CHAPTER 1

INTRODUCTION

Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. 90% of the world's data was generated in the last few years.

1.1 Big data :

Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks.

Big data involves the data produced by different devices and applications. given below are some of the fields that come under the umbrella of Big Data.

- **Black Box Data** : It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data** : Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** : The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data** : The power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data** : Transport data includes model, capacity, distance and availability of a vehicle.

- **Search Engine Data** : Search engines retrieve lots of data from different databases.

Fig 1.1 Big data

Thus Big Data includes huge volume, high velocity, and extensible variety of data.

The data in it will be of three types.

- **Structured data** : Relational data.
- **Semi Structured data** : XML data.
- **Unstructured data** : Word, PDF, Text, Media Logs.

1.1.2 Benefits of Big Data :

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

1.1.3 Big Data Technologies :

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real time and can protect data privacy and security.

1.1.3.1 Operational Big Data :

This include systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.

NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

1.1.3.2 Analytical Big Data :

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.

MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.

These two classes of technology are complementary and frequently deployed together.

	Operational	**Analytical**
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads

Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	MapReduce, MPP Database

1.1.3.3 Hadoop :

Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data.

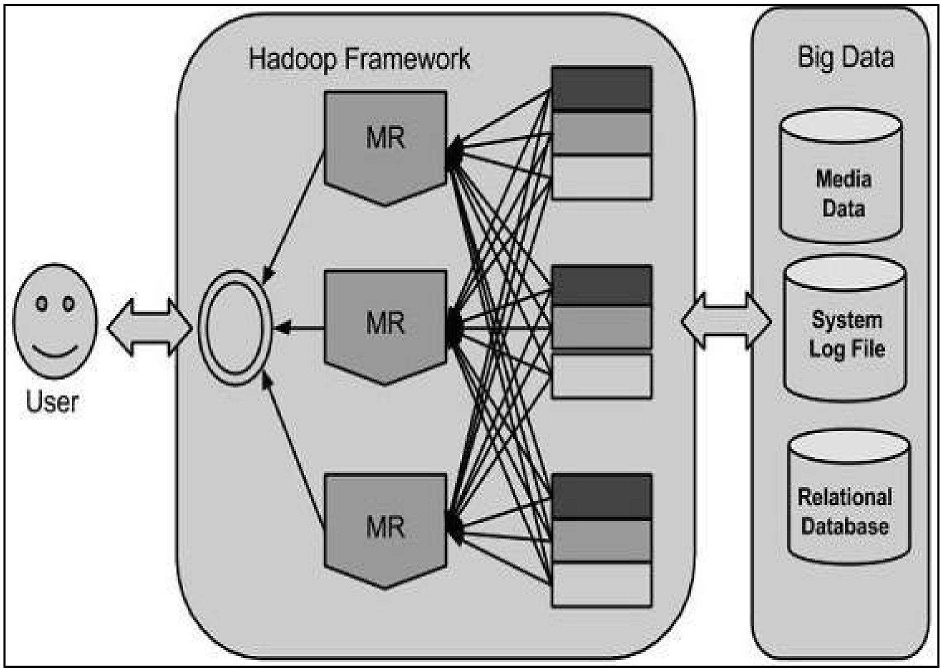


Fig 1.1.3.2 Hadoop framework

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop framework includes following four modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries provides filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
- **Hadoop Distributed File System (HDFS):** A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.

We can use following diagram to depict these four components available in Hadoop framework.

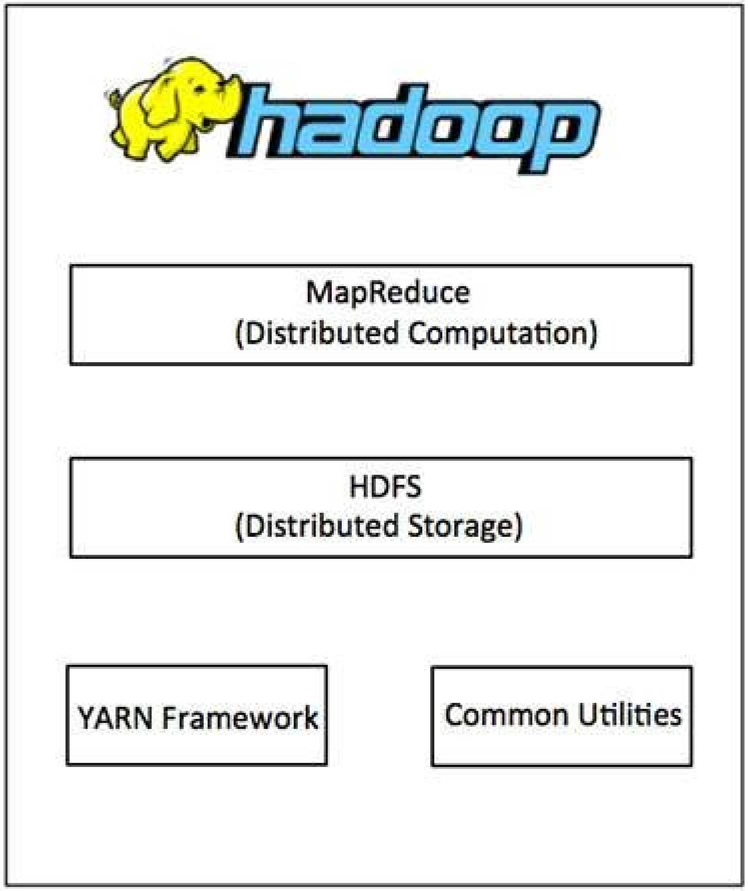


Fig 1.1.3.3 Hadoop Framework

Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.

MapReduce :

Hadoop **MapReduce** is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

The term MapReduce actually refers to the following two different tasks that Hadoop programs perform:

- **The Map Task:** This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples (key/value pairs).
- **The Reduce Task:** This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

Typically both the input and the output are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. The MapReduce framework consists of a single master **JobTracker** and one slave **TaskTracker** per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves TaskTracker execute the tasks as directed by the master and provide task-status information to the master periodically.

The JobTracker is a single point of failure for the Hadoop MapReduce service which means if JobTracker goes down, all running jobs are halted.

Hadoop Distributed File System :

Hadoop can work directly with any mountable distributed file system such as Local FS, HFTP FS, S3 FS, and others, but the most common file system used by Hadoop is the Hadoop Distributed File System (HDFS).

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner.

HDFS uses a master/slave architecture where master consists of a single **NameNode** that manages the file system metadata and one or more slave **DataNodes** that store the actual data.

A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode.

HDFS provides a shell like any other file system and a list of commands are available to interact with the file system.

About programming language R

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

R is free software distributed under a GNU-style copy left, and an official part of the GNU project called **GNU S.**.

The main features of R language are :

R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R:

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility ,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

As a conclusion, R is world’s most widely used statistics programming language. It's the # 1 choice of data scientists and supported by a vibrant and talented community of contributors. R is taught in universities and deployed in mission critical business applications.

1.1 Need of the Project

In the Survey Data Analytics, Surveys collect data from a targeted group of people about their opinions, behaviour or knowledge. Common types of surveys are written questionnaires, face–to–face or telephone interviews, focus groups and electronic(e-mail or Web site) surveys.

Surveys are commonly used with key stakeholders, especially customers and employees, to discover needs or assess satisfaction.

These surveys provide a high level of general capability in representing a large population. Due to the usual huge number of people who answers survey, the data being gathered possess a better description of the relative characteristics of the general population involved in the study. As compared to other methods of data gathering, surveys are able to extract data that are near to the exact attributes of the larger population.

Hence, the data collected from surveys are carefully scrutinised standardized. These data sets collected from surveys helps to represent the patterns related to human development, statistically. It would be easy to identify the problems associated to different areas and address those problems. Since ,these graphs project diversities we can identify better solutions adopted by other areas. So that the growth index would increase by improving the standard of living of people.

1.2 Data sets used to do the analysis:

For the project , Survey data analytics ,we refferedwww.ihds.com.

IHDS has been jointly organized by researchers from the University of Maryland and the National Council of Applied Economic Research (NCAER), New

Delhi, Funding for the second round of this survey is provided by the National Institutes of Health, grants R01HD041455 and R01HD061048. Additional funding is provided by The Ford Foundation.

The First Survey IHDS-I is conducted in 2004-05,The Second Survey IHDS-II is conducted in 2011-12. The Survey is conducted by Interviewing people directly from different areas from every state. Several questionnaires asked in the survey are useful to determine analyse several factors and the data is collected and made available for researchers to conduct researches on data.

The data sets were mad in different formats and categorized in several types are as follows.

- Household file: 36151-0002-Data.dta (N=42,152 households; variables=758)
- Individual file: 36151-0001-Data.dta (N=204,569 individuals; variables= 339)
- Eligible Woman file: 36151-0003-Data.dta (N=39,253 women; variables= 581)
- IHDS-I Tracking file: (N=208,485; variables=81)
- Non-resident family members: (N=4761 individuals; variables= 22)
- Birth history: (N= births; variables=)
- Medical facility: (N= medical facilities; variables=)
- Primary school: (N= schools; variables=)
- Village data: (N=1,410 villages; variables= 741)
- Wages and salaries: (N=64,289 jobs; variables=73

CHAPTER 2

ACQUAINTANCE WITH
DATA ANALYTICS TECHNIQUES

2.1 .Similar work conducted :

Similar work has been carried out by the researchers in the past. Some of these are listed below.

Reference 1:

WP05 : Desai Sonalde and Lester Andrist. 2007. “ Gender Scripts and Age at Marriage in India” Presented at the Annual Meetings of the Population Association of America. New York, N.Y. Demography 47(August): 667-687.

Abstract: This paper uses data from the newly collected India Human Development Survey, 2005 for over 40,000 households around the country to explore ways in which gender scripts regarding valuation of women's modesty and separation of male and female sphere shapes the decisions regarding age at marriage across different regions and different social classes.

Preliminary Results: This analysis is based on ever married women age 25-49. By age 25, more than 95 percent of Indian women are married. This analysis also states that women have less choices in selecting their spouse. There is gender bias in selecting the spouse. Even female are getting married before they reach eligible and acceptable age of marriage.

Reference 2:

WP03: Vanneman, Reeve, James Noon, MitaliSen, Sonalde Desai, and Abusaleh Shariff. 2006. “ Social Networks in India: Caste, Tribe, and Religious Variations.” Presented at the Annual Meetings of the Population Association of America. Los Angeles, CA.

Abstract:

They used the original data from a newly collected nationwide survey for 40,000 households in India . They find the expected hierarchy of Brahmins, high caste Hindus, other backward castes (OBCs), dalits, and tribals (adivasis) in access to these networks. Muslims are relatively low while other minority religions appear similar to high caste Hindus.

After controls for wealth, education, and other household characteristics, the advantages of Brahmins and the disadvantages of adivasis and Muslims remain substantial. However, the weak networks of OBCs and dalits are a consequence of their relative poverty and low education; compared to equivalent high caste Hindus, OBCs and dalits have nearly as good network access to these important institutions. In urban areas, dalits and adivasis do especially well, an effect we attribute to India's strong affirmative action policies.

Hence ,these are the other similar working papers related to our project . In our project ,we analysed the literacy rate in different states , child marriages based on caste and regions ,wages .

CHAPTER 3

SYSTEM STUDY AND ANALYSIS

3.1 User Requirements :

The Application itself is a report generation Program written in R language and developed using the R studio. To run the report generation programs, the user need to provide survey data that is collected either through offline or online surveys. The application programs takes the ordered data from user and produce the reports In a more readable graphic charts.

The application produces the standard reports and experts the standard data set include specific set of attributes, The users who want to use these reports are expected to have the basic understand on different pictorial representations like bar charts ,pie charts and line charts etc. The users also should have good knowledge on the basic statistical concepts like aggregate analysis, summarization concepts. The user is also expected to go through the user guide provided to understand the terminology used in the reports.

3.2 Major Users of Human Development Survey Analytics in India:

3.2.1 Non Profit Organisations(NPO):

A non-profit organization is often dedicated to furthering a particular social cause or advocating for a particular point of view. The valuable information that is represented in graphs and another patterns will be useful for non profit organisations like Swecha to showcase them to people and Government

3.2.2 Non Government Organisations(NGO) :

A non government organisation is dedicated to process the articles by research scholars. The Graphical information produced by this report will be useful for identifying different problems related to the people and society.

3.2.3 Government Organisations:

The information useful for the different organisations related to agricultural growth, Child education, GDP etc..

3.3 H/W & S/W Configuration:

The application entirely developed using R and R studio .R is a analytical and statistical programming language and R studio is an Interactive Digital Environment used for writing R scripts and draw plots and charts. The entire Application programs run on the R studio and R studio can be installed on any Windows ,Mac or Linux Machine.

3.3.1 R Installation In Windows :

- You can download the Windows installer version of R from R-3.2.2 for Windows (32/64 bit) and save it in a local directory.
- As it is a Windows installer (.exe) with a name "R-version-win.exe". You can just double click and run the installer accepting the default settings. If your Windows is 32-bit version, it installs the 32-bit version. But if your windows is 64-bit, then it installs both the 32-bit and 64-bit versions.
- After installation you can locate the icon to run the Program in a directory structure "R\R-3.2.2\bin\i386\Rgui.exe" under the Windows Program Files. Clicking this icon brings up the R-GUI which is the R console to do R Programming.

R Studio itself doesn't require a lot of computational power, so your requirements are going to be dependent on how you're using R. The number of cores, speed of the cores and the amount of RAM that you need is highly dependent on the work/analysis you will be doing. R itself is single threaded, and as such, you won't benefit from additional cores unless you are familiar with the various libraries that parallelize work and are then able to leverage multiple cores.

If you are new to R and data analysis, it is unlikely that you would use more than 1 of your cores and more than 1 GB of RAM for most of your analyses. However, if you intend to be analyzing larger data sets (>1GB) then it would be wise to invest in more RAM. Generally speaking, most people don't leverage the parallelization in R, and so you are better off with fewer cores that are faster than more cores that are slower.

3.4 Data Flow Diagrams :

The application programs in the system takes several types of datasets for analysis. The data that is collected through surveys and questionarries will be ordered in to data sets and be given with certain weight based on the accuracy measurement after cleaning the data . The cleaned data is processed with several algorithms and mathematical and statistical applications to produce or generate reports. The generated reports may be used for predicative analysis or for statistical analysis.

3.4.1 Data analysis Process :

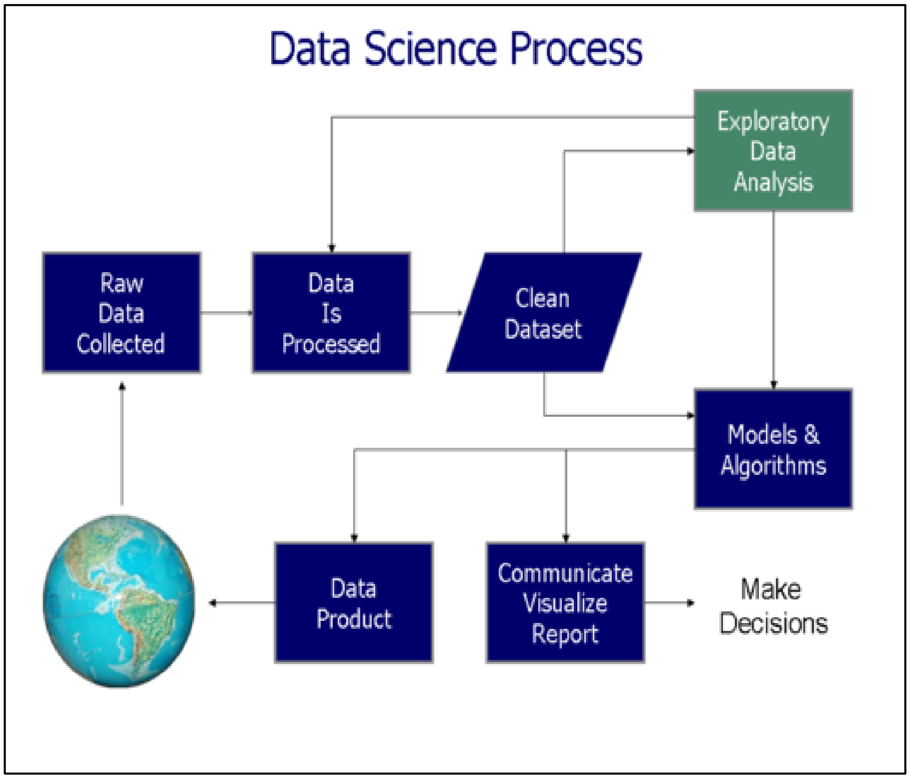


Fig 3.4.1 Data science process

CHAPTER 4

SYSTEM DESIGN

4.1 Generic Survey Process :

This section provides a general overview of the components of the survey system.

4.1.1Components :

Any specific survey must be seen as part of a broader framework of statistical collection. The specific details of this will be particular to an individual country, but some aspects of the system are common across many countries.

- Agriculture Data
- Educational Data
- Wage and Salary Data
- Marriage Data

4.1.2 Survey Methodology:

The survey was carried out in face-to-face interviews containing the following modules:

[1] An interview with a knowledgeable informant – typically the head of the household –regarding socio-economic condition of the household including income, employment educational status, consumption expenditure, and social capital.

2. An interview with an ever-married woman aged 15-49 regarding health, Education ,fertility, family planning, marriage, and gender relations in the household and community. Those ever-married women who were interviewed in

IHDS – I but were no longer eligible i.e. older than 49 years of age have also been interviewed.

3. An interview with youth in the households aged 15-18 years regarding Education ,employment, marriage, life skills, future planning, friendship and risky confidential behaviours.

4. Short reading, writing, and arithmetic knowledge tests were administered to all Available children aged 8-11and youth aged 15-18 in the household. These tests were developed in collaboration with researchers from Pratham, India, and were pretested to ensure comparability across languages

.

5. Height and weight measurement of children under age 5, aged 8-11, their mothers, and other available household members.

6. Facilities assessment of one government and one private primary school as well as a primary health care facility in the community.

7. Village questionnaire assessing employment opportunities and infrastructure facilities in the village.

The survey instruments were translated into 13 Indian languages and were administered by local interviewers.

4.2 The Stages of the Survey Process :

The statistical survey can be considered to fall into three parts all of which will be

discussed in this paper

- Planning and Design Phase
- Implementation and Analysis
- Dissemination and Archiving and Evaluation

4.2.1 Planning and Design Phase

The planning and design phase of a survey is critical to its success. Often this phase is not given sufficient attention. Four elements will be covered:

- Formulation of the statement of objectives;
- Selection of a survey frame;
- Determination of the sample design;
- Questionnaire design

4.2.2 Implementation and Analysis

The implementation and analysis phase is the core of the survey process. The stages of this phase are shown below. Quality control needs to take place at all stages during every operation. All too often some stages in this phase receive attention to the neglect of other areas. Data collection is very costly and can easily consume much of the resources. All too frequently data are successfully collected, but through poor planning the work is not completed to the final analysis stage. Sometimes weak quality control during either data collection or data capture results in a database full of problems. These problems can overwhelm the analyst during the estimation and analysis stages. A critical element of good survey planning is to distribute time and resources to ensure the proper balance to each operation.

The following sections will discuss each of these phases:

- Data collection (with special consideration to including data capture during collection)
- Data capture and coding

- Data Cleaning and Correction
- Editing and Imputation
- Estimation
- Data documentation Data analysis and presentation of survey results

4.2.3 Dissemination and Archiving

Four activities are identified as representing the third phase of the survey programme

- Data dissemination;
- Evaluation;
- Archiving

Once a survey has been completed the results of the survey need to the disseminated and archived. This phase is often seriously neglected. If the data collected are not analysed and disseminated, then all the resources expended on the collection will have been wasted.

Documentation and archiving of the survey are equally important. The value of statistical information can be greatly enhanced by ongoing use and further analysis. This additional use is very difficult without proper documentation. The database needs to be properly documented and archived not only for further use, but also to allow future survey takers to learn from the results of previous surveys and improve on them.

Evaluation is an extremely variable activity. National Censuses in developed countries have been subjected to a multitude of different techniques of evaluation into all their aspects. On the other hand even a small scale survey should be subject to some degree of critical assessment.

4.3 Questionnaire Design:

4.3.1 Introduction

A well designed questionnaire is essential to the success of a survey. A poorly Designed questionnaire results in many survey errors that could have been avoided. Sometimes questions and even complete questionnaires have proved to be unusable. Attention to the issues addressed below will help reduce these problems. There are excellent questionnaires. The questionnaire designer should never by hesitant to copy the good work of others. There are many examples of excellent questionnaires from which countries can draw.

4.3.2Consultation

Before undertaking a survey, the survey taker should carefully define the survey objectives and consult with users concerning the information needs. Many surveys are part of some broader system of statistics. A Census of Agriculture for example has it place in the overall scheme of National Statistics.

There are also international models for the content of Agricultural Censuses. The consultation process needs to balance these aspects while taking appropriate account of specific national needs. The survey producer is, however, in the final analysis responsible for the survey. Sometime data users may have specific interests that need to be placed in the broader context of the survey. And the survey specialist is responsible for limiting the total size of the questionnaire.

4.3.3 Considerations in Drafting the Questionnaires

Many factors need to be taken into account when designing a questionnaire. These are briefly outlined here:

- **Comparability** : Comparability of results with other surveys: One survey is part of a broader system of statistical information. Using the same questions as other surveys strengthens the overall information system.
- **Data Reliability**: Questions must be designed to facilitate responses. Cross Checks between different questions can improve the quality of responses
- **Non response**: Non response is a major problem in many surveys particularly if many questions are not relevant to respondents. Skips can be included to allow non-relevant questions to be by-passed. Care however needs to be taken to design skips in a way that avoid introducing additional complications
- **Interviewers**: Questions must be formulated so that they are clearly understood by the interviewer. The questions need to include sufficient wording so that the interviewer can ask a complete question of the respondent. Often interviewers will have to translate the questions into local languages. Too many words should also be avoided. Lots of examples or explanatory notes in a question can also create confusion.

• **Data processing**: Questionnaires often fail to include elements to assist the data entry process. The codes that will be entered should be organised so that the data entry can proceed easily. Numbers are easier to enter than words. Data entry staff should not have to simultaneously code questions and enter the data

• **Administrative requirements:** A questionnaire should include elements to facilitate the logistics and administration of the questionnaire. Standard geographic codes, interviewer references and other reference numbers will assist later management of the questionnaire. There is always a risk that questionnaires could fall apart. A code number recorded on every page will help correct this situation.

- **Review of Questionnaires:** Questionnaires need to be thoroughly reviewed by persons other than the survey team. Those doing the review should include both those likely to use the data and other independent experts. These independent experts should include both subject matter specialist who can assess the relevance of the content and experts of survey design. Many simple design errors could be avoided if a questionnaire is reviewed by the appropriate experts.

4.4 Doing the Analysis :

4..4.1 Approaches:

Data listings are readily produced by database and many statistical packages. They are generally on a case-by-case basis, so are particularly suitable in EDA as a means of tracking down odd values, or patterns, to be explored. For example, if material is in verbal form, such a listing can give exactly what every respondent was recorded as saying. Sorting these records – according to who collected them, say – may show up great differences in field workers‶ aptitude, awareness or approach. Data listings can be an adjunct to tabulation: in Excel, for example, the Drill Down feature allows one to look at the data from individuals who appear together in a single cell.

There is a place for the use of graphical methods, especially for presentational purposes, where simple messages need to be given in easily understood, and attention grabbing form. Packages offer many ways of making results bright and colourful, without necessarily conveying more information or a more accurate understanding. A few basic points are covered in the guide on Informative Presentation of Tables, Graphs and Statistics. Where the data are at all voluminous, it is a good idea selectively to tabulate most qualitative but numerically coded data i.e. the binary, nominal or ordered categorical types mentioned above. Tables can be very effective in presentations if stripped down to focus on key findings, crisply presented. In longer reports, a carefully crafted, well documented, set of cross-tabulations is usually an essential component of summary and comparative analysis, because of the limitations of approaches which avoid tabulation:-

- Large numbers of charts and pictures can become expensive, but also repetitive, confusing and difficult to use as a source of detailed information. With substantial data, a purely narrative full description will be so long-winded and repetitive that readers will have great difficulty getting a clear picture of what the results have to say. With a briefer verbal description, it is difficult not to be overly selective. Then the reader has to question why a great deal went into collecting data that merits little description, and should question the impartiality of the reporting.
- At the other extreme, some analysts will skip or skimp the tabulation stage and move rapidly to complex statistical modelling. Their findings are just as much to be distrusted! The models may be based on preconceptions rather than evidence, they may fit badly and conceal important variations in the underlying patterns.
- In terms of producing final outputs, data listings seldom get more than a place in an appendix. They are usually too extensive to be assimilated by the busy reader, and are unsuitable for presentation purposes.

4.4.2 One-Way Tables

The most straightforward form of analysis, and one that often supplies much of the basic information need, is to tabulate results, question by question, as “one-way tables�equot;. Sometimes this can be done using an original questionnaire and writing on it the frequency or number of people who ticked each box‟. Of course this does not identify which respondents produced particular combinations of responses, but this is often a first step where a quick and/or simple summary is required.

4.4.3 Cross-Tabulation :

Two-Way & Higher-Way Tables At the most basic level, cross-tabulations break down the sample into two-way tables showing the response categories of one question as row headings, those of another question as column headings. If for example each question has five possible answers the table breaks the total sample down into 25 subgroups. If the answers are subdivided e.g. by sex of respondent, there will be one three-way table, 5x5x2, probably shown on the page as separate two-way tables for males and for females. The total sample size is now split over 50 categories and the degree to which the data can sensibly be disaggregated will be constrained by the total number of respondents represented. There are usually many possible two-way tables, and even more three-way tables. The main analysis needs to involve careful thought as to which ones are necessary, and how much detail is needed. Even after deciding that we want some cross-tabulation with categories of question J as rows and question K as columns, there are several other decisions to be made:

- The number in the cells of the table may be just the frequency i.e. the number of respondents who gave that combination of answers. This may be rephrased as a proportion or a percentage of the total. Alternatively, percentages can be scaled so they total 100% across each row or down each column, so as to make particular comparisons clearer.
- The contents of a cell can equally well be a statistic derived from one or more other questions e.g. the proportion of the respondents falling in that cell who were economically-active women. Often such a table has an associated frequency table to show how many responses went in to each cell. If the cell frequencies represent small subsamples the results can vary wildly, just by chance, and should not be over-interpreted
- Where interest focuses mainly on one area of a two-way table it may be possible to combine rows and columns that we don't need to separate out, e.g. ruling party supporters vs. supporters of all other parties. This simplifies interpretation and presentation, as well as reducing the impact of chance variations where there are very small cell counts.
- Frequently we don't just want the cross-tabulation for all respondents. We may want to have the same table separately for each region of the country – described as segmentation – or for a particular group on whom we wish to focus such as AIDS orphans – described as selection.
- Because of varying levels of success in covering a population, the

response set may end up being very uneven in its coverage of the target population. Then simply combining over the respondents can misrepresent the intended population. It may be necessary to show the patterns in tables, sub-group by sub-group to convey the whole picture. An alternative, discussed in Part 1, is to weight up the results from the sub-groups to give a fair representation of the whole.

4.4.4 Tabulation & the Assessment of Accuracy :

Tabulation is usually purely descriptive, with limited effort made to assess the accuracy of the numbers tabulated. We caution that confidence intervals are sometimes very wide when survey samples have been disaggregated into various subgroups: if crucial decisions hang on a few numbers it may well be worth putting extra effort into assessing – and discussing – how reliable these are. If the uses intended for various tables are not very numerical or not very crucial, it is likely to cause unjustifiable delay and frustration to attempt to put formal measures of precision on the results.

Usually, the most important considerations in assessing the "quality" or "value" or accuracy of results are not those relating to „statistical sampling variation, but those which appraise the following factors and their effects:

- evenness of coverage of the target (intended) population
- suitability of the sampling scheme reviewed in the light of field experience and findings
- sophistication and uniformity of response elicitation and accuracy of field

recording

- efficacy of measures to prevent, compensate for, and understand non-response.
- quality of data entry, cleaning and metadata recording
- selection of appropriate subgroups in analysis

If any of the above factors raises important concerns, it is necessary to think hard about the interpretation of statistical measures of precision such as standard errors. A factor that has uneven effects will introduce biases, whose size and detect ability ought to be dispassionately appraised and reported with the conclusions. Inferential statistical procedures can be used to guide generalisations from the sample to the population, where a survey is not badly affected by any of the above. Inference addresses issues such as whether apparent patterns in the results have come about by chance or can reasonably be taken to reflect real features of the population. Basic ideas are reviewed in Understanding Significance: the Basic Ideas of Inferential Statistics.

More advanced approaches are described in Modern Methods of Analysis. Inference is particularly valuable, for instance, in determining the appropriate form of presentation of survey results. Consider an adoption study, which examined socioeconomic factors affecting adoption of a new technology. Households are classified as male or female headed, and the level of education and access to credit of the head is recorded. At its most complicated the total number of households in the sample would be classified by adoption, gender of household head, level of education and access to credit resulting in a 4-way table.

4.4.5 Profiles

Usually the questions as put to respondents in a survey need to represent atomic facets of an issue, expressed in concrete terms and simplified as much as possible, so that there is no ambiguity and so they will be consistently interpreted by respondents.

Basic cross-tabulations are based on reporting responses to such individual questions and are therefore narrowly issue-specific. A rather different approach is needed if the researchers ambitions include taking an overall view of individual, or small groups, responses as to their livelihood, say. Cross-tabulations of individual questions are not a sensible approach to people-centred or holistic summary of results.

Usually, even when tackling issues a great deal less complicated than livelihoods, the more important research outputs are complex molecules which bring together responses from numerous questions to produce higher-level conclusions described in more abstract terms. For example several questions may each enquire whether the respondent follows a particular recommendation, whereas the output may be concerned with overall compliance– the abstract concept behind the questioning. A profile is a description synthesising responses to a range of questions, perhaps in terms of a set of abstract nouns like compliance. It may describe an individual, cluster of respondents or an entire population.

One approach to discussing a larger concept is to produce numerous cross-tabulations reflecting actual questions and to synthesise their information content verbally. This tends to lose sight of the profiling element: if particular groups of respondents tend to reply to a range of questions in a similar way, this overall grouping will often come out only weakly. If you try to follow the group of individuals who appear together in one corner cell of the first cross-tab, you can't easily track whether they stay together in a cross-tab of other variables.

Another type of approach may be more constructive: to derive synthetic variables –indicators – which bring together inputs from a range of questions, say into a measure of compliance, and to analyse those, by cross-tabulation or other methods.

If we have an analysis dataset with a row for each respondent and a column for each question, the derivation of a synthetic variable just corresponds to adding an extra column to the dataset. This is then used in analysis just like any other column. A profile for an individual will often comprise a set of values of a suite of indicators.

4.4.6 Looking for Respondent Groups

Profiling is often concerned with acknowledging that respondents are not just a homogeneous mass, and distinguishing between different groups of respondents. Cluster analysis is a data-driven statistical technique that can draw out – and thence characterise – groups of respondents whose response profiles are similar to one another.

The response profiles may serve to differentiate one group from another if they are somewhat distinct. This might be needed if the aim were, say, to define target groups for distinct safety net interventions. The analysis could help clarify the distinguishing features of the groups, their sizes, their distinctness or otherwise, and so on. Unfortunately there is no guarantee that groupings derived from data alone will make good sense in terms of profiling respondents. Cluster analysis does not characterise the groupings; you have to study each cluster to see what they have in common. Nor does it prove that they constitute suitable target groups for meaningful development interventions.

Cluster analysis is thus an exploratory technique, which may help to screen a large mass of data, and prompt more thoughtful analysis by raising questions such as:-

- Is there any sign that the respondents do fall into clear-cut sub-groups?
- How many groups do there seem to be, and how important are their separations?
- If there are distinct groups, what sorts of responses do “typical” group Members give?

CHAPTER 5

RESULTS AND DISCUSSION

5.1 .Educational dropout rates:

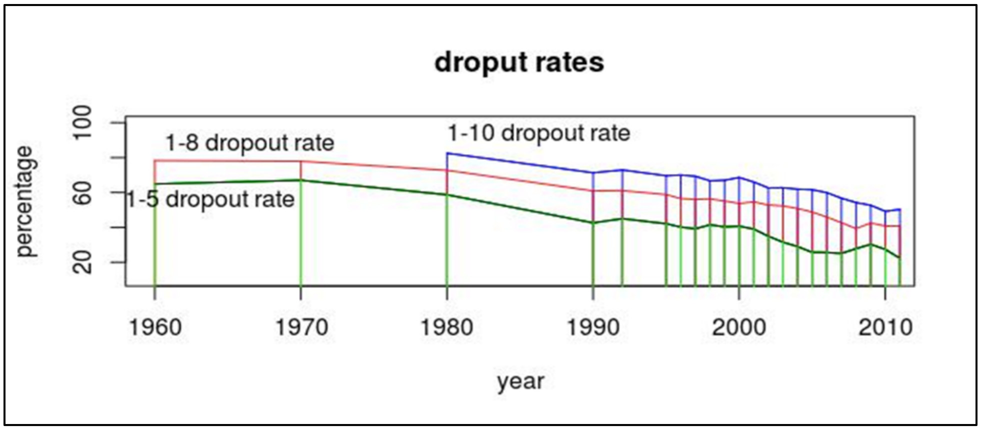


Fig 5.1 Educational dropout rates

The above fig 5.1 shows, educational dropouts from the year 1960-2010 . On the X-axis ,year were marked and on the Y-axis , percentage of dropouts in education were shown. Clearly ,in 1960 , the dropout rate from classes 1-5 is nearly 65% and from classes 1-8, it is nearly 80% . the dropout rate is same till 1970.

In 1975-1980 ,there is slight decrease in the dropouts in between classes 1-5 and classes 1-8 . In 1980 , there are nearly 83% dropouts in between classes 1-10 . then the curve is decreasing till 1990. In 1990 , there is slight increase in the dropouts in between the classes 1-5 and in between the classes 1-10. Then dropout rate is oscillating during the period 1995-2000. From 2001, the dropouts graph of classes 1-5 is gradually decreased to nearly 30% and then it is slightly increased in the year 2009 and the graph dropouts in between classes 1-8 is also slightly decreased compared to the previous year .but in the year 2010, there is a slight increase in the dropouts in between classes 1-5 and classes 1-8 and then they were decreased.

Conclusion : Dropout rates are steadily falling during the years 1990-2010 , it is a good sign . Due to the continues efforts of government and non-governmental organizations , everyone are aware of education benefits and willing to educate.

5.1.1 Girl dropouts rates:

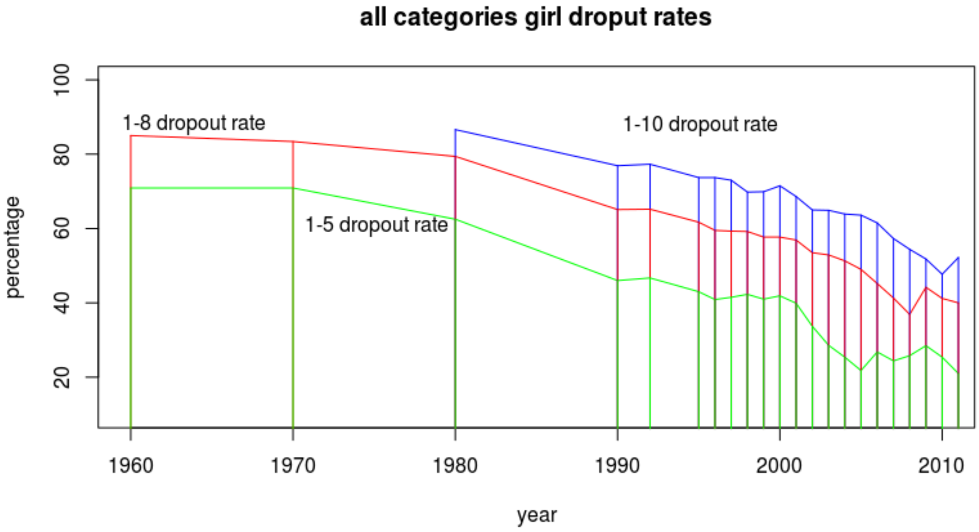


Fig 5.1.1 Girl dropout rates

The above fig 5.1.1 shows the girl educational dropouts .On the X-axis, years were marked and on the Y-axis ,percentage of dropouts were marked.

In 1960-70 , the dropout rate in classes 1-5 is nearly 70% and in classes 1-8 is nearly 83%. There is decrease in the dropouts rate in between the classes 1-5 in the year 1980. The dropouts in classes 1-10 is nearly more than 80%. The graph was decreasing till 1990. Then there is slight increase in the dropouts in 1992-93 and then decreased. The graph shows ups and downs till 2000. In 2005 ,there is a decrease in dropouts from classes 1-5 ,it is nearly 20 % decreased. In classes 1-8 and 1-10 ,the graph was slightly increased. In 2008, the dropouts from classes 1-8 were decreased ,nearly 10% compared to previous year. Then slightly increased in the year 2010.

Conclusion : The dropout rates are decreasing during the year 1990- 2010 .due to the continues efforts of government and non-governmental organizations , people are aware of female education benefits.

5.1.2 Caste /Girl dropouts :

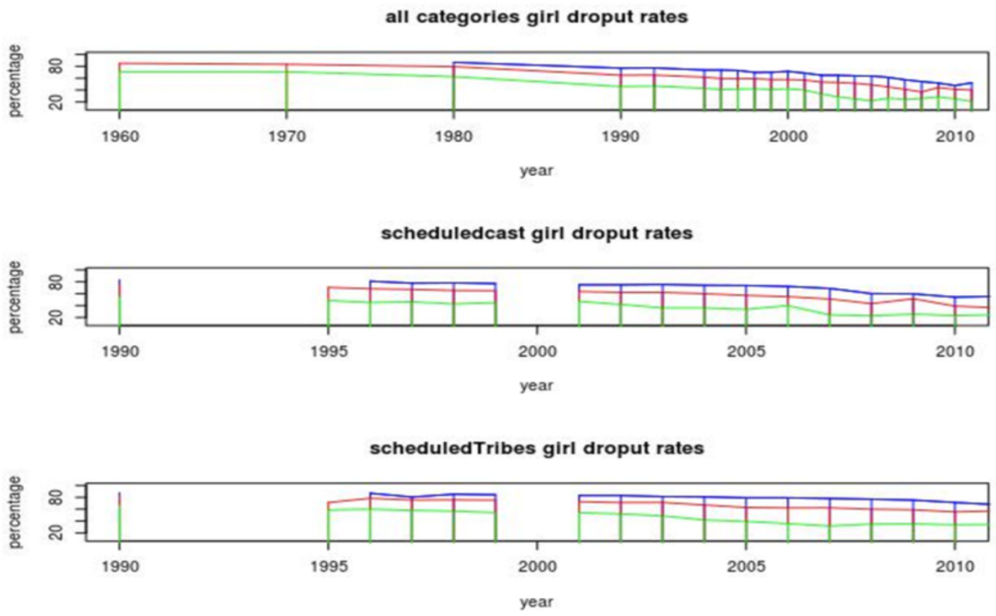


Fig 5.1.2 Caste wise girl dropout rates

The fig 5.1.2 shows the dropouts rates of girls in education based on caste. On the X-axis, years were marked and on the Y-axis ,percentage of dropouts were marked.

The second graph shows the scheduled caste dropouts. In 1990 , the dropouts in classes 1-5 is nearly 60% and in classes 1-8 is nearly 80%. There is no effective changes taken place in these graphs till 1995. In 1996 , the dropouts in the classes 1-10 are more than 80% .the dropout rates in the classes 1-5 increased in 2006 and then slightly decreased in 2007. From 2007 ,the dropouts in between classes 1-10 are slightly decreasing but there is slight increase in the dropouts in the classes 1-8 .

The third graph shows the scheduled tribes dropouts. In 1990 , the dropouts in classes 1-5 is nearly 70% and in classes 1-8 is nearly 83%.then decreased and increased slightly in 1996. In 1996, the dropouts in the classes 1-10 are more than 85% .From 2002, there is slight decrease in the dropouts graph of classes 1-5 that nearly 40%. But in 2007 ,the graph is slightly increasing to 50%. There is no noticeable changes taken place in dropouts in classes 1-10.

Conclusion : The dropout rates are decreasing during the year 2008-2010due to the continues efforts of government and non-governmental organizations , people are aware of female education benefit. They are implementing different programs to increase the female education. Now, people are aware of female education benefits so ,the dropout rates are decreasing.

5.1.3 State wise children distribution :

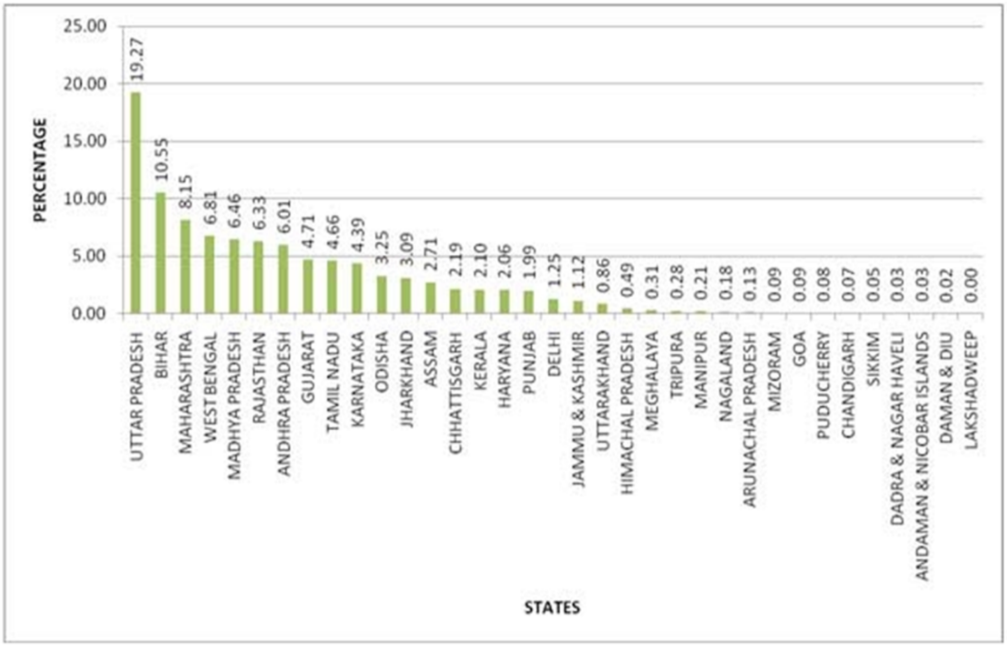


Fig 5.1.3 State wise children distribution

In the above fig 5.1.3, the state wise children distribution is shown .On the X-axis, states were marked and on the Y-axis ,percentage of Children was marked.

Out of 25 ,Uttar Pradesh contains 19.27 % of children . It has the highest population of children. Bihar contains 10.55% ,Maharashtra contains 6.81% of children. These are the top 3 states and Andhra Pradesh stood in seventh place with 6.01% of children. Daman & Diu and Lakshadweep stood in last two positions with 0.02% and 0.00 % of children.

Conclusion :

The state with highest children and growth rate is UtterPradesh. The state with lowest children and growth rate is Damen &Diu and Lakshadweep .

5.1.4 State wise literacy rate :

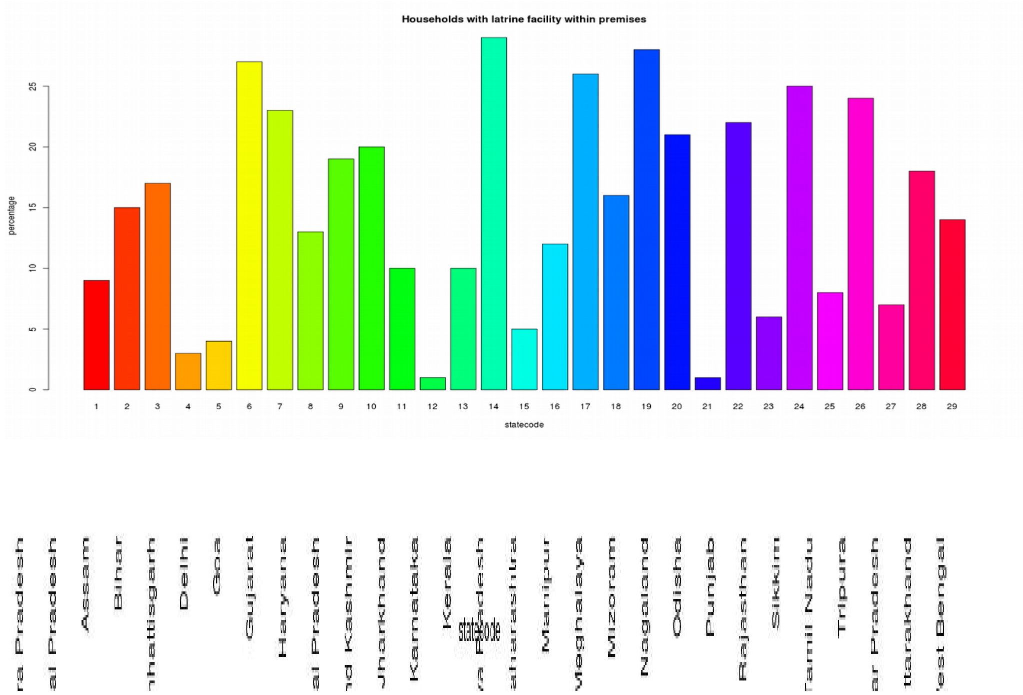


Fig 5.1.4 State wise literacy rate

The above fig 5.1.4 shows the state wise literacy rate .On the X-axis, State codes were marked and on the Y-axis ,percentage was marked.

The above graph states that Kerala stood first with the highest literacy rate with more than 92% followed by Delhi , Manipur , Mizoram in next three positions. Andhra Pradesh has 60% literacy rate. Orissa ,Jharkhand , Bihar occupied the last three positions. From this graph ,we can say that proper action should be taken to increase the literacy rate in India.

Conclusion :

The state with highest literacy rate is Kerala . The state with lowest literacy rate is Orissa .

5.1.5 State wise literacy rate(male):

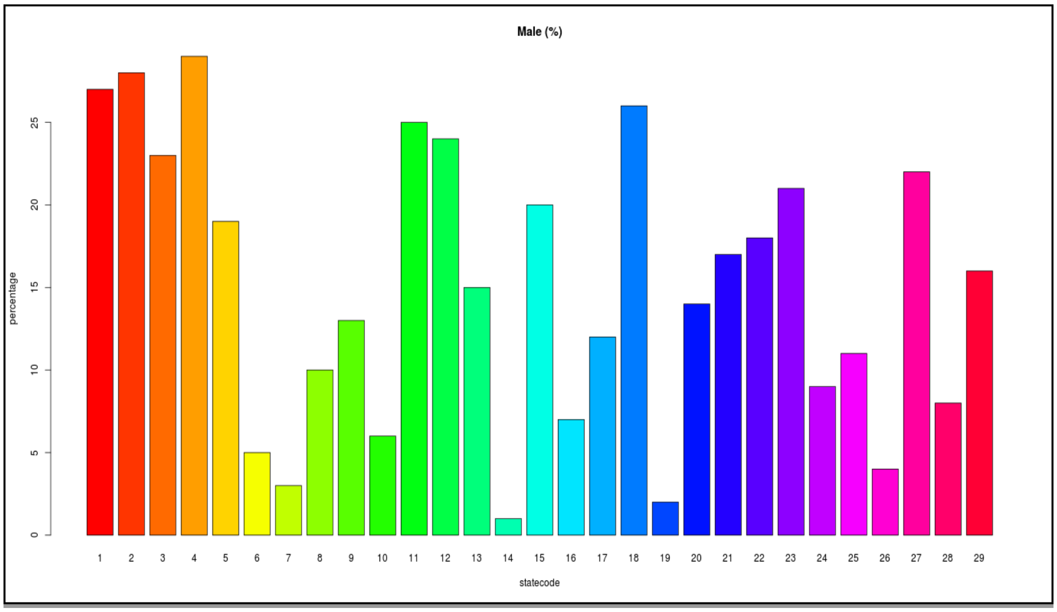


Fig 5.1.5 State wise male literacy rate

In the above fig 5.1.5 , the state wise literacy rate is shown .On the X-axis, State codes were marked and on the Y-axis ,percentage was marked.

The state codes are :

1. Arunachal Pradesh 2. Andhra Pradesh 3. Assam 4. Bihar
5.Chhatisgarh 6. Delhi 7. Goa 8.Gujarat 9. Haryana
10. Himachal Pradesh 11. Jammu & Kashmir 12.Jharkand 13.Karnataka
14.Kerala 15. Madhya Pradesh 16. Maharashtra 17. Manipur
18.Meghalaya 19. Mizoram 20. Nagaland 21. Orissa 22. Punjab
23. Rajasthan 24. Sikkim 25. Tamil Nadu 26. Tripura 27. Uttar Pradesh
28. Uttarakhand 29. West Bengal

The above graph states that Bihar stood first with the highest male literacy rate with more than 90% followed by Andhra Pradesh , Arunachal Pradesh , Meghalaya in next three positions. Goa, Tripura, Kerala, Madhya Pradesh , occupied the last three positions.

Conclusion: proper and effective measures should be taken to improve the male literacy rate by campaigning about education.

5.1.7 State wise literacy rate(Female):

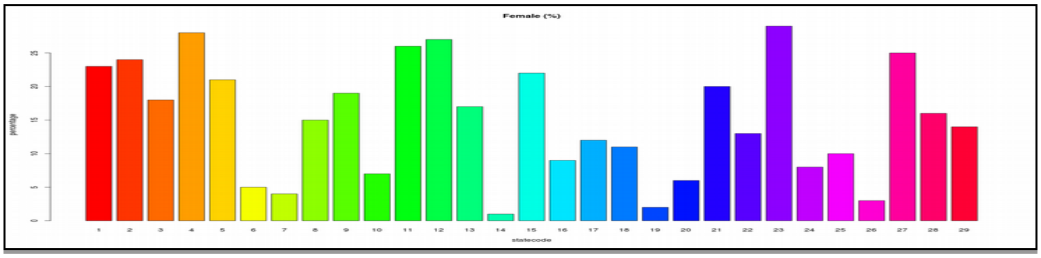


Fig 5.1.7 State wise female literacy rate

In the above fig 5.1.7, the state wise female literacy rate is shown .On the X-axis, State codes were marked and on the Y-axis ,percentage was marked.

The states codes are mentioned above. This graph states that Rajasthan Occupies first position . Mizoram ,Tripura occupies the last two positions . Andhra Pradesh has more than 80% of female literacy rate.

Conclusion :

Compared to male literacy rate female literacy rate is low in many states. Effective measures should be taken to increase the female literacy rate in India.

5.1.8 State wise gender gap in literacy rate:

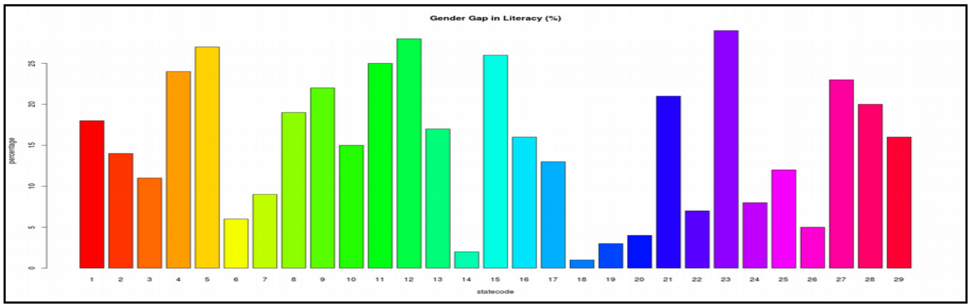


Fig 5.1.8 state wise gender gap in literacy rate

In the above fig 5.1.8 , the state wise literacy rate is shown .On the X-axis, State codes were marked and on the Y-axis ,percentage was marked.

The states codes are mentioned above. This graph states that Punjab has the highest gender gap in literacy rate which is more than 90%. It means that number of male educated is more than number of female educated . Followed by Jharkhand , Chhattisgarh . Andhra Pradesh has more than 50% gender gap in literacy rate. Kerala has the least gender gap .

Conclusion : Effective measures should be implemented to reduce the gender gaps in literacy rate and to increase the literacy rate.

5.2. Female marriages :

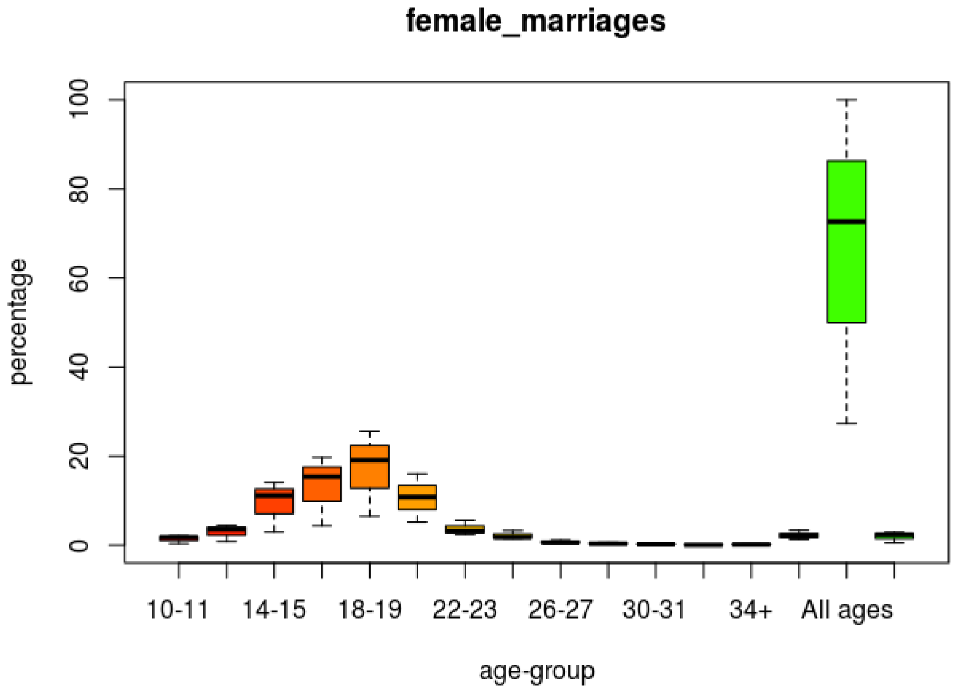


Fig 5.2 Female marriages

The above fig 5.2 shows the details about female marriages .On the X-axis, Age groups were marked and on the Y-axis ,percentage was marked.

From this graph ,we can say that in India , the most preferable age for marriage is in between 18-19. Nearly 20% of the people are getting married at this age . Then nearly 15% of the people are getting married at 16-17 age ,which is mostly seen in rural and urban areas. The most preferred and eligible age group for marriage is 21-26 ,is very low in both urban and rural areas. People are getting married at other age groups but in very low percent .

Conclusion: Proper action should be taken by the government to decrease the child marriages and people should be educated to avoid the child marriages .

5.2.1 Female marriages ages at Rural/Urban (14 - 21):

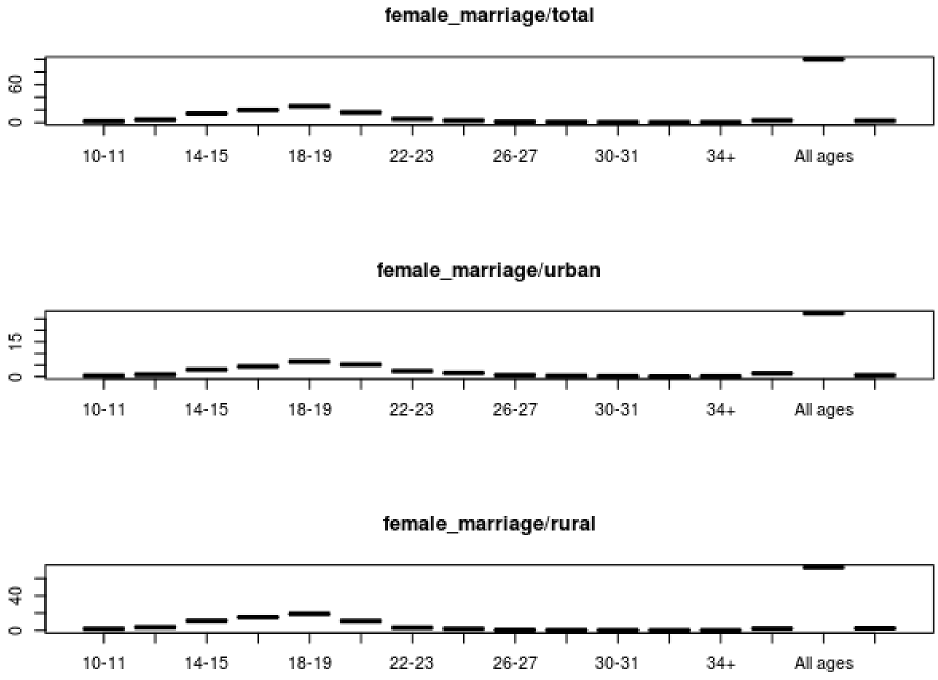


Fig 5.1.2 Female marriage ages at rural /urban areas

The above fig 5.1.2 shows the urban and rural female marriages. On the X-axis, Age groups were marked and on the Y-axis ,percentage was marked.

The second graph shows the female marriages at urban areas . The third graph shows the female marriages at urban areas .In both urban and rural areas

The preferred age for marriage is in between 14-21 . Mostly of them were child marriages .Females in these areas are getting married before they get major.

Conclusion: People should be aware about child marriages and their disadvantages in these areas to decrease the child marriages .

5.2.2 Child marriages in Rural/Urban areas(male):

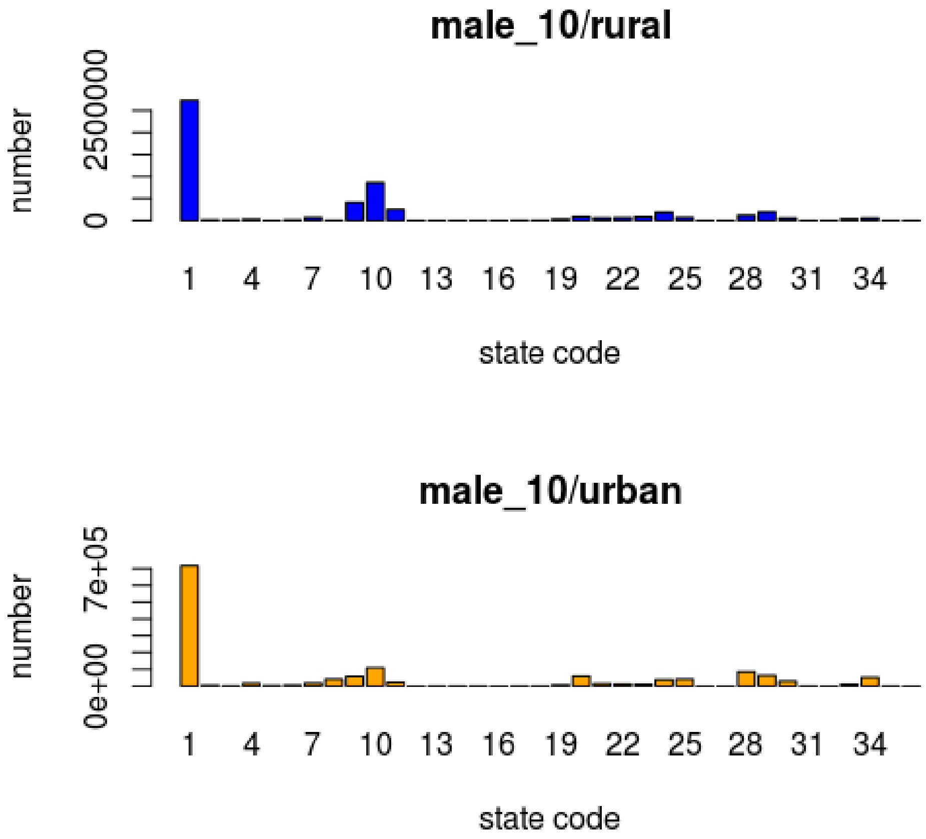


Fig 5.2.2 Male child marriages at rural /urban areas

The above fig 5.2.2 shows the urban and rural male child marriages. On the X-axis, State codes were marked and on the Y-axis ,number of marriages were marked.

The first graph shows the male child marriages at urban areas . The second graph shows the male child marriages at urban areas .In both urban and rural areas The state codes are mentioned in 5.1.3. this graph states that Rajasthan ,Uttar Pradesh ,Bihar areas have highest child marriages followed by Orissa and Jharkhand states.

Conclusion: People should be aware about child marriages and their disadvantages in these areas to decrease the child marriages . Proper action should be taken to avoid child marriages and female should get educated.

5.2.3 Child marriages in Rural/Urban (Female):

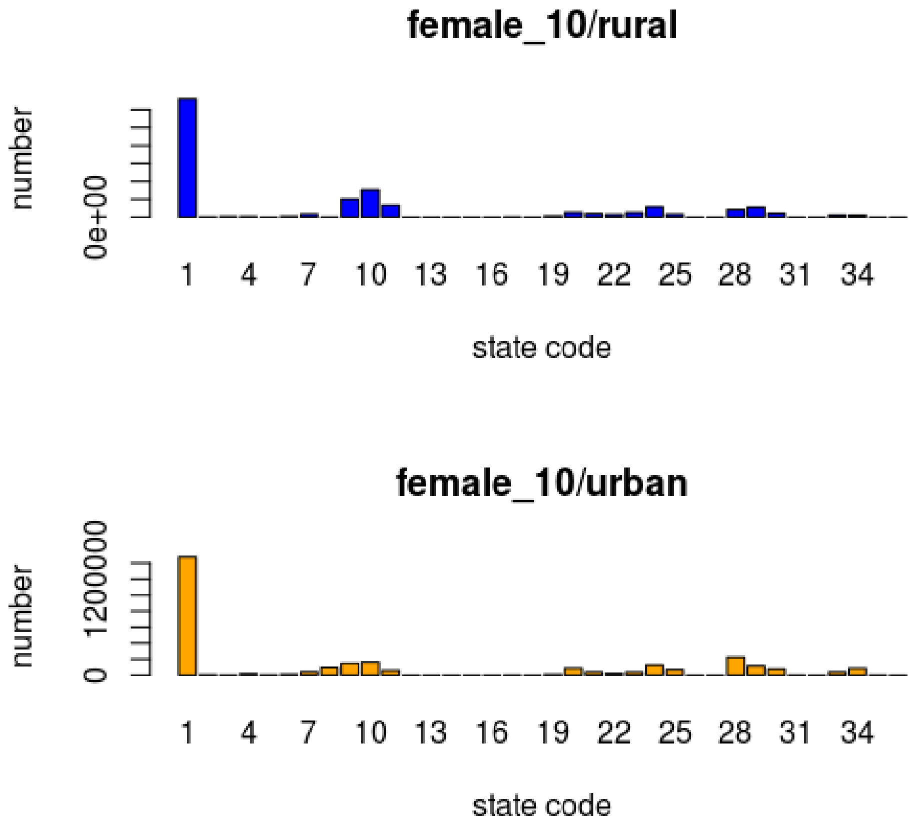


Fig 5.2.3 Female child marriages in rural/ urban areas

The above fig 5.2.3 shows the urban and rural female marriages. On the X-axis, State codes were marked and on the Y-axis ,number of marriages was marked. The second graph shows the female marriages at urban areas . The third graph shows the female marriages at urban areas .In both urban and rural areas

The state codes are mentioned in 5.1.3. this graph states that Rajasthan ,Uttar Pradesh ,Bihar areas have highest child marriages followed by Orissa and Jharkhand states.

Conclusion: Effective measures should be taken to control the child marriages by bringing in awareness in these areas and this happens when people get educated.

5.2 Wage Dependency in India:

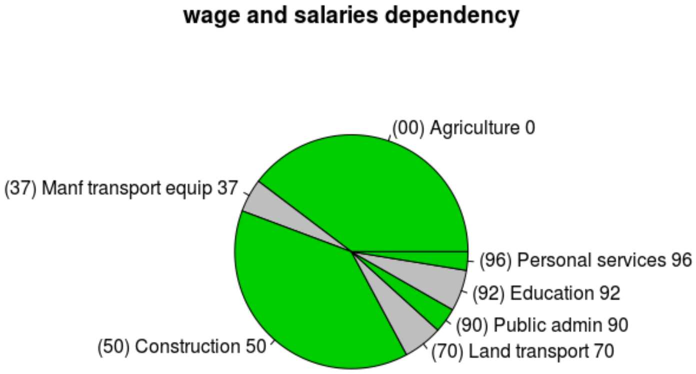


Fig 5.3 Wage dependency in India

The figure 5.3 shows the wage dependencies in India. In India, people mostly depend on agriculture for their livelihood and then on construction ,education, main transportation ,public administration.

In India ,mostly of the people work of daily wages than monthly salaries. This is due to lack of education ,unemployment ,lack of resources etc. Proper action should be taken to improve the standard of living of people .

We can see that very less people are depending on other education ,transportation, public administration.

Conclusion : As major people are depending upon agriculture ,people who are farming should be taught to use proper technology in in agriculture ,to yield more crops .so that ,they may gain more profits .then their standard of living also increases.

5.3.1.Working hours /day:

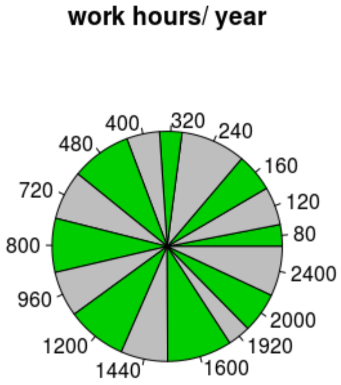


Fig 5.3.1 Working hours / year

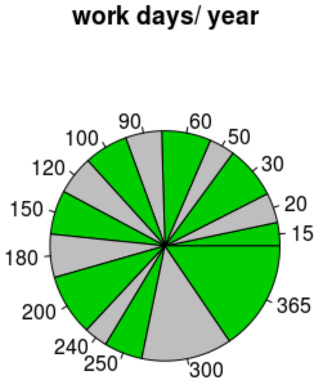


Fig 5.3.1.1 Working days/ year

The above fig 5.3.1 and fig 5.3.1.2 shows the number of working hour per day of the people in India. From these charts, we can observe that the people who are working for less than or equals to 300 hours(<=300) are more than the people who are working more than 300 hours(>300) .

This states that , in India most of the people work for daily wages. This shows that their standard of living is low . people are mostly depend on agriculture and construction for their living. This show that they have less chances to move to other professions .this happens because of their illiteracy, poverty, discrimination etc.

Conclusion: People should be taught with proper life skills to improve their standard of living. Proper measures should be taken to improve the employable skills in the people.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion :

The Info graphics and projections show the Diversities among different states in India like Education, Agricultural Performances and Growth rates, wages, child marriages etc.

The reports produced here will be useful for Government and Non-Government Organizations to know the statistics info graphically. This helps to identify the diversities easily so that appropriate measures can be taken . It makes ease to understand the problem and analysis can be done easily .so that ,it would be easy to address the problem . Most of the times many problems can not be identified but by this survey based analysis helps to find the unknown problems in a domain (like education ,medical ,security ,etc). so that the severity of the problem decreases and Few times this is helpful in finding the solution earlier .

In our project , we projected the diversities in different states of India. We analysed the reports produced . Based on our analysis , India has good growth rate in few areas like education, medical etc but lacking in many areas like avoiding child marriages , promoting female education etc. Today India has occupied 2nd position in population but many people are below poverty line . This is because lack of proper education , employment etc . If government take proper actions then soon The India ‘s standard of living increases .

6.1 Future work:

We can also analyse the big data by using Hadoop . It would be the most preferred choice because it provides the ability to store large scale data on HDFS (Hadoop Distributed File System). Even there are multiple solutions available in the market for analyzing this huge data like MapReduce, Pig and Hive.

Hadoop is most preferable because of its reliability , flexibility . It is economic and gives a scalable solution . so we can implement this type of huge data using hadoop.

For example , a college has large datasets like students records ,staff records, maintainance records etc . If we consider a student record , it has his/her student details, academic reports ,remarks , attendance etc . By analyzing these datasets, we can get the student status reports , behaviour etc like wise we can get each every student reports easily which helps to take proper actions at right time.

BIBLIOGRAPHY

[1] India Human Development Survey-2

http://ihds.info

- IHDS –India Human Development Survey
- The India Human Development Survey(IHDS) is a nationally representative, multi-topicsurveyof41,554householdsin1503 villagesand971urban neighbourhoods across India.
- The firstroundofinterviewswerecompletedin2004-5;data are Publicly available through ICPSR.
- A second round of IHDS re-interviewed mostofthesehouseholdsin2011-12
- (N=42,152)and data for same can be found here.
- IHDS has been jointly organized by researchers from the University of Maryland and the National Council of Applied Economic Research(NCAER), New Delhi.

[2] Open Government Data Platform India(ODG)

https://data.gov.in/

[3] NITI Aayog (National Institution for Transforming India, Government of India)

http://niti.gov.in/

[4] Child Line/1098Night&Day

http://www.childlineindia.org.in/

Appendix A

About programming language R

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

R is free software distributed under a GNU-style copy left, and an official part of the GNU project called **GNU S.**.

The main features of R language are :

R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R:

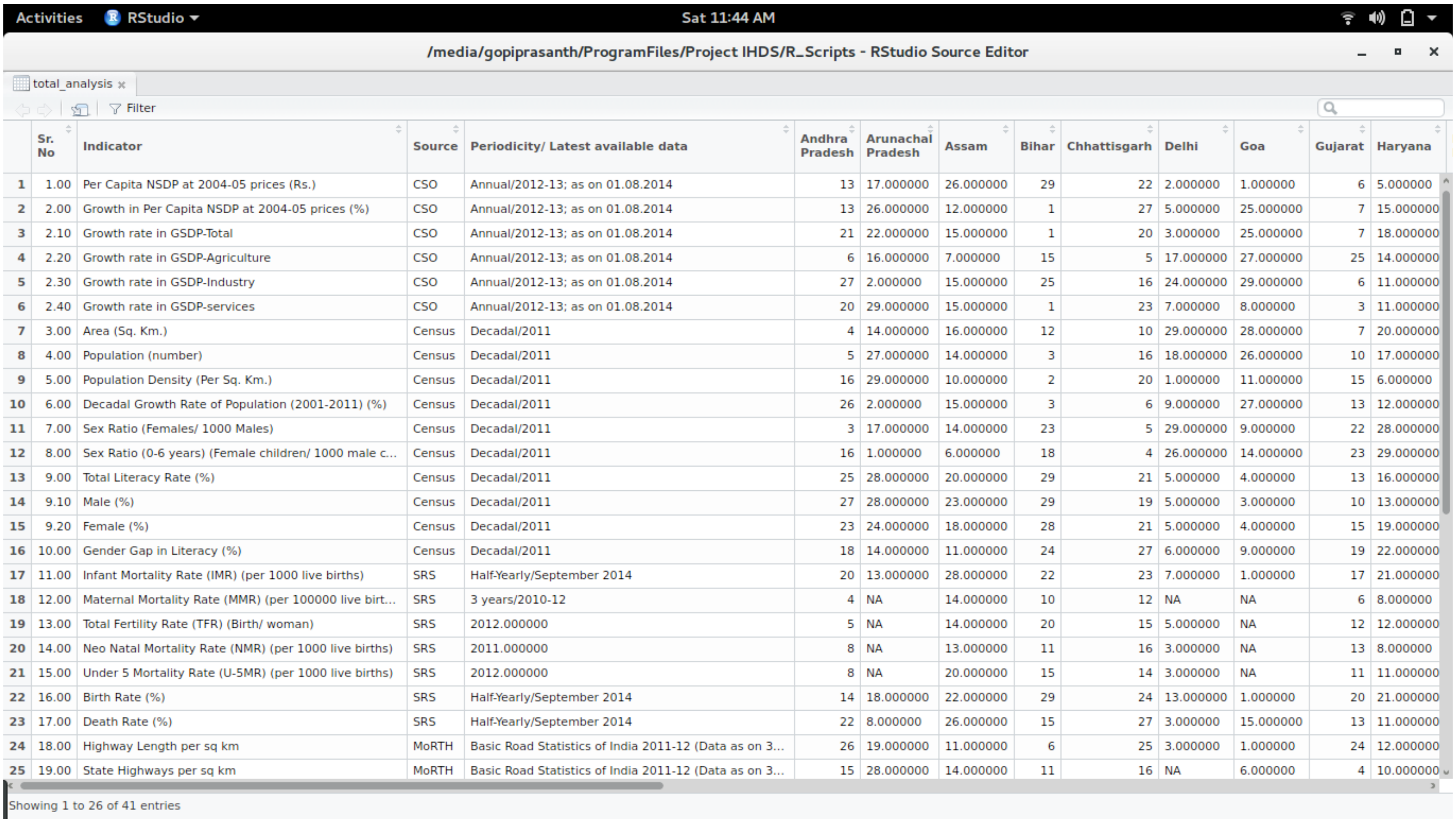
- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility ,

- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

As a conclusion, R is world’s most widely used statistics programming language. It's the # 1 choice of data scientists and supported by a vibrant and talented community of contributors. R is taught in universities and deployed in mission critical business applications.

Appendix B

Source Code

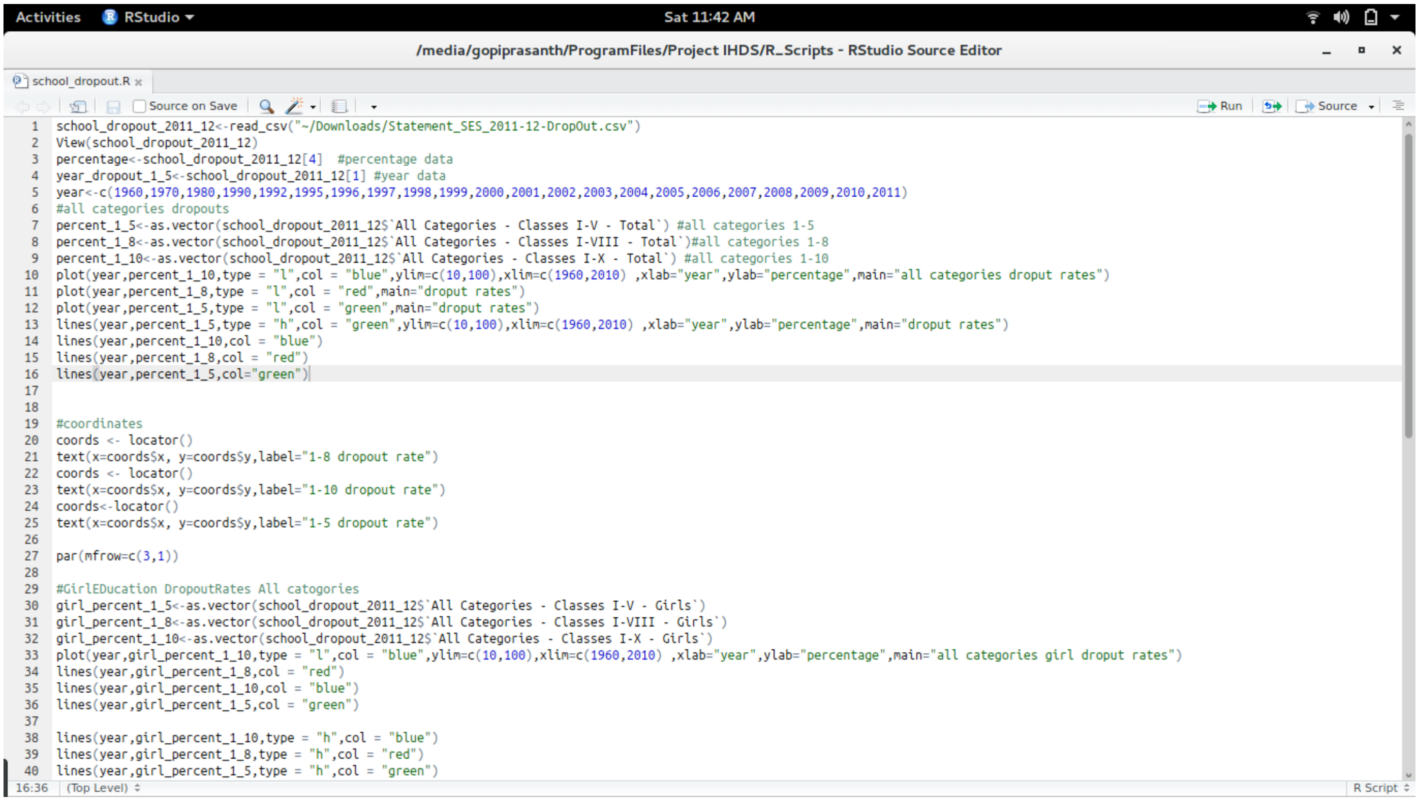


Activities
RStudio
Sat 11:44 AM
/media/gopiprasanth/ProgramFiles/Project IHDS/R_Scripts - RStudio Source Editor
total_analysis
Filter

	Sr. No	Indicator	Source	Periodicity/ Latest available data	Andhra Pradesh	Arunachal Pradesh	Assam	Bihar	Chhattisgarh	Delhi	Goa	Gujarat	Haryana
1	1.00	Per Capita NSDP at 2004-05 prices (Rs.)	CSO	Annual/2012-13; as on 01.08.2014	13	17.000000	26.000000	29	22	2.000000	1.000000	6	5.000000
2	2.00	Growth in Per Capita NSDP at 2004-05 prices (%)	CSO	Annual/2012-13; as on 01.08.2014	13	26.000000	12.000000	1	27	5.000000	25.000000	7	15.000000
3	2.10	Growth rate in GSDP-Total	CSO	Annual/2012-13; as on 01.08.2014	21	22.000000	15.000000	1	20	3.000000	25.000000	7	18.000000
4	2.20	Growth rate in GSDP-Agriculture	CSO	Annual/2012-13; as on 01.08.2014	6	16.000000	7.000000	15	5	17.000000	27.000000	25	14.000000
5	2.30	Growth rate in GSDP-Industry	CSO	Annual/2012-13; as on 01.08.2014	27	2.000000	15.000000	25	16	24.000000	29.000000	6	11.000000
6	2.40	Growth rate in GSDP-services	CSO	Annual/2012-13; as on 01.08.2014	20	29.000000	15.000000	1	23	7.000000	8.000000	3	11.000000
7	3.00	Area (Sq. Km.)	Census	Decadal/2011	4	14.000000	16.000000	12	10	29.000000	28.000000	7	20.000000
8	4.00	Population (number)	Census	Decadal/2011	5	27.000000	14.000000	3	16	18.000000	26.000000	10	17.000000
9	5.00	Population Density (Per Sq. Km.)	Census	Decadal/2011	16	29.000000	10.000000	2	20	1.000000	11.000000	15	6.000000
10	6.00	Decadal Growth Rate of Population (2001-2011) (%)	Census	Decadal/2011	26	2.000000	15.000000	3	6	9.000000	27.000000	13	12.000000
11	7.00	Sex Ratio (Females/ 1000 Males)	Census	Decadal/2011	3	17.000000	14.000000	23	5	29.000000	9.000000	22	28.000000
12	8.00	Sex Ratio (0-6 years) (Female children/ 1000 male c...	Census	Decadal/2011	16	1.000000	6.000000	18	4	26.000000	14.000000	23	29.000000
13	9.00	Total Literacy Rate (%)	Census	Decadal/2011	25	28.000000	20.000000	29	21	5.000000	4.000000	13	16.000000
14	9.10	Male (%)	Census	Decadal/2011	27	28.000000	23.000000	29	19	5.000000	3.000000	10	13.000000
15	9.20	Female (%)	Census	Decadal/2011	23	24.000000	18.000000	28	21	5.000000	4.000000	15	19.000000
16	10.00	Gender Gap in Literacy (%)	Census	Decadal/2011	18	14.000000	11.000000	24	27	6.000000	9.000000	19	22.000000
17	11.00	Infant Mortality Rate (IMR) (per 1000 live births)	SRS	Half-Yearly/September 2014	20	13.000000	28.000000	22	23	7.000000	1.000000	17	21.000000
18	12.00	Maternal Mortality Rate (MMR) (per 100000 live birt...	SRS	3 years/2010-12	4	NA	14.000000	10	12	NA	NA	6	8.000000
19	13.00	Total Fertility Rate (TFR) (Birth/ woman)	SRS	2012.000000	5	NA	14.000000	20	15	5.000000	NA	12	12.000000
20	14.00	Neo Natal Mortality Rate (NMR) (per 1000 live births)	SRS	2011.000000	8	NA	13.000000	11	16	3.000000	NA	13	8.000000
21	15.00	Under 5 Mortality Rate (U-5MR) (per 1000 live births)	SRS	2012.000000	8	NA	20.000000	15	14	3.000000	NA	11	11.000000
22	16.00	Birth Rate (%)	SRS	Half-Yearly/September 2014	14	18.000000	22.000000	29	24	13.000000	1.000000	20	21.000000
23	17.00	Death Rate (%)	SRS	Half-Yearly/September 2014	22	8.000000	26.000000	15	27	3.000000	15.000000	13	11.000000
24	18.00	Highway Length per sq km	MoRTH	Basic Road Statistics of India 2011-12 (Data as on 3...	26	19.000000	11.000000	6	25	3.000000	1.000000	24	12.000000
25	19.00	State Highways per sq km	MoRTH	Basic Road Statistics of India 2011-12 (Data as on 3...	15	28.000000	14.000000	11	16	NA	6.000000	4	10.000000

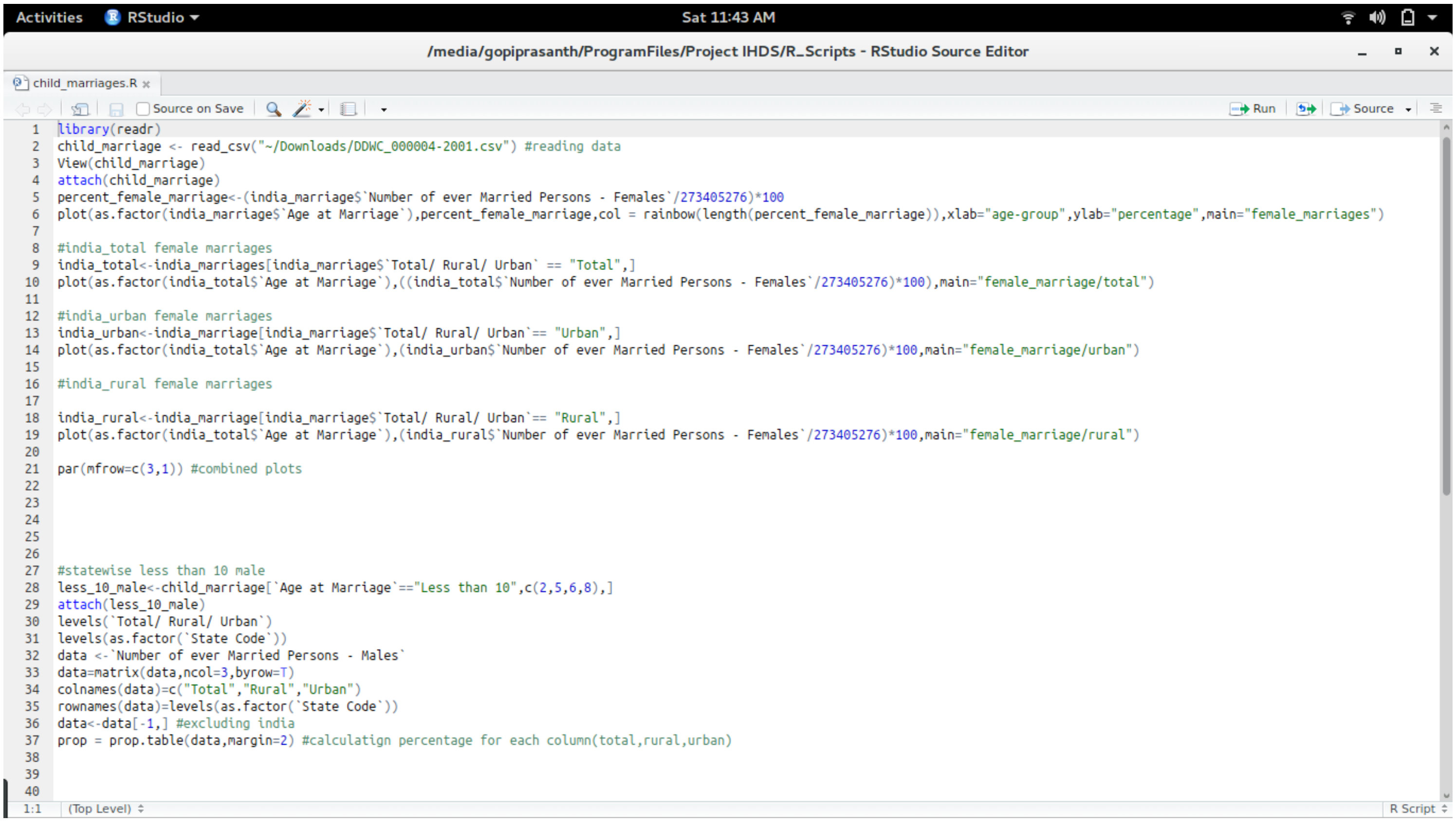
Showing 1 to 26 of 41 entries

Fig (a): Sample data set



```r
school_dropout_2011_12<-read_csv("~/Downloads/Statement_SES_2011-12-DropOut.csv")
View(school_dropout_2011_12)
percentage<-school_dropout_2011_12[4]  #percentage data
year_dropout_1_5<-school_dropout_2011_12[1] #year data
year<-c(1960,1970,1980,1990,1992,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011)
#all categories dropouts
percent_1_5<-as.vector(school_dropout_2011_12$`All Categories - Classes I-V - Total`) #all categories 1-5
percent_1_8<-as.vector(school_dropout_2011_12$`All Categories - Classes I-VIII - Total`)#all categories 1-8
percent_1_10<-as.vector(school_dropout_2011_12$`All Categories - Classes I-X - Total`) #all categories 1-10
plot(year,percent_1_10,type = "l",col = "blue",ylim=c(10,100),xlim=c(1960,2010) ,xlab="year",ylab="percentage",main="all categories droput rates")
plot(year,percent_1_8,type = "l",col = "red",main="droput rates")
plot(year,percent_1_5,type = "l",col = "green",main="droput rates")
lines(year,percent_1_5,type = "h",col = "green",ylim=c(10,100),xlim=c(1960,2010) ,xlab="year",ylab="percentage",main="droput rates")
lines(year,percent_1_10,col = "blue")
lines(year,percent_1_8,col = "red")
lines(year,percent_1_5,col="green")


#coordinates
coords <- locator()
text(x=coords$x, y=coords$y,label="1-8 dropout rate")
coords <- locator()
text(x=coords$x, y=coords$y,label="1-10 dropout rate")
coords<-locator()
text(x=coords$x, y=coords$y,label="1-5 dropout rate")

par(mfrow=c(3,1))

#GirlEDucation DropoutRates All catogories
girl_percent_1_5<-as.vector(school_dropout_2011_12$`All Categories - Classes I-V - Girls`)
girl_percent_1_8<-as.vector(school_dropout_2011_12$`All Categories - Classes I-VIII - Girls`)
girl_percent_1_10<-as.vector(school_dropout_2011_12$`All Categories - Classes I-X - Girls`)
plot(year,girl_percent_1_10,type = "l",col = "blue",ylim=c(10,100),xlim=c(1960,2010) ,xlab="year",ylab="percentage",main="all categories girl droput rates")
lines(year,girl_percent_1_8,col = "red")
lines(year,girl_percent_1_10,col = "blue")
lines(year,girl_percent_1_5,col = "green")

lines(year,girl_percent_1_10,type = "h",col = "blue")
lines(year,girl_percent_1_8,type = "h",col = "red")
lines(year,girl_percent_1_5,type = "h",col = "green")
```

Fig (b) Implementation sample code



```r
library(readr)
child_marriage <- read_csv("~/Downloads/DDWC_000004-2001.csv") #reading data
View(child_marriage)
attach(child_marriage)
percent_female_marriage<-(india_marriage$`Number of ever Married Persons - Females`/273405276)*100
plot(as.factor(india_marriage$`Age at Marriage`),percent_female_marriage,col = rainbow(length(percent_female_marriage)),xlab="age-group",ylab="percentage",main="female_marriages")

#india_total female marriages
india_total<-india_marriages[india_marriage$`Total/ Rural/ Urban` == "Total",]
plot(as.factor(india_total$`Age at Marriage`),((india_total$`Number of ever Married Persons - Females`/273405276)*100),main="female_marriage/total")

#india_urban female marriages
india_urban<-india_marriage[india_marriage$`Total/ Rural/ Urban`== "Urban",]
plot(as.factor(india_total$`Age at Marriage`),(india_urban$`Number of ever Married Persons - Females`/273405276)*100,main="female_marriage/urban")

#india_rural female marriages

india_rural<-india_marriage[india_marriage$`Total/ Rural/ Urban`== "Rural",]
plot(as.factor(india_total$`Age at Marriage`),(india_rural$`Number of ever Married Persons - Females`/273405276)*100,main="female_marriage/rural")

par(mfrow=c(3,1)) #combined plots





#statewise less than 10 male
less_10_male<-child_marriage[`Age at Marriage`=="Less than 10",c(2,5,6,8),]
attach(less_10_male)
levels(`Total/ Rural/ Urban`)
levels(as.factor(`State Code`))
data <-`Number of ever Married Persons - Males`
data=matrix(data,ncol=3,byrow=T)
colnames(data)=c("Total","Rural","Urban")
rownames(data)=levels(as.factor(`State Code`))
data<-data[-1,] #excluding india
prop = prop.table(data,margin=2) #calculatign percentage for each column(total,rural,urban)



```

Fig(c) Implementation sample code