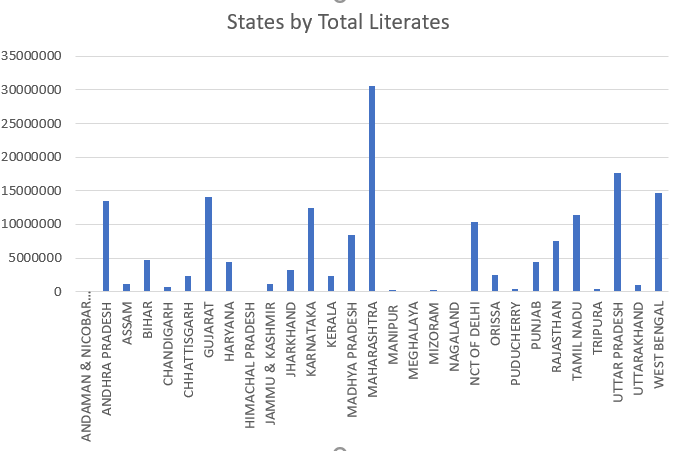
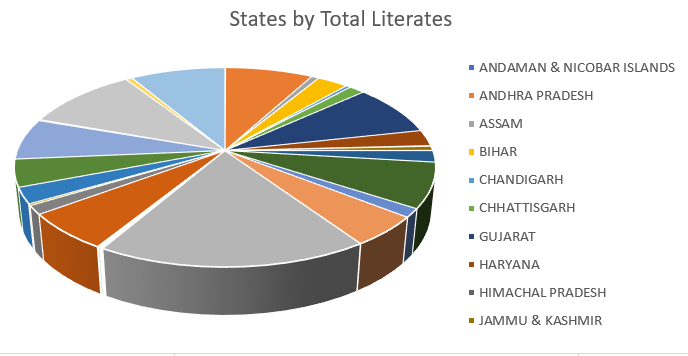


**Female graduate rate is also high in northern and few northeaster states except Kerala**

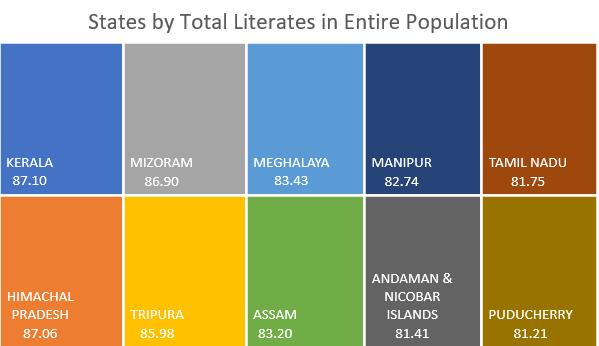
**4B. State by Female total Graduates**



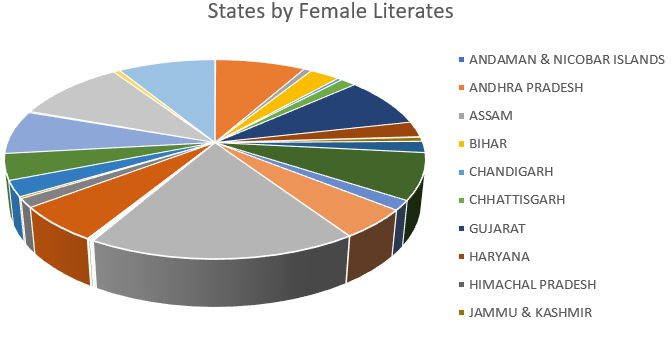
**Just like the male population rate, male graduate rate is also high in northern states of the country**

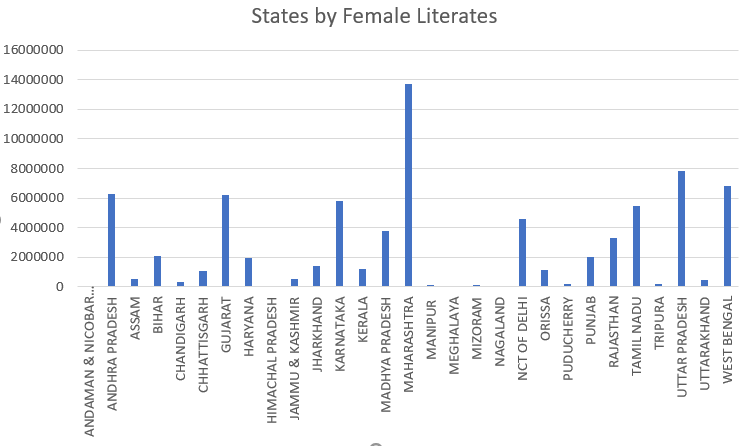
**5. State By literacy rate**

**Kerala has maximum literacy rate after that Mizoram.**

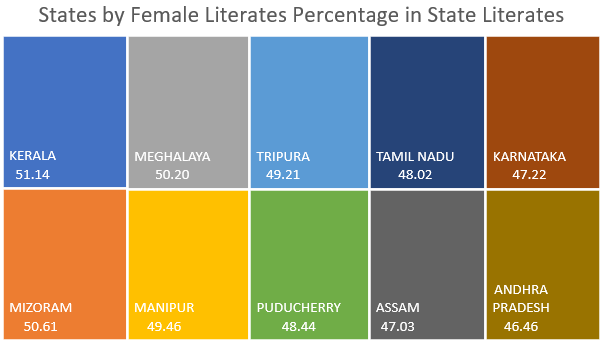


**5A. State by female literacy**

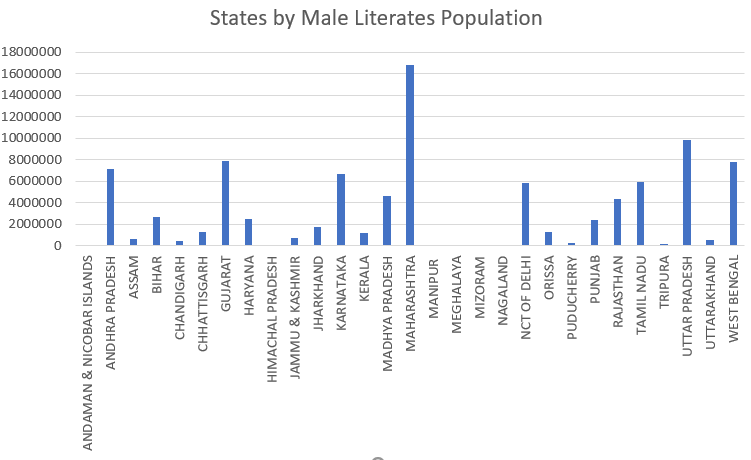
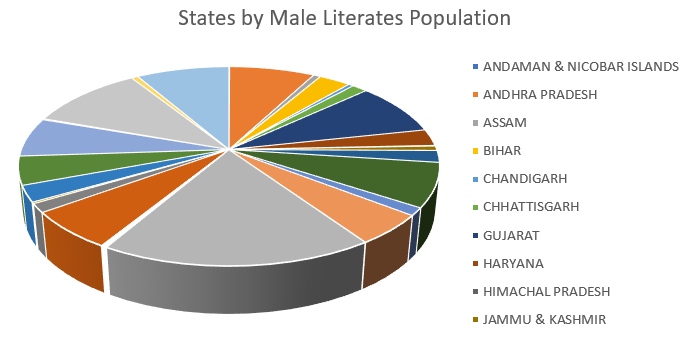




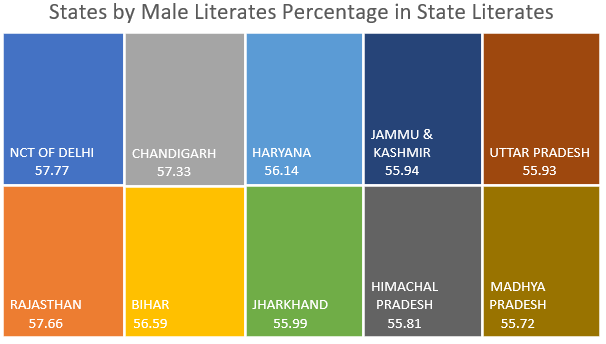
**Only Kerala, Mizoram and Meghalaya have more women literacy percentage.**



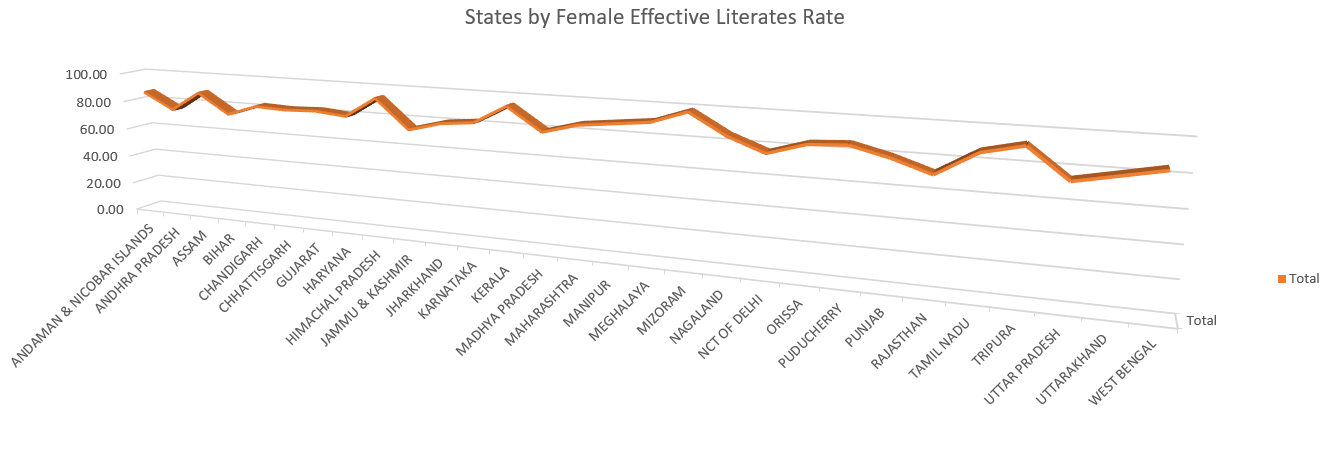
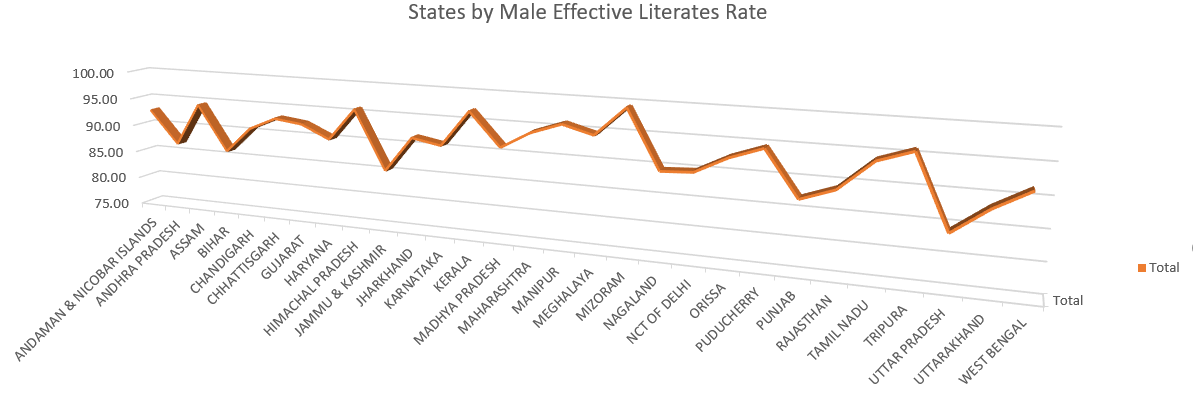
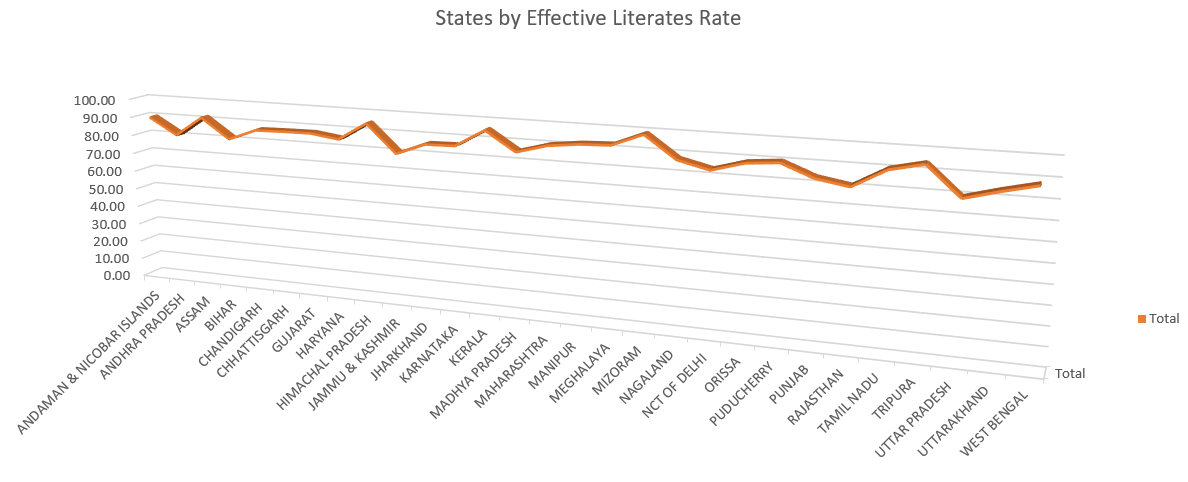
**5B. State by male literacy**



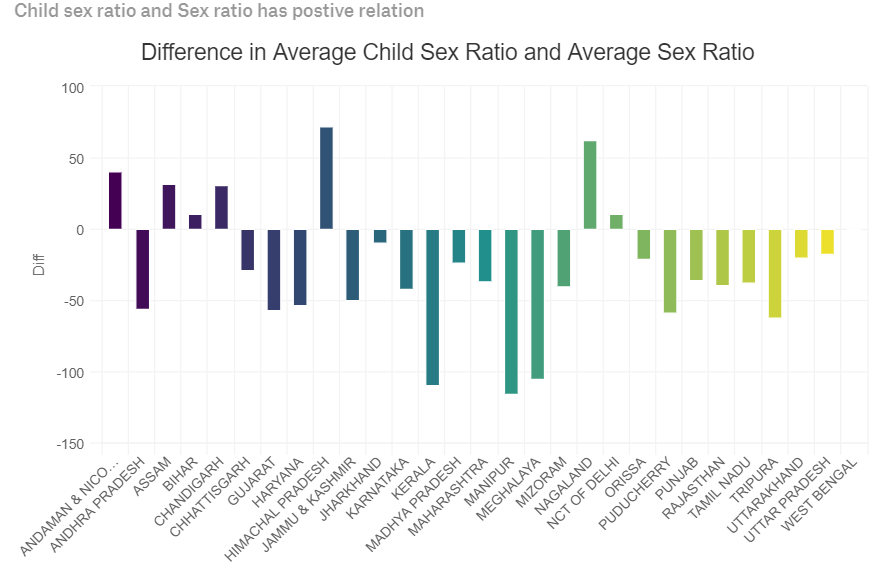
**NCT of Delhi,Rajashtan and Chandigarh is sharing maximum male literates.**



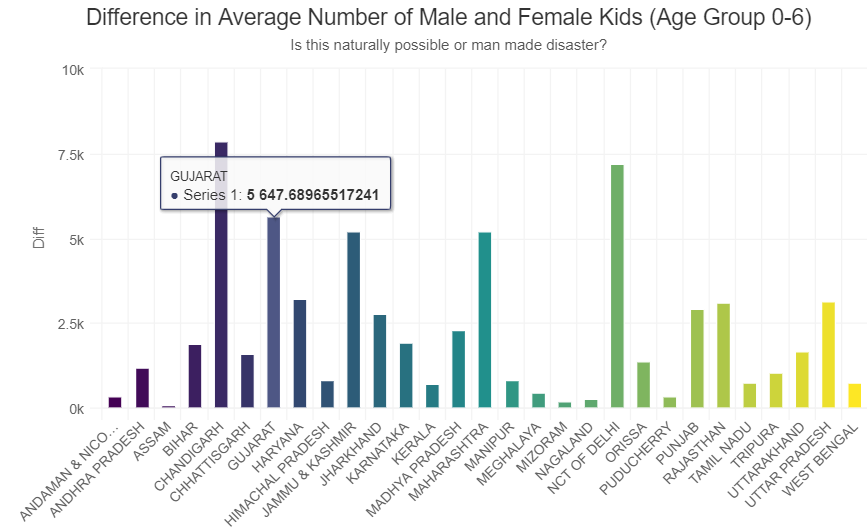
**5C. State by Effective Literacy rate**



**6.** Relation between Average child sex ration and average sex ratio

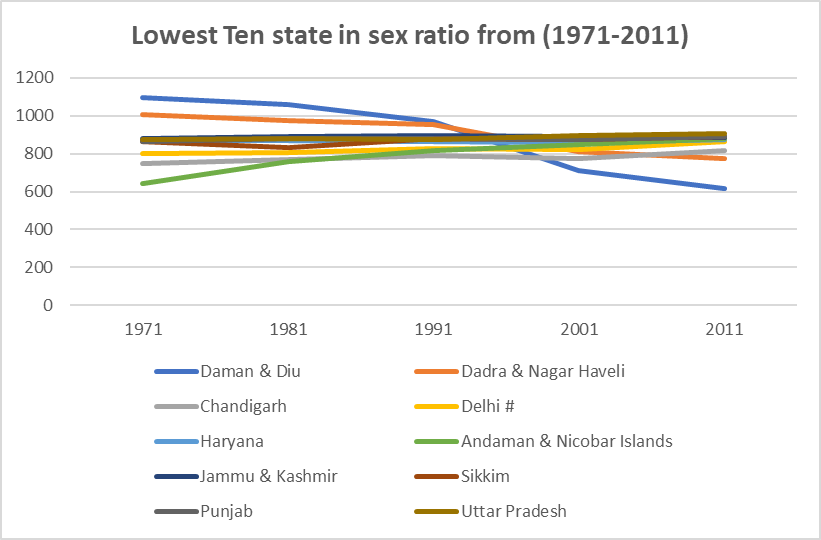


**7. Difference in Average Number of Male and Female Kids(Age Group 0-6)**



**7. Top Ten state in sex ratio from (1971-2011)**

**7. Lowest Ten state in sex ratio from (1971-2011)**

****

**-:- List of Analysis with results -:-**

**-:- Future scope -:-**

A census is the procedure of systematically acquiring and recording information about the members of a given population. Census is conducted after every 10 years. In this report I have derived all the records of census 2001-2011. I have calculated the basic details of the country such as literacy rate, total populations, sex ratio, state, population under 0-6 years, total number of male graduates, total number of female graduates, location and many more.

These are used in many fields such as-

* **Utility in Administrative and Policy:** The population census provides the basic data for administrative purposes. One of the most basic of the administrative uses of census data is in the demarcation of constituencies and the allocation of representation on governing bodies. Detailed information on the geographic distribution of the population is indispensable for this purpose.
* **Utility of Census data for Research Purposes:** The population census provides indispensable data for scientific analysis and appraisal of the composition, distribution and past and prospective growth of the population. The changing patterns of urban-rural concentration, the development of urbanised areas, the geographic distribution of population according to occupation and education, the sex and age structure of population, social and economic characteristics of population are the questions of scientific interest which are of importance both to research and practical problems of industrial and commercial growth and management.
* **Utility of Census data in Business and Industry:** The census data has many important uses for individuals and institutions in business and industry. It is very difficult to make a full assessment of the multiplicity of ways in which trade and business make use of the census data.
* **Utility of Population Census to other types of Censuses:** Certain information collected as part of a population census, or incidental to it, can be most useful in conducting and/or utilizing the results of housing, agricultural or establishment censuses taken at about the same time or near about as the population census.
* **Utility of Population Census to Electoral Rolls:** Some countries have taken advantage of the enumeration for a population census to collect, at the same time, information needed for the establishment of electoral rolls. This procedure is not generally advisable because of the deleterious effect the secondary purpose might have on the quality of the census results. It increases the burden on the enumerator and it may tempt some respondents deliberately to falsify their replies to some census questions (e.g., on age or citizenship) in order to appear eligible for placement on the electoral roll.
* **Census as frame for Sample Surveys:**  The rapidity of current changes in the size and other characteristics of populations and the demand for additional detailed data on social and economic characteristics which are not appropriate for collection in a full-scale census, have brought about the need for continuing programmes of intercensal sample surveys to collect current and detailed information on many topics which are usually investigated at ten-year intervals in the population censuses.
* **Utility of Census data in Planning:** The census data is indispensable for social and economic planning of the Country. The Planning Commission utilises the Census data on the distribution of population by age, sex classified by rural and urban regions, cities, town areas and social groups to analyse the growth of consumer demand and savings in the process of development. The census data also prove useful in national income estimates and estimates on differential personal incomes in rural and urban areas and the composition of rural and urban consumption of groups of goods and services and income elasticity co-efficient. An analysis of areas of different population size with different characteristics certainly serves as a basis for Government plans and investigations in basic social capital. The data on economic activity and educational levels of the individual as collected in the census is very important for manpower planning. The housing needs can also be accurately estimated by using the census data on population. Besides all these, the census data can prove very useful in the formulation of policies on education, health, agriculture, food and development of road, rail transport etc. In a nut-shell, it can be said that the census data is extremely useful for all types of planning.
* **Utility of population census to civil registration and vital statistics:** Census data serve as denominators for the computation of vital rates, especially rates specific for characteristics normally investigated only at the time of the census. Conversely, census results, time-adjusted by vital and migration statistics, can provide estimates of the future size, distribution and other characteristics of the population of the total country and subnational areas. Further, census data on fertility can provide a bench-mark check on the reliability of current birth statistics. It is consequently desirable that procedures for the collection of census data, vital statistics and migration statistics be closely co-ordinated with regard to coverage, concepts, definitions, classifications and tabulations.

**-:- References -:-**

Samidha Paroha, 2015. How do ETL tool work.

[https://www.quora.com/How-do-ETL-tools-work]. Accessed November 10,2018.

April 14 2018. Mathematical Functions.

[https://www.w3resource.com/excel/excel-mathematical-function.php]. Accessed November 10,2018.

June 19,2016. Population Census 2011.

[https://www.bankexamstoday.com/2016/02/population-census-2011-at-a-glance-pdf-version.html]. Accessed November 11,2018.

July 10,2017. Volatile Function

[https://exceljet.net/glossary/volatile-function]. Accessed November 11,2018.

Prologue

[http://censusindia.gov.in/Census\_And\_You/prologue\_census\_and\_you.aspx]. Accessed November 11,2018.

**-:- Bibliography -:-**

1. [**https://exceljet.net/glossary/volatile-function**](https://exceljet.net/glossary/volatile-function)
2. [**http://www.etldatabase.com/etl-process/**](http://www.etldatabase.com/etl-process/)
3. [**https://www.kaggle.com/zed9941/top-500-indian-cities/home**](https://www.kaggle.com/zed9941/top-500-indian-cities/home)
4. [**https://www.quora.com/How-do-ETL-tools-work**](https://www.quora.com/How-do-ETL-tools-work)
5. [**www.w3resource.com**](http://www.w3resource.com)
6. [**www.censusindia.gov.in**](http://www.censusindia.gov.in)