Exploratory Data Analysis

(Using R programming)

SUBMITTED TO:

Prof. VASUDHA BHATNAGAR

SUBMITTED BY:

SAKSHI (144)

Prog_Category_Wise_Education.csv

About the Dataset

- The dataset has been taken from https://data.gov.in/ .
- It consists of category wise number of students (Male and Female) enrolled in different programmes/courses.
- It contains 189 rows and 17 columns.
- It consists of following attributes - Programme, TotalMale, TotalFemale, SCMale, SCFemale, STMale, STFemale, OBCMale, OBCFemale, PWDMale, PWDFemale, MuslimMale, MuslimFemale, MinorityMale, MinorityFemale, ForeignMale, ForeignFemale.
- Last row of the data, contains the category wise total number of students enrolled in all the programmes.
- There are some rows with NA values in the data too.

Purpose

The purpose of choosing this dataset is to analyze the variation of category-wise enrollment of students in different programmes and visualize those using different plots. Visualization is used to highlight the useful information.

Plots Used for Visualization

- Scatter plot
- Bar plot
- Pie Chart / 3-D Pie Chart
- Histogram
- Box plot
- Time series plot
- Density Plot

Reading Data

```
df=read.csv("F:\\Maths\\data.csv",header = TRUE)
```

View Data

```
View(head(df,10))
```

	Programme	TotalMale	TotalFemale	SCMale	SCFemale	STMale	STFemale	OBCMale	OBCFemale
1	Acharya-Acharya	13213	8675	680	743	194	142	1938	2699
2	Alankar-Alankar	2457	1805	518	291	3	0	91	55
3	A.N.M.-Auxiliary Nurse and Midwife	155	6983	13	1101	32	565	54	1981
4	Ayurvedacharya-Ayurvedacharya	8281	8256	388	489	238	316	1091	1428
5	Ayurveda Vachaspati-Ph.D in Ayurveda	72	46	1	4	0	2	7	3
6	B.A.-Bachelor of Arts	3583157	3928613	608013	608713	261274	235488	1137716	1312400
7	B.A. B.Ed.-Bachelor of Arts, Bachelor of Education	1545	2212	240	172	222	210	390	652
8	B.Agri.-Bachelor of Agriculture	48656	15118	6868	1626	1953	836	18539	4970
9	B.A.(Hons)-Bachelor of Arts (Honors)	421153	461644	60330	54389	30393	35254	87478	89551
10	B.A. L.L.B.-Bachelor of Arts, Bachelor of Law or Laws	16062	10124	2028	982	528	383	2905	1586

PWDMale	PWDFemale	MuslimMale	MuslimFemale	MinorityMale	MinorityFemale	ForeignMale	ForeignFemale
218	5	228	24	70	26	51	0
1	1	359	277	0	0	0	0
0	8	0	26	0	219	0	0
14	13	86	105	8	17	25	29
0	0	0	0	0	0	2	1
8657	12694	166826	220540	34235	50226	1586	1037
60	83	90	246	105	75	0	0
55	21	627	351	119	162	43	2
1228	793	43519	41049	5744	8335	169	260
57	27	1017	623	319	280	76	69

Data dimensions

```
dim(df)
```

```
[1] 189  17
```

Data description

```
View(summary(df))
```

	Var1	Var2	Freq
1		Programme	Length:189
2		Programme	Class :character
3		Programme	Mode :character
4		Programme	NA
5		Programme	NA
6		Programme	NA
7		Programme	NA
8		TotalMale	Min. : 1
9		TotalMale	1st Qu.: 240
10		TotalMale	Median : 1207
11		TotalMale	Mean : 150042
12		TotalMale	3rd Qu.: 9503
13		TotalMale	Max. :14178978
14		TotalMale	NA
15		TotalFemale	Min. : 2
16		TotalFemale	1st Qu.: 210
17		TotalFemale	Median : 1043
18		TotalFemale	Mean : 122568
19		TotalFemale	3rd Qu.: 8256
20		TotalFemale	Max. :11582708
21		TotalFemale	NA

22		SCMale	Min. : 0.0
23		SCMale	1st Qu.: 20.0
24		SCMale	Median : 111.5
25		SCMale	Mean : 18910.2
26		SCMale	3rd Qu.: 1059.0
27		SCMale	Max. :1777555.0
28		SCMale	NA's :1
29		SCFemale	Min. : 0
30		SCFemale	1st Qu.: 17
31		SCFemale	Median : 72
32		SCFemale	Mean : 15417
33		SCFemale	3rd Qu.: 744
34		SCFemale	Max. :1449185
35		SCFemale	NA's :1
36		STMale	Min. : 0.0
37		STMale	1st Qu.: 5.0
38		STMale	Median : 31.5
39		STMale	Mean : 6568.9
40		STMale	3rd Qu.: 290.8
41		STMale	Max. :617477.0
42		STMale	NA's :1

43		STFemale	Min. : 0.0
44		STFemale	1st Qu.: 4.0
45		STFemale	Median : 23.0
46		STFemale	Mean : 5358.0
47		STFemale	3rd Qu.: 207.8
48		STFemale	Max. :503649.0
49		STFemale	NA's :1
50		OBCMale	Min. : 0
51		OBCMale	1st Qu.: 41
52		OBCMale	Median : 246
53		OBCMale	Mean : 46090
54		OBCMale	3rd Qu.: 2324
55		OBCMale	Max. :4332479
56		OBCMale	NA's :1
57		OBCFemale	Min. : 0
58		OBCFemale	1st Qu.: 38
59		OBCFemale	Median : 160
60		OBCFemale	Mean : 39536
61		OBCFemale	3rd Qu.: 1609
62		OBCFemale	Max. :3716356
63		OBCFemale	NA's :1

64		PWDMale	Min. : 0.0
65		PWDMale	1st Qu.: 0.0
66		PWDMale	Median : 2.0
67		PWDMale	Mean : 394.1
68		PWDMale	3rd Qu.: 22.0
69		PWDMale	Max. :37245.0
70		PWDMale	NA
71		PWDFemale	Min. : 0.0
72		PWDFemale	1st Qu.: 0.0
73		PWDFemale	Median : 1.0
74		PWDFemale	Mean : 319.3
75		PWDFemale	3rd Qu.: 9.0
76		PWDFemale	Max. :30174.0
77		PWDFemale	NA
78		MuslimMale	Min. : 0
79		MuslimMale	1st Qu.: 1
80		MuslimMale	Median : 19
81		MuslimMale	Mean : 6159
82		MuslimMale	3rd Qu.: 234
83		MuslimMale	Max. :582058
84		MuslimMale	NA

85		MuslimFemale	Min. : 0
86		MuslimFemale	1st Qu.: 1
87		MuslimFemale	Median : 17
88		MuslimFemale	Mean : 5447
89		MuslimFemale	3rd Qu.: 160
90		MuslimFemale	Max. :514719
91		MuslimFemale	NA
92		MinorityMale	Min. : 0.0
93		MinorityMale	1st Qu.: 0.0
94		MinorityMale	Median : 14.0
95		MinorityMale	Mean : 2328.7
96		MinorityMale	3rd Qu.: 126.2
97		MinorityMale	Max. :218900.0
98		MinorityMale	NA's :1
99		MinorityFemale	Min. : 0.0
100		MinorityFemale	1st Qu.: 1.0
101		MinorityFemale	Median : 17.5
102		MinorityFemale	Mean : 2709.7

103		MinorityFemale	3rd Qu.: 162.8
104		MinorityFemale	Max. :254715.0
105		MinorityFemale	NA's :1
106		ForeignMale	Min. : 0.0
107		ForeignMale	1st Qu.: 0.0
108		ForeignMale	Median : 3.0
109		ForeignMale	Mean : 225.6
110		ForeignMale	3rd Qu.: 25.0
111		ForeignMale	Max. :21321.0
112		ForeignMale	NA
113		ForeignFemale	Min. : 0.0
114		ForeignFemale	1st Qu.: 0.0
115		ForeignFemale	Median : 2.0
116		ForeignFemale	Mean : 134.2
117		ForeignFemale	3rd Qu.: 16.0
118		ForeignFemale	Max. :12684.0
119		ForeignFemale	NA

Handling missing values

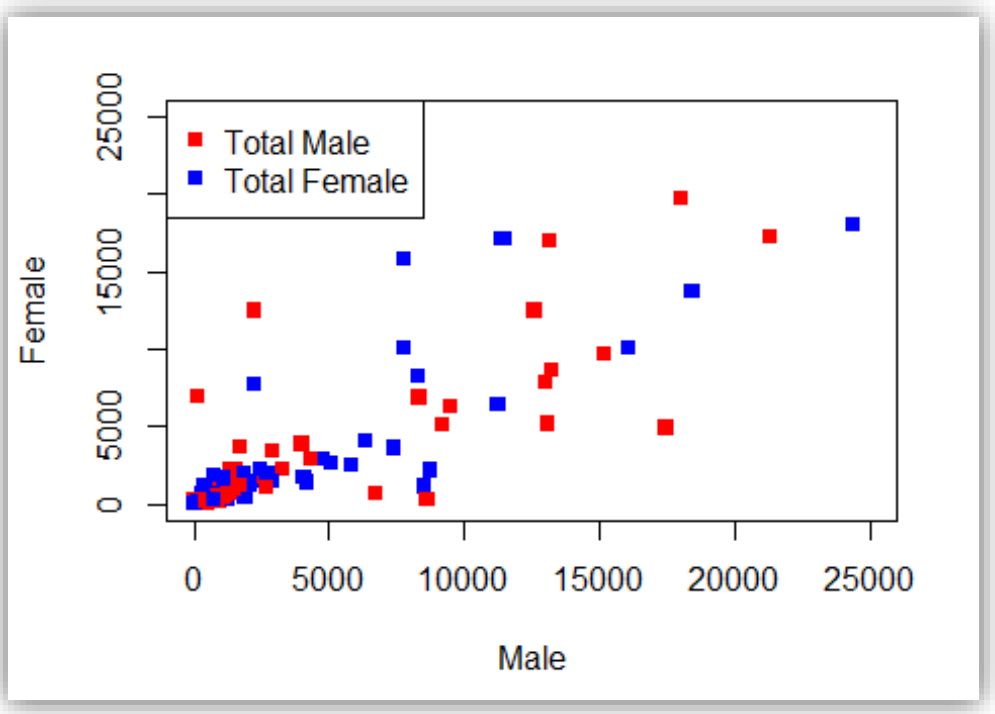
```
#find out how many NA values are there in the whole dataset
print(sum(is.na(df)))
[1] 8
#Replacing NA values with 0
df[is.na(df)]=0
#After Replacing, Total No. of NA values are there in the whole dataset
print(sum(is.na(df)))
[1] 0
```

Scatter Plot

```
#To display scatter plot of Total Male and Total Female in different programmes

plot(df$TotalMale,df$TotalFemale,xlab="Male",ylab="Female",col=c("red","blue"),pch=c(15,15),xlim=c(0,25000),ylim=c(0,25000))

legend("topleft",c("TotalMale","TotalFemale"),col=c("red","blue"), pch=c(15, 15))
```



Description

Scatter plot is used to observe how the attribute values are distributed in the dataset.

Interpretation

It is observed that most of the values of "TotalMale" and "TotalFemale" are lying in the range [0, 5000] for different "Programme".

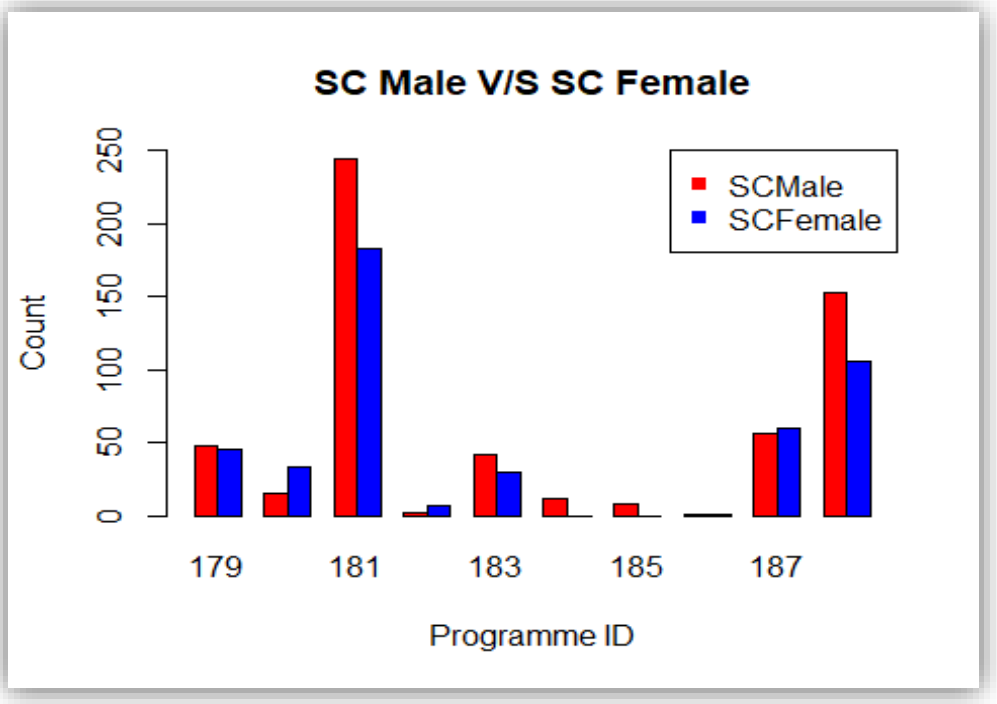
Bar Plot

```
#Bar plot comparing SC-Male and Sc-Female in different programmes
subdata= tail(df,11)

#To remove last row from consideration
subdata=subdata[-nrow(subdata),]

barplot(t(subdata[c('SCMale','SCFemale')]),xlab="ProgrammeID",ylab="Count", beside = TRUE, col=c("red","blue"), ylim=c(0,250), main='SC Male V/S SC Female')

legend("topright", c("SCMale","SCFemale"),col=c("red","blue"),pch=c(15,15))
```



Description

Bar plot is used to compare attribute values for one or more attributes in the data.

Interpretation

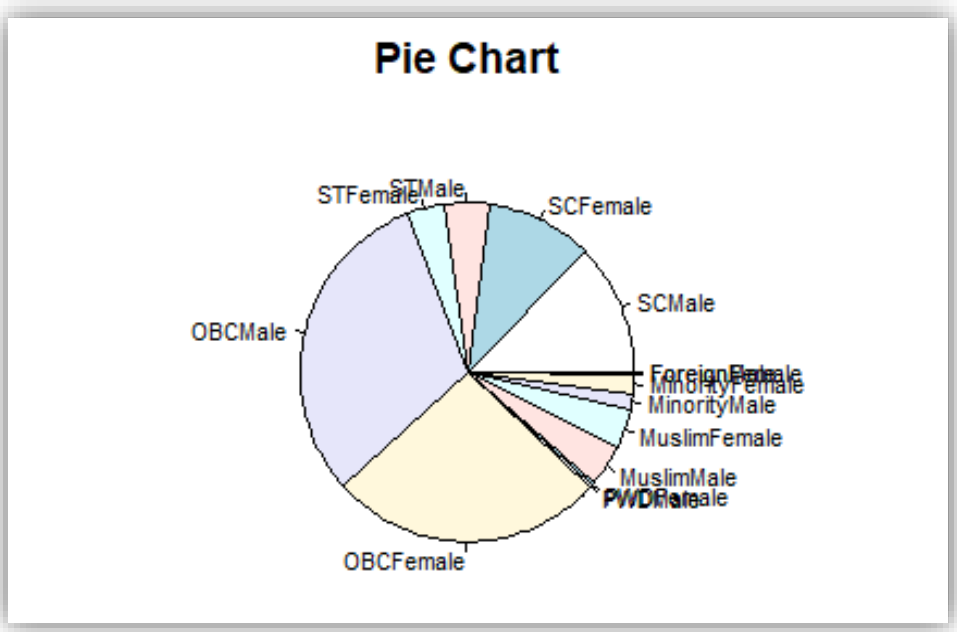
Here, we are comparing number of enrollments of "SCMale" and "SCFemale" in 10 programmes. From the plot, we observed that number of female enrollment is less as compared to number of male enrollments except for the programmes with ID=180, 182 and 187. That means those are programmes in which female students are less interested in.

Pie Chart

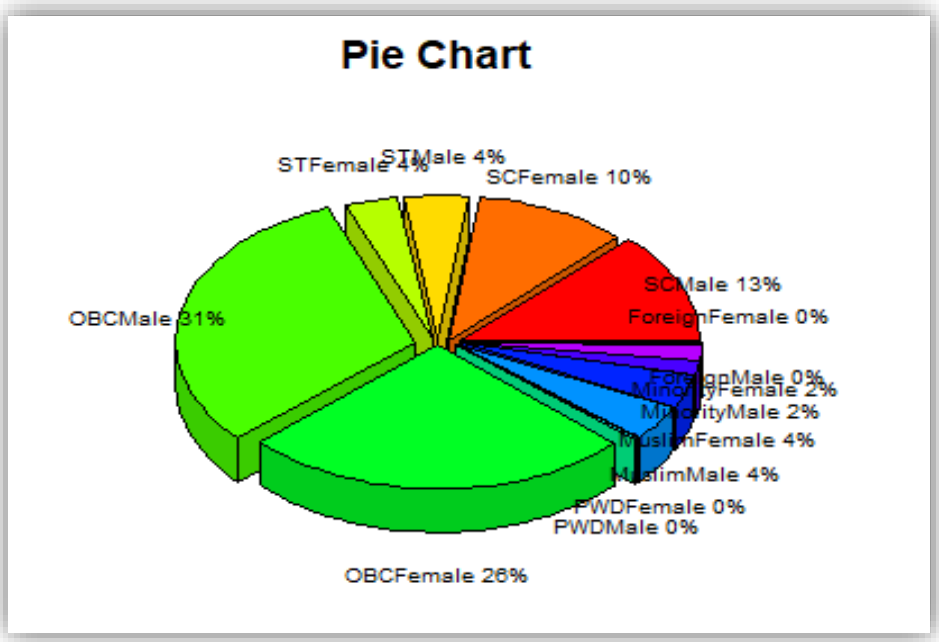
```
#Pie Chart
slices = df[nrow(df),4:17]
lbls=names(df[0, 4:17])
pie(x=t(slices), labels = lbls, main="Pie Chart",radius=0.8,cex=0.7)
```

3-D Pie Chart

```
library(plotrix)
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep=" ") # add % to labels
pie3D(x=t(slices),labels=lbls,main="PieChart",explode=0.1,theta=pi/3,radius=1,labelcex=0.6,)
```



(PIE CHART)



(3-D PIE CHART)

Interpretation

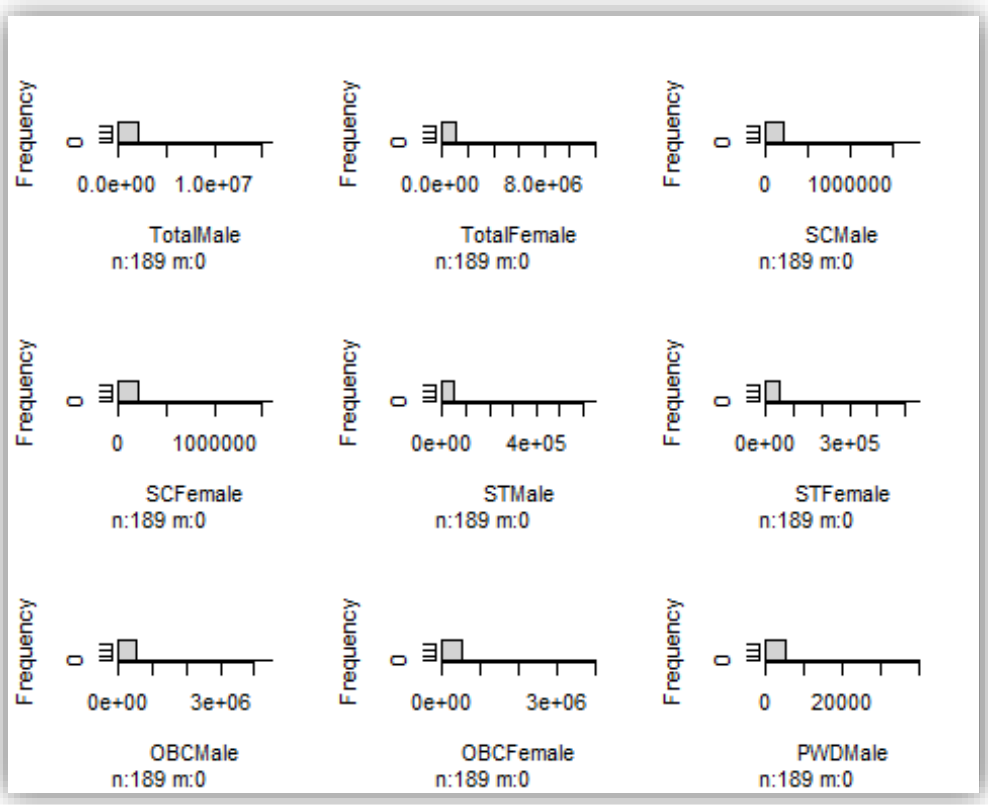
The size of each slice shows the category-wise proportion of students enrolled in all programmes. It is observed that –

- The maximum number of enrollments are for "OBCMale" (31%) and "OBCFemale" (26%).
- The minimum number of enrollments are for "PWDMale" (0%), "PWDFemale" (0%), "ForeignMale" (0%), "ForeignFemale" (0%)

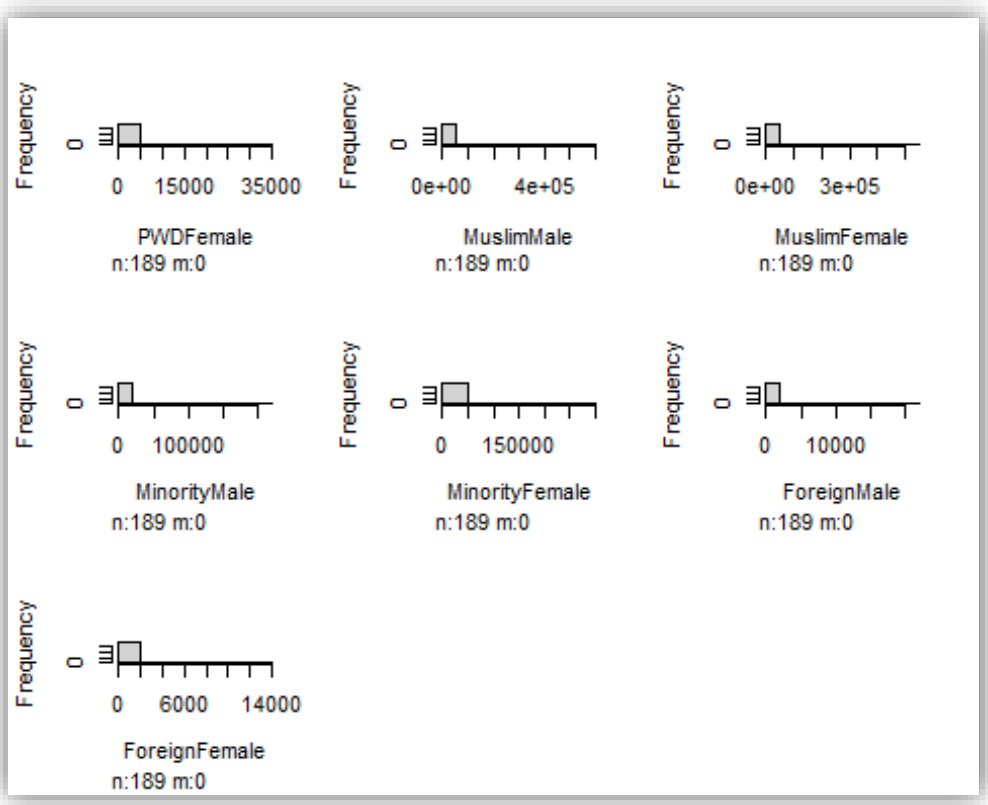
NOTE - All percentages are round figures **

Histogram

```
# histogram of all attributes
library(Hmisc)
d1=df[1:189,2:10]
#print(d1)
hist.data.frame(d1)
```




```
d2=df[1:189,11:17]
#print(d2)
hist.data.frame(d2)
```



Description

The function tries to compute the maximum number of histograms that will fit on one page, then it draws a matrix of histograms. If there are more qualifying variables than will fit on a page, the function waits for a mouse click before drawing the next page.

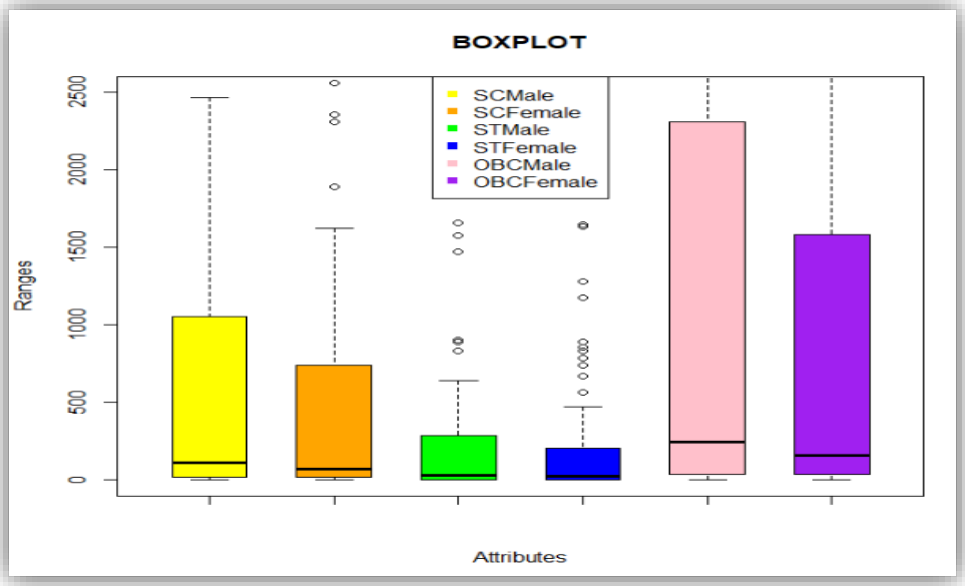
Interpretation

The data is right – skewed.

A right-skewed distribution: A right-skewed distribution is also called a positively skewed distribution. In a right-skewed distribution, a large number of data values occur on the left side with a fewer number of data values on the right side.

Box Plot

```
# plots with time series
boxplot(df$SCMale,df$SCFemale,df$STMale,df$STFemale,df$OBCMale,df$OBCFemale,
main="BOXPLOT", ylab="Ranges", ylim=c(0,2500), xlab="Attributes", boxwex=0.6,
col=c("yellow","orange","green","blue","pink","purple"))
legend("top",c("SCMale","SCFemale","STMale","STFemale","OBCMale","OBCFemale"),col=c("yellow","orange","green","blue","pink","purple"),pch=15)
```



Description

A boxplot gives a nice summary of one or more numeric variables. A boxplot is composed of several elements:

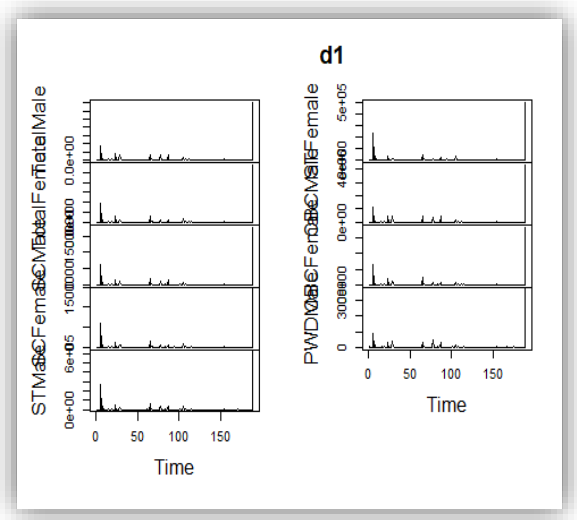
- The line that divides the box into 2 parts represents the median of the data. If the median is 10, it means that there are the same number of data points below and above 10.
- The ends of the box shows the upper (Q3) and lower (Q1) quartiles. If the third quartile is 15, it means that 75% of the observation are lower than 15.
- The difference between Quartiles 1 and 3 is called the interquartile range (IQR)
- Dots (or other markers) beyond the extreme line shows potential outliers.

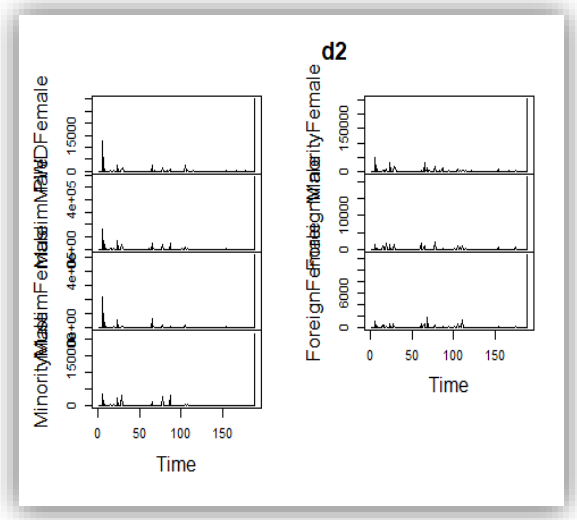
Interpretation

- "STMale" and "STFemale" have lowest number of enrollments in all the programmes.
- It is also observed that the half of the attribute values are closer to 0 in all the boxplots.

Plots with time series

```
# plots with time series
plot.ts(d1)
# plots with time series
plot.ts(d2)
```





Description

Time series plots are an excellent way to begin the process of understanding what sort of process might have generated the data of interest. Traditionally, time series have been plotted with the observed data count on the y-axis and time on the x-axis.

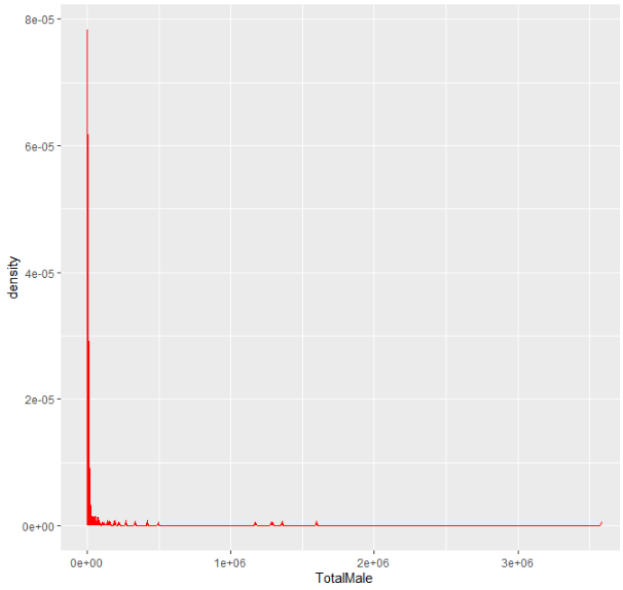
Note that our data is stored in R as a dataframe object, but for the plot the class is transformed to a more user-friendly format for dealing with time series. Fortunately, the ts() function will do just that, and return an object of class **ts** as well.

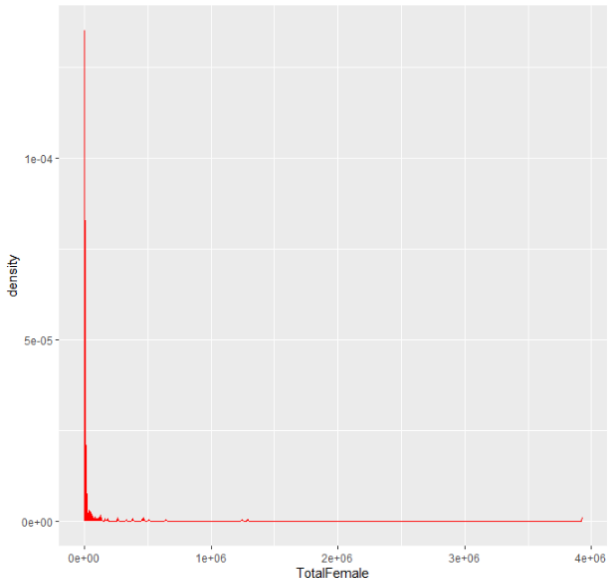
Interpretation

It can observed the values for each attribute is increasing and then decreasing while moving from first row to last row.

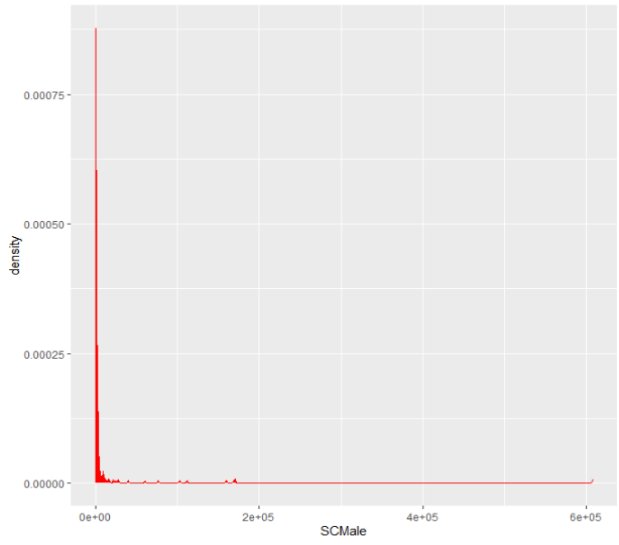
Density Plots

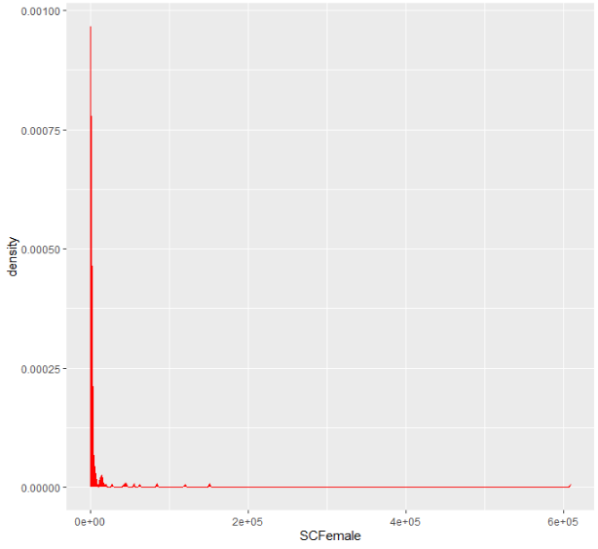
```
ggplot(data=df,aes(x=TotalMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=TotalFemale))+geom_density(fill='red',color='red')
```



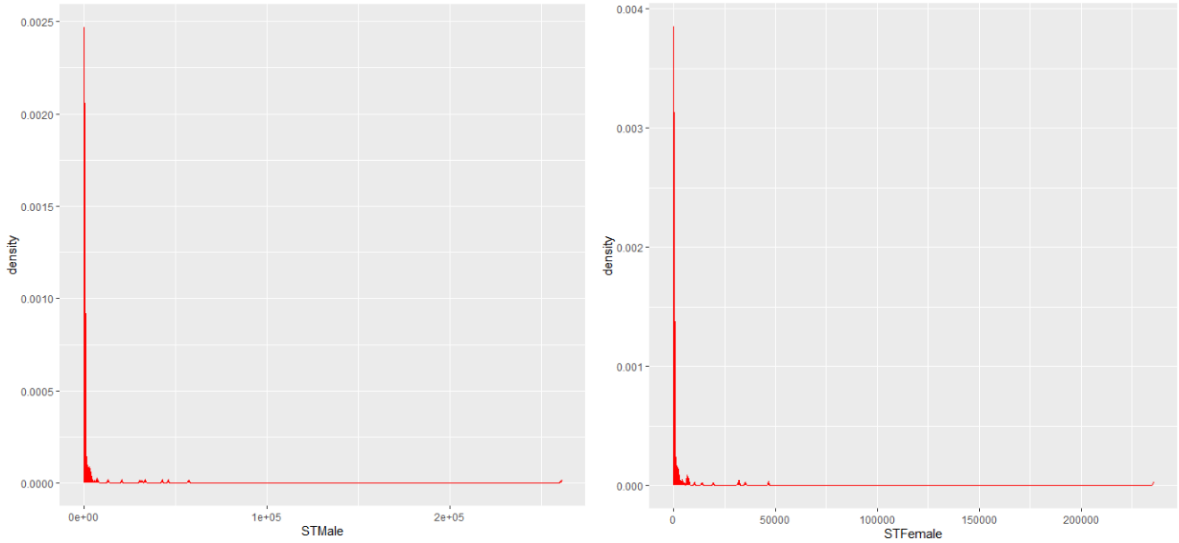



```
ggplot(data=df,aes(x=SCMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=SCFemale))+geom_density(fill='red',color='red')
```

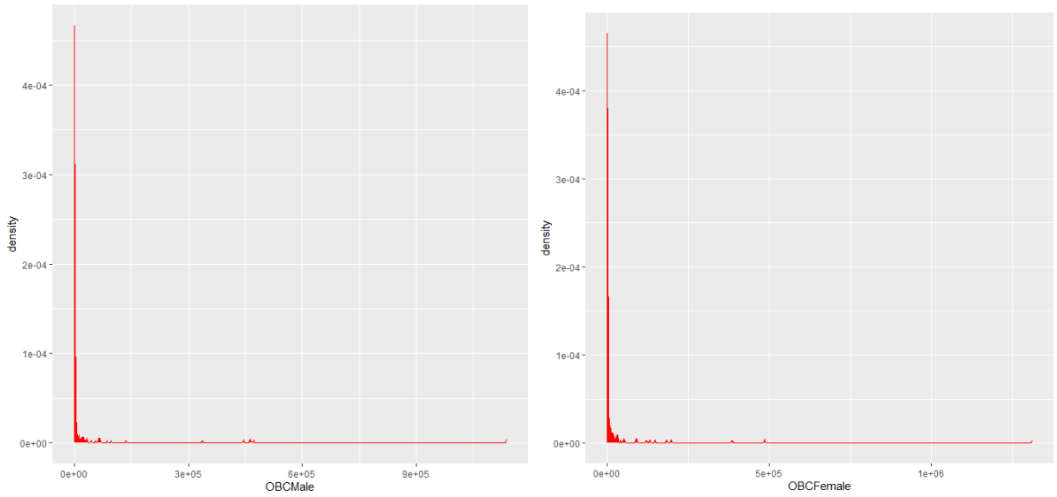




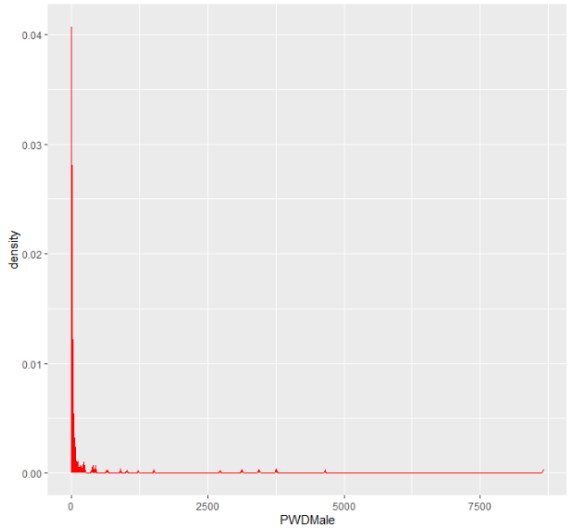

```
ggplot(data=df,aes(x=STMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=STFemale))+geom_density(fill='red',color='red')
```

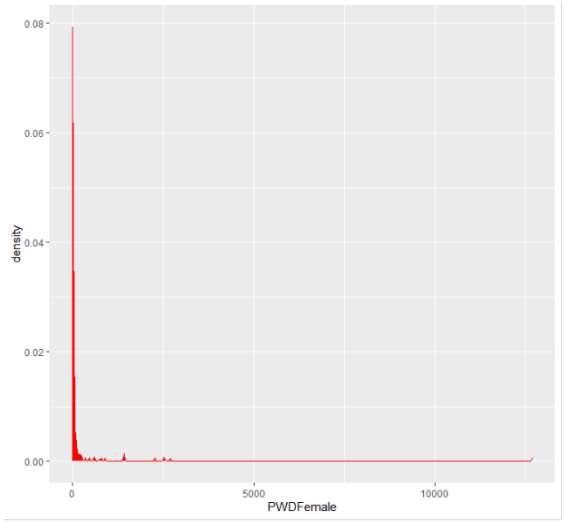



```
ggplot(data=df,aes(x=OBCMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=OBCFemale))+geom_density(fill='red',color='red')
```

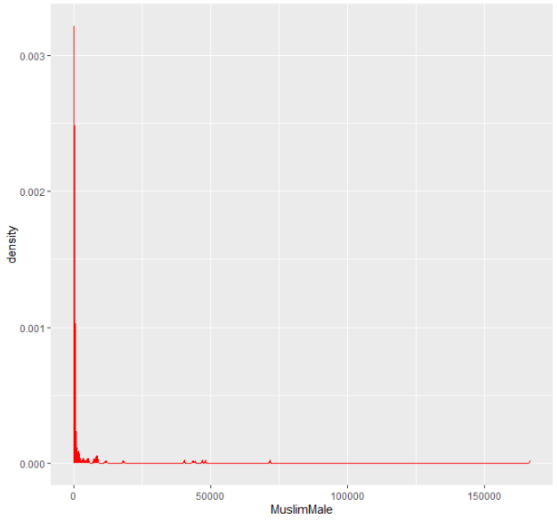


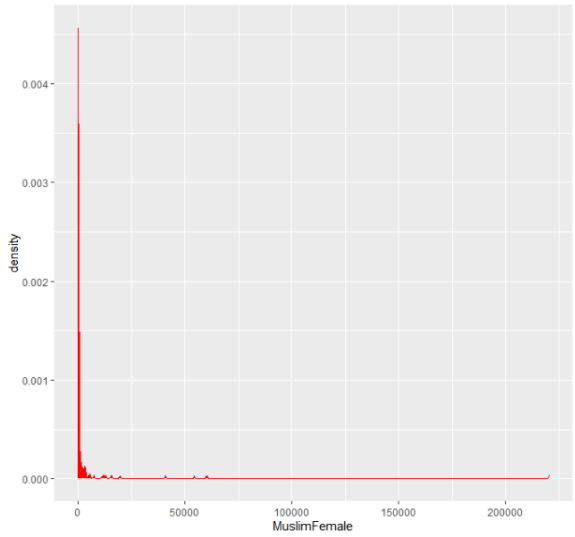

```
ggplot(data=df,aes(x=PWDMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=PWDFemale))+geom_density(fill='red',color='red')
```



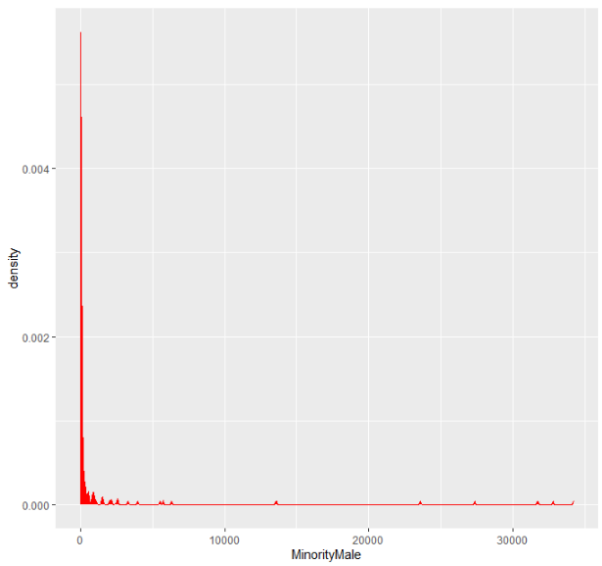


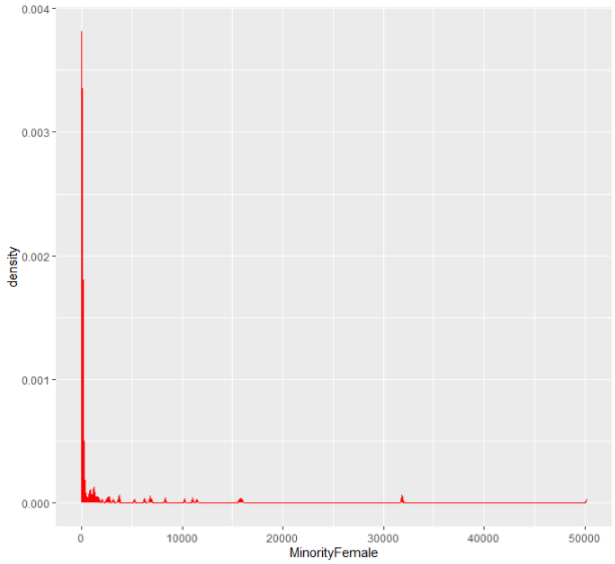

```
ggplot(data=df,aes(x=MuslimMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=MuslimFemale))+geom_density(fill='red',color='red')
```



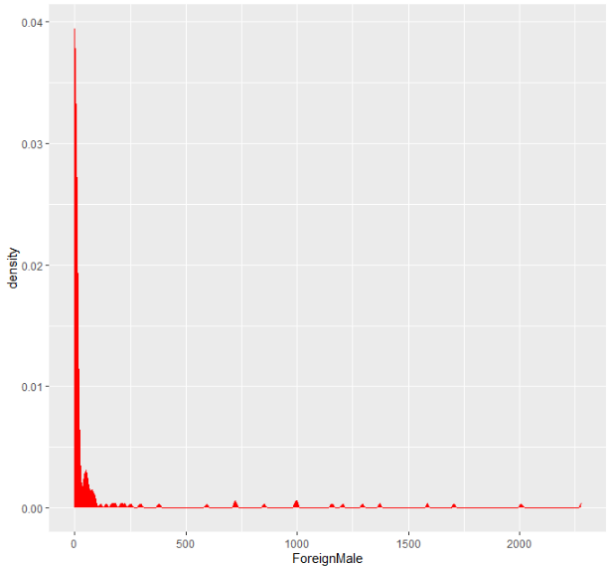


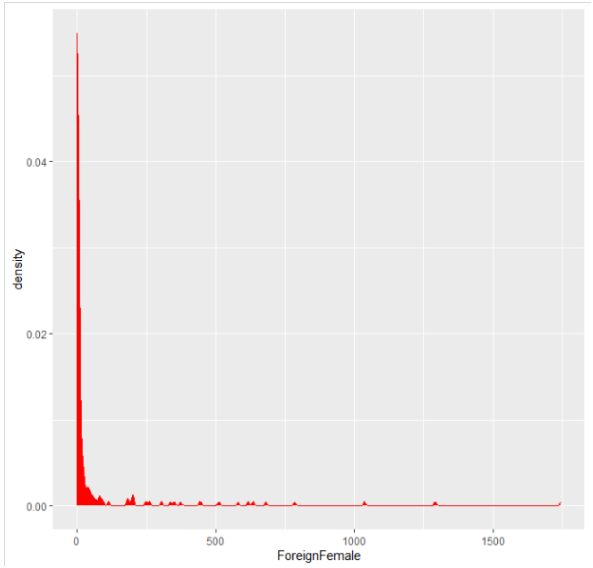

```
ggplot(data=df,aes(x=MinorityMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=MinorityFemale))+geom_density(fill='red',color='red')
```






```
ggplot(data=df,aes(x=ForeignMale)) +geom_density(fill='red',color='red')
ggplot(data=df, aes(x=ForeignFemale))+geom_density(fill='red',color='red')
```





Description

A density plot is a representation of the distribution of a numeric variable. It shows the probability density function of the variable.

It is a smoothed version of the histogram and is used in the same concept.

Interpretation

All attributes in our data are Right skewed.