**ANALYSIS OF TRENDS IN DISTRIBUTION OF POPULATION AND LITERACY RATES IN INDIAN CITIES**

|  |  |
| --- | --- |
| **CS656//NJIT/Prof.Moshiur Rahman** |  |
| **Developed By: Seetha Veezhinathan** |  |
| **Version: 1.0.0** |  |
| **Date Submitted: 05/13/2017** |  |

1. **INTRODUCTION**

* Creating a visualization of the population distribution and literacy rate scattered in the cities of India. This project aims at giving a statistical view of how the population and literacy rates are present and graduates present in India.
* Using RStudio I have implemented this.
* The Data source was collected from – Census 2011.
* Using R Programming Language, I have incorporated the same.
* It is an open source programming language which is used for statistical computing and graphics. It incorporates all the standard statistical tests, models, and analysis, as well as providing a comprehensive language for managing and manipulating data.
* One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

1. **PACKAGES USED**

* **DPLYR**
* The dplyr package makes these steps fast and easy:
  + By constraining your options, it simplifies how you can think about common data manipulation tasks.
  + It provides simple “verbs”, functions that correspond to the most common data manipulation tasks, to help you translate those thoughts into code.
  + It uses efficient data storage backends, so you spend less time waiting for the computer.
* Dplyr aims to provide a function for each basic verb of data manipulation:
  + filter() (and slice())
  + arrange() – used to sort a to sort a variable in descending order
  + select() - keeps only the variables you mention
  + (and rename())
  + distinct()
  + mutate()
  + summarise() - Summarise multiple values to a single value
* **GGPLOT2**
  + ggplot2 is a data exploration and visualization package written in R.
  + It describes the theoretical division of graphs into semantic components.
  + This approach of handling elements of a graph separately and building the features up in a series of layers allows for unmatched versatility and control.
* **HIGHCHARTER**
  + This function creates a Highchart chart using htmlwidgets.
  + The widget can be rendered on HTML pages generated from R Markdown, Shiny, or other applications.
  + The main features of this package are:
    - Various chart type with the same style: scatters, bubble, line, time series, heatmaps, treemap, bar charts, networks.
    - Chart various R object with one function. With hchart(x) you can chart: data.frames, numeric, histogram, character, density, factors, ts, mts, xts, stl, ohlc, acf, forecast, mforecast, ets, igraph, dist, dendrogram, phylo, survfit classes.
    - Support Highstock charts. You can create a candlestick charts in 2 lines of code. Support xts objects from the quantmod package.
    - Support Highmaps charts. It’s easy to create choropleths or add information in geojson format.
    - Piping styling.
    - Themes: you configurate your chart in multiples ways. There are implemented themes like economist, financial times, google, 538 among others.
    - Plugins: motion, drag points, fontawesome, url-pattern, annotations.
* **MAPS**
  + Create high quality maps that may be shaded or projected in a variety of ways.
  + Additional arguments may be used for coloring in counties, states, or countries (col, fill), looking at different map projections (projection), among many other options.
* **SCALES**
  + Graphical scales map data to aesthetics, and provide methods for automatically determining breaks and labels for axes and legends.
  + One of the most difficult parts of any graphics package is scaling, converting from data values to perceptual properties.
  + The inverse of scaling, making guides (legends and axes) that can be used to read the graph, is often even harder!
  + The idea of the scales package is to implement scales in a way that is graphics system agnostic, so that everyone can benefit by pooling knowledge and resources about this tricky topic.
* **DT (DATA TABLES)**
  + This function creates an HTML widget to display rectangular data (a matrix or data frame) using the JavaScript library DataTables.
  + The data.table R package provides an enhanced version of data.frame that allows you to do blazing fast data manipulations.
  + It works well with very large data files.
  + Can behave just like a data frame.
  + Offers fast subset, grouping,  update,  and joins.
  + Makes it easy to turn an existing data frame into a data table.

1. **TOOLS USED**

* **RStudio**

1. **TECHNOLOGIES USED**

* **Analysis done using R Programming**
* **RMarkdown –** used to generate HTML pages which can be rendered using shiny/RMarkdown applications.
* **Knitr HTML -** knitr is an engine for dynamic report generation with R.It is a package in the statistical programming language R that enables integration of R code into LaTeX, LyX, HTML, Markdown, AsciiDoc, and reStructuredText documents.

1. **DATA SET DESIGN**

****

**Data of 500 Cities with population more than 1 Lac by Census 2011**

|  |  |
| --- | --- |
| 'name\_of\_city' | Name of the City |
| 'state\_code' | State Code of the City |
| 'state\_name' | State Name of the City |
| 'dist\_code' | District Code where the city belongs ( 99 means multiple district ) |
| 'population\_total' | Total Population |
| 'population\_male' | Male Population |
| 'population\_female' | Female Population |
| '0-6\_population\_total' | 0-6 Age Total Population |
| '0-6\_population\_male' | 0-6 Age Male Population |
| '0-6\_population\_female' | 0-6 Age Female Population |
| 'literates\_total' | Total Literates |
| 'literates\_male' | Male Literates |
| 'literates\_female' | Female Literates |
| 'sex\_ratio' | Sex Ratio |
| 'child\_sex\_ratio' | Sex ratio in 0-6 |
| 'effective\_literacy\_rate\_total' | Literacy rate over Age 7 |
| 'effective\_literacy\_rate\_male' | Male Literacy rate over Age 7 |
| 'effective\_literacy\_rate\_female' | Female Literacy rate over Age 7 |
| 'location' | Lat,Lng |
| 'total\_graduates' | Total Number of Graduates |
| 'male\_graduates' | Male Graduates |
| 'female\_graduates' | Female Graduates |

1. **IMPLEMENTATION**

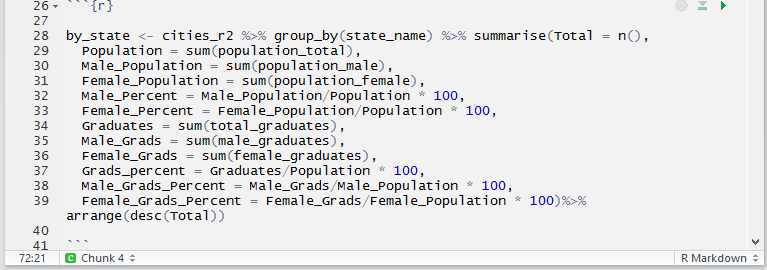
There are 27 scenarios implemented totally for the analysis of population and literacy rates using the Census 2011 Data Set. All the scenarios are explained as follows.

In general we are using **Hchart()** function of the HIGHCHARTER package to generate all the graphs. This is accomplished with the help of GGPLOT2 which is used for generating graphs.

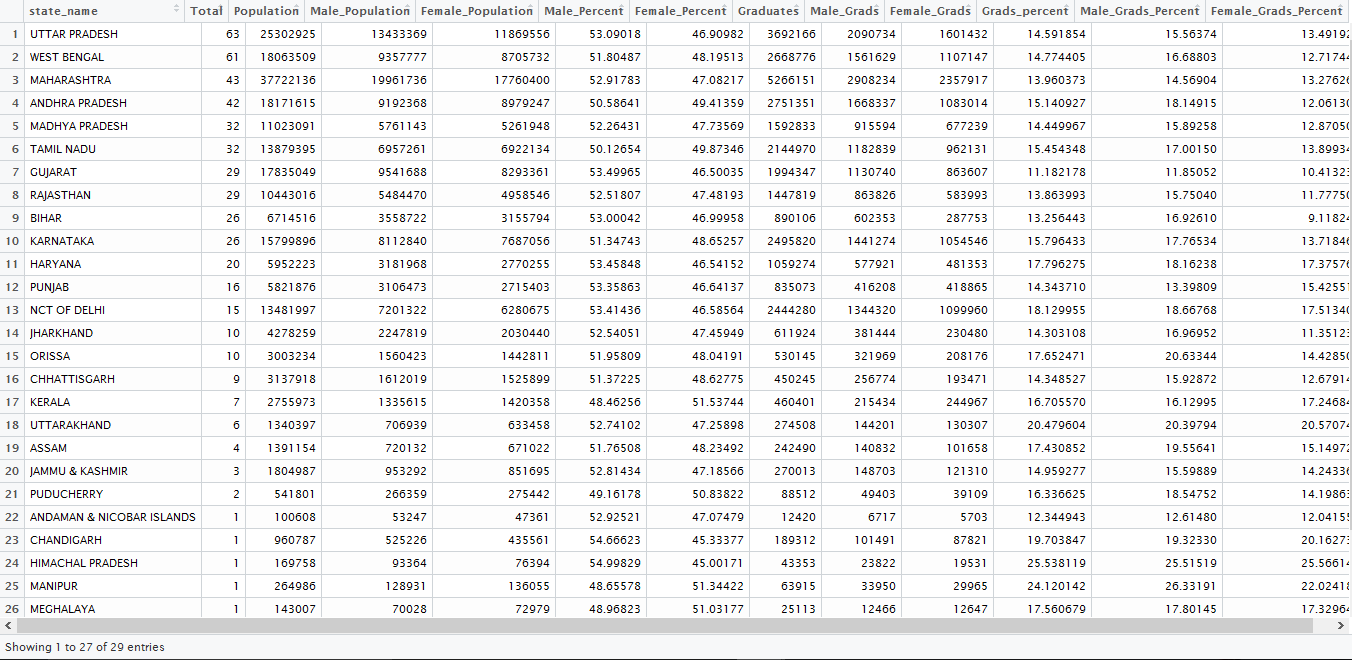
**Finding Top 500 Cities of the Country**

The Data manipulations are done for the data set and stored in a variable ‘by\_state’.

Attached below the snapshot of the code for ‘by\_State’ variable data manipulation:



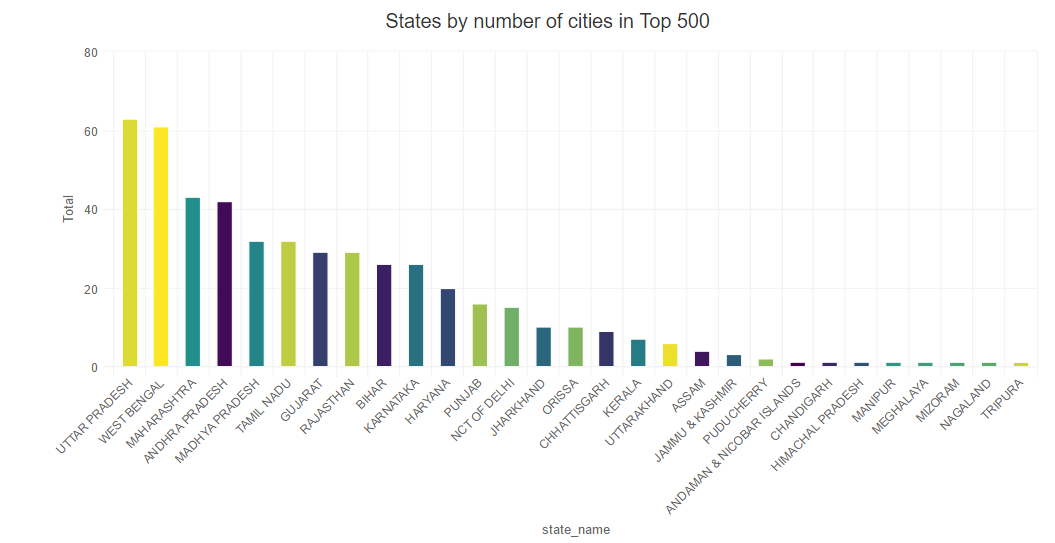
Snapshot of the result of the variable with total values for all the specified columns:

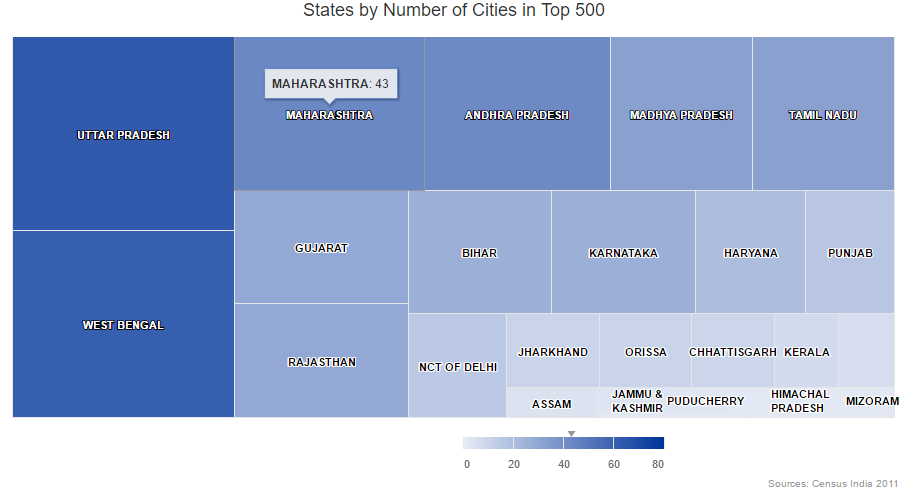


1. **States by number of cities in Top 500**

* In this scenario, we are considering the number of cities in each state from the data set.
* The states which has the highest number of cities in Top 500 cities are displayed followed by the states which has the least number of cities in Top 500.
* The output is displayed both in the COLUMN and TREEMAP view.
  + - X Coordinate contains the ‘State\_name’ variable and
    - Y Coordinate contains the ‘Total’.
* We can infer from the output that Uttar Pradesh has more number of cities in Top 500

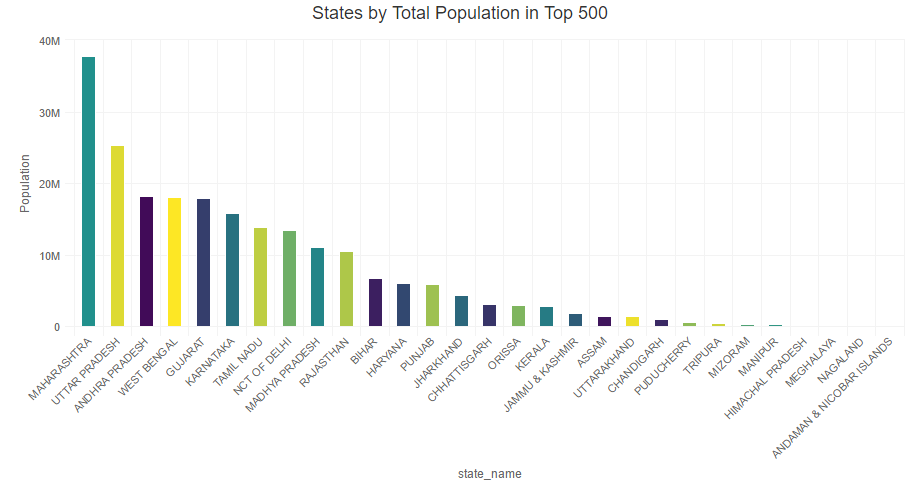
(total of 63 cities) and Tripura has the least number of cities ( total up to 1).





1. **States by Total Population in Top 500 Cities**

* In this scenario, we are considering the states with the highest to lowest population from the data set.
* The states which has the largest population in Top 500 cities are displayed followed by the states which has the lowest population.
* The output is displayed both in the COLUMN and TREEMAP view.
  + - X Coordinate contains the ‘State\_name’ variable and
    - Y Coordinate contains the ‘Population’.
* We can infer from the output that Maharashtra has highest population and Manipur has the lowest population.

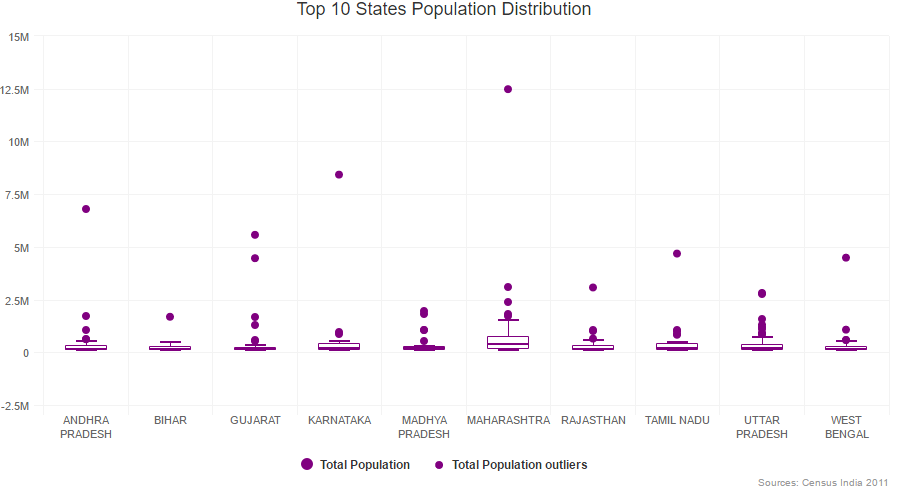


A picture containing screenshot

Description generated with very high confidence

1. **Top 10 States Population Distribution**

* In this scenario, the Top 10 states from the variable ‘by\_State’ is considered and its population distribution is provided in the form of **HCboxplot.**
* From the top 10 states we generate the Boxplot based upon the highest population for the 10 states.
* We can infer that Maharashtra has the highest population among the top 10 states of the country and Karnataka has the lowest population.



1. **States by High Male Population Percentage**

* In this scenario, we are considering the states with the highest male population percentage.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘State\_name’ variable and
    - ‘Value’ parameter contains ‘Male\_Percent’.
* We can infer from the output that Himachal Pradesh has highest Male population percentage and Kerala has the lowest male population percentage.

**A picture containing screenshot

Description generated with very high confidence**

1. **States by High Female Population Percentage**

* In this scenario, we are considering the states with the highest female population percentage.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘State\_name’ variable and
    - ‘Value’ parameter contains ‘Female\_Percent’.
* We can infer from the output that unlike Male Population percent Kerala has highest female population percentage and Himachal Pradesh has the lowest female population percentage.

**A picture containing screenshot

Description generated with very high confidence**

1. **States by Total Graduate Percentage in Entire Population**

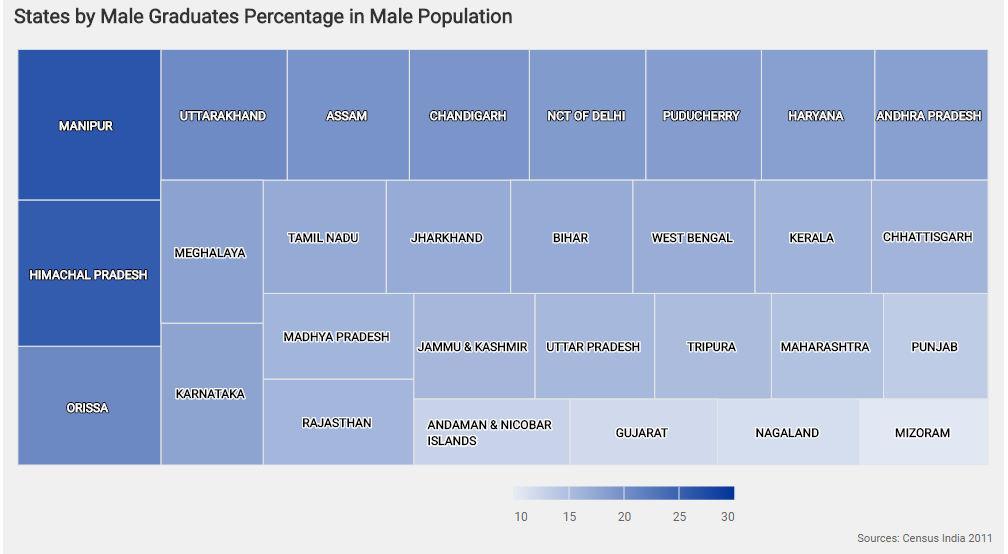
* In this scenario, we are considering the Total Graduate percentage in the entire population.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘State\_name’ variable and
    - ‘Value’ parameter contains ‘Grads\_percent’.
* We can infer from the output that Himachal Pradesh has highest Total Graduates in entire population percentage and Mizoram has the lowest graduates in entire population.

**A picture containing screenshot

Description generated with very high confidence**

1. **States by Male Graduate Percentage in Male Population**

* In this scenario, we are considering the Male Graduate percentage in the entire male population.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘State\_name’ variable and
    - ‘Value’ parameter contains ‘Male\_Grads\_Percent’.
* We can infer from the output that Manipur has highest male graduates in entire male population percentage and Mizoram has the lowest male graduates in entire population.

****

1. **States by Female Graduate Percentage in Female Population**

* In this scenario, we are considering the Female Graduate percentage in the entire female population.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘State\_name’ variable and
    - ‘Value’ parameter contains ‘Female\_Grads\_Percent’.
* We can infer from the output that Himachal Pradesh has highest female graduates in entire female population percentage and Mizoram has the lowest female graduates in entire population.

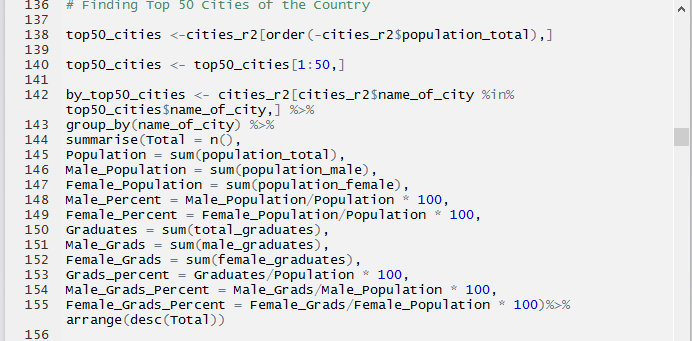
**A picture containing screenshot

Description generated with very high confidence**

**Finding Top 50 Cities of the Country**

The Data manipulations are done for the data set and stored in a variable ‘top50\_cities’.

Attached below the snapshot of the code for ‘top50\_cities’ variable data manipulation:



1. **Top 50 Cities by Population**

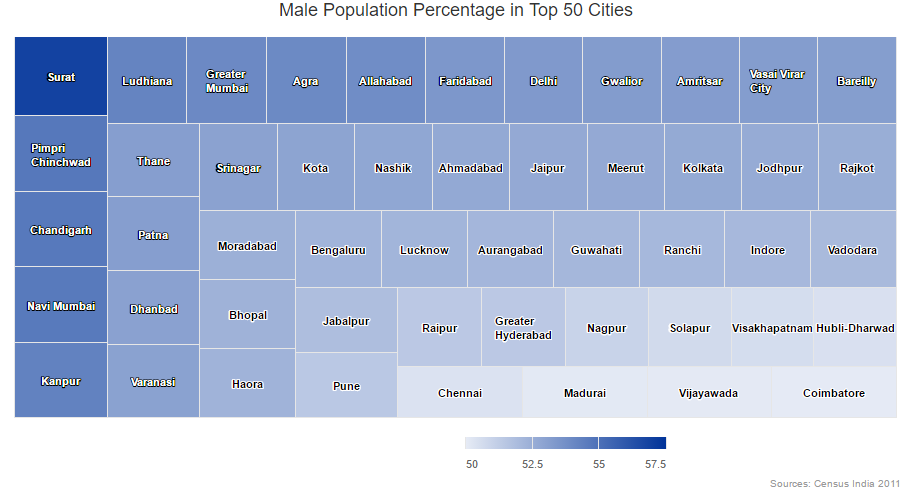
* In this scenario, we are considering the Top 50 cities by the population in them.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘name\_of\_city’ variable and
    - ‘Value’ parameter contains ‘Population’.
* We can infer from the output that Mumbai has the highest population and Moradabad has the lowest population in the top 50 cities.

**A picture containing screenshot

Description generated with very high confidence**

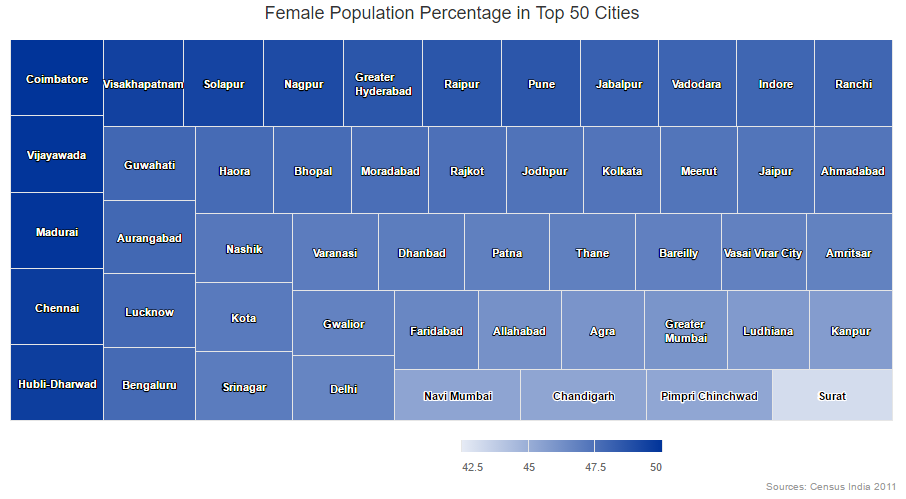
1. **Male Population Percentage in Top 50 Cities**

* In this scenario, we are considering the Male Population Percentage in Top 50 Cities.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘name\_of\_city’ variable and
    - ‘Value’ parameter contains ‘Male\_Percent’.
* We can infer from the output that Surat has the highest male population and Coimbatore has the lowest male population in the top 50 cities.



1. **Female Population Percentage in Top 50 Cities**

* In this scenario, we are considering the Female Population Percentage in Top 50 Cities.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘name\_of\_city’ variable and
    - ‘Value’ parameter contains ‘Female\_Percent’.
* We can infer from the output that Coimbatore has the highest female population and Surat has the lowest female population in the top 50 cities.



1. **Total Graduate Percentage in Top 50 Cities**

* In this scenario, we are considering the Total Graduate Percentage in Top 50 Cities.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘name\_of\_city’ variable and
    - ‘Value’ parameter contains ‘Grads\_percent’.
* We can infer from the output that Lucknow has the highest number of total graduates in Top 50 cities and Surat has the lowest number of total graduates in Top 50 cities.

**A picture containing screenshot

Description generated with very high confidence**

1. **Male Graduate Percentage in Top 50 Cities**

* In this scenario, we are considering the Male Graduate Percentage in Top 50 Cities.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘name\_of\_city’ variable and
    - ‘Value’ parameter contains ‘Male\_Grads\_Percent’.
* We can infer from the output that Allahabad has the highest number of male graduates in Top 50 cities and Surat has the lowest number of male graduates in Top 50 cities.

**A picture containing screenshot

Description generated with very high confidence**

1. **Female Graduate Percentage in Top 50 Cities**

* In this scenario, we are considering the Female Graduate Percentage in Top 50 Cities.
* The output is displayed in the TREEMAP view.
  + - X Coordinate contains the ‘name\_of\_city’ variable and
    - ‘Value’ parameter contains ‘Female\_Grads\_Percent’.
* We can infer from the output that Pune has the highest number of female graduates in Top 50 cities and Surat has the lowest number of female graduates in Top 50 cities.

**A picture containing screenshot

Description generated with very high confidence**

1. **Male Literacy rate versus Female Literacy Rate**

* Mizoram has the highest Male Literacy and Female literacy rate in the country followed by Kerala.
* Effective literacy rate of male increases as effective literacy rate of female increases. It has positive correlation

**A close up of a map

Description generated with high confidence**

1. **Male Graduate rate versus Female Graduate Rate**

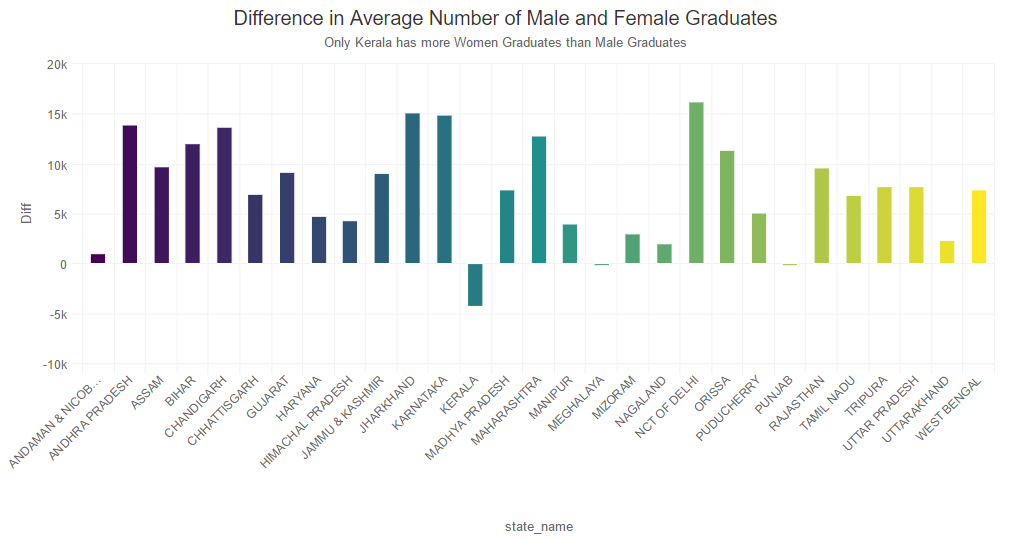
**A picture containing sky

Description generated with very high confidence**

* Himachal Pradesh has the highest Male Graduate rate and Female Graduate rate in the country followed by Mizoram.
* Effective graduate rate of male increases as effective graduate rate of female increases. It has positive correlation

1. **Difference in Male and Female Graduate Rates**

* In this scenario, we are considering the average number of male and female graduates.
* The output is displayed in the COLUMN view.
  + - X Coordinate contains the ‘state\_name’ variable and
    - ‘Value’ parameter contains Difference of the male and female graduates.
* We can infer from the output that Delhi has the highest number of male graduates and Kerala has the Highest number of female graduates.

****

1. **Total Population Distribution in Top 500 Cities**

* 75% of the 500 Cities have less than 350K Population.
* Biggest city in Top 500 in terms of population is 124 times the smallest one.



A picture containing wall, indoor

Description generated with high confidence

1. **Total Population Distribution (0-6) Age Group in Top 500 Cities**

* 75% of the 500 cities have less than 38k population.
* biggest city in top 500 in terms of population (0-6 age group) is 184 times the smallest one.

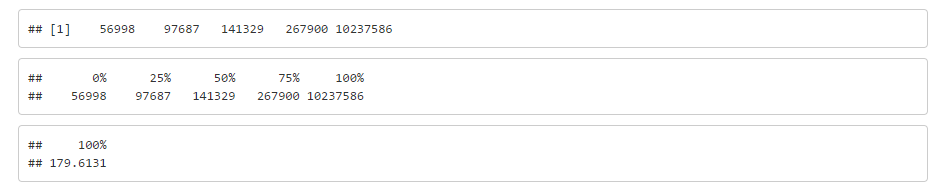


A picture containing wall

Description generated with very high confidence

1. **Distribution of Total Literates in Top 500 Cities**

* 75% of the 500 cities have less than 275k literates.
* Biggest city in top 500 in terms of number of literates is 180 times the smallest one.



A picture containing wall, sky

Description generated with high confidence

1. **CONCLUSION**

The analysis for Population distribution across top 500 cities and for the top 50 cities are performed. Also, we have generated graphs for variations in Male and Female literacy rates and Graduate rates across the top 500 Indian Cities.