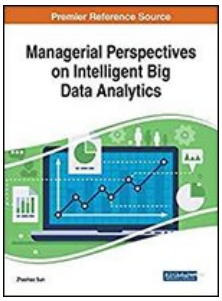

Managerial Perspectives on Intelligent Big Data Analytics

by Zhaohao Sun

Chapter 8: Census Data Analysis and Visualization Using R Tool: A Case Study

Veena Gadad,
Rashtreeya Vidyalaya College of Engineering,
India

Sowmyarani C. N.,
Rashtreeya Vidyalaya College of Engineering,
India

ABSTRACT

As a result of increased usage of internet, a huge amount of data is collected from variety of sources like surveys, census, and sensors in internet of things. This resultant data is coined as big data and analysis of this leads to major decision making. Since the collected data is in raw form, it is difficult to understand inherent properties and it becomes just a liability if not analyzed, summarized, and visualized. Although text can be used to articulate the relation between facts and to explain the findings, presenting it in the form of tables and graphs conveys information effectively. Presentation of data using tools to create visual images in order to gain more insights into data is called as data visualization. Data analysis is processing and interpretation of data to discover useful information and to deduce certain inferences based on the values. This chapter concerns usage of R tool and understanding its effectiveness for data analysis and intelligent data visualization by experimenting on data set obtained from University of California Irvine Machine Learning Repository.

INTRODUCTION

R is an open source programming language whose main purpose is to deliver an user friendly way to perform data analysis, statistics and data visualization. The survey performed by IEEE spectrum on "The top programming languages of 2017" (Cass, 2018) tells that the R language is on the sixth position and python on first position among top 48 programming languages used by data scientists for analysis. As of June 2018, R ranks 10^{th} in TIOBE index, a measure of popularity of programming languages (TIOBE The software Quality Company, 2018)The reason that R is used popularly is:

1. R is a open source programming language- There is no limit with respect to subscription costs or license management. The libraries of the language are freely accessible.
2. R is best statistical analysis tool- Data can be accessed in variety of format and many operations can be performed on the data with several functionalities useful for modern statistician. "dplyr" and "ggplot2" are examples for data manipulation and plotting.
3. R provides best data visualization tools to create graphs, bar charts, multi panel lattice charts, scatter plots and custom designed graphics.
4. R has consistent online support- The language being open source has a loyal support from statisticians, scientists and engineers.

Big data has potential to revolutionize the operational and strategic impacts, however there is paucity of empirical research. (Wamba, S. F, 2015). Big data as it is difficult to describe data without performing data analysis and visualization. In this article the important features of R to manage big data are discussed, relevant examples are articulated by carrying out experiments on UCI repository adult data set. The following are the main objectives of this article:

1. Performing descriptive analysis for quantitative describing or summarizing the properties of the collected data. This includes examining the mean, standard deviation, minimum, maximum and median for numeric data or frequency of observation for nominal data.
2. Intelligent data visualization for descriptive analysis with graphs like histograms, scatter plots and QQ plots.
3. Exploratory data analysis is used to understand the properties and find patterns in the data set with visual methods (R, pp. 10-50). R provides number of functions useful for exploratory data analysis like box plot, histograms, scatter plot, violin plot etc‥
4. To perform statistical tests to perform statistical inferences and to draw some conclusions about the data. R provides functions to determine p-value and alpha to test the null hypothesis.

5. Generation of dynamic documents using R Markdown and R programming language.

ORGANIZATION OF THE PAPER

The initial part of the paper presents managerial perspective of big data, description of the dataset used to understand the R tool, existing proprietary and open source tools to perform data analysis and visualization. In the rest of paper, usage of important R - libraries are discussed which can be used to perform effective data analysis and data visualization using census data as part of case study. Generation of reports using R markdown is discussed in the last part of the paper.

MANAGERIAL PERSPECTIVE OF BIG DATA

In digital era, huge amount of data is collected through surveys. The intension of any survey is to perform statistics, derive implications and to make decisions. Carrying out surveys is a reliable method to get feedback directly from the source/individuals/data owners. Some of survey examples are: customer satisfaction survey, employee survey, product survey, market research survey, website feedback survey, real time data from sensors etc. The data collected should be managed and utilized for analysis. The steps involved in managing any survey data remains the same as shown in Figure 1.

- **Planning:** This is a crucial stage of the entire process, various issues are to be addressed such as objectives of data collection, determining the type of data collection, sample design and sample unit size.
- **Data Collection:** Once the planning phase is completed and the decision on what data had to be collected is decided, either a software or hardware are deployed to collect the data. The questionnaire needs to be prepared to collect the data appropriately.
- **Processing:** The data collected is in the raw form, it has to be processed which involves data cleaning and transformations as per the requirements. R tool provides various functions like separate, gather, merge etc. for tidying the messy data.
- **Analysis:** The data just becomes liability if not analysed appropriately. R provides library dplyr. This provides various functions to perform data analysis.
- **Dissemination:** Since the data is collected for a purpose it should be disseminated as per the requirements of the concerned user.
- **Evaluation:** The process ends with evaluation of the processed data and user consultation.

DATASET DESCRIPTION

In this paper, analysis of census data which is obtained from University of California Irvine (UCI) machine learning repository (C. Blake, C. Merz, 1998) is used. The dataset is multivariate with 15 variables.

The total number of records are 32561, out of which variables type_ employer has 1836 records, occupation has 1843 records and country has 583 records. The description of the attributes in the data set is shown in table 1.

BACKGROUND

Big data has variety of data structures and the number of attributes is much larger. Many tools both commercial and open source are used by industries to exploit the hidden structure of the data. In this section popularly used commercial and open source data analysis tools their specific use and advantage is discussed.

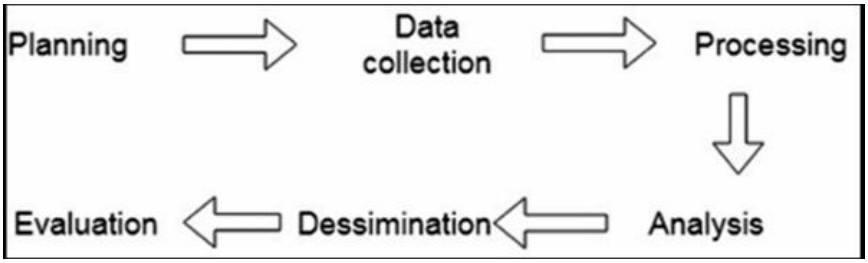

Figure 1: Managerial perspective of big data

Table 1: Description of adult data set

Sl.No	Attribute Name	Data type	Description	Possible Values

Table 1: Description of adult data set

Sl.No	Attribute Name	Data type	Description	Possible Values
1	age	Numeric	Age of the individual	10,11,12……39,40,41.
2	workclass	categorical	class of work	"Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3	fnlwgt	Continous	Final weight determined by census org	Numeric
3	education	Ordered Factor	Education of the individual	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
4	education-num	continous	Number of years of education	Numeric
5	maritial-status	categorical	marital status of the individual	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
6	occupation	categorical	Occupation of the individual	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
8	relationship	categorical	Present relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
9	race	categorical	Race of the individual	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
10	gender	categorical	gender of the individual	Male, Female
11	capital-gain	Continuous	Capital gain made by the individual	Numeric
12	capital-loss	Continuous	Capital loss made by the individual	Numeric
13	hours-per-week	Continuous	Average number of hours spent by the individual on work	Numeric.
14	native-country	categorical	Native country of the individual	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad & Tobago, Peru, Hong, Holand-Netherlands
15	income level	categorical	income level of an individual	"<=50k",">=50k"

Commercial Big Data Analysis Tools

SAS (Statistical Analysis System) is a suite of software tool that facilitates data analysis, reporting, data mining, predictive modelling, business intelligence, data management etc. It supports powerful visualization and interactive dashboards (SAS - https://www.sas.com/en_in/software/access.html). SAS can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis on it. It also provides a graphical point-and-click user interface for non-technical users and more advanced options through the SAS language. The other advantages of SAS are ability to handle large databases, easy to debug, tested algorithms, data security etc. The disadvantage associated with this tool is that it is not an open source software therefore algorithms used in SAS are not made public for common use. Other disadvantages include cost, lack of graphical representation, needs to learn SAS language etc.

Another such software is Tableau (Tableau - https://www.tableau.com/products/prep) It is a tool for visually analysing and

exploring the data, it can combine multiple databases easily. Tableau queries relational databases, OLAP cubes, cloud databases, and spreadsheets and then generates a number of graph types. It has mapping functionality and is able to plot latitude, longitude coordinates and connect to spatial files. The products can also extract data and store and retrieve from its in-memory data engine. The tool does not require any complex scripting and after analysis the reports can be shared easily by publishing the reports to Tableau server. Other advantages of the software are ease of implementations, quickly create interactive visualizations, can handle large amounts of data, use of other scripting languages in tableau etc, however the limitations are that the software can be used only on static data extracts and the libraries are not open source. Tableau does not provide the feature of automatic refreshing of reports with the scheduling.

Qlik View is another commercial data analysis tool that is used for searching, visualizing and analyzing the data with in depth insight on the data (Qlik View- https://www.qlik.com/us/products/qlikview). It provides better value to existing data stores with clean and simple user interface. The feature of in-memory data processing, gives superfast result to the users also the aggregations are calculated on the fly and data is compressed to 10% of original size. The software has data volume limitations and hence may not be suitable for big data analysis. The other limitations include it cannot be OLTP (OnLine Transaction Processing Tool).

Splunk- is a proprietary software platform used to search, analyze and visualize the machine generated data gathered from websites, applications, sensors, devices etc. It is also used for system performance analysis, monitor business metrics and provides a dashboards to visualize and analyze results. (Splunk -https://www.splunk.com/en_us/resources.html). Splunk captures, indexes, and correlates real-time data in a searchable repository from which it can generate graphs, reports, alerts, dashboards, and visualizations. The main advantage of the tool is its real time data processing, the input data can be in any format (.csv, json or other), can configure the tool to give alerts/ event notifications, create knowledge objects for operational intelligence. The main limitation of the software is it is not open source software.

Open Source Big Data Analysis Tools

There are few open source Big data analysis tools such as R it is an open source language developed for data analysis, it is used mainly for statistical analysis and data visualization. R is open source and easy to install locally and has rich libraries to serve the purpose(R -https://www.r-project.org). The ggplot2 library provides visualization functionality, dplyr library provides necessary functions to perform data analysis. R also has libraries for linear and non linear modelling, time series analysis, clustering and graphical representation. It provides the most comprehensive statistical analysis package that incorporates all standard statistical tests, models and analysis.

Another data analysis tool is Python, itis a multi purpose programming language that has support for data analysis, visualization a machine learning and building models (Eubank, 2015). It provides functionalities through libraries like NumPy and pandas. The package Pandas makes importing and analysing the data much easier. NumPy is a general purpose array-processing package. It provides high performance multidimensional array object and tools for working with these arrays.

Google Fusion Tables is another open source tool used for data analysis, mapping and large dataset visualization(Google Fusion Tables-https://sites.google.com/site/fusiontablestalks/home). It is the web service provided by google for data management. Fusion can filter and summarize many rows, combining multiple databases, generate single visualization that includes sets of data. Data are stored in multiple tables that Internet users can view and download. Google fusion table is designed such that it can handle hundreds and thousands of tables with diverse schemas, sizes and query load characteristics. (Gonzalez, H., Halevy, A., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., & Shen, W, 2010) The limitation of the tool is its scalability (Balakrishnan, 2017).

Table 2 and Table 3 summarizes the commercial data analysis tools and open source data analysis tools.

Table 2: Commercial big data analysis tools

Sl.No	Name	Specific Use	Advantage	Disadvantage
1	SAS (Statistical Analysis System)	Statistical Analysis, Business Intelligence, data mining, predictive modelling, data management.	Easy to learn and use. Ability to handle large data base. Tested algorithms. SAS GUI.Data Security	SAS is not open source. Cost inefficient, lack of graphical representation.
2	Tableau	Data discovery and data exploration. Data visualization	Get connected to different data sources from files and server. Quickly create data visualization, can handle large amount of data.	The software can be used only on static data extracts and the libraries are not open source. It does not provide the feature of automatic refreshing of reports with the scheduling.
3	Qlikview	Data discovery and decision making. Data analysis and visualization	Flexibility in access, fast analysis, quick time to value, user centric interactivity.	Requires trained developer, not open source.

Table 2: Commercial big data analysis tools

Sl.No	Name	Specific Use	Advantage	Disadvantage
4	Splunk	Analyze machine generated data.	It can pull data from multiple systems in real time. The input data can be in any format. The tool can be configured to generate alerts and notifications.	Not an open source tool

Table 3: Open source big data analysis tools

Sl.No	Name	Specific Use	Advantage
1	R	Statistical analysis Data visualization	Built in data analysis functionality. Core libraries are well maintained by the CRAN and key programmers Open source.
2	Python	Prototyping, visualization and data analysis on small and medium sized data sets. Machine learning and building tools.	Easy to learn and use Less code and large work. Open source Libraries for handling large multi dimensional arrays and matrices.
3	Google Fusion Tables	Summary with simple aggregate statistics, variety of charts.	Scalability

Growing Popularity of Open Source Tools

Recently open source software, applications and projects are commonly used than the closed source software also known as proprietary software. In proprietary software only the author has a legal copy and he only can modify, distribute, inspect or alter the software. Some of the proprietary softwares are Apple iOS mobile operating system, Microsoft Office suite, Adobe Photo shop etc. To use such softwares the end user must purchase the software and agree to some specific terms set forth by the owner. Open source software can be altered, shared by anyone. Open source software is equal to or more capable than professional proprietary software. The benefits and popularity of open source software is extremely high mainly because the software is free to use, modify and distribute. This software are more secured and accessible to everyone. The main advantage of such software tools is, it is free from complex licensing issues and do not need anti-piracy measures such as product activation and serial number. Some of the examples of open source softwares are Libre Office, Linux operating system and its flavours, VLC media player, Android mobile operating system etc. With Big data it is possible to enhance decision making and organizational perspective with the help of many available tools, technologies and management as discussed in (Storey, 2017).

MAIN FOCUS OF THE CHAPTER

The main aim of the article is to describe important libraries in R that can be used to perform effective data analysis, data visualization and generate reports. The applications of relevant functions is discussed taking census data as a case study. Various graphs are demonstrated using examples.

DATA MANAGEMENT LIFE CYCLE

The data management life cycle consists of various phases from data collection to generation of information from the data, figure 2 depicts various stages of data management systems.

1 Data Collection

Data collection is a process of gathering information in a systematic fashion which subsequently allows for data analysis to be performed on the collected information. In this first step the data is collected from various sources to get complete and accurate picture of an area of interest, the data is stored in form of a tables. Accurate data collection is essential for maintaining integrity of research, making business decisions and ensuring quality assurance. A formal data collection process is necessary as it ensures that the data gathered are both defined and accurate and that subsequent decisions based on arguments embodied in the findings are valid (Sapsford, R., & Jupp, V, 2006)

2 Over Viewing the Data

Collected data is over viewed to understand the number and nature of variables, number of rows/columns in the data. This can be done by using various commands in R shown in table 4.

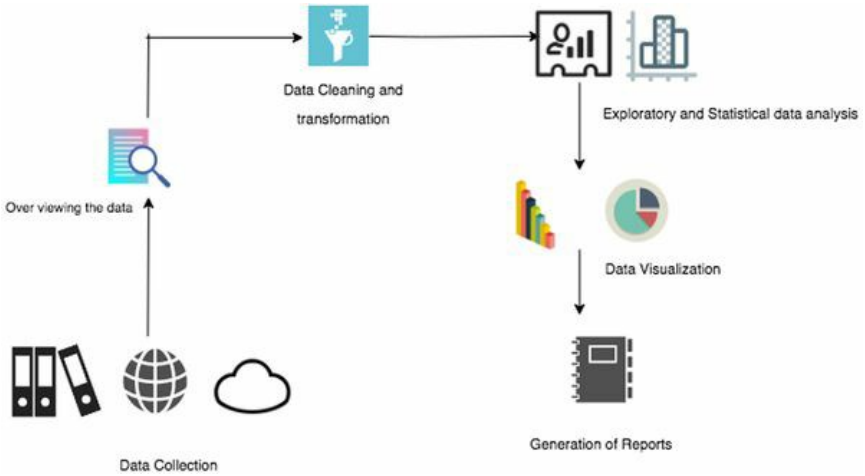

Figure 2: Data management life cycle

Table 4: Over viewing data using R

Sl.No	Command Name	Usage	Description
1.	dim() Example: dim(db.adult) 32561 db.adult is the data frame.	dim(x) dim(x) =value where x is an object and value is either NULL or numeric vector	Retrieve or sets number of observations of an object
2	names()	names(x) or names(x)= value	Retrives or sets names of an object
3	head() and tail()	head(x), tail(x)	Displays first/last few rows of the object
4	str()	str(x)	Compactly displays the internal structure of the object x
5	levels()	levels(x)	provides access to the factors of the categorical attributes.
6	factor()	factor(x)	Used to encode a vector as a factor
7	length()	length(x)	To retrieve length of vectors and factors.

3 Data Cleaning and Transformations

This process deals with detecting bad or missing values and shaping the data for efficient analysis. Before proceeding with any type of analysis the collected data should be cleaned by detecting and removing corrupt or inaccurate records from the data set. The bad values are NA, NaN or Inf, there can be invalid values like negative number for age attribute or values beyond the range. In R the function complete. cases() on the data frame returns TRUE for the observation where there is no NA's otherwise FALSE.

For Example:

```
TotalRows=nrow(db.adult)
CompleteRows=sum(complete.cases(db.adult))
TotalRows/CompleteRows
```

Based upon this value decision may be take to either drop or not to drop the invalid /missing observation. In R the process of transforming and mapping data from one form to another form with the intent of making it more appropriate and valuable for analysis is called data wrangling (Boehmk, 2016). It can be carried out using the functions in library tidyr and dplyr. The tidyr package makes it easy to "tidy " the data, so that it becomes appropriate to manage, visualize and model. tidyr provides four main functions for tidying the messy data (Wickham H, 2014). The description of these functions is discussed in table 6. dplyr package is used for data manipulation. The functions provided are used for performing exploratory data analysis and manipulation (Wickham. H., Francois. R, 2015). The description and use of functions of dplyr is presented in table 5.

4 Exploratory and Statistical Analysis

Given any data set the variables can be classified as numeric and categorical. The numeric variables can further be classified as continuous and discrete. The categorical variables can be further classified as ordinal and nominal variables. The figure 3 explains the possible ways of performing analysis on the data.

Table 5: Description of functions in dplyr library

Sl.No	Function Name	Description	Usage
1.	glimpse	Provide summary of each column of the dataset.	glimpse(dataset)
2.	summary	If the data is numeric or integer, the summary distribution of the column including minimum and maximum, mean is displayed	summary(dataset)
3.	Filter	Returns the rows that satisfy certain condition	filter(dataframe, condition)
4.	Summarise	Used to summarise multiple value into a single value.	summarize(dataframe, expression)
4.	Group By	Used to group the data by one or more variables.	group_by(dataset, grouping attributes)
5.	Count	Used to tally the observations based on a group.	count(dataframe, expression)
6.	Arrange	For sorting data in ascending or descending order (ascending is default)‥	arrange(dataframe, col_name)
7.	select	used to take a subset of a data frame by columns	select(Dataframe,col_name)

Table 6: Description of functions in tidyr

Sl.No	Function	description	usage
1	gather	Takes multiple columns, and gathers them into key-value pairs: it makes "wide" data longer	gather(variable name/s)
2	separate	Separate one column into several.	separate(variable name/s)
3	spread	Takes two columns (a key-value pair) and spreads them in to multiple columns, making "long" data wider.	spread(variable name/s)
4	unite	unite several columns into one	unite(variable name/s)

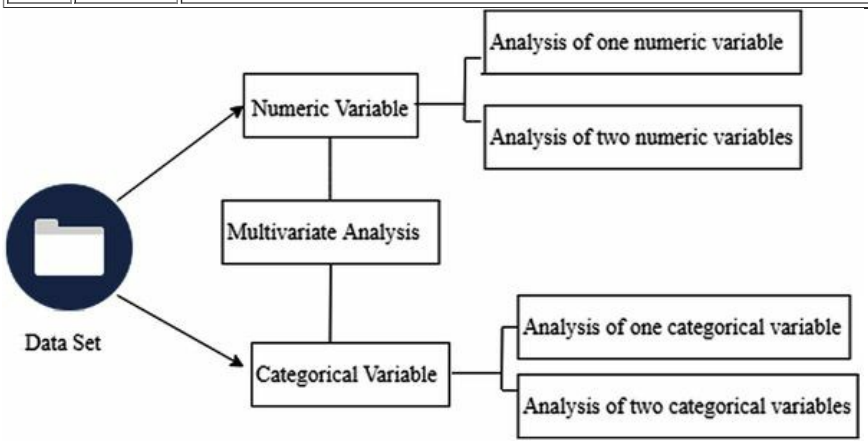

Figure 3: Variable analysis

Analysis of One Numeric Variable

This includes finding mean, standard deviation, median, central tendency, dispersion and finding five number summary. Most of the functions used for this analysis are from basic packages that are loaded by default in R. Table 7 lists the functions and their usage to perform analysis of numeric variables.

Analysis of One Categorical Variable

This can be done using frequency tables and using proportion. The useful functions are listed in the table 8.

Analysis of Two Categorical Variables

A contingency table also called as cross tabulation is often used to analyze the relationship between two or more categorical variables. In R xtabs function is used to construct contingency table. For example, the contingency table for work class and gender is given below.

Table 7: Functions for analysis of numeric variable

Sl.no	Function	Description	Usage

Table 7: Functions for analysis of numeric variable

Sl.no	Function	Description	Usage
1	mean()	Calculates the arithmetic mean of the numeric data	mean(data)
2	median()	Calculates the midpoint of frequency distribution of observed values or quantities.	median(data)
3	sd()	Calculates standard deviation of the data	sd(data)
4	range()	Calculating measures of dispersion	range(data)- Gives the minimum and maximum value.
5	cv()- defined in raster package	Computing coefficient of variation- standard deviation/mean	cv(data)
6	summary()	overall description of the data- Minimum, 1st Quartile, Median, Mean, 3rd Quartile, Maximum value of the data	summary(data)
7	describe()- defined in psych package	overall description of data including standard deviation, sample size, skewness and kurtosis.	describe(data)

Table 8: Useful functions for analyzing categorical variables

Sl.No	Function	Description	Usage
1	table()	To determine occurrences of each type of categorical variable	mytable=table(data)
2	prop.table()	To determine the proportions	prop.table(mytable)
3	margin.table()	To compute the sum of table entries for a given index.	margin.table(x, margin = NULL) x- data object margin- index number
4	ftable()	To generate multidimensional tables based on 3 or more categorical variables	mytable =table(A, B, C)) ftable(mytable)

```
xtabs(~ workclass + sex, data=adultfull)
  workclass              sex
                      Female  Male
  Federal-gov          309   634
  Local-gov            824  1243
  Private             7642 14644
  Self-emp-inc         126   948
  Self-emp-not-inc     392  2107
  State-gov            484   795
  Without-pay            5     9
```

The table shows the number of observations that are associated with work class and gender combinations. The association can also be performed on more than two categorical variables. Most commonly used approach is to use graphical summaries to understand the relationship between two categorical variables, this will be discussed in intelligent visualization section.

Analysis of Two Numeric Variables

This is to determine the association between the pairs of numerical variables in a sample. The analysis is also called as bivariate associations. Pearson's Correlation coefficient is used to determine association between the variables (Benesty J., 2009). The mathematical formula for Pearson's correlation coefficient is:

$$rxy = \frac{1}{N-1}\left(\sum_{i=1}^{N} \frac{Xi - Xmean}{Sx}\frac{Yi - Ymean}{Sy}\right)$$

Xi and Yi are two different variables in a sample. Sx and Sy denote the standard deviation of each sample. Xmean and Ymean denote the sample mean. N is sample size. The coefficient is designed to summarize the strength of linear association, if the result is 0 then the two variables are not correlated, if the value is +1 or -1 they are related. In R cor is used used to compute Pearson's correlation coefficient.

For example, the Pearson correlation coefficient between age and hours per week is 0.068, this indicates that they are not correlated.

Analysis of One Numeric Variable Grouped by a Categorical Variable or Multivariate Analysis

Box plot is commonly used approach to explore the relationship between categorical-numerical variables. This is discussed in

detail in intelligent visualization section.

Examples Using Census Data to Understand Exploratory and Descriptive Analysis

This type of analysis is performed to describe or characterize the data prior to any complex analysis or modelling. For example in the case study computing the mean, calculation of average of distribution, determining the range to characterize its variability constitutes descriptive analysis. Using R tool computing the mean, standard deviation, minimum, maximum and median for numeric data or frequency of observation for nominal data is much easier. This section discusses the descriptive analysis on census data using R tool

Inspecting the Census Data Set

1. ls(adultfull) – Lists all the variables in the dataset.

```
"age"            "capital_gain"   "capital_loss"   "education"      "edunum"
"fnlwt"          "hours_week"    "marital_status" "native_country" "occupation"
"race"           "relationship"   "salary"         "sex"     "workclass"
```

2. str(adultfull)- Structure of the dataset

```
 $ V1:int 39 50 38 53 28 37 49 52 31 42…
 $ V2: Factor w/ 9 levels " ?"," Federal-gov",··: 8 7 5 5 5 5 5 7 5 5…
 $ V3:int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449
…
 $ V4: Factor w/ 16 levels " 10th"," 11th",··: 10 10 12 2 10 13 7 12 13 10…
 $ V5:int 13 13 9 7 13 14 5 9 14 13…
 $ V6: Factor w/ 7 levels " Divorced"," Married-AF-spouse",··: 5 3 1 3 3 3 4 3
5 3…
 $ V7: Factor w/ 15 levels " ?"," Adm-clerical",··: 2 5 7 7 11 5 9 5 11 5…
 $ V8: Factor w/ 6 levels " Husband"," Not-in-family",··: 2 1 2 1 6 6 2 1 2 1
…
 $ V9: Factor w/ 5 levels " Amer-Indian-Eskimo",··: 5 5 5 3 3 5 3 5 5 5…
 $ V10: Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2…
 $ V11: int 2174 0 0 0 0 0 0 0 14084 5178…
 $ V12: int 0 0 0 0 0 0 0 0 0 0…
 $ V13: int 40 13 40 40 40 40 16 45 50 40…
 $ V14: Factor w/ 42 levels " ?"," Cambodia",··: 40 40 40 40 6 40 24 40 40 40
…
 $ V15: Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2…
```

The dataset contains 6 numeric variables and 9 categorical/factor variables. Mean, median, standard deviation and variance can be determined on numeric variables. The numeric variables can be extracted from the large dataset using subset() and select function. Let the resultant dataset be stored in test. data variable.

```
test.data=subset(adultfull,select=c("age","fnlwt","edunum","capital_
gain","capital_loss","hours_week"))
str(test.data)
 $ age       : num  39 50 38 53 28 37 49 52 31 42…
 $ fnlwt     : num  77516 83311 215646 234721 338409…
 $ edunum    : num  13 13 9 7 13 14 5 9 14 13…
 $ capital_gain: num  2174 0 0 0 0…
 $ capital_loss: num  0 0 0 0 0 0 0 0 0 0…
 $ hours_week:num  40 13 40 40 40 40 16 45 50 40…
```

Computing Mean of Variables

```
mean(test.data$age) -38.4379 – The average age of the population is 38 yrs.
mean(test.data$hours_week)40.93124 – The average working hours is 41 hrs.
mean(test.data$capital_gain) -1092.008 – The average capital gain of the popu-
lation is $1092
mean(test.data$capital_loss) -88.37249-The average capital loss of the popula-
tion is $88
mean(test.data$edunum) -10.12131 – The average education number is 10.

To understand the relation between edunum and education group_by function of
dplyr library is used as follows.
A=adultfull[,c("edunum","education")]
library(dplyr)
```

```
B=group_by(unique(A),"education")
```

Variable B now has mapping of edunum and education. This is shown in the table 9

Average edunum 10 shows that the minimum education of the population is some college.

Computing Standard Deviation of the Variables

Standard deviation is a number that indicates how measurements of a population are spread out from the mean. Lower this value means most of the numbers are close to mean, higher value indicates the spread of the numbers. In R sd(VAR) function is used to determine the standard deviation. For example:

```
sd(test.data$age) -13.13
sd(test.data$hours_week) -11.97
sd(test.data$capital_gain) -7406.346
sd(test.data$capital_loss) -404.29
```

Table 9: Mapping of edunum and education variables

edunum	Education
1	Preschool
2	1^{st}-4^{th}
3	5^{th}-6^{th}
4	7^{th}-8^{th}
5	9^{th}
6	10^{th}
7	11^{th}
8	12^{th}
edunum	**Education**
9	HS-grad
10	Some College
11	Assoc-voc
12	Assoc-acdm
13	Bachelors
14	Masters
15	Prof-School
16	Doctorate

Computing Median of the Variables

In a population the mean of the variables is influenced by the outliers. Median is another way to measure the centre of the numeric variable. In R the function median(VAR) is used to compute the median.

For example

```
median(test.data$age) -37
median(test.data$hours_week) -40
median(test.data$capital_gain) -0
median(test.data$capital_loss) -0
```

Summary of the Observation

In R the summary of the observation is obtained by using a function fivenum(VAR). It is the most useful function as it gives the five number summary- minimum value, 1^{st} Quartile, median, 3^{rd}Quartile and maximum value. A quartile describes division of observations into four defined intervals based upon values of data. A boxplot gives the graphical representation of median and

quartiles explained later.

5 Intelligent Data Visualization Using R

Bigdata means large and complex datasets, processing them with traditional data processing applications is difficult. Data analysis followed by intelligent data visualization yeilds a better aesthetic to Bigdata that is easily understood by the analyst. Any visualization presented, has to represent quantitative and qualitative information of the data, data graphics(also called as data visualization) instruments the former and information graphics instruments the later. These presents intense as sophisticated information on certain subjects in a planned and comprehensible manner (Dur, 2014). In R, ggplot2 library provides most elegant graphics framework to design for any type of data graphics. The graphs can be created on both univariate and multivariate numerical and categorical data in a straightforward manner. ggplot2 is based on the grammar of graphics, every graph is built from few components like data set, set of geoms(visualization marks that represent the data points) and the coordinate system. To display values, map variables in the data to visual properties of the geom(aesthetics) like size, color and x and y locations. The main features of ggplot2 are: it is consistent underlying grammar of graphics, plot specification at a high level of abstraction, theme system for polishing plot appearance, mature and complete graphics system. This library does not support 3-dimensional graphics, Graph-theory type graphs igraph package and interactive graphics. The useful functions contained in the library are described in the table 10.

The ggplot() function of ggplot2 library that creates a basic plot object requires dataframe as the argument and this data frame has all necessary features to generate a plot. The aesthetic mappings are specified by using aes() inside ggplot like X and Y axis that are respective variables from the dataset. The graph is then added with layers, scales and facets.

A layer (is also called as geom) is added using geom_function that combines data, aesthetic mapping, geometric object, statistical transformation and position adjustment. (Prabhakaran, 2017) illustrates many graphs with a primary purpose, some of them are as listed in table. The scales control the details of how data values are translated to visual properties. Facets generates a multiples of different subset of the data.

Usage of Proposed Managerial Perspective for Census Data Analysis With Examples

1. **Workclass analysis for gender based on age group:** The various workclasses are: Federal-gov, Local-gov,Private, Self-Employed, State-gov, Without pay. The figure 4 shows gender analysis with their age in various workclasses.

Inference from figure 4:

- Maximum population works for private workclass
- Female and male population equally work in local government.
- There are people below age 15 and above 70 who work without pay.
- Female population is less in self-employed workclass.

Table 10: Description of useful functions in ggplot library

Sl.no	Function	Description	Usage
1	ggplot()	Create basic plot object that will display something	ggplot(data=df, aes(x=xcol_name,y=ycol_name))
2	geom_point()	Creates scatter plot on top of blank ggplot using a geom layer.	geom._point(col="steelblue", size=3) // Specifies color and size of the points. geom._point(aes(col="col_name"),size=3) //Set the points to reflect categories in another column.
3	ggtitle()	To add titles and lables to the chart	ggtitle("Main title", subtitle="Sub title Name")
4	xlab()	To add label to x axis	xlab("x-axis label")
5	ylab()	To add label to y axis	ylab("y-axis label")
6	coord_cartesian()	To change X and Y axis text and its location	coord_cartesian(xlim=c(0,max_limit), ylim=0,max_limit)) // To set the limit.
7	geom._encircle()	used to encircle certain group of points or region in the chart, to draw the attention to those peculiar cases.	

Table 11: Various types of graphs in R using ggplot()

Sl. No	Primary purpose	Types of Graphs
1.	Correlation between two variables	Scatter plot, Jitter plot, Bubble plot, Box plot

Table 11: Various types of graphs in R using ggplot()

Sl. No	Primary purpose	Types of Graphs
2.	Deviation	Diverging bar, lollipops, dot graph and slope chart
3.	Ranking	Bar chart, Dot Plot, slope chart etc.
4.	Composition	Pie chart, Bar chart, tree map etc‥
5.	Distribution	Histogram, boxplot, density plot, violin plot, etc‥

2. **Workclass analysis for age based on gender:** The figure 5 shows the age of male and female population working for different workclasses.

   Inference from figure 5:

   - Maximum population is engaged in working between age group 25 to 30.
   - There is gradual decrease in working population after age of 50.
   - There are male and female population working even after age of 75.

3. **Total working hours based on age:** The figure 6 shows how many people with different age group work for how many hours per week.

   Inference from figure 6:

   - Age group 30-40 works for maximum hours per week.
   - There is gradual decrease in the working hours for age group 50-75.
   - There are people at age 80 who works for few hours per week.

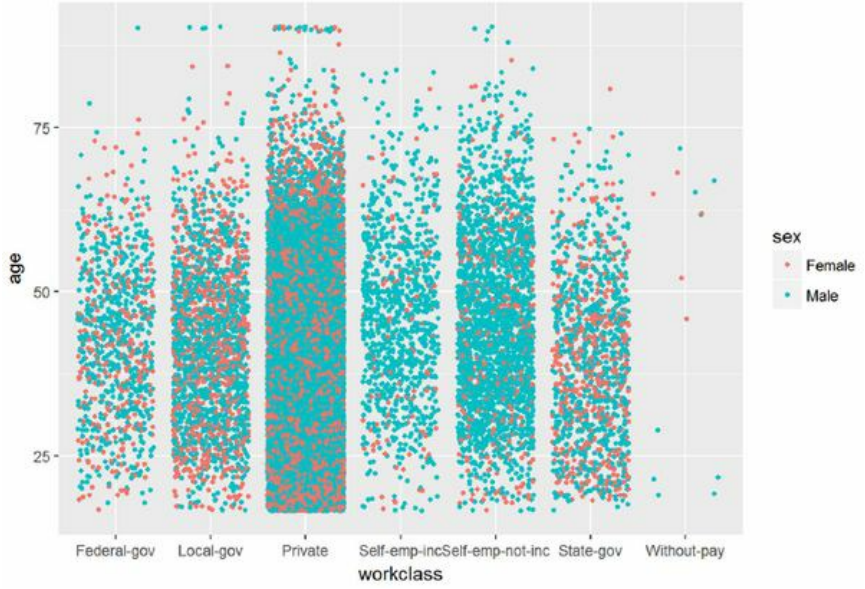

Figure 4: Workclass analysis for gender based on age group

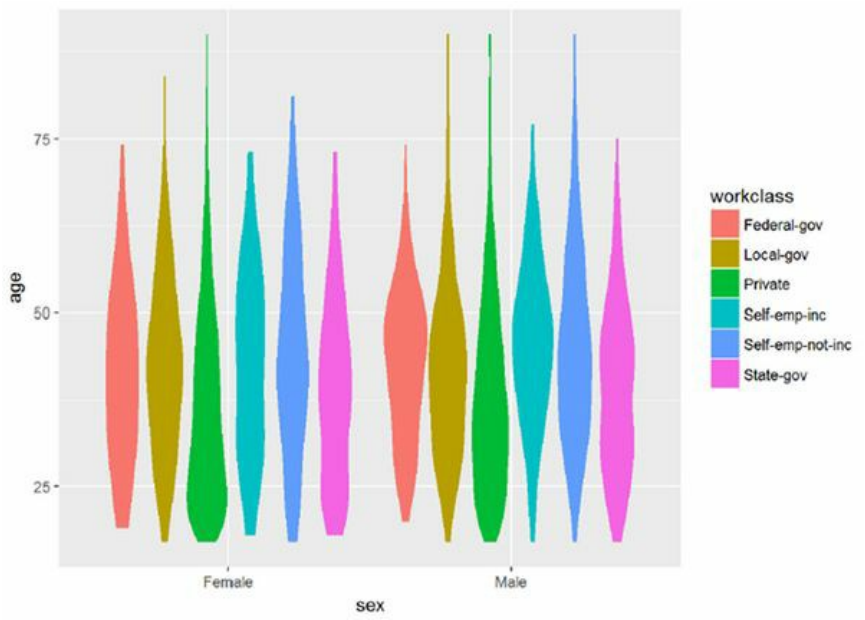

Figure 5: Workclass analysis for age based on gender

4. **Analysis of population with different race in different workclass based on age group:** The figure 7 shows different races of people and their age working for different workclass.

   Inference from figure 7:

   - All types of races are into federal govt and private workclass.
   - Americ-Indian-Eskimos are not involved in self-employment.

5. **Analysis of salary earned by people under different age group:** The figure 8 shows the salary earned by different age group.

   The following inferences can be drawn from figure 8:

   - 40% of the population between age group 30-40 earn salary >=50k.
   - 20% of the population between age group 10-20 earn salary <=50k.
   - 70% of the population between 65-75 earn salary >=50k.

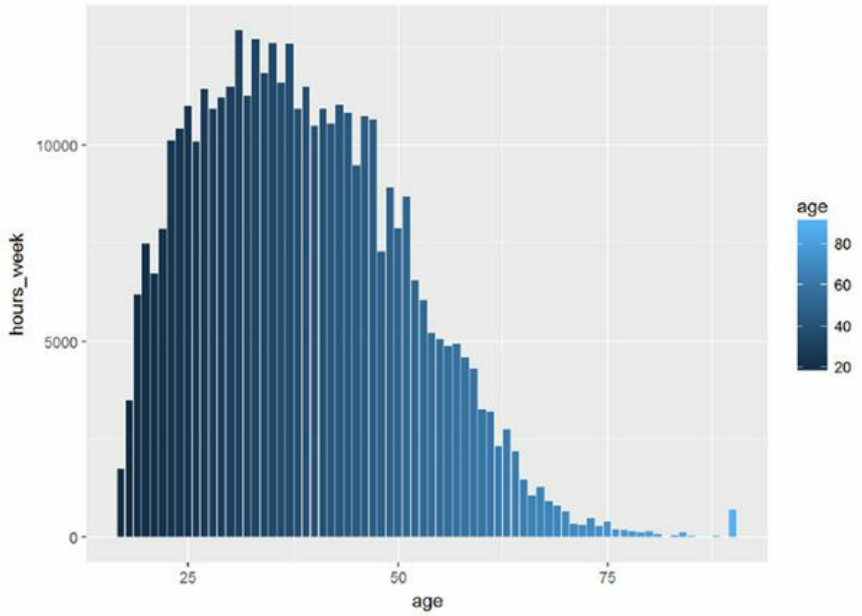

Figure 6: Total working hours based on age

6. **Analysis of Income based on qualification:** The figure 9 shows the visualization.

   Inference from figure 9:

   - 76% of people earn salary >=50k and 24% of people earn <=50k with doctorate.
   - 50% of people with bachelor and master earn equally.

6 Generation of Reports Using R

Using R Mark down and R programming language dynamic documents can be created. It supports dozens of static and dynamic output formats including HTML, PDF, MS word, HTML5 slides, scientific articles, websites etc. An R Markdown document is written in markdown (an easy-to-write plain text format) consists of chunks of embedded R code with.rmd extension. R Markdown uses markdown syntax. It provides an easy way of creating documents.

The documents produced by R markdown can be converted into many other file types. The documents that R Markdown provides are fully reproducible and support a wide variety of static and dynamic output formats. It is also possible to analyze the data into high quality documents, reports and presentations using R Markdown. There are a large number of tasks that can be done using R Markdown:

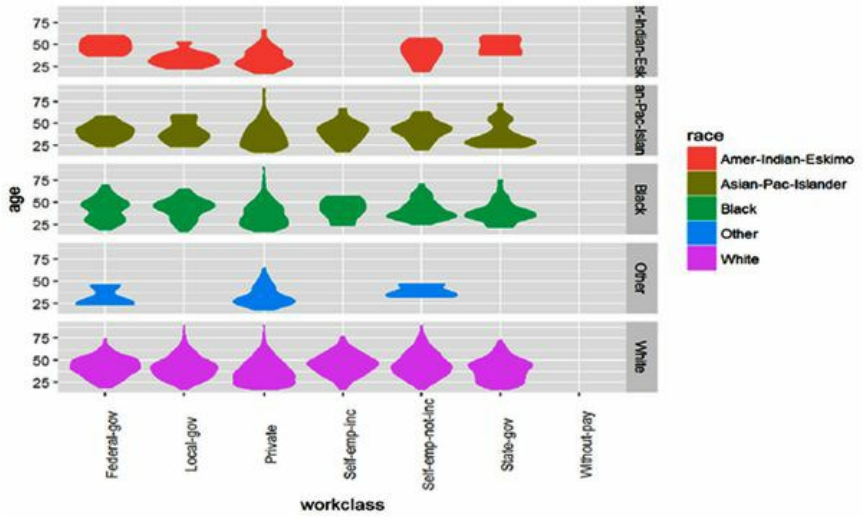

Figure 7: Different race in different workclass based on age

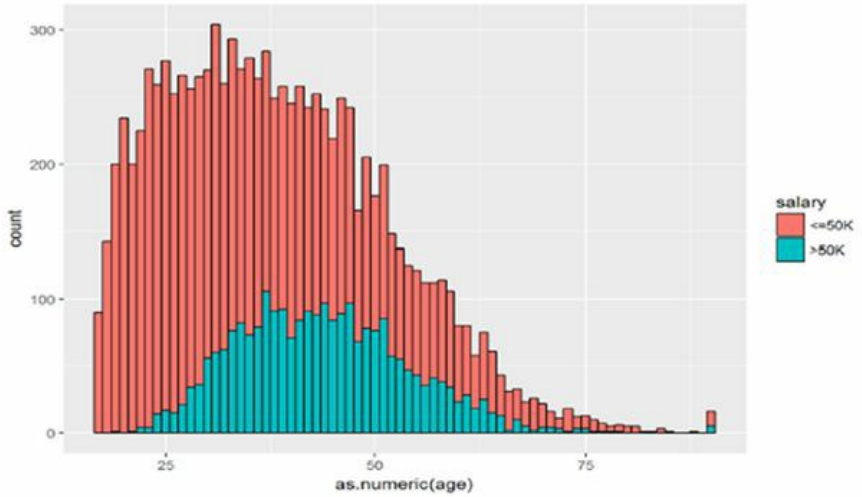

Figure 8: Analysis of salary earned under different age group

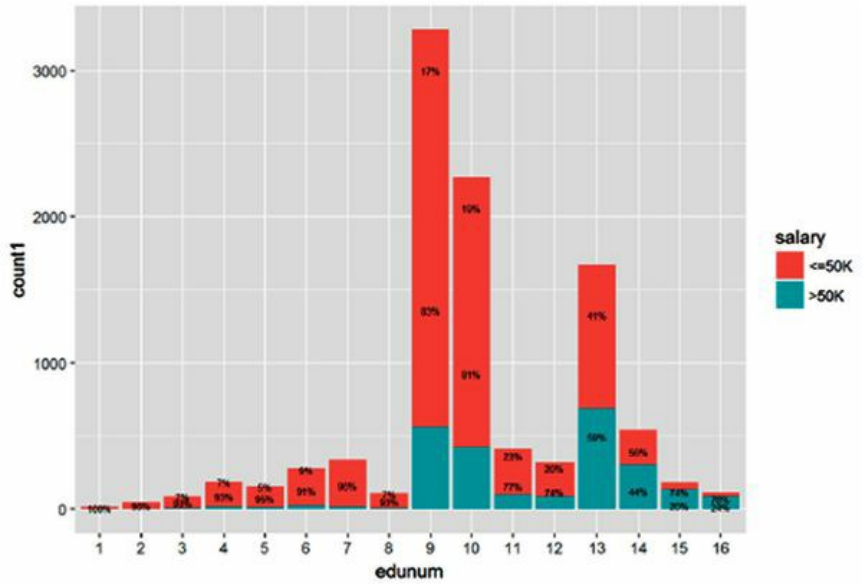

Figure 9: Income level analysis based on qualification

- Create a report in different formats like as PDF, HTML, or Word.
- Design notebooks with code snippet.
- Write journal articles.
- Write books of multiple chapters.
- Design attractive websites and blogs.

R Markdown file contents are discussed below.

The Header Section

At the top of any R Markdown file there is always the header section. The header section includes a title, author, date and the file type that appears in the output.

```
---
title: "BOOK CHAPTER"
author: "Veena Gadad"
date: "4 November 2018"
output: html_document
---
```

Code Section

Code to be included in.rmd document should be enclosed between three backwards apostrophes ``` for example:

```
```{r}
summary(adult)
```
```

Within the brackets{r, code}, rules can be assigned for the code chunk using code instructions. Some of the instructions are:

1. eval: When set to TRUE the code executed and the results included in the output.
2. echo: When set to TRUE the code displayed alongside the results.
3. warning: When set to TRUE the warning messages are displayed in the output. 4.error: When set to TRUE error messages displayed in the output.

Inserting Figures

To insert the figures in the output, there are instructions to set the figure dimensions, the instructions can be inserted as:

```
```{r,fig.width=5, fig.height=5,echo=FALSE}
plot(cars)
```
```

Inserting Tables

Using code instructions in R Markdown it is easy to print the contents of a data frame by enclosing the name of the data frame in a code chunk‥

```
```{r,echo=FALSE}
dataframe
```
```

A better solution is to use the table formatting function kable() in the knitr package. For example,

```
```{r,echo=FALSE}
library(knitr)
kable(dataframe,digits=1)
```
```

Formatting Text

Commands in Markdown can also be used to change the appearance of the output file, some of the common formatting commands are:

```
To include header of style1: # header 1
To include header of style1 header 2: ## header 2
To include header of style1 header 3: ### header 3
To include header of style1 header 4: # header 4
To include bold text: **bold** text
To include italics text: *italics* text
To include code text: `code` text
To inlcude a link: [link](www.rvce.edu.in)
To include a picture: R Studio Logo ![R Studio Logo](img/R Studio Logo.png)
To include LaTeX equation syntax: $A = \pi \times r^{2}$
```

Compiling an R Markdown File

The file in R Markdown can be transformed in two ways:

1. **knit-** This is a compiler that runs the part of R code in the document and appends the results of the code to the document, so that the document can consists of the graphs by actually running the R code. R Markdown file contains the code it needs to make its own graphs, tables and numbers, also the document can be updated by re-knitting it.
2. **convert-** Using "pandoc" it is possible to transform the R Markdown file into any new format like HTML, PDF or MS word, preserving the code results and formatting in the original.rmd file.

CONCLUSION

Data analysis and visualization are essential to understand and present the data systematically, understand the hidden patterns. As Big data is getting bigger and bigger in this digital age proper data management tools become very much essential. A systematic representation of data helps in understanding the marketing strategies, comparisons of results, decision making, estimating the targets etc. Among many of the available open source data analysis tools, R is one such open source tool using which the data management can be done efficiently. The article discusses various libraries and functions to carry out data analysis, intelligent visualization and generation of reports using R. The dplyr and ggplot2 libraries of R provides number of features using which any data can be analysed. Using R markdown and R programming language dynamic documents can be created. As a part of future work, this work can be extended to understand the libraries and functions in R to perform statistical analysis, preserving the privacy of the data and to apply data analysis using R for the data sets stored in cloud environment.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

About us: Splunk Software. (n.d.). Retrieved from *https://www.splunk.com/en_us/resources.html*

Balakrishnan, S. J.-A. (2017). *Google Fusion Tables*. *Encyclopedia of GIS*, 788-797.

Benesty, J. C. J. (2009). *Pearson Correlation Coefficient*. In *Noise Reduction in Speech Processing*. Springer Topics in Signal Processing (Vol. 2). Berlin: Springer. doi:10.1007/978-3-642-00296-0_5

Björn Berg, S. G. (n.d.). Retrieved from Qlik Technologies, Inc.: *https://www.qlik.com/us/products/qlikview*

Blake & Merz. (1998). *UCI repository of machine learning databases*. Academic Press.

Boehmk, B. C. (2016). *Data Wrangling with R*. Springer. doi:10.1007/978-3-319-45599-0

Cass, S. (2018). *The 2017 Top Programming Languages*. Retrieved from *https://spectrum.ieee.org/computing/software/the-2017- top-programming-languages*

Dur, B. I. (2014). *Data visualization and infographics in visual communication design education at the age of information*. *Journal of Arts and Humanities*, 5, 39–50.

Eubank, N. (2015). *Data Analysis in Python*. Retrieved from *http://www.data-analysis-in-python.org/*

Gentleman, R. I. (n.d.). *About R. Introduction to R*. Retrieved from *https://www.r-project.org/*

Gonzalez, H., Halevy, A., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., & Shen, W. (2010, June). *Google fusion tables: data management, integration and collaboration in the cloud*. *In Proceedings of the 1st ACM symposium on Cloud computing* (pp. 175-180). ACM. 10.1145/1807128.1807158

Google. (n.d.). *Tutorials*. Retrieved from *https://sites.google.com/site/fusiontablestalks/home*

Institute, S. (n.d.). *Products & Solutions A- Z*. Retrieved from SAS/ACCESS® Software: *https://www.sas.com*

Pat Hanrahan, C. C. (n.d.). *Combine, shape, and clean your data for analysis with Tableau Prep*. Retrieved from *https://www.tableau.com/products/prep*

Prabhakaran, S. (2017). *Top 50 ggplot2 Visualizations - The Master List*. Retrieved from r-statistics.co: *http://r-statistics.co/Top50-Ggplot2-Visualizations- MasterList-R-Code.html*

R, P. R. (n.d.). *Understanding statistics in the behavioral sciences*. Cengage Learning.

Sapsford, R., & Jupp, V. (Eds.). (2006). *Data collection and analysis*. Sage. doi:10.4135/9781849208802

Storey, V. C., & Song, I. Y. (2017). *Big data technologies and management: What conceptual modeling can do*. *Data & Knowledge Engineering*, 108, 50–67. doi:10.1016/j.datak.2017.01.001

TIOBE The software Quality Company. (2018, June). Retrieved from *https://www.tiobe.com/tiobe-index/*

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). *How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study*. *International Journal of Production Economics*, 165, 234–246. doi:10.1016/j.ijpe.2014.12.031

Wickham, H. (2014). *Tidy data*. *Journal of Statistical Software*, 59(10), 1–23. doi:10.18637/jss.v059.i10 PMID:26917999

Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). *dplyr: A grammar of data manipulation*. *R package version 0.4*, 3.