**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

**INCOME BASED DATA ANALYSIS ON US CENSUS**

Submitted by

Krishna Chaitanya S 1MS12CS047

Nirupam K N 1MS12CS068

Sharath R 1MS12CS140

Meghana B S 1MS12CS056

*In partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu/), **May 2015**

# Abstract

The US Census Bureau conducts the American Community Survey generating a massive dataset of with millions of datapoints. The rich dataset contains detailed information of approximately 3.5 million households about who they are and how they live including ancestry, education, work, transportation, internet use and residency. The objective of the project is to reveal unique insights into the lives of the population using Statistical analysis. The main focus of this paper is to concentrate on bringing out unique insights into the expenses and the economic status of the citizens. Among all the attributes, demotic graphs and conclusions will be drawn and classification into various economic categories will be done.

#### Contents

***Declaration i***

***Abstract ii***

***List of Figures***

***List of Tables***

1. **INTRODUCTION 1**
2. **LITERATURE SURVEY 2**
3. **SOFTWARE REQUIREMENT SPECIFICATIONS**

3.1 External Interface Requirements 3

3.1.1 User Interfaces

3.1.2 Hardware Interfaces

3.1.3 Software Interfaces

3.1.4 Communication Interfaces

3.2 Functional Requirements 4

3.2.1 Functional Requirement 1.1

3.2.2 Functional Requirement 1.2

3.2.3 Functional Requirement 1.3

3.2.4 Functional Requirement 1.4

3.2.5 Functional Requirement 1.5

3.2.6 Functional Requirement 1.6

3.3 Software System Attributes 6

3.3.1 Reliability

3.3.2 Availability

3.3.3 Security

3.3.4 Portability

3.3.5 Maintainability

3.3.6 Performance

3.4 Performance Requirements 7

3.5 Database Requirement 8

3.6 Design Constraints 8

3.7 Other Requirements 8

1. **DESIGN**

4.1 Modules Description 9

4.2 Architecture Design 11

4.3 Flow Diagram 12

1. **IMPLEMENTATION**

5.1 Tools and Technology Used 13

5.2 Information about the implementation of Modules with Snapshots 13

5.3 Optimization 21

1. **CONCLUSION & SCOPE FOR FUTURE WORK 22**
2. **REFERENCES**

**1. INTRODUCTION**

“*A large income is the best recipe for happiness I ever heard of”* quotes the famous English novelist Jane Austen. Income is a primary concern that dictates the standard of living and economic status of an individual. Taking into account, its importance and impact on determining a nation’s growth, this paper aims at presenting meaningful insights which can be used to serve as the basis for many wiser decisions by the nation’s administrators. The dataset for the conclusion derived, consists of over 3.5 million data points and is about 4.5GB in size. It consists of many attributes describing the population such as age, income, degree, internet usage, domains of work, expenditure, occupation, fuel cost, transportation method, Internet access details and so on. The idea is to implement statistical analysis and various algorithms such as K-means algorithm, Naïve Bayes Classifier and others to arrive at crucial information of value.

Modules have been created to deal with the huge datasets and to bring out unique insights into the lifestyle of the population. The module involving the classification into economic classes is extended further to predict to which category of economic class people belong, on the basis of factors such as state, number of vehicles owned, employment history and so on. The above prediction is done using Naive- Bayes classifier.

**2. LITERATURE SURVEY**

Increasing social inequalities in Income and Wealth throughout the globe have refocused attention on social class as key determinant. Robert E. Fay and Roger A. Herriot explore an adaptation of the James-Stein estimator applied to sample estimates of income for small places (i.e., population less than 1,000) from the 1970 Census of Population and Housing. The adaptation incorporates linear regression in the context of unequal variances. The new estimates for these small places now form the basis for the Census Bureau's updated estimates of per capita income for the General Revenue Sharing Program.

N. Krieger, D. R. Williams, and N. E. Moss, in their paper, discuss concepts and methodologies concerning, and guidelines for measuring, social class and other aspects of socioeconomic position (e.g. income, poverty, deprivation, wealth, education). These data is collected at the individual, household, and neighborhood level, to characterize both childhood and adult socioeconomic position. Guidelines for linking census-based socioeconomic measures are presented, as are recommendations for analyses involving social class, race/ethnicity, and gender. Suggestions for research on socioeconomic measures are provided, to aid monitoring steps toward social equity in health.

Studies have also shown that poverty and income are powerful predictors of homicide and violent crime, hypothesizing the effect of growing gap between the poor and the rich. This growing gap can be mediated through social cohesion, social capital etc. The problem of income allotment disparity between urban households and rural households has already become a new problem in China.

**3. SOFTWARE REQUIREMENT SPECIFICATION**

**3.1 External Interface Requirements**

* 1. User Interfaces

The user provides the input CSV file for the python based application. The user runs the python program from the terminal. Graphical representations include pie charts, maps and plots.

* 1. Hardware Interfaces

Minimum Requirements: The application will run on a laptop with min 4GB RAM, 64 bit configuration, and runs on Linux Platform.

Recommended Requirements: 8GB RAM/40GB Hard Disk/ I7 processor.

* 1. Software Interfaces

The software is executed by running the python program on the terminal. **Hadoop**: Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware .**Java 1.7**: Java is used to work with Apache Hadoop. **Python**: Python is a widely used general-purpose, high-level programming language and it is used for passing and editing the dataset.  **R:** R is a programming language which has well defined libraries for data mining algorithms. R scripts are executed in RGUI

* 1. Communication Interfaces

Internet connection and a web browser are required in order to make use of several functions and to be executed such as searching, viewing and downloading.

**3.2 Functional Requirements**

3.2.1 Functional Requirement 1.1 - Gender distribution in occupation

|  |  |
| --- | --- |
| Use case name | Gender distribution in occupation |
| Input | 1. Occupation of each person in the database  2. Total number of occupations and name |
| Algorithm used | Map-Reduce algorithm |
| Expected Output | The percentage distribution of men and women in the workforce. |

3.2.2 Functional Requirement 1.2 - Education - Salary relationship

|  |  |
| --- | --- |
| Use case name | Education - Salary Relationship |
| Input | 1. Education degree of each person 2. Field of work 3. Place of stay |
| Algorithm used | Naive Bayes Classifier |
| Expected Output | 1. Average salary a person can expect based on his education level. 2. Classes of income for the entire population |

3.2.3 Functional Requirement 1.3 - English proficiency

|  |  |
| --- | --- |
| Use case name | English proficiency |
| Input | 1. Place of stay 2. Place of birth 3. Nationality 4. Estimated skill of each person |
| Algorithm used | Map-Reduce algorithm |
| Expected Output | 1. Proficiency in English in percentage and a scale of 0-4 2. Bitmap output representing the proficiency |

\

3.2.4 Functional Requirement 1.4 - Classification into economic classes

|  |  |
| --- | --- |
| Use case name | Classification into economic classes |
| Input | 1. Household 2. Supplementary 3. Income 4. Place of Stay |
| Algorithm used | K – means |
| Expected Output | Economic classes hierarchy of people |

3.2.5 Functional Requirements 1.5 – Benford’s Law of Income

|  |  |
| --- | --- |
| Use case name | Benford’s Law |
| Input | Income of each individual |
| Algorithm used | Map Reduce Algorithm |
| Expected Output | The percentage distribution according to the main significant digit of all incomes and comparison with the theoretical law |

3.2.6 Functional Requirement 1.6 – Internet Usage among children

|  |  |
| --- | --- |
| Use case name | Internet Usage among children |
| Input | 1. State they are currently living in. 2. Number of children 3. Do they have Internet access or not. |
| Algorithm used | Statistical Analysis |
| Expected Output | Heatmaps are generated. |

**3.3 Software System Attributes**

3.3.1 Reliability

The software will meet all of the functional requirements without any fail and graphically display the appropriate data.

3.3.2 Availability

The program is available 24x7.Provided we need browser.

3.3.3 Security

Our code do not need Internet, so it is not vulnerable to attacks . All the computations are performed locally on the machine.

3.3.4 Portability

This software is designed to run on any Operating system. Provided it has Python and Hadoop for processing data.

3.3.5 Maintainability

The code should be written clearly and well documented. It is user friendly and easy to use.

3.3.6 Performance

It depends on computer hardware specifications like RAM, GPU etc. Since we are handling huge data, it should be parallel processed.

**3.4 Performance Requirements**

In this project, we deal with a database with over 3.5 million data points. Hence it is crucial for us to give importance to Time complexity. Computer should not be shutdown in between and Hadoop should not be stopped. The given data should be normalized and should contain all the data correctly.

We try to focus on minimizing the time taken for computations in different cases:

a. Given datasets.

b. Required attributes with null tuples.

c. Required attributed without null tuples.

d. After code optimization.

**3.5 Database Requirement**

Since this project uses over a rich dataset that contains detailed information of approximately 3.5 million households about who they are and how they live including ancestry, education, work, transportation, Internet use, and residency etc. i.e., over 3.5 million datapoints with over 232 attributes and is about 4.5 GB in size. We do not need any conventional database to store and process the data. We are using Hadoop Distributed File System for storing and processing of data.

**3.6 Design Constraints**

Data taken is only for one year.

Hadoop is slow compared to Apache Spark.

Hadoop is heavy resource consumer. Minimum RAM required = 4 GB

Minimum hard disk required = 20 GB. Minimum process required = i3

**3.7 Other Requirements**

We should have knowledge about statistical analysis and how to implement it in graphs.

**4. DESIGN**

**4.1 Modules Description**

This project deals with implementation of 6 modules.

1. Gender distribution in Occupation

In this module, we focus primarily on bringing out the gender distribution in occupations, i.e., the percentage distributions of men and women in the workforce.

2. Education – Salary relationship using Naïve Bayes Classifier

This module uses Naïve Bayes Classifier to predict the average salary a person can expect based on his education level. Here we predict a person’s income based on his education degree, field of work, place of stay and other parameters where the total income of the entire application will be segregated into a few classes of income.

3. English proficiency with Bitmap

In this module, we focus on accessing the knowledge of the participants in the English language and represent same using Bitmap image and explore the spoken English language proficiency among the people who participated in the survey based on place of stay, place of birth, nationality etc. On a scale of 0-4, with 0 being the lowest and 4 being the highest.

4. Hierarchy of Economic classes using K-means and income prediction using Naïve Bayes Classifier

This module uses K-means to classify the people who participated in the survey into various economic classes based on various parameters such as household, supplementary, total, other income etc. Based on the number of clusters given as input, categorization of the Economic class hierarchy is done. Using this as a basis, we use Naïve Bayes classifier to classify the people among various classes of Income (taken from the output of K-means) and are checked with their original economic class to which they belong. The below output table shows the efficiency. The elements along the principal diagonal are rightly predicted and the rest are the cases of misprediction.

5. Benford’s Law on Income

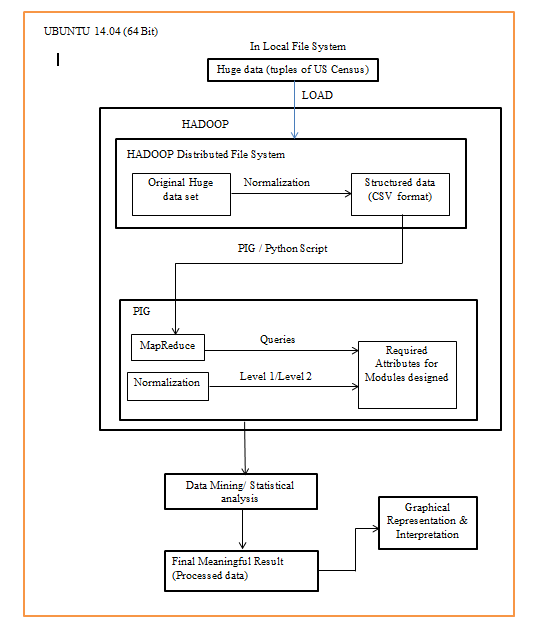
Benford’s Law, also called the First-Digit Law, and is a phenomenological law about the frequency distribution of leading digits in many real-life sets of numeric data. The law states that in many naturally occurring collections of numbers the small digits occur disproportionately often as leading significant digits. Here, Benford’s law has been used to obtain a distribution of US income.

6. Percentage of Children without Internet Access

Internet which has added to one of the needs of life might not help the survival of life directly. Internet keeps one connected with the rest of the world. It is really essential to know about the current activities happening throughout the globe. With the advent of technology, the gathering of information has been possible through Internet just by sitting within the four walls of the room. Children of the present world are exposed to this technical privilege. Right from the young age, they can feed their curiosity with right thoughts through Internet. Internet usage might also have several complications associated with it, such as undesired maturity at a small age, hindrance to physical and social development and so on. However with a proper parental or guardian guidance these hurdles can be resolved. In the light of these scenarios, a module is presented depicting the usage of Internet by children across various parts of the country. The bitmap images are used to show the degree of usage across all states with a color scale. This module might not directly fit into Income domain, but might indirectly hint the economic conditions across the country.

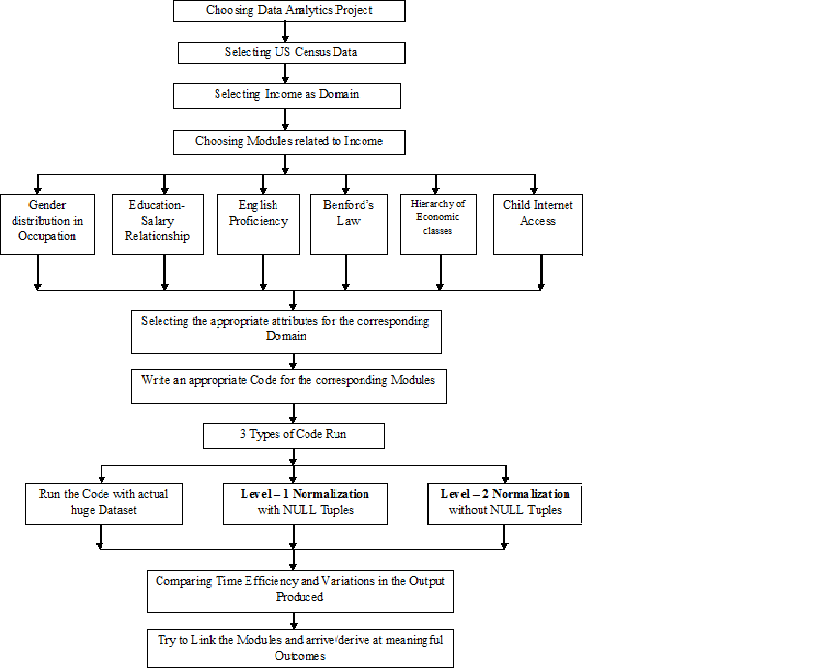
**4.2 Architecture Design**

The architectural design is the design of the entire software system; it gives a high-level overview of the software system, it provides information on the decomposition of the system into modules (classes), dependencies between modules, hierarchy and partitioning of the software modules.



**Figure 1 – System Architecture.**

**4.3 Flow Diagram**



**Figure 2 – Flow Diagram**

**5. IMPLEMENTATION**

**5.1** **Tools and Technologies Used**

The software is executed by running the python program on the terminal.

**Hadoop**: Apache Hadoop is an open-source software framework written in Java for distributed storage and processing of very large data sets on the computer clusters built from the commodity hardware.

**Java 1.7**: Java is used to work with Apache Hadoop.

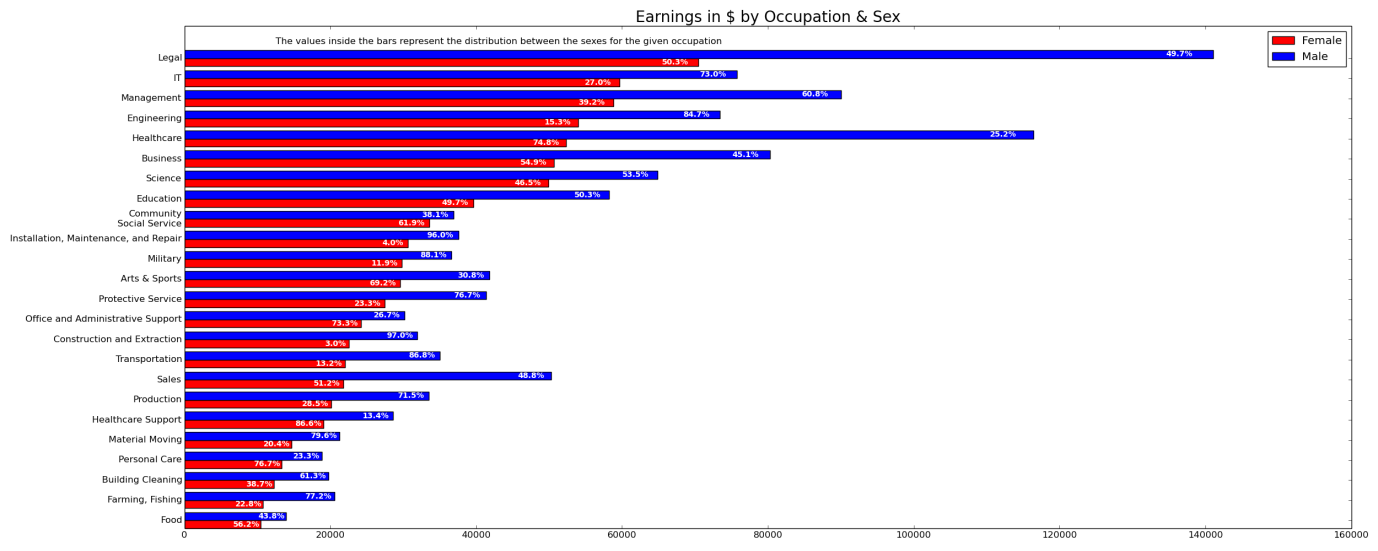
**Python**: Python is a widely used general-purpose, high level programming language and it is used for passing and editing the dataset.

**R**: R is a programming language which has well defined libraries for data mining algorithms. R scripts are executed in RGUI.

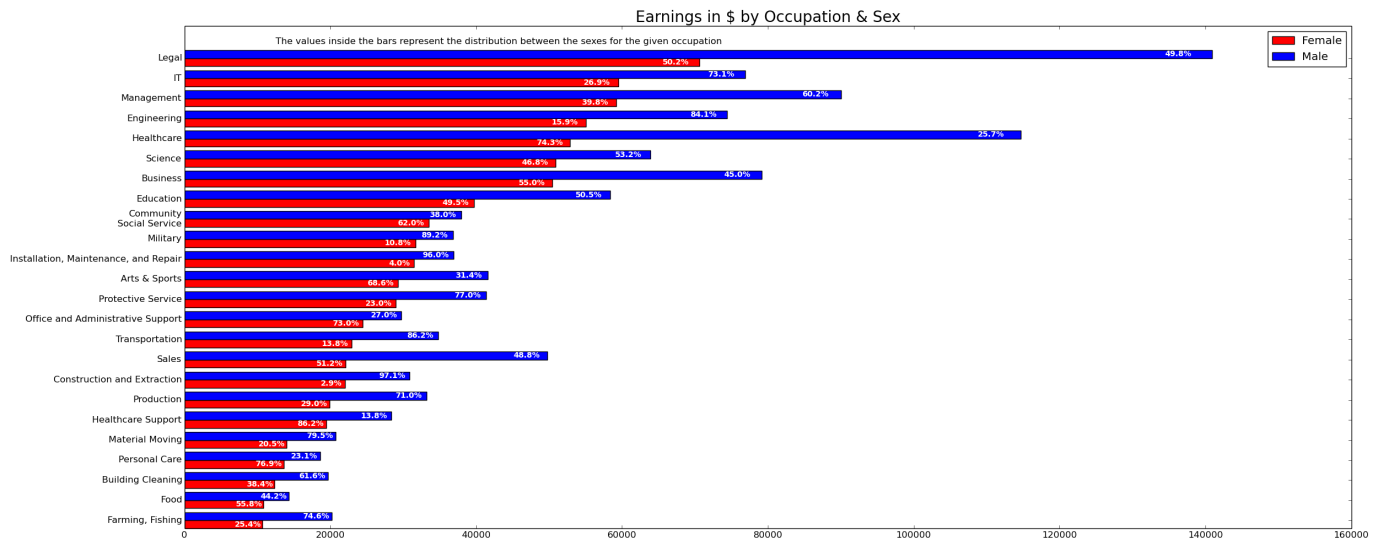
**PIG**: Pig is a data flow platform for writing Hadoop operations in a language called Pig Latin. Apache is a high-level procedural language platform developed to simplify querying large data sets in Apache Hadoop and MapReduce. It provides SQL like interface to process data on Hadoop and thus helps the programmer focus on business logic and help increase productivity. It supports a variety of data types and the use of user-defined functions (UDFs) to write custom operations in Java, Python and JavaScript.

**5.2** **Information about the implementation of Modules with Snapshots**

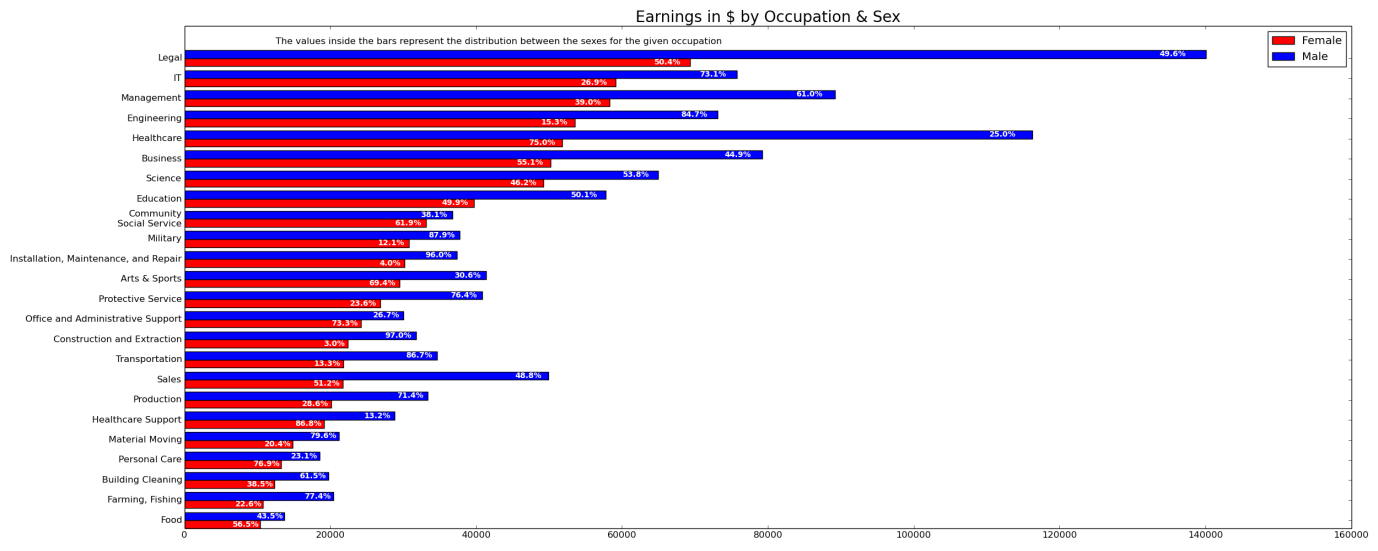
1. Gender distribution in Occupation
   1. Choose the required attributes for the module such as PUMA (Public User Micro data Area code), ST (State Code), PERNP (Total Person’s Earning), OCCP (Occupation), SEX (1 – Male, 2 – Female).
   2. Choose the different occupations like IT, Legal, Education, Farming etc.
   3. Calculation:
      1. Count the number of males and females in each occupation and their earning in that occupation.
      2. Calculate the percentage of men and women working in the occupation
   4. Plot the Graph.



**Figure 3 – Gender distribution in Occupation on Given Dataset**

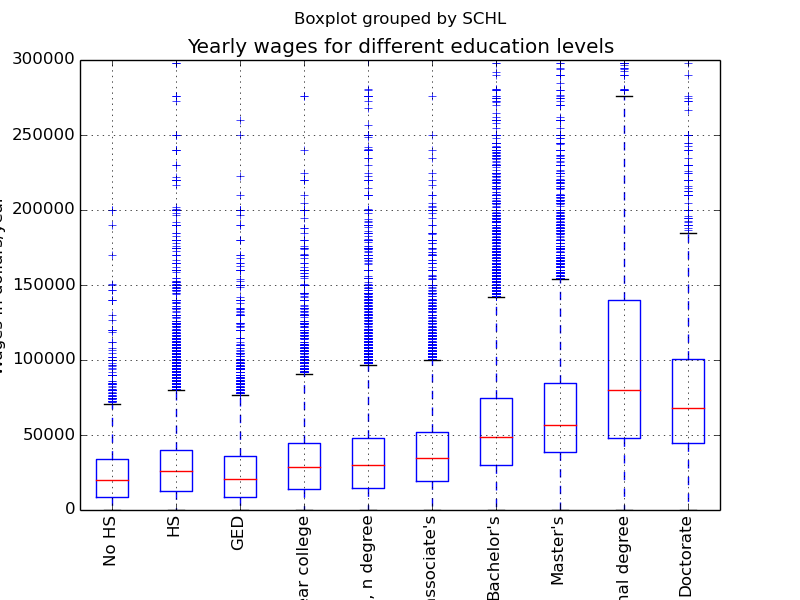


**Figure 4 – Gender distribution in Occupation on Level 1 Normalization**

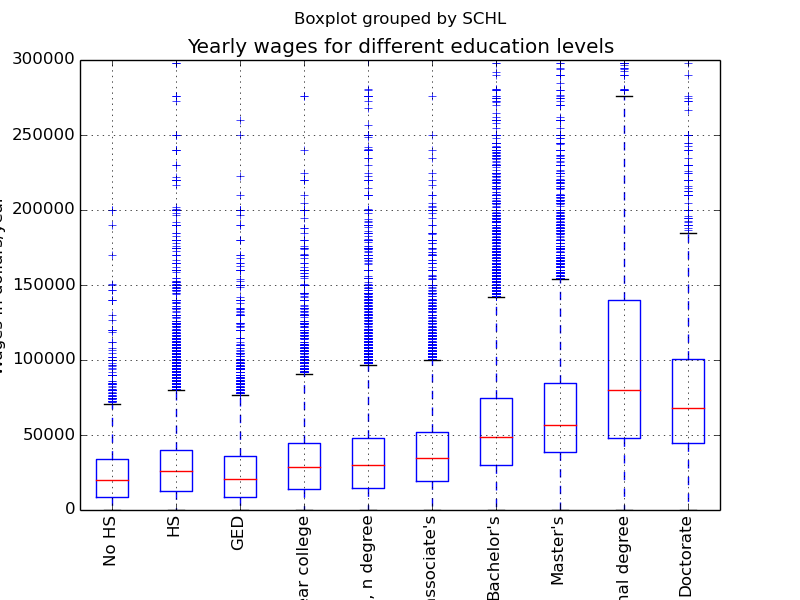


**Figure 5 – Gender distribution in Occupation on Level 2 Normalization**

1. Education – Salary relationship using Naïve Bayes Classifier
   1. Choose the required attributes for the module such as SCHL (Educational Attainment), WKL (When Last worked), AGEP (Age) and WAGP (Income from past 12 months).
   2. Choose different levels of education like No schooling, Schooling, Bachelor’s degree, Master’s degree etc.
   3. Condition: Schooling with greater than grade 11 & When Last Worked within the past 12 months & Age between 30 to 40.
   4. Calculation:
      1. Find minimum salary for each level of schooling.
      2. Find maximum salary for each level of schooling.
      3. Sum of all the salary in each level of school and compute its average.
   5. Plot the Graph.

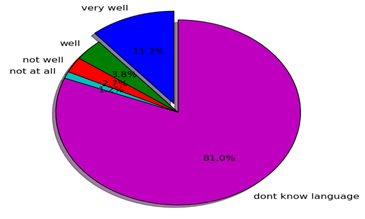


**Figure 6 – Education Salary relationship for given data**

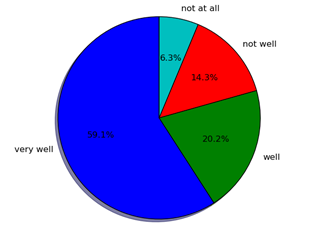


**Figure 7 – Education Salary relationship for data without NULL tuples**

1. English Proficiency
   1. Choose the required attributes for the module.
   2. The various classes of English Proficiencies such as not well, very well, well etc are captured.
   3. Count the number of people falling under each category.
   4. Find the percentage of each category and depict it on a pie chart.

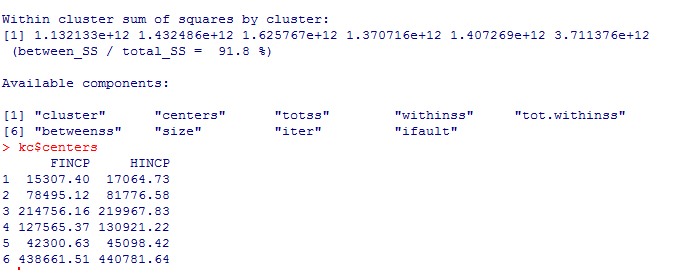


**Figure 8 – English Proficiency level for given data**

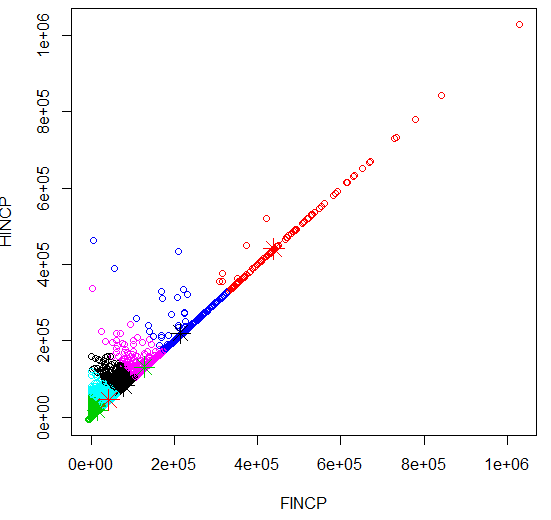


**Figure 9 – English Proficiency level for data without NULL tuples**

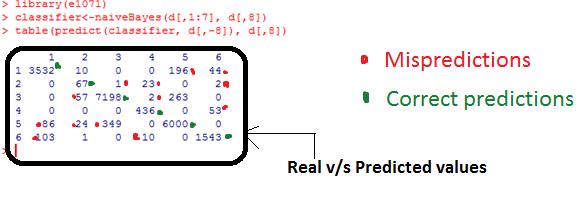
1. Hierarchy of Economic classes using K-means and income prediction using Naïve Bayes Classifier
   1. Choose the required attributes for the module such as FINCP, HINCP, VEH, RNTP, ST, FES, WKEX.
   2. Apply K – means algorithm by specifying the number of clusters required.
   3. Plot the graph with HINCP as y-axis and FINCP as x-axis to determine the clusters.
   4. The output of K – means i.e., cluster centers are given as input to Naïve Bayes Classifier for predictions and mis-predictions.



**Figure 10 – Clusters Center**

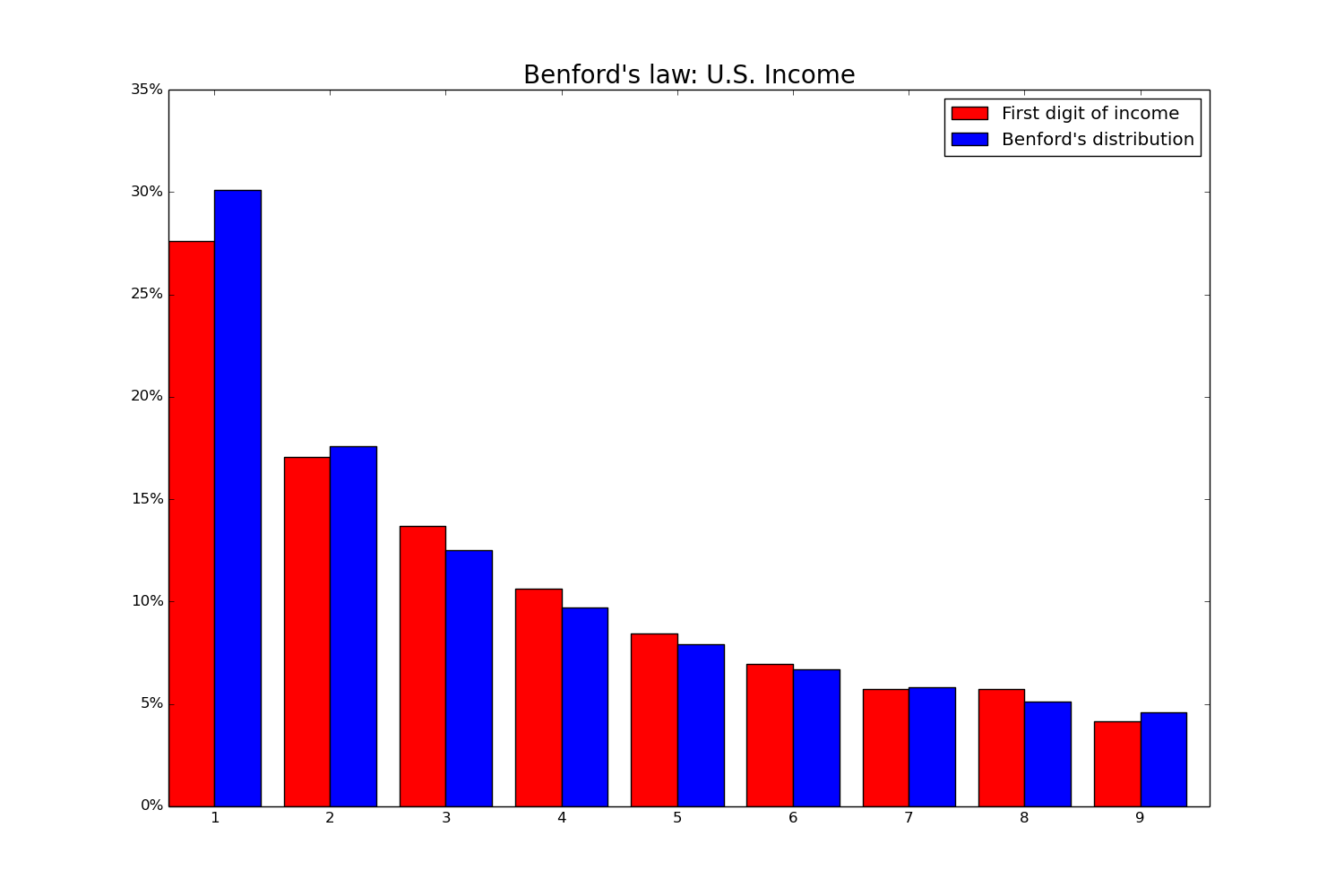


**Figure 11 - Graph depicting the clusters with different colors.**

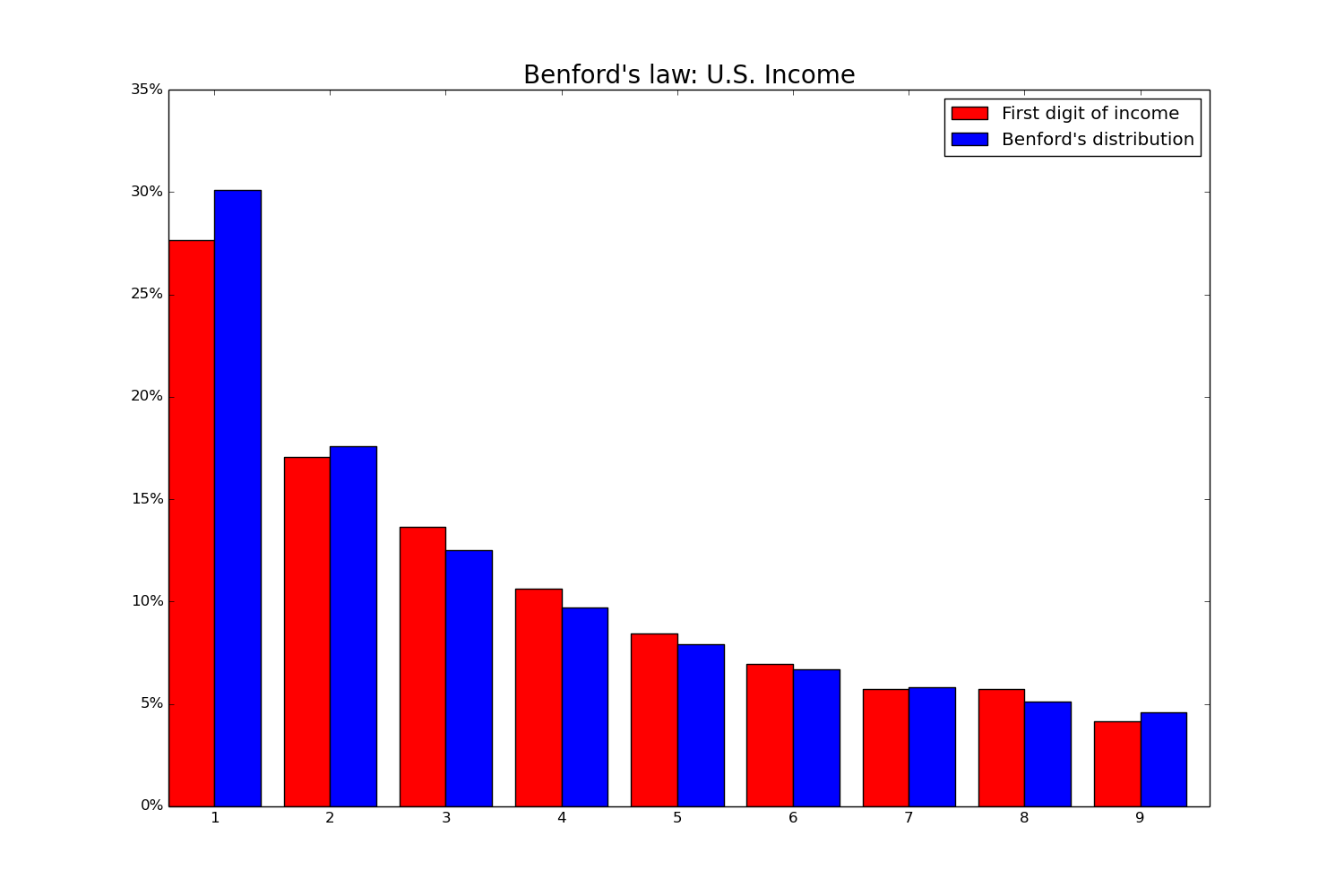


**Figure 12 – Output of Classifier with Correct predictions and Mispredictions.**

1. Benford’s Law
   1. Choose the required attributes for their module as in PUMA (Public Use Micro data Area code), ST (State Code), and PINCP (Total Person’s Income).
   2. Calculate the occurrence of most significant bit in the given Income.
   3. Run step (b) for 2 loops – Theoretical and Actual.
   4. Compare the number of occurrences of MSB in theoretical vs. Actual Income.

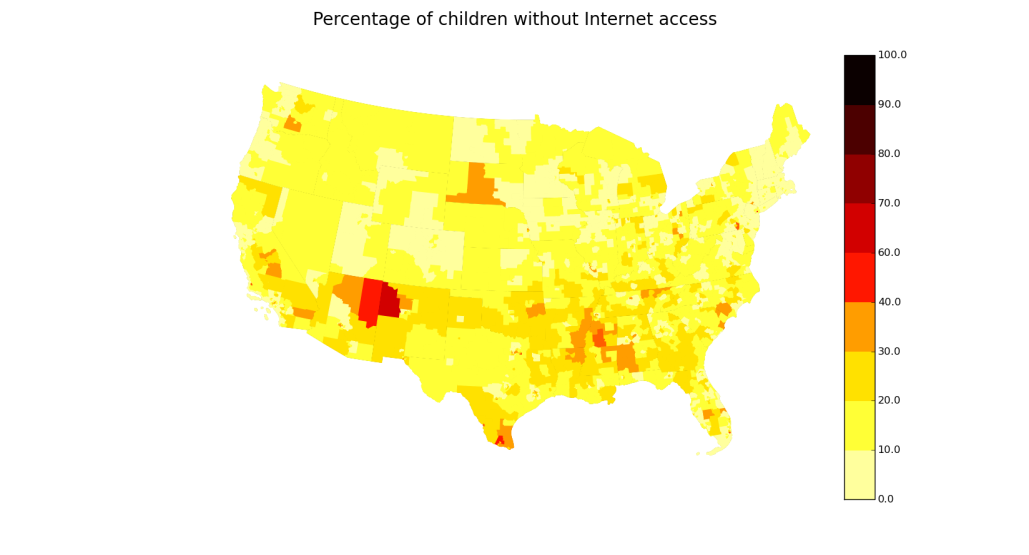


**Figure 13 – Benford’s Law applied to given data**

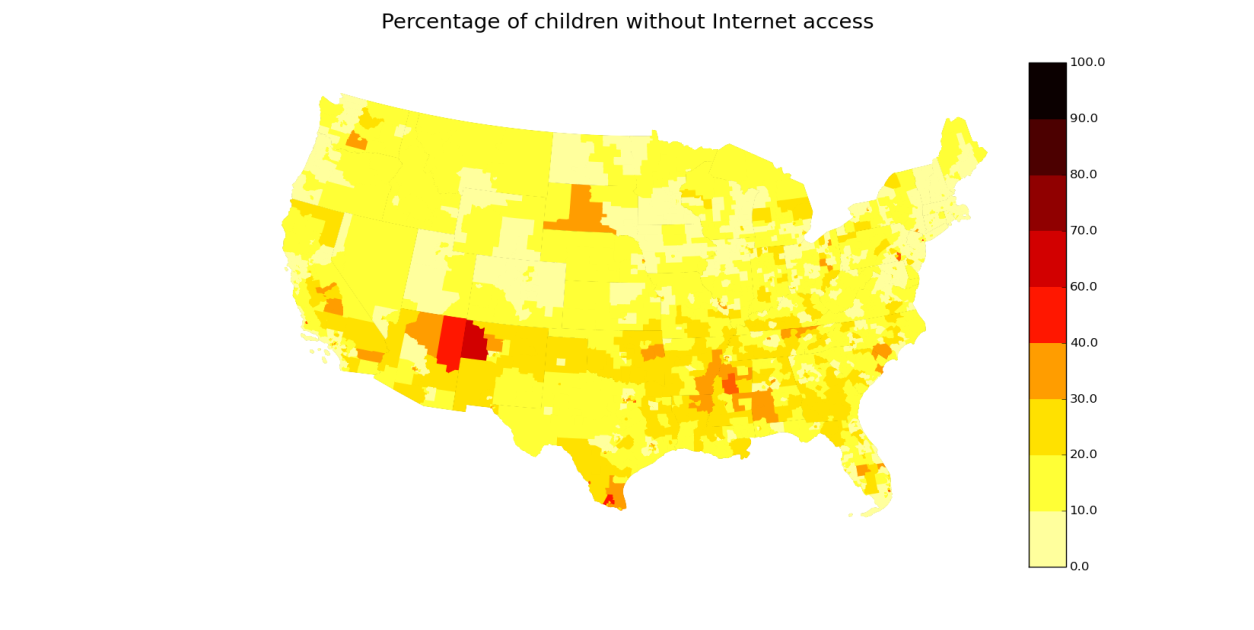


**Figure 14 – Benford’s Law applied to normalized data**

1. Percentage of Children without Internet Access
   1. Choose the required attributes for the module (ST, PUMA, NRC and ACCESS).
   2. Group them by ST, PUMA and ACCESS and find the total number of children with Internet access for a corresponding state.
   3. Get the number of children without Internet access by subtracting result obtained in (b) from total children in a state.
   4. Plot the Heatmap accordingly using shapefiles given.



**Figure 15 – Heatmap for Children without Internet access (Given data)**

**Figure 16 – HeatMap for children without Internet access (normalized data)**

**5.3 Optimization**

|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **Given Dataset**  **(in s)** | **Level – 1 Normalization (with NULL tuples)**  **(in s)** | **Level – 2 Normalization (without NULL tuples)**  **(in s)** |
| Gender distribution in Occupation | 60.023703 | 28.935224 | 17.782561 |
| Benford’s Law | 63.656673 | 13.593124 | 13.455113 |

**Table 1 – Time taken to execute the module for given and normalized dataset**

The modules are executed with different inputs (i.e. Given Dataset, Data after normalization with NULL tuples, Data after Normalization without NULL tuples), and the time efficiency in each case is noted down and is depicted in the above table. The time is measured in terms of seconds.

Without normalization, the percentage of women in farming and fishing sector was found to be 77.4%. After normalization, the result for the same sector was found to be 74.6%, showing a decrease of a 3.6%. The result of this analysis is depicted in Figure 17 – without normalization and Figure 18 – after normalization.

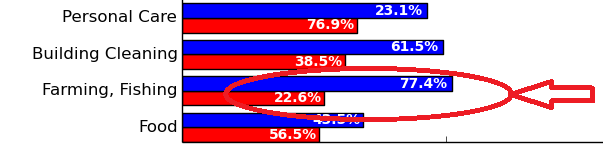


Figure 17 - Gender distribution in occupation without normalization

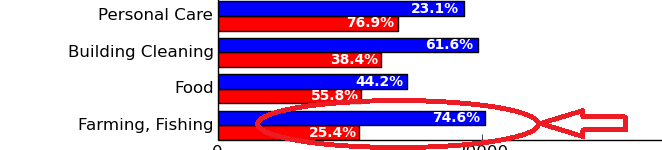


Figure 18 - Gender distribution in occupation after normalization

**6. CONCLUSION AND SCOPE FOR FUTURE WORK**

The income is a pivotal factor which a country needs to focus on. The meaningful insights provided by the above modules produce considerable number of factors which gives the picture of the economic clout. Based on the economic rungs, the major decisions on the parliamentary budget and others can be banked upon. In the light of bringing higher efficiencies, improved algorithms for prediction and classification can be focused. The deeper delineation in case of graphs or histograms might provide much scope for deriving insights.

A wealth of insights into the lives of the population represented by the US Census data has been derived using statistical analysis and a variety of algorithms. These insights can be used to influence future economic reforms in the country, and can also be used to affect changes in the work sector and at various education levels. Further work can be done to build upon the insights that have been uncovered in this paper.

**7. REFERENCES**

1. <http://rayli.net/blog/data/top-10-data-mining-algorithms-in-plain-emglish/> - Web URL on Data Mining Algorithms.
2. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html> - Web URL on K-means Algorithm.
3. <https://plot.ly/python/> - Web URL on plotly package Implementation.
4. [www.annualreviews.org/doi/abs/10.1146/annurev.publhealth.18.1.341](http://www.annualreviews.org/doi/abs/10.1146/annurev.publhealth.18.1.341) - Measuring Social Class in US Public Health Research: Concepts, Methodologies, and Guidelines by N. Krieger, D. R. Williams and N. E. Moss.
5. [www.sciencedirect.com/science/article/pii/S0277953698000975](http://www.sciencedirect.com/science/article/pii/S0277953698000975) - Social capital, income inequality and firearm violent crime by Bruce P. Kennedy, Ichiro Kawachi, Deborah Prothrow-Stith, Kimberly Lochner, Vanita Gupta.
6. [www.tandfonline.com/doi/abs/10.1080/01621459.1979.10482505](http://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10482505) - Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.
7. https://en.wikipedia.org/wiki/Apache\_Hadoop: Hadoop Wikipedia.
8. Inequality in income and mortality in the United States – Paper published on Analysis of mortality and potential pathways.
9. <https://en.wikipedia.org/wiki/Python_(programming_language)> - Python Wikipedia.
10. <https://cwiki.apache.org/confluence/display/PIG/PigTutorial> - Pig tutorials.
11. <http://www.orzota.com/pig-tutorialfor-beginners/> - Detailed Pig definition and description.
12. <http://www.r-statistics.com/2013/08/k-means-clustering-from-r-in-action> - webpage on k-means clustering using R.
13. <http://www.rdatamining.com/examples/association-rules> - webpage on Naïve Bayes Classifier using R.
14. [https://en.wikipedia.org/wiki/Benford%27s\_law](https://en.wikipedia.org/wiki/Benford's_law) - Benford’s law Wikipedia page.
15. Arno Berger and Theodore P Hill, Benford’s Law Strikes Back – No Simple Explanation in Sight for Mathematical Gem, 2011 – Benford’s law explanation.