SHIV SHAKTI
International Journal in Multidisciplinary and Academic Research (SSIJMAR)
Vol. 7, No. 2, April 2018 (ISSN 2278 – 5973)

Data Analysis of Indian Cities and States

Top 500 cities

T. Raghavendra Gupta (*Author*)
Department of Computer Science and Engineering
Hyderabad Institute of Technology and Management
Hyderabad, India
E-mail:raghavendraguptat.cse@hitam.org

Maddimsetty Aishwarya (*Author*)
Department of Computer Science and Engineering
Hyderabad Institute of Technology and Management
Hyderabad, India
E-mail: settyaishwarya@gmail.com

Sai Teja (*Author*)
Department of Computer Science and Engineering
Hyderabad Institute of Technology and Management
Hyderabad, India
E-mail:teja_indy@rediffmail.com

Priyanshi Alle (*Author*)
Department of Computer Science and Engineering
Hyderabad Institute of Technology and Management
Hyderabad, India
E-mail: priyanshialle@gmail.com

Ganesh Varma (*Author*)
Department of Computer Science and Engineering
Hyderabad Institute of Technology and Management
Hyderabad, India
E-mail:alluri.ganesh46@gmail.com

***Abstract*—Data analysis project using python, based on the information that is collected and exploring it for any insights.Performing analysis on census data of indian cities. Data handling, data analysis, data visualisation, valuation of the data.**

Keywords—Analysis of data,plotting graphs.

I. INTRODUCTION (*HEADING 1*)

The indian census is the most credible source of information on demography(population,economic activity, literacy rate and education,housing and household,urbanization and mortality etc.) and many other sociocultural and demographic data since 1872. A census is contrasted with sampling in which information is obtained from subset of population; main population estimate is updated by such intercensal estimates. Modern census data are used for research, business marketing, and planning, and as a baseline for designing sample surveys by providing a sampling frame such as an address register. Census counts are necessary to adjust samples to be representative of a population by weighting them as is common in opinion polling. Stratification requires knowledge of the relative sizes of different population strata which can be derived from census enumerations. In few countries, the census provides the official counts used to apportion the number of elected representatives to regions. In other cases, a carefully chosen random sample can provide more accurate information than attempts to get a population census.

II. PREREQUISITES

A. Collecting the data set.

Data analysis project using python, based on the information that is collected and exploring it for any insights. Complex study of data parameters. Data from census 2011. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. An example of this type is the data sets collected by space agencies performing experiments with instruments aboard space probes. Data sets that are so

large that traditional data processing applications are inadequate to deal with them are known as bid data. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum.

Data collection methods

Surveys, interviews and focus groups are primary instruments for collecting information. Today, with help from Web and analytics tools, organizations are also able to collect data from mobile devices, website traffic, server activity and other relevant sources, depending on the project.

Big data and data collection

Big data describes voluminous amounts of structured, semi-structured and unstructured data collected by organizations. But because it takes a lot of time and money to load big data into a traditional relational database for analysis, new approaches for collecting and analyzing data have emerged. To gather and then mine big data for information, raw data with extended metadata is aggregated in a data lake. From there, machine learning and artificial intelligence programs use complex algorithms to look for repeatable patterns.

B. Cleaning of the data.

The data set that is collected which is in the csv format, is cleaned. Cleaning involves: removing the nulls,redundancies. Checking the formats of content. Smoothening and removing any inconsistencies involved. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records). Some data cleansing solutions will clean data by cross checking with a validated data set. A common data cleansing practice is data enhancement, where data is made more complete by adding related information.

Process

Data auditing: The data is audited with the use of statistical and database methods to detect anomalies and contradictions: this eventually gives an indication of the characteristics of the anomalies and their locations. Several commercial software packages will let you specify constraints of various kinds (using a grammar that conforms to that of a standard programming language, e.g., JavaScript or Visual Basic) and then generate code that checks the data for violation of these constraints. This process is referred to below in the bullets "workflow specification" and "workflow execution." For users who lack access to high-end cleansing software, Microcomputer database packages such as Microsoft Access or File Maker Pro will also let you perform such checks, on a constraint-by-constraint basis, interactively with little or no programming required in many cases.

Workflow specification: The detection and removal of anomalies is performed by a sequence of operations on the data known as the workflow. It is specified after the process of auditing the data and is crucial in achieving the end product of high-quality data. In order to achieve a proper workflow, the causes of the anomalies and errors in the data have to be closely considered.

Workflow execution: In this stage, the workflow is executed after its specification is complete and its correctness is verified. The implementation of the workflow should be efficient, even on large sets of data, which inevitably poses a trade-off because the execution of a data-cleansing operation can be computationally expensive.

Post-processing and controlling: After executing the cleansing workflow, the results are inspected to verify correctness. Data that could not be corrected during execution of the workflow is manually corrected, if possible. The result is a new cycle in the data-cleansing process where the data is audited again to allow the specification of an additional workflow to further cleanse the data by automatic processing.

Good quality source data has to do with “Data Quality Culture” and must be initiated at the top of the organization. It is not just a matter of implementing strong validation checks on input screens, because almost no matter how strong these checks are, they can often still be circumvented by the users. There is a nine-step guide for organizations that wish to improve data quality

Declare a high level commitment to a data quality culture

Drive process reengineering at the executive level

Spend money to improve the data entry environment

Spend money to improve application integration

Spend money to change how processes work

Promote end-to-end team awareness

Promote interdepartmental cooperation

Publicly celebrate data quality excellence

Continuously measure and improve data quality

Others include:

Parsing: for the detection of syntax errors. A parser decides whether a string of data is acceptable within the allowed data specification. This is similar to the way a parser works with grammars and languages.

Data transformation: Data transformation allows the mapping of the data from its given format into the format expected by the appropriate application. This includes value conversions or translation functions, as well as normalizing numeric values to conform to minimum and maximum values.

Duplicate elimination: Duplicate detection requires an algorithm for determining whether data contains duplicate representations of the same entity. Usually, data is sorted by a key that would bring duplicate entries closer together for faster identification.

Statistical methods: By analyzing the data using the values of mean, standard deviation, range, or clustering algorithms, it is possible for an expert to find values that are unexpected and thus erroneous. Although the correction of such data is difficult since the true value is not known, it can be resolved by setting the values to an average or other statistical value. Statistical methods can also be used to handle missing values which can be replaced by one or more plausible values, which are usually obtained by extensive data augmentation algorithms.

III. THE STAGES OF ANALYSIS.

The general data analysis of any dataset involves a number a steps like Analysing literacy rate of the cities and further state. Top 10 cities with most number of literates alive.Male and female literacy rates.Analysing effective literacy rate, graduates.Correlation between urbanisation and literacy.Sex ratio Which states have most of its male population in urban areas.Which states have most of its younger population in urban areas. Top 10 states.

Data analysis covers everything from reading the source methodology behind a data collection to creating a data visualization of the statistic you have extracted. All the steps in-between include deciphering variable descriptions, performing data quality checks, correcting spelling irregularities, reformatting the file layout to fit your needs, figuring out which statistic is best to describe the data, and figuring out the best formulas and methods to calculate the statistic you want. These steps and many others fall into three stages of the data analysis process: ***evaluate, clean, and summarize***.

C. Male and Female literacy rate:

This is the most common analysis taken in any set of people. Male and female literacy rates gives us insights on a lot of correlated factors like zones with highest literacy rates, females being more in number in education and employment sector and thus further steps can be taken. Although this was a greater than six fold improvement, the level is below the world average literacy rate of 84%. The 2011 census, indicated a 2001–2011 decadal literacy growth of 9.2%, which is slower than the growth seen during the previous decade. An old 1990 study estimated that it would take until 2060 for India to achieve universal literacy at then-current rate of progress. One of the main factors contributing to this relatively low literacy rate is usefulness of education and availability of schools in vicinity in rural areas. There is a shortage of classrooms to accommodate all the students in 2006–2007. Such inadequacies resulted in a non-standardized school system where literacy rates may differ. Such inadequacies resulted in a non-standardized school system where literacy rates may differ. Absolute poverty in India has also deterred the pursuit of formal education as education is not deemed of as the highest priority among the poor as compared to other basic necessities.

Literacy rate of India in 2011 is 74.04%. The Male literacy rate is 82.14% and Female literacy rate is 65.46% according to Census 2011.

Among the Indian states, Kerala has the highest literacy rate 93.91% and then Mizoram 91.58%.

Among the Union Territories, Lakshadweep has the highest literacy rate of 92.28%.

Bihar has the lowest literacy rate in India with 63.82% .

The Male literacy is highest in Lakshadweep 96.11% and Kerala 96.02%.

The Female literacy is highest in Kerala 91.98% and Mizoram 89.40%.

Lowest male literacy is in Bihar 73.39%.

Lowest female literacy is in Rajasthan 52.66%.

	States/UT	Total Literates	Male Literates	Female Literates	Total Literacy Rate	Male Literacy rate	Female Rate
	INDIA	77,84,54,120	44,42,03,762	33,42,50,358	74.04	82.14	65.46
01	Jammu & Kashmir	72,45,053	43,70,604	28,74,449	68.74	78.26	58.01
02	Himachal Pradesh	51,04,506	27,91,542	23,12,964	83.78	90.83	76.60
03	Punjab	1,89,88,611	1,06,26,788	83,61,823	76.68	81.48	71.34
04	Chandigarh #	8,09,653	4,68,166	3,41,487	86.43	90.54	81.38
05	Uttarakhand	69,97,433	39,30,174	30,67,259	79.63	88.33	70.70
06	Haryana	1,69,04,324	99,91,838	69,12,486	76.64	85.38	66.77
07	NCT of Delhi #	1,27,63,352	72,10,050	55,53,302	86.34	91.03	80.93
08	Rajasthan	3,89,70,500	2,41,84,782	1,47,85,718	67.06	80.51	52.66
09	Uttar Pradesh	11,84,23,805	7,04,79,196	4,79,44,609	69.72	79.24	59.26
10	Bihar	5,43,90,254	3,27,11,975	2,16,78,279	63.82	73.39	53.33
11	Sikkim	4,49,294	2,53,364	1,95,930	82.20	87.29	76.43
12	Arunachal Pradesh	7,89,943	4,54,532	3,35,411	66.95	73.69	59.57
13	Nagaland	13,57,579	7,31,796	6,25,783	80.11	83.29	76.69
14	Manipur	18,91,196	10,26,733	8,64,463	79.85	86.49	73.17
15	Mizoram	8,47,592	4,38,949	4,08,643	91.58	93.72	89.40
16	Tripura	28,31,742	15,15,973	13,15,769	87.75	92.18	83.15
17	Meghalaya	18,17,761	9,34,091	8,83,670	75.48	77.17	73.78
18	Assam	1,95,07,017	1,07,56,937	87,50,080	73.18	78.81	67.27
19	West Bengal	6,26,14,556	3,45,08,159	2,81,06,397	77.08	82.67	71.16
20	Jharkhand	1,87,53,660	1,11,68,649	75,85,011	67.63	78.45	56.21
21	Orissa	2,71,12,376	1,53,26,036	1,17,86,340	73.45	82.40	64.36
22	Chhattisgarh	1,55,98,314	89,62,121	66,36,193	71.04	81.45	60.59
23	Madhya Pradesh	4,38,27,193	2,58,48,137	1,79,79,056	70.63	80.53	60.02
24	Gujarat	4,19,48,677	2,39,95,500	1,79,53,177	79.31	87.23	70.73
25	Daman & Diu	1,88,974	1,24,911	64,063	87.07	91.48	79.59
26	Dadra & Nagar Haveli	2,28,028	1,44,916	83,112	77.65	86.46	65.93
27	Maharashtra	8,25,12,225	4,62,94,041	3,62,18,184	82.91	89.82	75.48
28	Andhra Pradesh	5,14,38,510	2,87,59,782	2,26,78,728	67.66	75.56	59.74
29	Karnataka	4,10,29,323	2,28,08,468	1,82,20,855	75.60	82.85	68.13
30	Goa	11,52,117	6,20,026	5,32,091	87.40	92.81	81.84
31	Lakshadweep	52,914	28,249	24,665	92.28	96.11	88.25
32	Kerala	2,82,34,227	1,37,55,888	1,44,78,339	93.91	96.02	91.98
33	Tamil Nadu	5,24,13,116	2,83,14,595	2,40,98,521	80.33	86.81	73.86
34	Puducherry	9,66,600	5,02,575	4,64,025	86.55	92.12	81.22
35	Andaman & Nicobar Islands #	2,93,695	1,64,219	1,29,476	86.27	90.11	81.84

Literacy rate Male and Female India 2011 Census

D. Sex ratio and correlation between the areas:

This is the next most pursued analysis through census datasets, as it is important to understand the correlation between these two factors as the government can understand the patterns based on male and female birth and mortal rate and then come up with any rescual plans to protect the zones from these subjects. Sex ratio means number of female population per thousand of male population. The sex composition of a population is generally stated in terms of sex ratio i.e. number of female per thousand of males. It depicts the current condition with respect to status of girl child, gender discrimination, infanticides and feticides. Present sex composition of child population determines the future vital events such as marriage rate, labour force, age structure, birth and death, migration, etc. According to 2011 census, the sex ratio of Satara District is 988 females per thousand males among the tehsils, Javali records the highest sex ratio of 1068 female per thousand males and Mahabaleshwar has the lowest sex ratio of 937 female per thousand males. The area which is higher in literacy rate and sex ratio that region would be higher in human development index.

The major cause of the decrease of the female birth ratio in India is considered to be the violent treatments meted out to the girl child at the time of the birth. The Sex Ratio in India was almost normal during the phase of the years of independence, but thereafter it started showing gradual signs of decrease. Though the Sex Ratio in India has gone through commendable signs of improvement in the past 10 years, there are still some states where the sex ratio is still low and is a cause of concern for the NGO organizations. One of the states which is showing a decreasing trend in the population of women 2011 and is a cause of concern is Haryana. The state of Haryana has the lowest rate of sex ratio in India and the figure shows a number of 877 of females to that of 1000 of males.

There are also states such as Puducherry and Kerala where the number of women is more than the number of men. Kerala houses a number of 1084 females to that of 1000 males. While Puducherry and Kerala are the only two states where the number of female is more than the number of men, there are also states in India like that of Karnataka, Andhra Pradesh and Maharashtra where the sex ratio 2011 is showing considerable signs of improvement. Some facts related to the Sex Ratio in India follows, the main cause of the decline of the sex ration in India is due to the biased attitude which is meted out to the women. The main cause of this gender bias is inadequate education. Pondicherry and Kerala houses the maximum number of female while the regions of Daman and Diu and Haryana have the lowest density of female population.

E. Urbanization and young population moving uptowns:

The third category includes the set of data analysis that gives us an idea about the cities that are urbanised . We could come up with the analysis that in last 15 years the younger gen moving into urban areas has increased on a very huge scale. because of the increasing opportunities in the urban cities.

Measurement of the degree of urbanisation in a country like India is considered very important. Various measures are being used for the purpose. As per the first simple method we observed that the total urban population in India in 1981 was a little less than one fourth of the total population in comparison to that of one-ninth in 1921 and one-sixth in 1951.

The second method, i.e., the urban-rural growth differential (URGD) method also revealed that the growth rates of both rural and urban population are very close to each other at present.

Third method showing the growth of urban population reveals that as the total population of the country rose by about three times since 1921 but the total urban population of the country increased by about six-times. Thus all the methods observed more or less same results.

If we compare degree of urbanisation in India with that of developed countries then we can find that India is lagging far behind the high-income countries. In 1985, the proportion of urban population to total population was 92 per cent in U.K., 86 per cent in Australia, 76 per cent in Japan, and 74 per cent in U.S.A. as against only 25 per cent in India.

In India, towns are classified into six different classes. From the census data, it has been observed that in Class I town (having a population more than 1 lakh) the proportion of urban population concentration has increased from 25.7 per cent in 1901 to 60.4 per cent in 1981. Thus there is an increasing trend towards huge concentration of population in the bigger towns.

In Class II and Class III towns together, the proportion of urban population remained almost constant at the level of 26 to 28 per cent during the period 1901-81. But in the remaining Class IV, Class V and Class VI towns together, the relative proportion of urban population concentration declined sharply from 47.2 per cent in 1901 to only 13.6 per cent in 1981.

Besides continuation of urbanisation process, a number of Class II towns have been transformed into a Class I town and the number of Class I towns has thus increased from 74 in 1951 to 216 in 1981.

Accordingly, the total population of Class I towns also increased from 273 lakhs in 1951 to 943 lakh in 1981 showing an increase of nearly 245 per cent. During the same period, the number of Class II towns has increased from 95 to 270 and that of Class III towns increased from 330 to 739 in 1981.

Total population of Class II and Class III towns increased from 330 to 739 in 1981. Total population of Class II and Class III towns increased by 130 per cent, i.e., from 97 lakh in 1951 to 224 lakh in 1981. While the number of class IV towns has increased from 85 lakh to 149 lakh, the number of Class V and class VI towns and their total population declined sharply during the same period.

Again the number of big cities with million plus population has increased from 12 in 1981 to 27 in 2001 and their total population also increased from 42.1 million in 1981 to 73.0 million in 2001. As per 2001 census the size of population of four-cities of India are 11.9 million for Mumbai, 4.58 million for Kolkata, 9.8 million for Delhi and 4.2 million in Chennai.

IV. IMPLEMENTATION

The resulting analysis further can be trained and using machine learning algorithms predictive analysis can be done. This data analysis of census data is used by the governments and NGOs for coming up with suitable solutions.

ACKNOWLEDGMENT *(Heading 5)*

An endeavor of a long period can be successful only with the advice of many well-wishers.

We would like to thank my internal supervisor Mr. T. Raghavendra Gupta for our technical guidance, constant encouragement and enormous support provided to us for carrying out our project work.

We also want to express our sincere gratitude to all my family members and my friends for their individual care and everlasting moral support.

REFERENCES

[1] Census 2011

http://censusindia.gov.in/2011-prov-results/paper2/data_files/India2/Table_2_PR_Cities_1Lakh_and_Above.xls

[2] Google Geocoder for Location Fetching.

[3] Graduation Data Census 2011

http://www.censusindia.gov.in/2011census/C-series/DDWCT-0000C-08.xlsx

[4] https://books.google.co.in/books?hl=en&lr=&id=qYSViFRNMlwC&oi=fnd&pg=PT23&dq=references+for+data+analysis&ots=UbYHNDNOsm&sig=3GzkeUhhPNOb2GWCVqFO6IafR88#v=onepage&q=references%20for%20data%20analysis&f=false

[5] https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2014.963405?journalCode=uasa20