ERRORS IN CENSUS AGE DATA AND ITS RELATIONSHIP WITH LITERACY RATE:A STUDY FOR SOME STATES OF INDIA

PROJECT REPORT

Submitted by

SAMPRITI DUTTA

EXECUTIVE SUMMARY

The objective of this project is to analyze the quality of age data in some states of India namely Assam, Odisha and Kerala and in overall India based on census data of 2011 and making a useful comparison among them. To satisfy the objective , certain indices such as Myer's Blended index , UNJS and Dependency Ratio have been calculated to measure the age heaping ,overall bias and errors in the reported age data and to understand the pressure of the productive population there. The linear relationship between the literacy rate and the different indices is being found with the help of the correlation coefficient .

INTRODUCTION

Age Distribution is one of the most important pieces of data, that are received from census. Censuses are foundational data sources for population studies, the accuracy of which determines the validity and reliability of demographic resource. Throughout the procedure of census errors may occur on many occasions such as; age reporting and recording, especially in developing countries like India, particularly in certain specific population groups (Registrar General of India,2008).

Usually,age data suffers from problems like the 'age-under enumeration' and ' age distortions' due to liking for certain ages e.g. digit like '0' and '5' as preferred more compared to digits like '1' or '9' in societies having low literacy rate(commonly known as digit preference).In a society where vital registration is under developed, a young person whose has not reached the minimum legal age for marriage may report an older age for marital registration. Due to similar reasons , some young people may report older ages in applications to join the military , and some older adults may exaggerate their ages for the legitimacy of retirement welfare. Unintentional misreporting is mainly random, and thus , it will not cause a severe bias in results. However , intentional misreporting would lead to 'age-heaping', a situation in which the population at a certain age or an age ended with certain digit significantly outnumbers population at adjacent ages or ages ended with other digits .

Mason and Cape (1987) says that there are four main causes of age misreporting in any censuses or surveys .Ignorance of actual ages, miscommunication between interviewers and informants ,distortions of ages , errors in recording or processing. All through the data quality of age reporting is also affected by census enumerators and the household registration management and achieving, these errors are often systematical biases and thus may not generate 'age heaping' in practice.

2.Methodology

2.1. Data Collection:

The age data were collected according to the purpose of the project from the website of the censusindia.The link is being given in the acknowledgement part.

2.2 Methodology:

2.2.1. Literacy Rates:

For the analysis of literacy rates of India and the different states the formula has been used for the age groups '15-19','20-24',……,

Literacy rate = (No. of males/ No. of females)*100;

2.2.2. Myer's Blended Index:

Age heaping and digit preference were measured by Myer's blended index and United Nation's Joint score(UNJS).Age ratio score and Sex ratio score were also calculated .

Myer's blended index is calculated for the age above 23 up to 72 and shows the excess deficit of people in ages ending in any of the 10 digits expressed as percentages. It is based on the assumption that the population is equally distributed among the different ages.The steps in the calculation of Myer's blended index are as follows:

- Sum of populations ending in each digit over the whole range is calculated (e.g.,23,3,…..;24,34,….)
- Ascertain sum excluding the first population combined in step 1 (e.g.,24,34,……;25,35,……)
- Weight the sums in step 1 and 2 and the results to obtain an blended population(e.g,weights 1 and 9 for digit 0,weights 2 and 8 for digit 1,….)
- Convert it into percentages.
- Take the deviation of each percentage in step 4 from 10.0, which is the expected value of each percentage .
- A summary index of preference of for all terminal digits is derived as one half of the sum of the deviations from 10.0,each without regard to signs.

2.2.3. UNJS:

The United Nation's joint score is calculated to measure tendentious bias developed by the UN population Division makes use of age distribution in 5-year age intervals separately for males and females and consists of 3 components namely average sex ratio score(S), average male age ratio score(M),average female age ratio score(F).The components are defined as follows:

- ARS of males and femalesare calculated for age groups and are defined here as the ratio of the population in a given age group and to one-third the sum of the population in that age group and in the preceding and following groups multiplied by 100.The age ratio is expressed for a 5-year age group as follows:

$$\text{ARS for } {}_5P_x = {}_5P_x * 10/(\ 1/3[{}_5P_{x-5} + {}_5P_x + {}_5P_{x+5}])$$

,where ${}_5P_x$ is the population between the age group x to x+5.

- Sex ratio score (SRS) for different age groups are here defined as the average of the successive difference differences irrespective of signs of sex ratio of each age group.

Hence ,the UNJS is then defined as

$$UNJS = 3S + M + F$$

2.2.4. Dependency ratio:

The Dependenc ratio is calculated to compare the number of dependent individuals by age to the total population and indicates unemployment. It is the ratio between the number of dependents(anyone above and below the working age) and the number of independent in the potential labor force.The dependency ratio is expressed as follows;

Dependency Ratio=$({}_0P_{14}+{}_{69}P_{100})/{}_{15}P_{65}$.

All the above measures have been measured and compared through different diagrams of India, Assam, Kerala, Odisha and the measures of Bihar, West Bengal, Punjab, Andhra Pradesh, Uttar Pradesh, Jammu and Kashmir have been collected from my project mates.

3. Results

3.1. Values of Differene Indices and Correlation Ratios:

3.1.1Table No. 1:

Location	Literacy Rate	UNJS	Myer's Blended Index	Dependency Ratio
India	69.3	10.19	32.75	51.89868
Assam	61.49	12.47863	71.7962	58.8324
Kerala	84.31	9.905252	49.31006	46.61169
Odisha	63.89	6.155075	123.5998	53.49377

3.1.2. Table No. 2:

Location	Literacy Rate	UNJS	Myer's Blended Index	Dependency Ratio
Andhra Pradesh	62.18	10.8306	40.2903	48.5433
Bihar	55.41	5.2117	44.9283	81.7518
Jammu and Kashmir	62.72	21.368	32.7617	58.77014
Punjab	73.03	7.1809	26.6403	47.9831
Uttar Pradesh	62.46	22.88675	42.10209	63.81085
West Bengal	73.27	11.63085	28.83624	43.76348

[Footnote: The data of Table No.2 has been collected from my project mates Madhushree Ash and Nivedita Bakshi.]

3.1.3. Table No.3:

Title: Values Of Correlation Coefficient Of The Indices With Literacy Rates

Myer's Blended Index	UNJS	Dependency Ratio
-0.211	-0.139	-0.741

3.2. Diagrammatic Representation and Explanation:

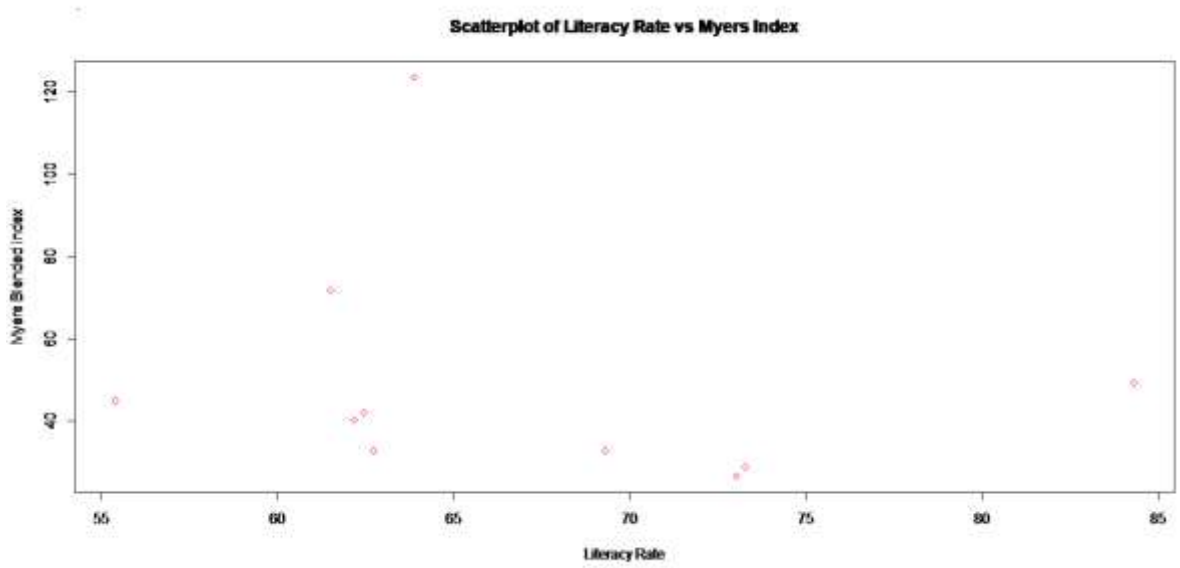

Fig.1. Scatterplot of Literacy Rates and Myer's Blended Index

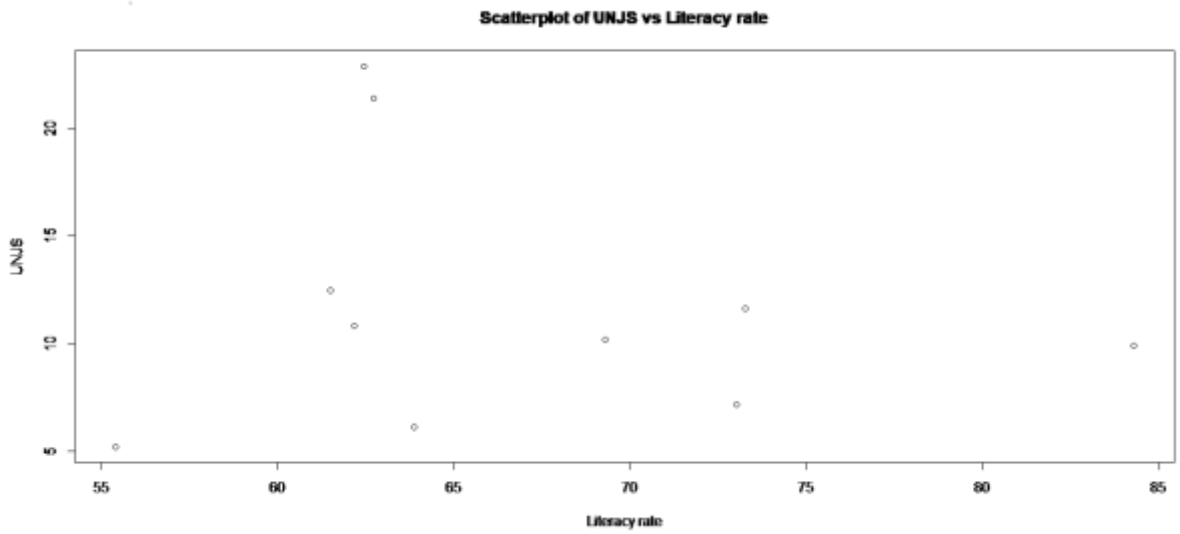

Fig.2.Scatterplot of UNJS and literacy rates

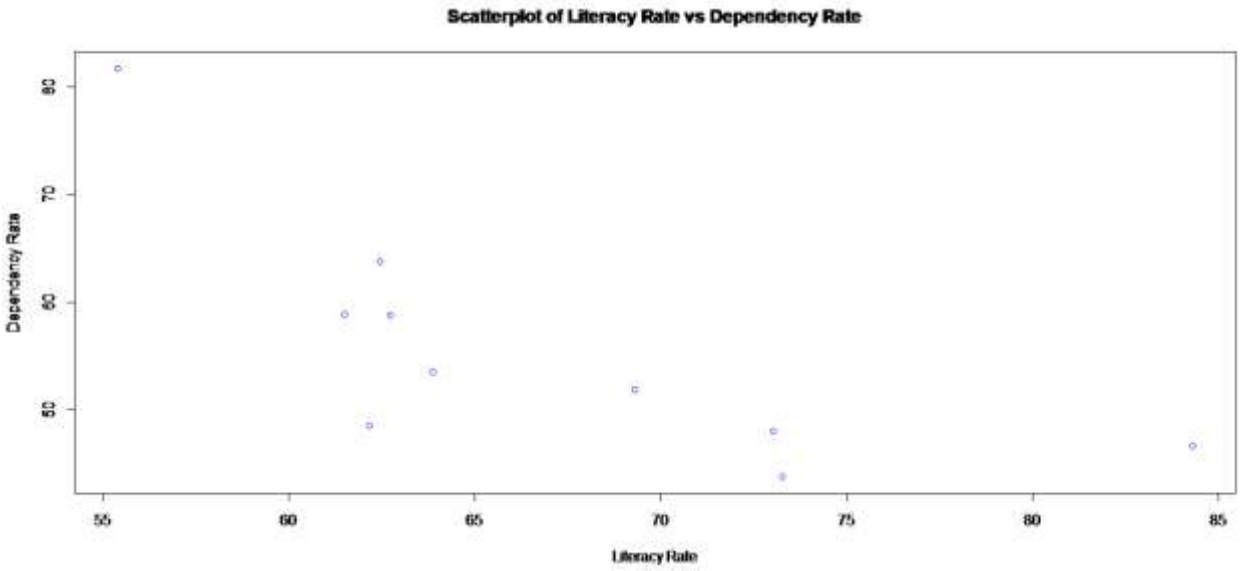

Fig.3.Scatterplot of dependency ratio and literacy rates

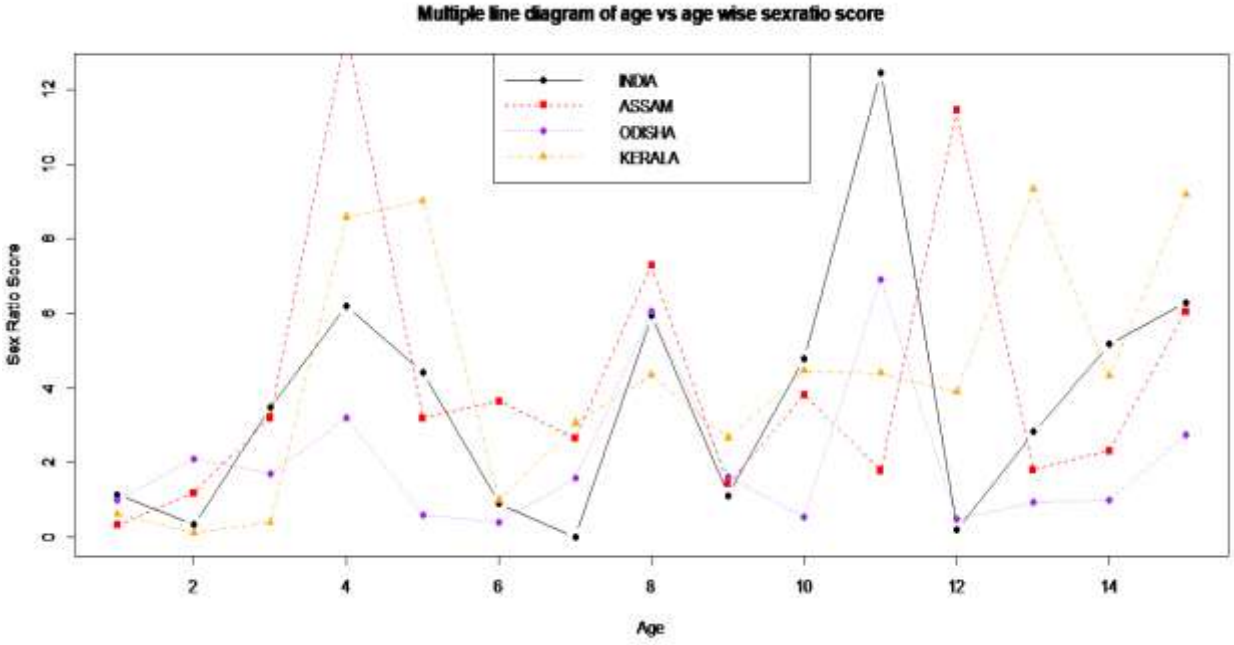

Fig.4. Multiple line diagram of age wise sex ratio score of India, Assam, Odisha and Kerala

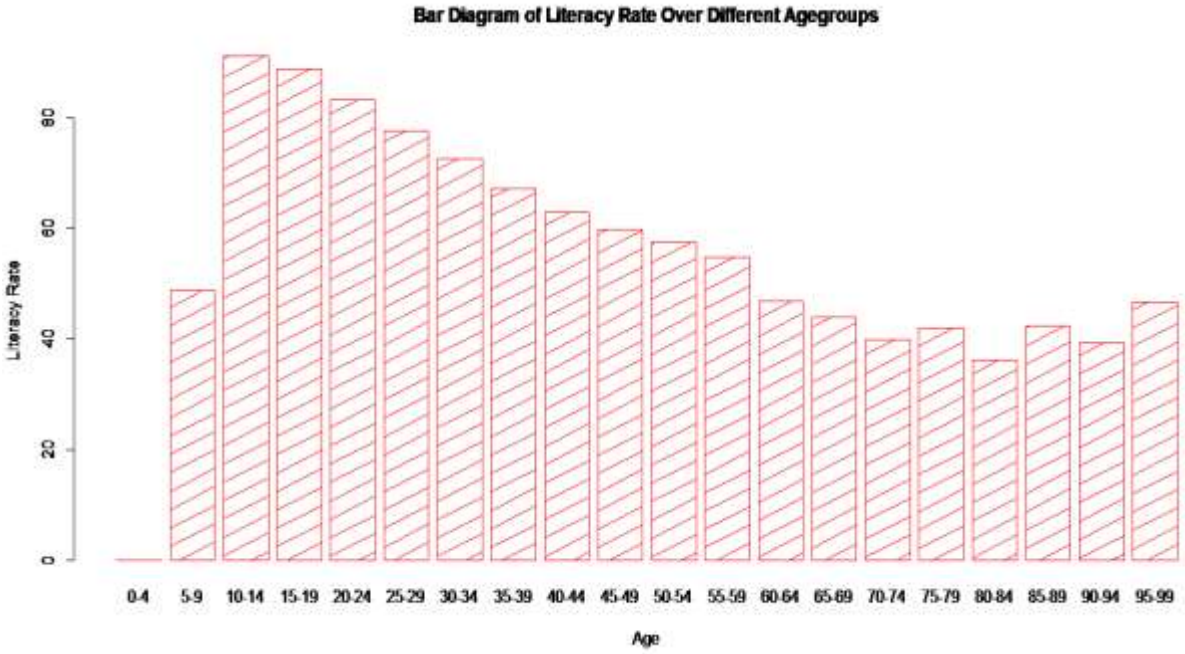

Fig.5. Bar diagram of age wise literacy rates of India

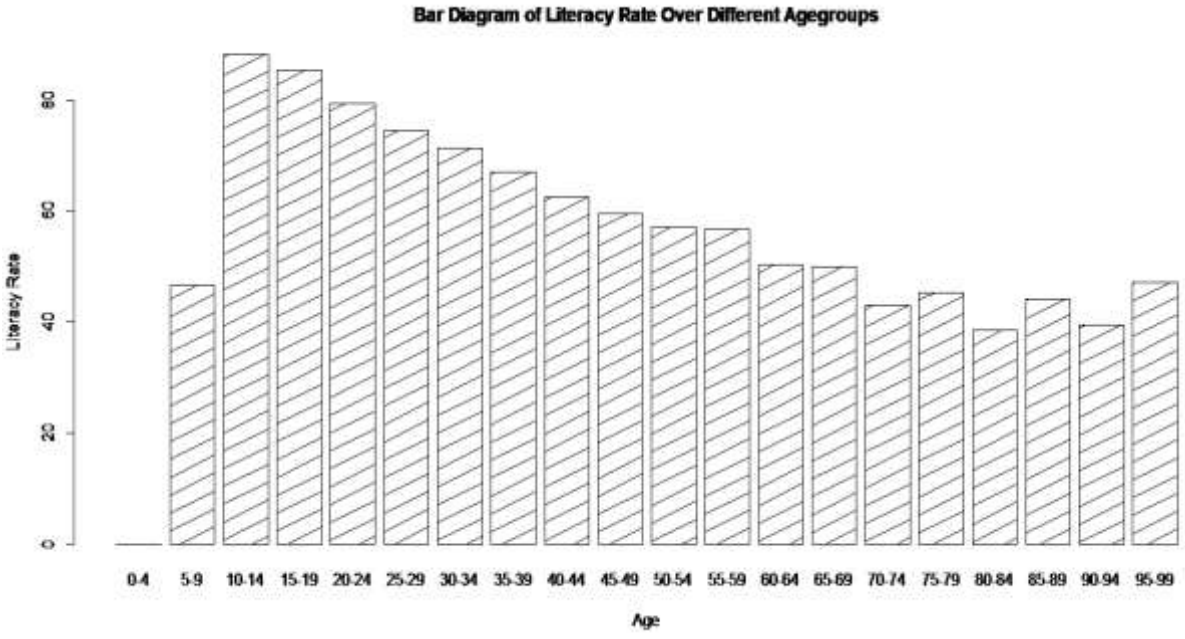

Fig.6. Bar diagram of age wise literacy rates of Assam

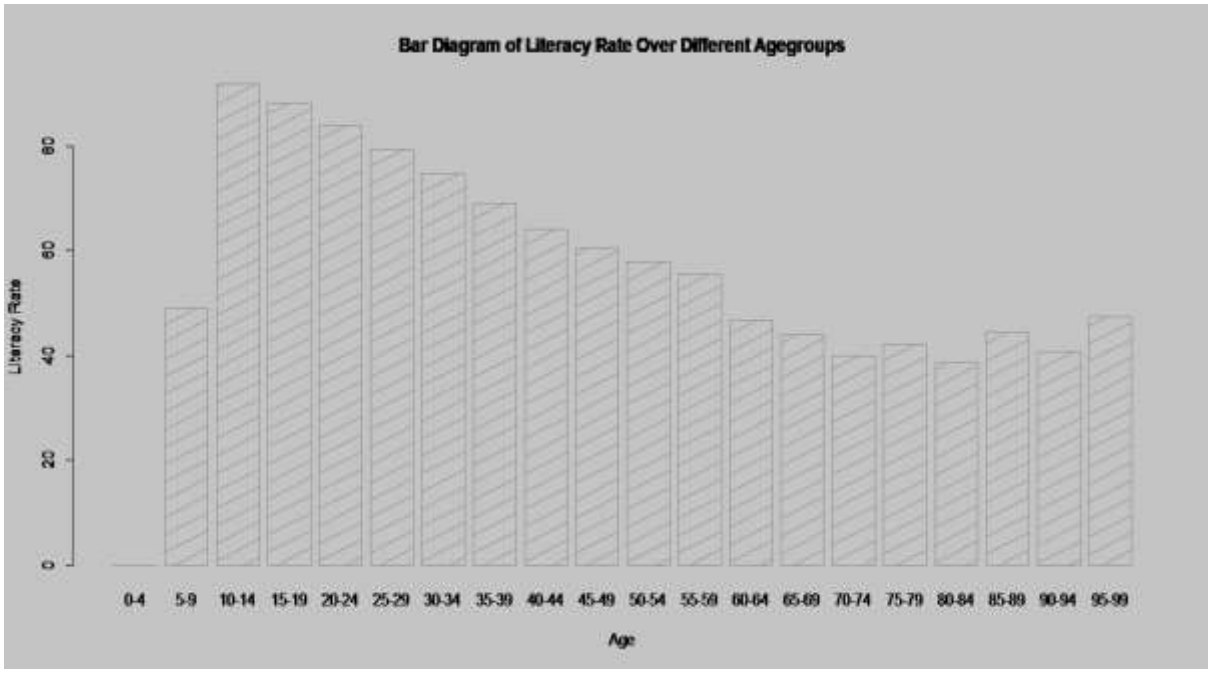

Fig.7.Bar diagram of age wise literacy rates of Odisha

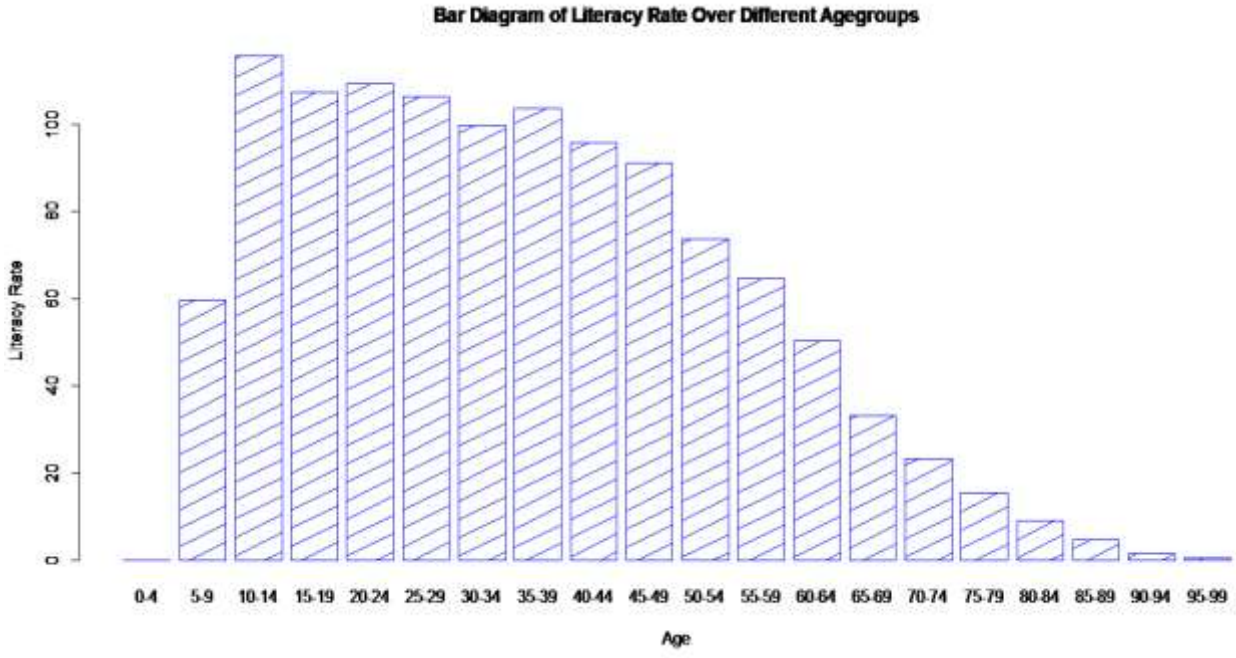

Fig.8. Bar diagram of age wise literacy rates of Kerala

3.3.Explanation:

- The table no.1 shows that, the Myer's Blended Index ,calculated for the states Assam and Kerala ,based on the data of Indian Census 2011,remain within its range of 0-90, such as 71.79 for Assam and 49.31 for Kerala. Whereas the same of Odisha is 123.59. Being a quite high value for Odisha indicates there is a greater tendency of age heaping in data and consequently a low quality of data. For India and the other states the quality of age reporting data is moderate.

- From table no.1,data from UNJS Index as measured by the age ratio score and sex ratio score indicates that the data quality for Assam, Odisha and Kerala are quite good ,the index value of them are respectively 12.47, 6.15, 9.9 being within the desired range of 0-19.However,the UNJS index value of India is 10.18931 ,which also remains well within the range .

- Comparing the dependency ratios from table no.1 it can be said that Assam has a high dependency ratio(58.8324%) compared to the other two states and India and the dependency ratio of Kerala (46.61169%) is quite low , which indicates that there are sufficient people working who can support the dependent population. Whereasa high value of it points to more financial stress on working people and the overall economy faces a greater burden in supporting the rest of the population .

- Finally the Literacy Rates (table no.1), according to the 2011 census of India, the average literacy rate of India stands at 69.30256%.The literacy rate of Kerala(84.31%) is higher than India and Odisha, Assam , which points out the existence of an effective primary education system.

- In Fig.1 ,it is clearly visible from the scatter plot of Myer's Blended Index and Literacy rates of different states of India ,that the points atfirst at a high position and then decreases for the rest part randomly .The correlation coefficient (-0.211) also indicates a weak association between them, which shows that the literacy does not effect much in the age heaping linearly.

- In Fig.2, the scatter plot of UNJS and Literacy Rates of different states of India shows that the points are scattered randomly and there is no increasing or decreasing pattern at all. The correlation coefficient(-0.139) between them is almost nearly zero ; i.e. the literacy does not have any linear relationship with the UNJS.

- In Fig.3 the scatterplot of Dependency Ratio and Literacy rates of different states shows an approximate negative association (-0.741) between them .The correlation coefficient indicates that , greater the literacy rates ,lower the dependency ratio ,i.e.in

a population consisting large number of literate persons ,there are sufficient people working who can support the dependent population .

- In Fig.4,it is evident that the lines of age wise sex ratio score of India, Assam, Odisha, and Kerala are different from each other. There is no monotonic trend in either the four lines and they are showing some random manner(some where increases and some where decreases) in the different sex ratio scores for different age groups .
- From Fig.5 to Fig.8 ,the bar diagrams showing literacy rates of different age groups from 0-4 to 95-99 of India , Asam , Odisha and Kerala are represented.
  In Fig.5,It can be seen that the bar graphs has a zero value of literacy rate at the age group 0-4 and has its highest value at the age group 10-14 .Then the bars show a decreasing trend upto the age group 65-69 and after that the values of literacy rates for the rest of the age groups (70-74 to 95-99) do not differ much from each other .

  In Fig.6 and Fig.7 the picture is more or less similar .Though for Kerala in Fig.8 the diagram is someway different from the rests. The values for literacy rates is the highest for the age group 10-14 for all and for Kerala the rates are almost same upto the group 45-49 and thereafter it is decreasing.

  A noticeable part of the bar graphs lies in the fact that the age group 10-14 possesses the highest literacy rate compared to the other age groups which suggests the existence of an effective primary education system or literacy programs that have enabled a large proportion of children population of acquire the ability of using words, despite of the state they belong to.

4. Conclusion:

The main objective of the project lies on the study of the relationship between age data errors of census with the literacy rates .From the value of the correlation ratio , it is clearly visible that there is not any extent of association of Myer's Index and UNJS with Literacy Rate but there exists some negative association between Dependency Ratio and Literacy Rate.

The secondary objective is to compare the indices(Myer's Blended Index,UNJS, Dependency Ratio) of different states with India. In that regard , the values of Myer's Blended Index of the states Assam, Odisha and Kerala are quite higher than that of India. The UNJS of the states and India do not differ much . The Dependency Ratios of Assam and Odisha are little higher than that of India ,whereas for Kerala it is a bit lower. For Literacy Rates of Assam , Odisha and India are close but that of Kerala is higher .

6. Reference:

Website links:

1. https://censusindia.gov.in/
2. https://ndpublisher.in/

Research papers:

1. Womack, Alpren, Martineau, Jambai, Singh, Kaiser, Redd, April 2020. Quality of age data in the Sierra Leone Ebola database.

2. Manish Singh, January 2016. Understanding digit preferences in India Modified Whipple Index: An analysis of 6forty districts of India.

3. Danan Gu, Qiushi Feng, July 2019. Use of the average age ratio method in analyzing age heaping in censuses: The case of China.

4. Narendra Kumar, Naresh Kumar, Ritu Rani, December 2016. Gender Disparity in Literacy: Districts level evidence from selected states of India

5. UNESCO Institute for Statistics, Jose Pessoa, 2008, Montreal. Guidelines and methodology for the collection, processing and dissemination of international literacy data.

6. Yusuf Bello, January 2017. Age-sex accuracy index chart for monitoring distribution of patients

7. Akhilesh Yadav, Minakshi Vishwakarma, Shekhar Chauhan, September 2019. The quality of age data: Comparison between two recent Indian censuses 2001-2011.

8. Bupe B Bwalya, Million Phiri, Cynthia Mwansa, 2015. Digit preference and its implications on population projections in Zambia: Evidence from the census data.