ISSN:0975-9646

Dashboard using Data Analytics and Statistical Modeling

Sunad M[1], Aditya Naik[2], Mrinal Panda[4], Swarnim Suman[4],Radhika K R[5]

Computer Science and Engineering Department, BMS Institute of Technology
Bangalore, Karnataka

***Abstract*— Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the science to verify or disprove existing models or theories. There is large amount of data which are either structured (RDBMS) or non-structured (Multimedia content). Using the large amount of unprocessed data available, various analytics are done, and displayed on an interactive and effective dashboard which provides key insights into the domain scenario of the company, aka, iShippo.**

***Keywords*— Put your keywords here, keywords are separated by comma.**

I. INTRODUCTION

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the science to verify or disprove existing models or theories. Data analytics is distinguished from data mining by the scope, purpose and focus of the analysis. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.

Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information". Dashboards often provide at-a-glance views of KPIs (key performance indicators) relevant to a particular objective or business process.

India is a home to a large number of handicraft and small scale industries. A major part of rural sector development depends on the upliftment of these industries.[1]

Technology plays a vital role in the improvement of this sector and iShippo provides this with an interactive platform for buying and selling of goods. The analytics done in this paper will provide key insights for the better functioning of the company.

II. SCOPE AND MOTIVATION

The Table 1.1 gives the distribution of handicrafts units for household and non-household items. Despite this according to the national commission for enterprises in the unorganized sector (NCEUS) states that 77 percent of total population live on an income less than 20 Rs per day. The irony is that 50 percent of the huge number constitutes 92 percent of the rural artisans, which implies that 422.7 million people work and are supporting dependents with income less than 20 Rs per day without job or social security (2007) thus the market is primed for a logistics provider which targets the specific demographic group so as to promote handicraft industry in India. [2]

TABLE I
Distribution of Handicraft Units and Artisans by Sector(Percent)

Handicraft Sector	# of Units	# of Artisans
Household Sector	97.96	96.27
Non-Household Sector	2.04	3.73
All India Total	100.0 (1455056)	100 (4761186)
Note: Figures in Parentheses are Absolute Numbers		

III. PROJECT OVERVIEW

An Overview of the project is given by the overall process, modules of the system and the architecture of the system.

A. The Overall Process of the System.

Data is collected from necessary sources based on the need for analysis.

The given data is cleaned of all NA and redundant values to give a clean and usable data.[3]

Data Transformation is carried out.[3,4]

The Pincode information is converted into Latitude/Longitude pairs using a Python script.

We create Heat maps with the Lat/Long data, using R and Leaflet.[3,4,11]

Key analysis is made based on the relationships established.

An RShiny dashboard is designed as the front-end for displaying the analysis done.

B. Modules of the system

Data Cleaning Module: R script which is used to clean all the relevant data(Buyers, Sellers and the logistic partners‘ information). After cleaning, each cleaned dataset is stored in a separate file.[4]

Login module: A Security feature on the bootstrap platform which allows only authorized users to login and view the sensitive information on the Dashboard.

Analytics module: Every Analysis made on each of the datasets which are to be displayed onto the dashboard are segregated onto their own individual R scripts.[3,4]

RShiny Module: The dashboard created will call upon each of the relevant R scripts to get the appropriate work done on a real time basis.

The Python Module: This module converts the Pincode information to the necessary Lat/Long Pairs.

The Leaflet Module: With the help of the front end, and the relevant Data Cleaning Module and Python Module, the necessary heat maps are generated and displayed onto the dashboard.[11]

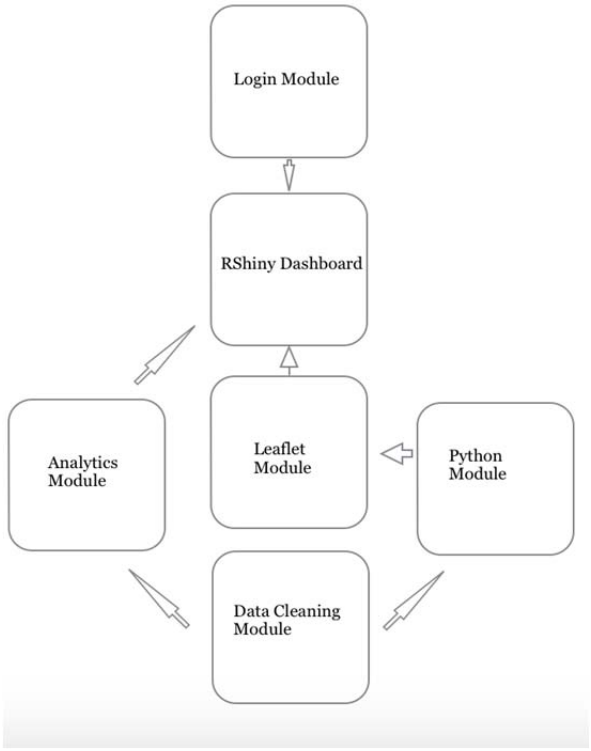

Fig. 1 Modules of the System

C. Architecture of the System

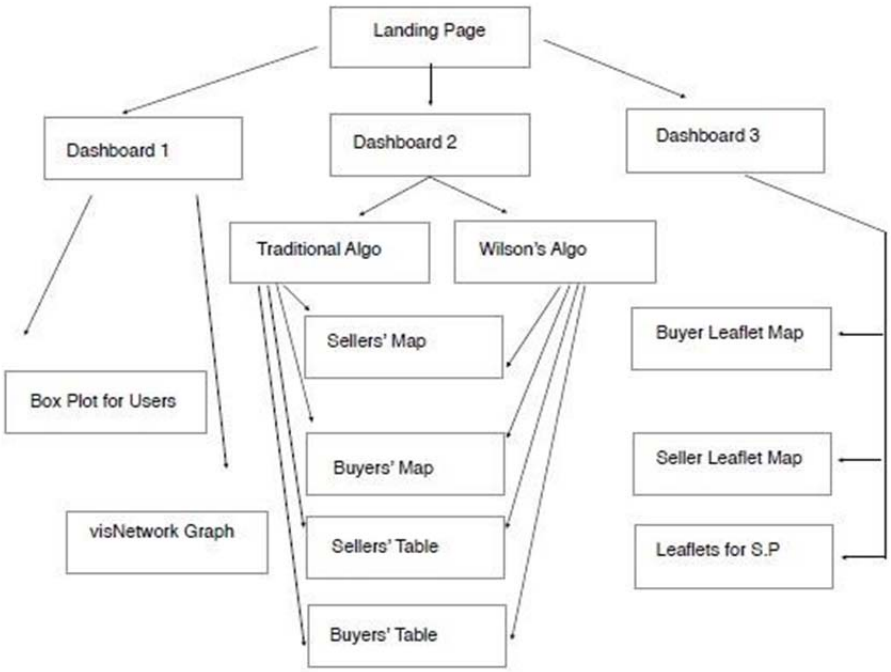

Fig. 2 Architecture of the Analytics' Dashboard

D. Use Case Diagram

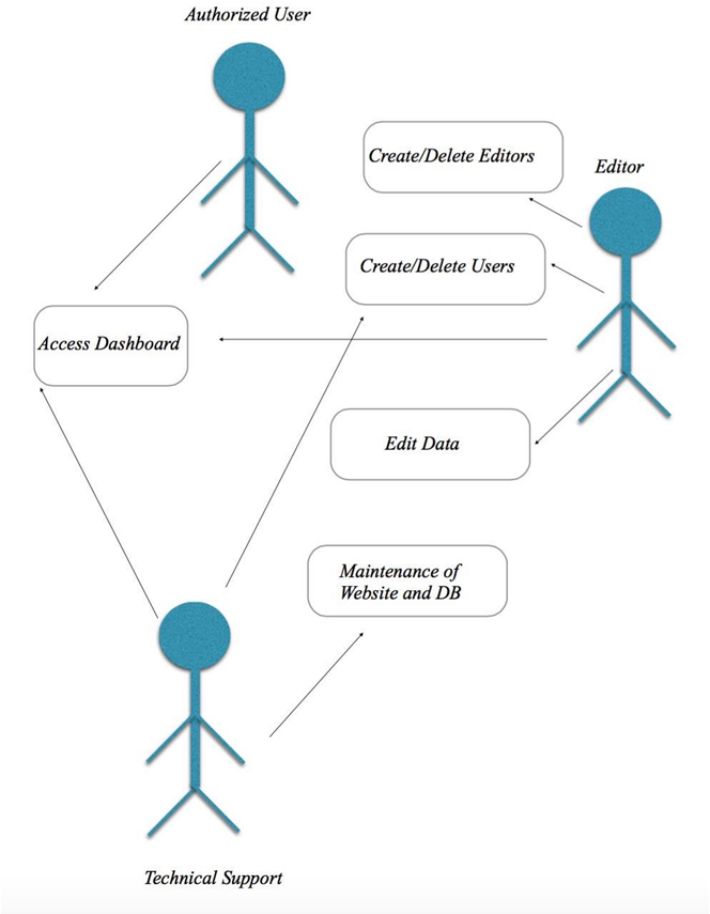

Fig. 3 Use Case Diagram of the System

IV. Wilson Score Confidence Method

In statistics, a binomial proportion confidence interval is a confidence interval for a proportion in a statistical population. It uses the proportion estimated in a statistical sample and allows for sampling error. There are several formulas for a binomial confidence interval, but all of them rely on the assumption of a binomial distribution. In general, a binomial distribution applies when an experiment is repeated a fixed number of times, each trial of the experiment has two possible outcomes (labeled arbitrarily success and failure), the probability of success is the same for each trial, and the trials are statistically independent.

A simple example of a binomial distribution is the set of various possible outcomes, and their probabilities, for the number of heads observed when a (not necessarily fair) coin is flipped ten times. The observed binomial proportion is the fraction of the flips which turn out to be heads. Given this observed proportion, the confidence interval for the true proportion innate in that coin is a range of possible proportions which may contain the true proportion. A 95% confidence interval for the proportion, for instance, will contain the true proportion 95% of the times that the procedure for constructing the confidence interval is employed. Note that this does not mean that a calculated 95% confidence interval will contain the true proportion with 95% probability.[10]

The need to balance the proportion of positive ratings with the uncertainty of a small number of observations is done using Wilson Score Confidence Method.

Here, Score = Lower bound of Wilson score confidence interval for a Bernoulli parameter.

$$\left(\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n]/n}\right)/(1 + z^2_{\alpha/2}/n)$$

Here $\hat{p}$ is the *observed* fraction of positive ratings, $z_{\alpha/2}$ is the (1-α/2) quantile of the standard normal distribution, and n is the total number of ratings.[6]

We use this, to determine the proportion of female sellers state-wise across India, by taking into account the number of observations, i.e., number of Sellers' or Buyers' This is represented graphically as shown in the figure.[7]

Observe fig. 4 and fig. 5, which gives the distribution of female sellers' on the iShippo platform, using the traditional algorithm(number of female sellers divided by the total number of sellers) and the Wilson Score Confidence Method respectively.

We can clearly observe that Karnataka has the highest proportion of Female Sellers' in India by using the Wilson Score Confidence method.

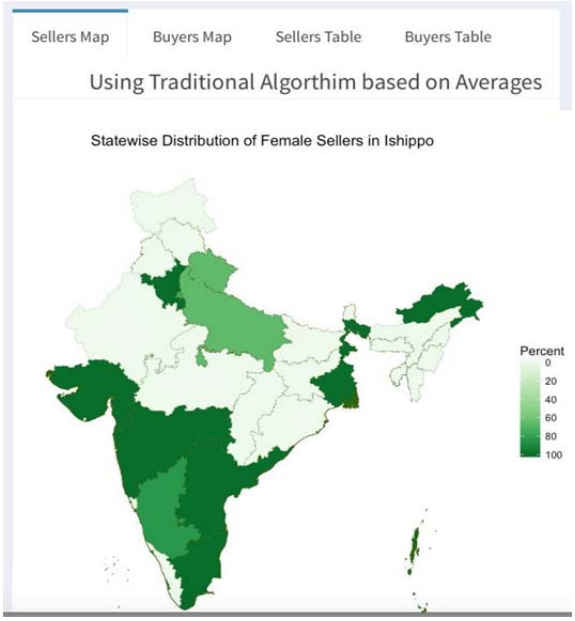

Fig. 4 Traditional Algorithm Implementation

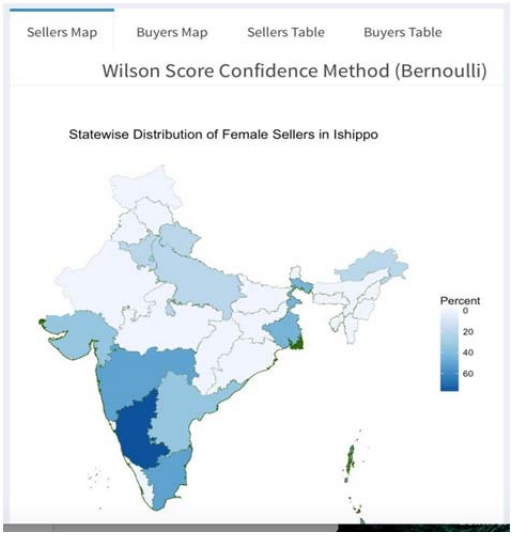

Fig. 5 Wilson Score Confidence Implementation

V. Box Plots

In descriptive statistics, a box plot or boxplot is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box and- whisker diagram. Outliers may be plotted as individual points. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacing‘s between he different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers.[5,8]

The top and bottom lines of the rectangle are the 3rd and 1st quartiles (Q3 and Q1), respectively. The length of the rectangle from top to bottom is the interquartile range (IQR).[8,9]

The line in the middle of the rectangle is the median (or the 2nd quartile, Q2). [8,9]

The top whisker denotes the maximum value or the 3rd quartile plus 1.5 times the interquartile range (Q3 + 1.5*IQR), whichever is smaller.[8,9]

The bottom whisker denotes either the minimum value or the 1st quartile minus 1.5 times the interquartile range (Q1 – 1.5*IQR), whichever is larger. A nice addition to add to box plots is notches. According to Chambers et al. (Page 62, 1983), the 2 medians are significantly different with 95% confidence if the notches of 2 box plots do not overlap.

From the Boxplots, we can derive that:
For Sellers’: Median Age: Male: 38, Female: 33
First Quartile: Male: 31 Female: 25
Third Quartile: Male: 46 Female: 42
Outliers: None
For Buyers‘:
Median Age: Male: 37, Female: 40
First Quartile: Male: 35, Female: 36
Third Quartile: Male: 42, Female: 45
Outliers: Female: 79

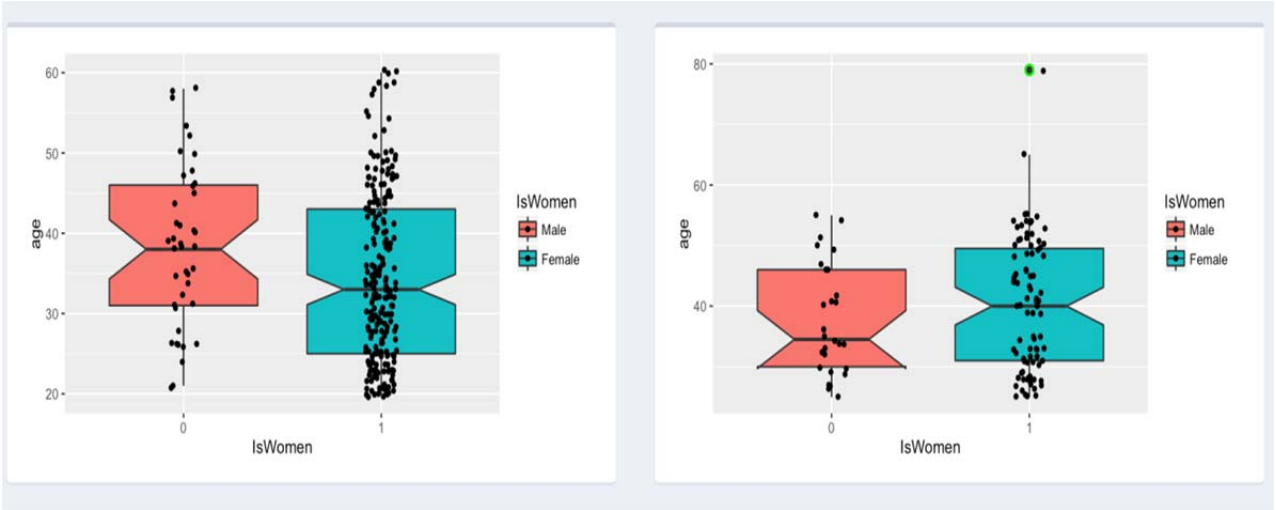

Fig. 5 Box plots depicting age vs. sex for buyers‘ and sellers‘ respectively

VI. VISNETWORK GRAPH

visNetwork is an R package for network visualization, using vis.js javascript library. The graph shows the correlation between the various Buyers', Sellers' and their corresponding Delivery Partner (Delivery, Parceled and Ecom Express). In this snapshot, the seller 197 sells his product to various Buyers(Buyer97, Buyer14, etc.). Also, majority of his sales are handled by Ecom Express.

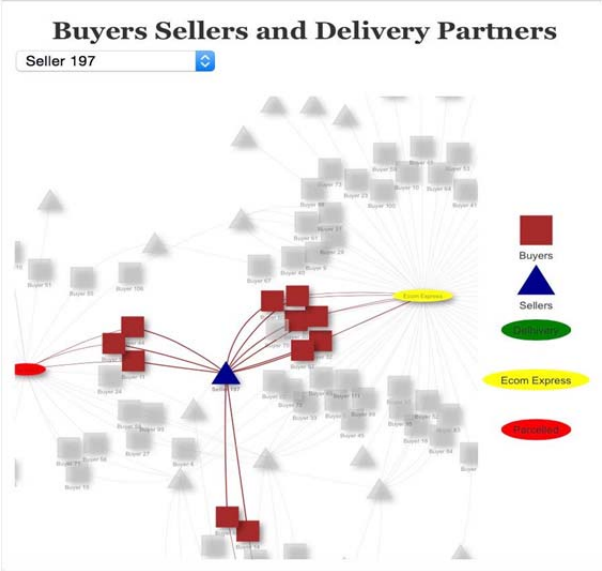

Fig. 6 visNetwork Graph

VII. LEAFLET HEATMAP FOR USERS' AND DELIVERY PARTNERS

Leaflet is one of the most popular open-source JavaScript libraries for interactive maps.This R package makes it easy to integrate and control Leaflet maps in R.[11]

Features:
Interactive panning/zooming
Compose maps using arbitrary combinations of:
Map tiles
Markers
Polygons
Lines
Popups
GeoJSON
Create maps right from the R console or RStudio
Embed maps in knitr/R Markdown documents and Shiny apps.
Easily render spatial objects from the sp package, or data frames with latitude/longitude columns.
Use map bounds and mouse events to drive Shiny logic
Distribution

It is observed from the Fig. 7 that the distribution of Buyers' are restricted primarily to Karnataka, and specifically to Southern Karnataka. Delhivery provides service to most parts of India, but relatively, it is lacking in the state of Andhra Pradesh.

Heat maps are made by using the Leaflet library in R.

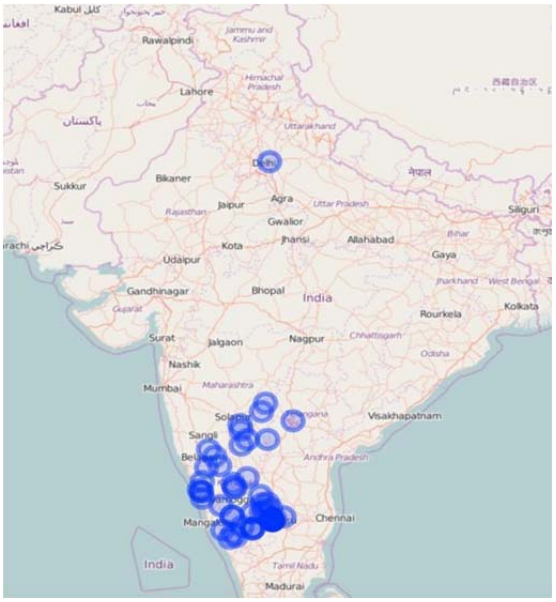

Fig. 7 Leaflet Heat Map for Buyers

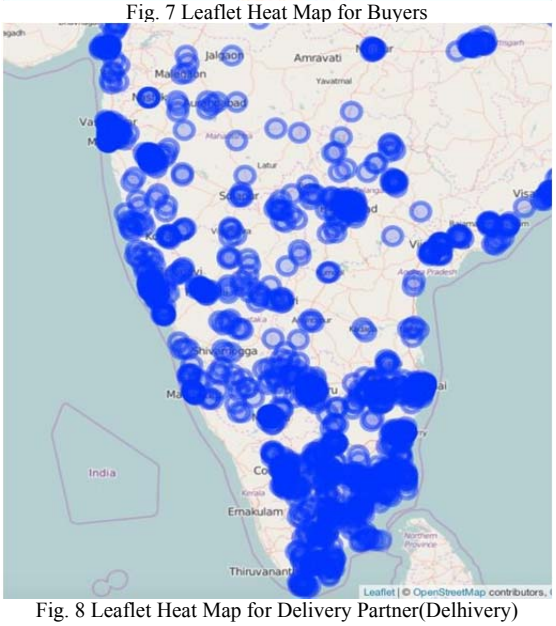

Fig. 8 Leaflet Heat Map for Delivery Partner(Delhivery)

VIII. RSHINY DASHBOARD

Dashboards give signs about a business letting the user know something are wrong or something is right. The corporate world has tried for years to come up with a solution that would tell them if their business needed maintenance or if the temperature of their business was running above normal. Dashboards typically are limited to show summaries, key trends, comparisons, and exceptions.

Shiny is a new package from RStudio that makes it incredibly easy to build interactive web applications with R.

Shiny applications are automatically ―live in the same way that spreadsheets are live. Outputs change instantly as users modify inputs, without requiring a reload of the browser. Shiny user interfaces can be built entirely using R, or can be written directly in HTML, CSS, and JavaScript for more flexibility.

Depicting the entire Dashboard, the following figure shows the various options and info graphs displayed on the dashboard which gives it an ergonomic edge.

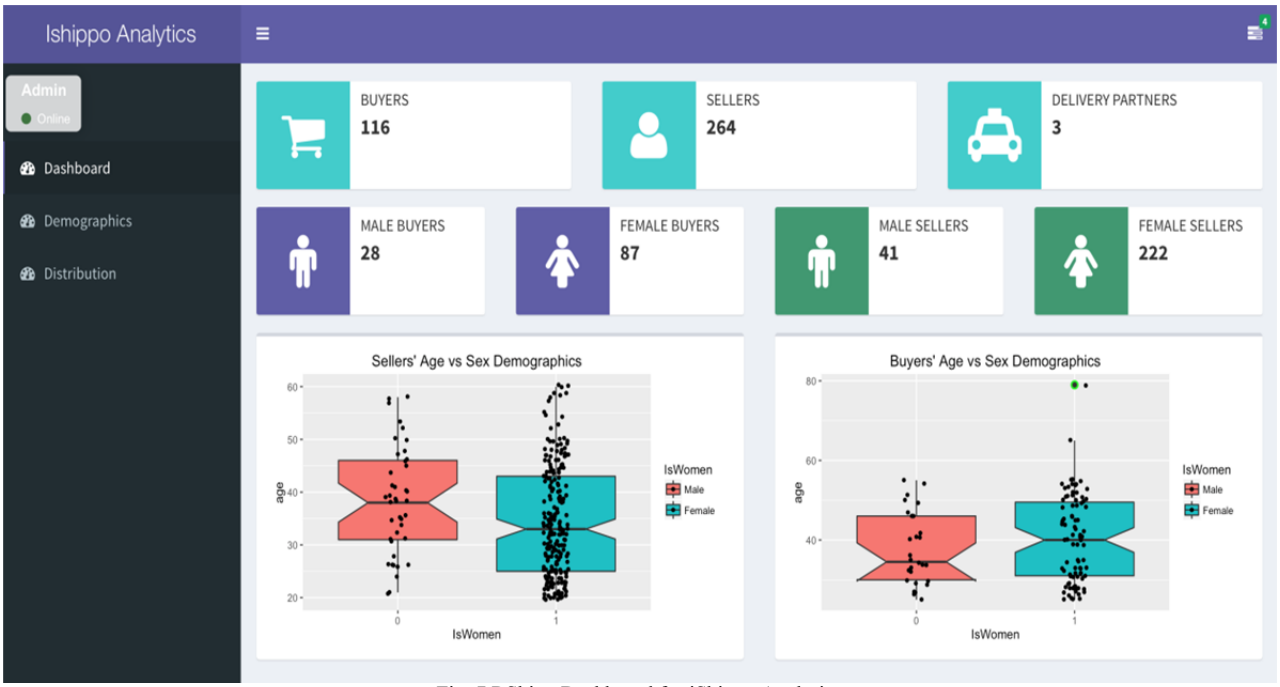

Fig. 7 RShiny Dashboard for iShippo Analytics

IX. Conclusions and Future Enhancements

With the development of this system, which recognizes and represents the hot spots in a city, there can be an exponential rise in productivity through targeted marketing and strategy implementation.

The interactive Dashboard will also provide a one-stop solution for all the demographic information of the users of iShippo.

The visNetwork Graph developed in this system provides key insights into which sellers and indeed selling and to which particular buyers. Also, conclusions can be drawn onto which delivery partner actually the most effective, thus, is ensuring there is no wastage of resources.

By comparing the Traditional method of averages vs. the Wilson Score Confidence method for Bernoulli Parameters, it is observed that Karnataka has a higher Ranking using the Wilson algorithm as compared to the traditional algorithm.

The box plots provide an effective method as to visualize the median age, 1st quartile age and the third quartile age of the Users of iShippo.

Proposed future enhancements include the inclusion of a login page and the possibility that new users can be added only by the Editors, i.e., the CTO of iShippo.

Also, an option to include the third gender i.e., transgender population as part of the demographic study shown, is a possible future enhancement.

Acknowledgment

We acknowledge Dr. Mohan Babu G.N, Principal, BMS Institute of Technology & Management and Dr. Thippeswamy G, Head of Department, Department of Computer Science and Engineering, for their co-operation and encouragement.

We acknowledge our project guide, Mrs. Radhika K R, Dept. of CSE, BMSIT&M and the members of the P.A.R.C Committee, BMSIT&M.

We also would like to thank Mr. Vasudeva S Tenkilaya, Vice President, iShippo, and Mr. Sarath Holigi, Product Manager, iShippo, for their continuous support and opportunity to work on this project.

References

1. Dr I. Satya SunDaram (Jan 2012) Handicrafts: Vast Untapped potential. Available:http://www.efymag.com/admin/issuepdf Handicrafts_Jan12.pdf
2. S.S. Solanki 2008 Science and Technology Available: http://www.nistads.res.in/indiasnt2008/t6rural/t6rur5.htm
3. J H Maindonald (2008 January 19) Using R for Data Analysis and Graphics Introduction, Code and Commentary Centre for Mathematics and Its Applications, Australian National University. Available: https:// cran.r-project.org/doc/contrib/usingR.pdf
4. Roger D Pheng (2015) R Programming for Data Science Available: http://leanpub.com/rprogramming
5. Brys, G., Hubert, M. & Struyf, A. (2004),A robust measure of skewness‘, Journal of Computational and Graphical Statistics 13(4), 996–1017.
6. http://www.evanmiller.org/how-not-to-sort-by-average-rating.html
7. Chambers, J., Cleveland, W., Kleiner, B. & Tukey, P. (1983),Graphical methods for data analysis‘, The Wadsworth Statistics/Probability Series. Boston, MA: Duxury.
8. Esty, W. W. & Banfield, J. D. (2003),The box-percentile plot‘, Journal of Statistical Software.
9. Kampstra, P. (2008), Beanplot: A boxplot alternative for visual comparison of distributions‘, Journal of Statistical Software 28, 1–9. Code Snippet 1.
10. Dellarocas, C., "Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems", in Proceedings of ICIS, (2000).
11. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer, New York, 2009. URL http://had.co.nz/ggplot2/book. [p144]