International Journal of Medical and Health Research
ISSN: 2454-9142
Received: 19-01-2019; Accepted: 23-02-2019
www.medicalsciencejournal.com
Volume 5; Issue 3; March 2019; Page No. 116-121

A graphical visualisation of cancer data utilizing R statistical software

Immad A Shah[1], Shakeel A Mir[2], Imran Khan[3], Nageena Nazir[4]
[1-4] Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, Jammu and Kashmir, India

Abstract
Medical science is a highly important and an advanced field that employs a variety of quantitative and qualitative methods. There is a widespread evidence of extensive application of statistical methods and software in medical research. R software is now being extensively used in health and medical fields for analyzing, interpreting and visualising medical data and has also contributed significantly in generalizing from the experiences of individual patients to the population at large.

Keywords: medical data, cancer, multivariate analysis, graphics, packages, R software

1. Introduction
Statistical software help researchers design studies, analyze data from medical experiments, decide what data to collect, help interpret the results of medical research. Such software help researchers make sense of the data collected to decide whether a treatment is working or to find factors that contribute to diseases. Medical statisticians with the help of statistical tools and software design and analyse studies to identify the real causes of health issues as distinct from chance variation. Besides the general analysis Multivariate data graphical visualization, as a specific type of information visualization, is also being carried out using statistical software and is an active research field with numerous applications in diverse areas ranging from science, medical communities in which the correlations between many attributes are of vital interest (Immad, 2018). The more appropriate term for multivariate data visualization should be multidimensional multivariate data visualization (Hoffman, 2001).

Both clinical and statistical reasoning are crucial to progress in medical science. Clinical researchers must generalize from the few to many and combine empirical evidence with theory. Empirical knowledge in both medical and statistical sciences is generated from observations and data. Medical theory is based upon established biology and hypotheses while as statistical theory is derived from mathematical and probabilistic models. (Piantadosi, 2005). To establish a hypothesis requires both a theoretical basis in biology and statistical support for the hypothesis, based on the observed data and the theoretical statistical model.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices. It is a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities. R is a relatively new and freely available programming language and software environment for statistical computing and graphics. The name is partly based on the (first) names of the two R authors (Robert Gentleman and Ross Ihaka), and concept being partly taken from the name of the Bell Labs language 'S'. It compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS.

2. Material and Methods
In the present study secondary data on the number and proportion of cancers by site ICD-10 and method of diagnosis in Delhi for the year 2012 obtained from Indian council of Medical Research database (ICMR,2012) has been used. The data set comprised of 51 cancer sites and 6 methods of detection, three of which viz. DCO, Unknown have been assigned to Others group. R was downloaded from www.r-project.org and was used to visualise the data and generate various plots.

3. Results
The first thing, to analyse the multivariate data is to read it into R, and to plot the data. The data can read into R using the read. Table ( ) function. To read in the top portion of the data set the function head ( ) is used which displays just the upper few observations from the entire data set.

> head (data)

The result of this function is shown below:

```
> ICD10=read.table("clipboard",header=T)
> head(ICD)
```

	ICD010	Site	Microscopic	X0ray	Imaging	Clinical	Others
1	C00	Lip	29	1	21	0	
2	C010C02	Tongue	606	10	41	2	
3	C030C06	Mouth	643	16	41	3	
4	C070C08	SalGland	68	2	5	0	
5	C09	Tonsil	145	3	4	1	
6	C10	Oropharynx	101	5	6	0	

This gives a clear understanding of how our dataset looks like. First and the second column comprises of the various ICD's and sites of cancer in males respectively. The subsequent columns comprise of the various cancer detection methods. Once the multivariate data set is read into R, the next step is usually to make a plot of the data.

3.1 Bar plot
A bar chart represents data in rectangular bars with length of

the bar proportional to the value of the variable. R uses the function *ggplot()* function from the ggplot2 package to create bar charts. *"ggplot2"* provides a unified interface and set of options, and special cases required in base graphics. To load the *"ggplot"* package in R the following syntax is used:

```
> library("ggplot2", lib.loc="~/R/win-library/3.4")
```

R can draw both vertical and Horizontal bars in the bar chart. In bar chart each of the bars can be given different colors. The basic syntax to create a bar plot using ggplot2 package in R is:

```
> ggplot(ICD, aes(x=Site, y=Microscopic)) + geom_bar(stat = "identity")
```

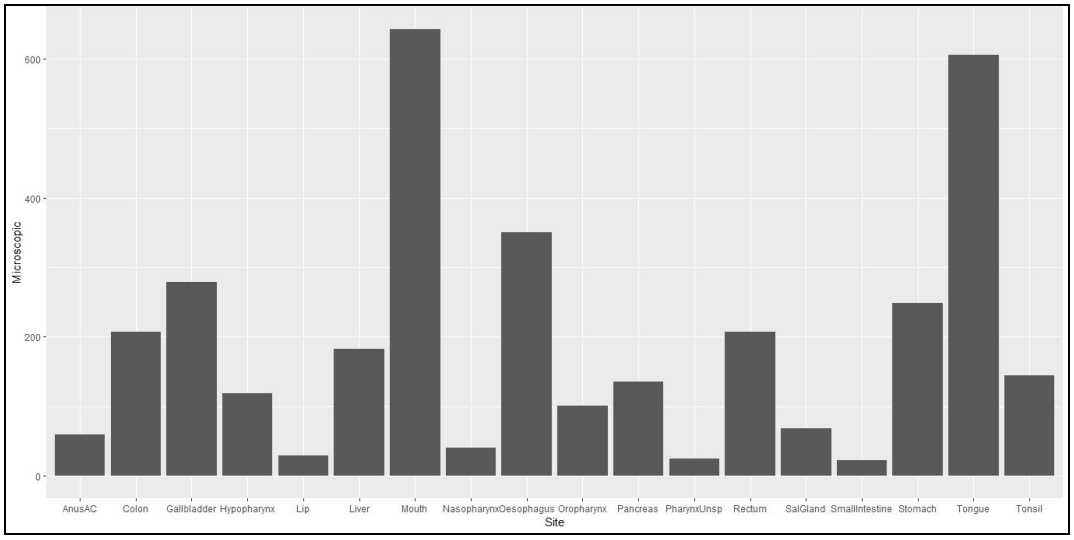

Fig 1: Bar plot.

The x-axis of the plot comprises of the various cancer sites in males and the y-axis comprise of the number of males who go for microscopic detection.
In the above code, we are using the fill attribute in the geom_bar() function to give the bar plot a color.

```
> ggplot(ICD, aes(x=Site, y=Microscopic)) + geom_bar(stat = "identity",fill="aquamarine4")
```

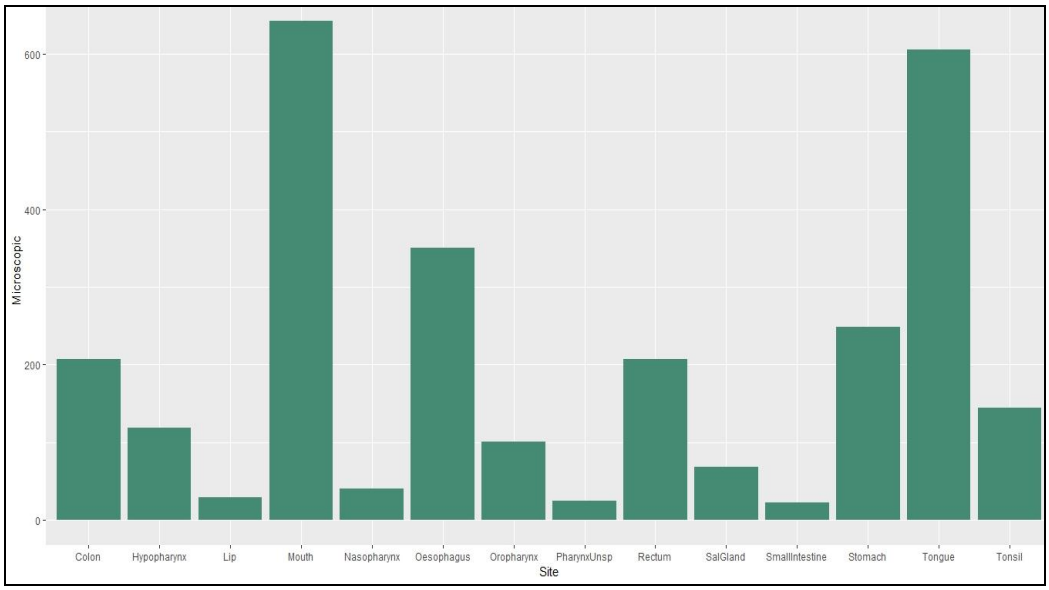

Fig 2: Coloured Barplot

3.2 Horizontal Barplot

A barplot with bars plotted horizontally is known as a Horizontal Barplot. To create the horizontal barplot the data on total number of cancers by site and method of diagnosis for both males and females have been used. The syntax for obtaining a horizontal plot is shown as under:

```
> data=read.table("clipboard",header=T)
> head(data)
```

	Site	Males	Females
1	Lip	51	19
2	Tongue	659	193
3	Mouth	703	213
4	SalGland	75	39
5	Tonsil	153	25
6	Oropharynx	112	13

```
> Total=data
#For males
> x <- Total[1:8, "Males"]
> par(mai=c(1,2,1,1))
> barplot(x, names.arg = c(
"Lip","Tongue","Mouth","SalGland","Tonsil","Oropharynx
","Nasopharynx","Hypopharynx"), horiz = T,las=1)
```

```
#For females
> x <- Total[1:8, "Females"]
> par(mai=c(1,2,1,1))
> barplot(x, names.arg = c(
"Lip","Tongue","Mouth","SalGland","Tonsil","Oropharynx
","Nasopharynx","Hypopharynx"), horiz = T,las=1)
```

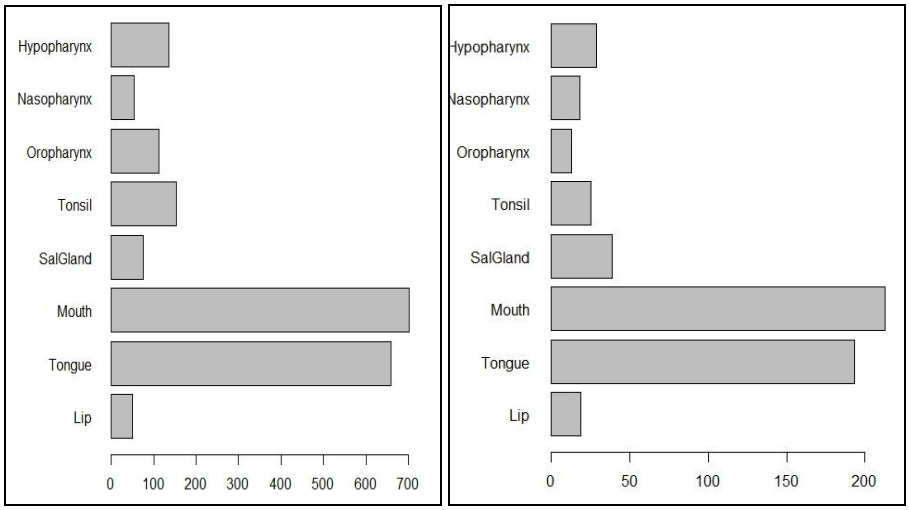

Fig 3: Horizontal Barplot

To specify the plot names the following commands are used:

```
> barplot(x, names.arg = c(
"Lip","Tongue","Mouth","SalGland","Tonsil","Oropharynx
","Nasopharynx","Hypopharynx"),main="males", horiz =
T,las=1)
> barplot(x, names.arg = c(
"Lip","Tongue","Mouth","SalGland","Tonsil","Oropharynx
","Nasopharynx","Hypopharynx"),main="females", horiz =
T,las=1)
```

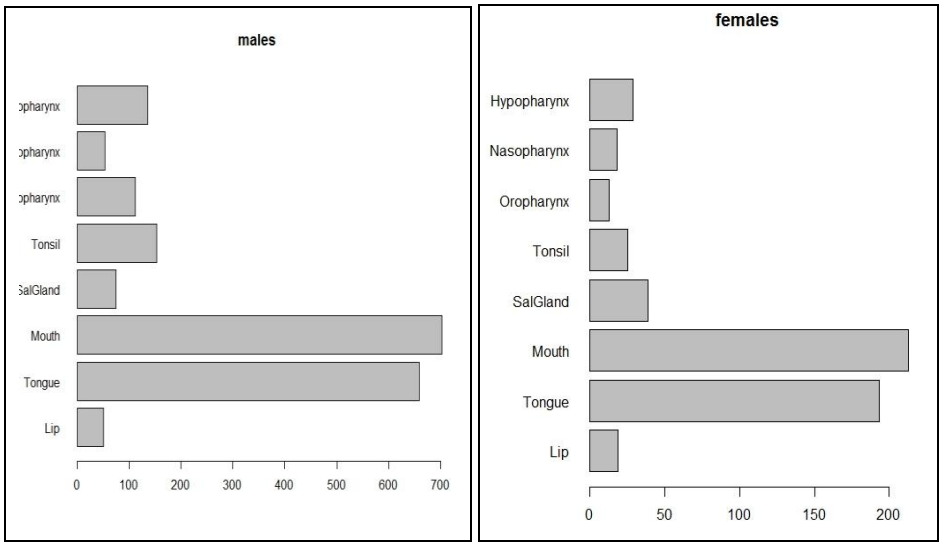

Fig 4

3.3 Stacked barplot

These plot types are a specific sort of barplot. Like grouped barplot, they display a numerical value for several entities, organised into groups and subgroups. To create a stacked barplot the data on Population by Five Year Age Group and Gender: 2012 in Delhi has been used. The syntax for creating stacked plots in R is:

```
> Age=read.table("clipboard",header=T)
> head(Age)
```

	AgeGroup	Sex	Total
1	00-04	M	718477
2	05-Sep	M	812427
3	Oct-14	M	884776
4	15-19	M	921666
5	20-24	M	953086
6	25-29	M	882401

```
> library("ggplot2", lib.loc="~/R/win-library/3.4")
> install.packages("Rcpp")
```

Error in install.packages : Updating loaded packages
> install.packages("Rcpp")

> ggplot(data=data, aes(x=AgeGroup, y=Total, fill=Sex)) + geom_bar(stat="identity")

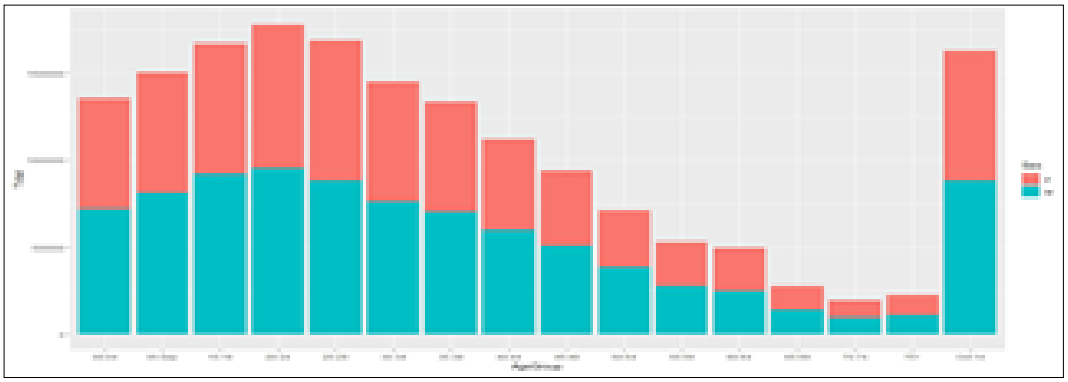

Fig 5: Stacked Barplot (a)

To create the side ways stacked barplot the following syntax is used:

> ggplot(data=data, aes(x=AgeGroup, y=Total, fill=Sex)) + geom_bar(stat="identity", position=position_dodge())

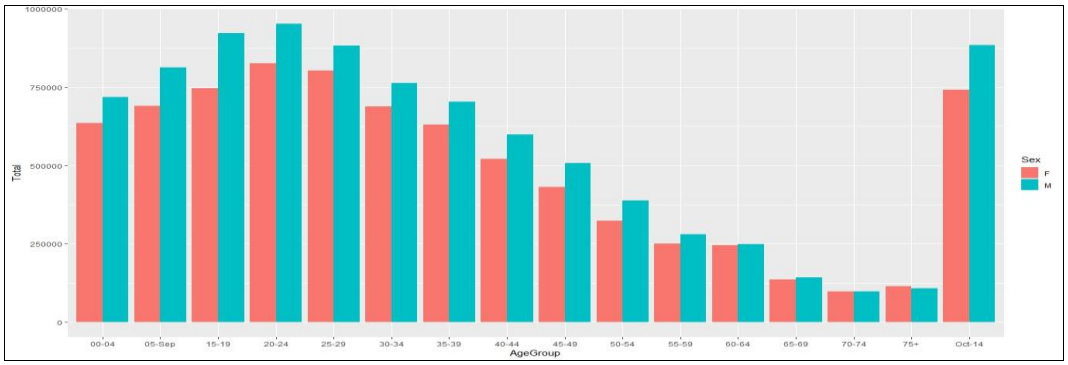

Fig 6: Stacked Barplot (b)

Also to make a color custom barplot the following syntax is used:

> ggplot(data=data, aes(x=AgeGroup, y=Total, fill=Sex)) + geom_bar(stat="identity", position=position_dodge(),colour="black") +
+ scale_fill_manual(values=c("#999999", "#E69F00"))

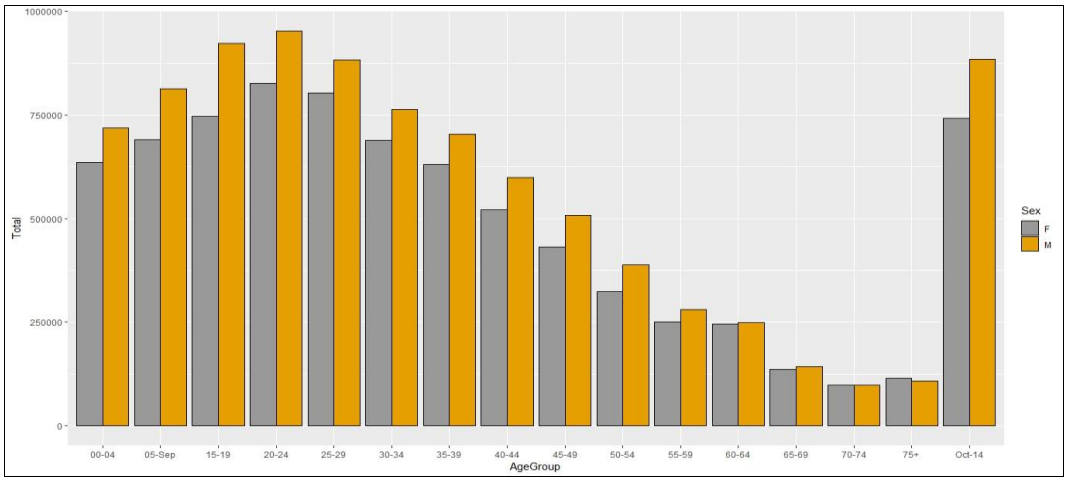

Fig 7: Coloured Stacked Barplot (b)

3.4 Line Plot

To draw multiple lines, the points must be grouped by a variable; otherwise all points will be connected by a single line. In this case, the data has been grouped by sex. The following syntax is used to create a line plot:

> ggplot(data=Age, aes(x=AgeGroup, y=Total, group=Sex, colour=Sex)) + geom_line(size=1.5) +
+ geom_point(size=3, fill="white") +
+ scale_shape_manual(values=c(22,21))

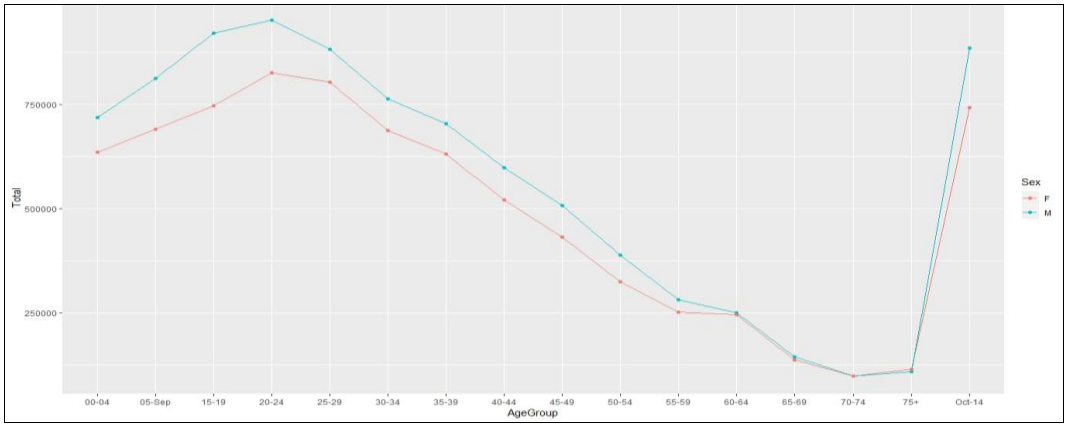

Fig 8: Line plot

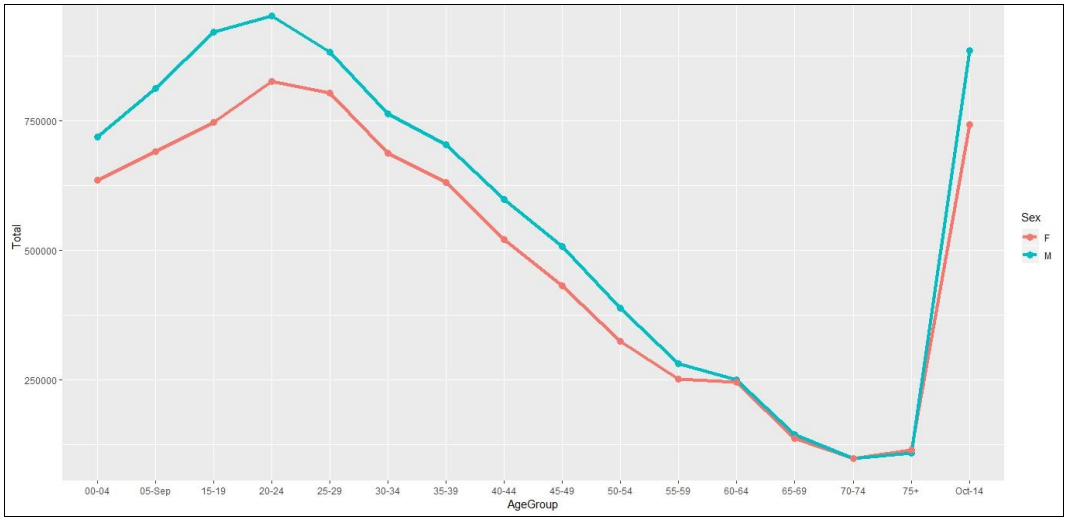

Fig 9: Line plot with thicker lines

3.5 Stacked line graph

To understand how the graph is made, it's useful to see how the data is structured. There are 51 cancer sites and for each site we have the number of males with undergo microscopic and clinical detection. To obtain a stacked bar graph the following command is used:

```
> ggplot(ICD10, aes(x=Microscopic, y=Clinical, fill=Site))+geom_bar(stat="identity")+guides(fill=guide_legend(reverse=TRUE))
```

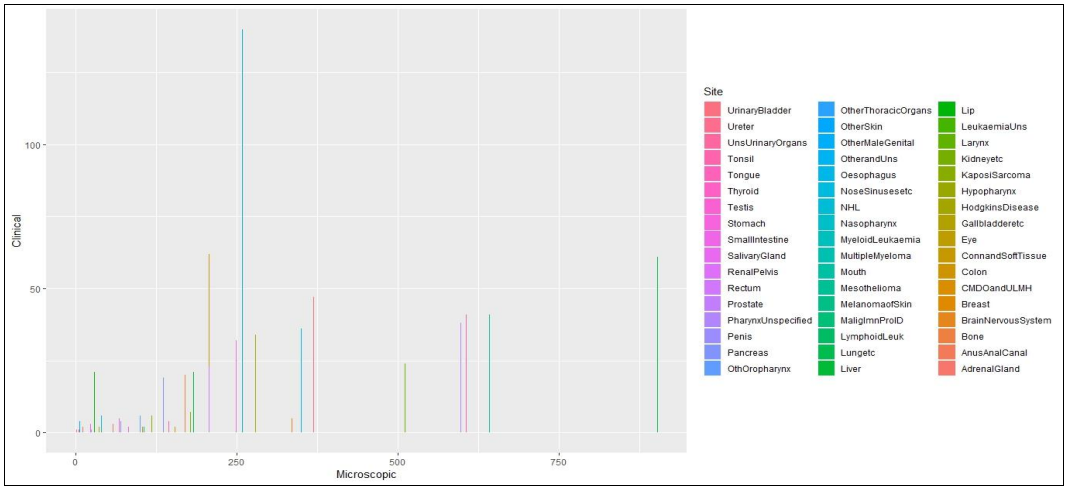

Fig 6: Stacked Bar Plot.

4. Conclusion

Multivariate data visualization is strongly motivated by the many situations while trying to obtain an integrated understanding of the data distributions and investigate the inter-relationships between different data attributes. The primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics and plots. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about

data and evidence. Such an effective visual display tool is demanded to facilitate users to identify, locate, distinguish, categorize, cluster, rank, compare, associate or correlate the underlying data. Graphics can effectively complement statistical data analysis in various ways. R commands mentioned can be helpful to researchers for data visualisation.

5. References

1. Frequently asked questions on R. Kurt Hornik. Available from: http:// cran. r-project. org/doc/FAQ/R-FAQ.html#Why-is-R-names-R003f.
2. Core Team R. A language and enivironment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org, 2013.
3. Sha. IA, Bhat O, Yousuf S. Graphical Visualisation of Multivariate dataset. International conference on Recent researches and innovations in Science, Management, Education and Technology, 2018, (ICRRISMET) ISBN- 987-93-87793-16-3.
4. Piantadosi S. Clinical Trials: A Methodologic Perspective, 2005, Second Edition. DOI: 10. 1002/ 0471740136
5. Hoffman PE, Grinstein GG. A Survey of Visualizations for High-Dimensional Data Mining (Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers, 2001, 47-82.
6. Indian Council of Medical Research Database, www.ncdirindia.org/NCRP/ALL_NCRP_REPORTS, 2012, Delhi_Ann.pdf)