Exploratory Data Analysis Report

2019-03-15

Exploratory Data analysis (EDA)

Analyzing the data sets to summarize their main characteristics of variables, often with visual graphs, without using a statistical model.

1. Overview of the data

Understanding the dimensions of the dataset, variable names, overall missing summary and data types of each variables

```
# Overiview of the data
ExpData(data=data,type=1)
# Structure of the data
ExpData(data=data,type=2)
```

Overview of the data

Descriptions <fctr>	**Obs** <fctr>
Sample size (Nrow)	1000
No. of Variables (Ncol)	9
No. of Numeric Variables	2
No. of Factor Variables	3
No. of Text Variables	2
No. of Logical Variables	2
No. of Date Variables	0
No. of Zero variance Variables (Uniform)	0
%. of Variables having complete cases	55.56% (5)
%. of Variables having <50% missing cases	44.44% (4)

1-10 of 12 rows Previous **1** 2 Next

Structure of the data

S.no	Variable Name	Variable Type	% of Missing	No. of Unique values

9 rows

Target variable

Summary of categorical dependent variable

1. Variable name - **Died**
2. Variable description - ****

```
##    Died Frequency Descriptions
## 1 FALSE       465          Died
## 2  TRUE       535          Died
```

2. Summary of numerical variables

Summary of all numerical variables

Summary statistics when dependent variable is categorical **Died**. Summary statistics will be splitted into category level

```
ExpNumStat(data,by="GA",gp=Target,Qnt=seq(0,1,0.1),MesofShape=2,Outlier=TRUE,round=2)
```

Vname <fctr>	Group <fctr>	TN <dbl>	n... <dbl>	nZero <dbl>	n... <dbl>	NegInf <dbl>	PosInf <dbl>	NA_Value <dbl>	Per_of_Missing <dbl>
Height_cm_	Died:All	1000	0	0	1000	0	0	0	0.00
Height_cm_	Died:FALSE	465	0	0	465	0	0	0	0.00
Height_cm_	Died:TRUE	535	0	0	535	0	0	0	0.00
IQ	Died:All	1000	0	0	898	0	0	102	10.20
IQ	Died:FALSE	465	0	0	422	0	0	43	9.25
IQ	Died:TRUE	535	0	0	476	0	0	59	11.03

6 rows | 1-10 of 34 columns

3. Distributions of Numerical variables

- Box plots for all numerical variables vs categorical dependent variable - Bivariate comparision only with categories
- Quantile-quantile plot(Univariate)
- Density plot (Univariate)
- Box plot (univariate and Bivariate)

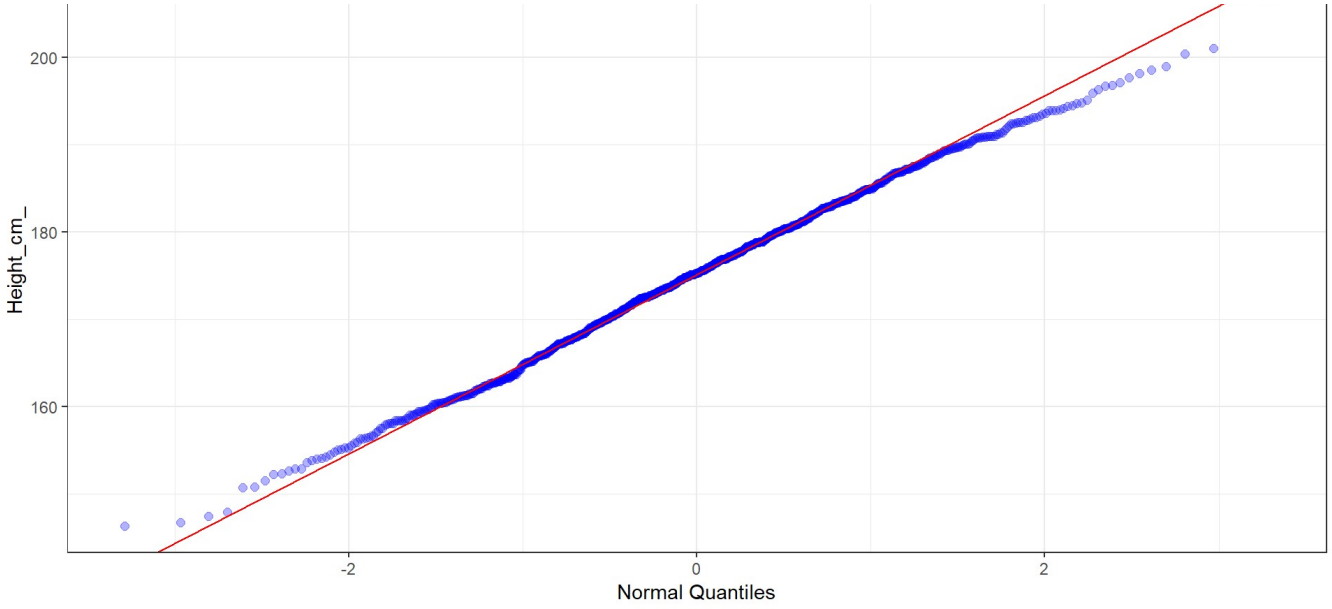
Quantile-quantile plot for Numerical variables - Univariate

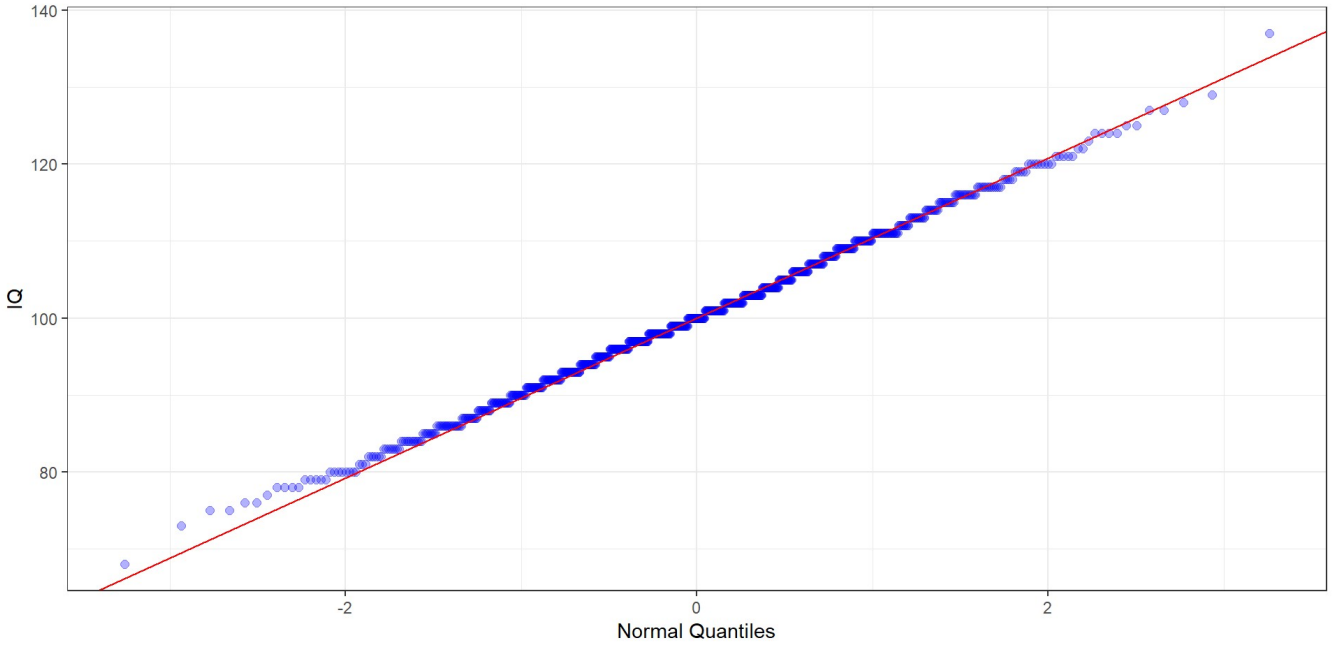
Quantile-quantile plot for all Numerical variables

```
ExpOutQQ(data,nlim=4,fname=NULL,Page=c(2,2),sample=sn)
```

```
## $`0`
```

page 1 of 1



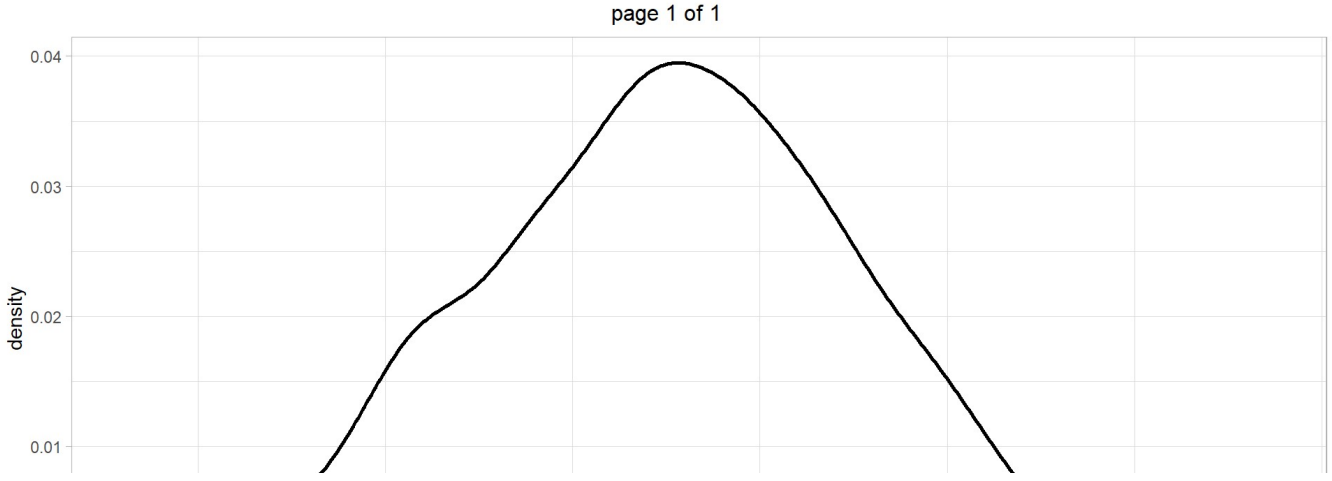


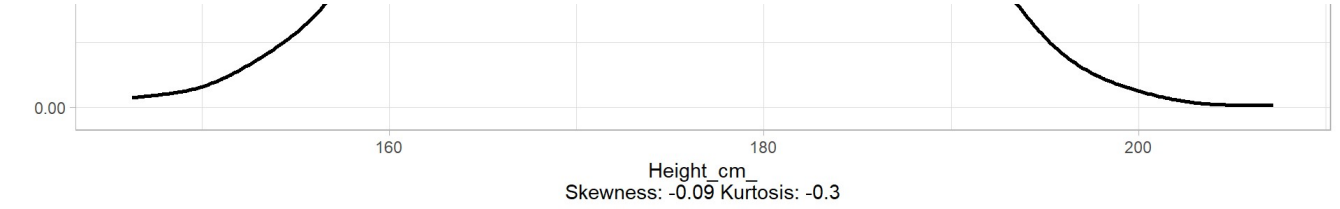
Density plots for Numerical variables - Univariate

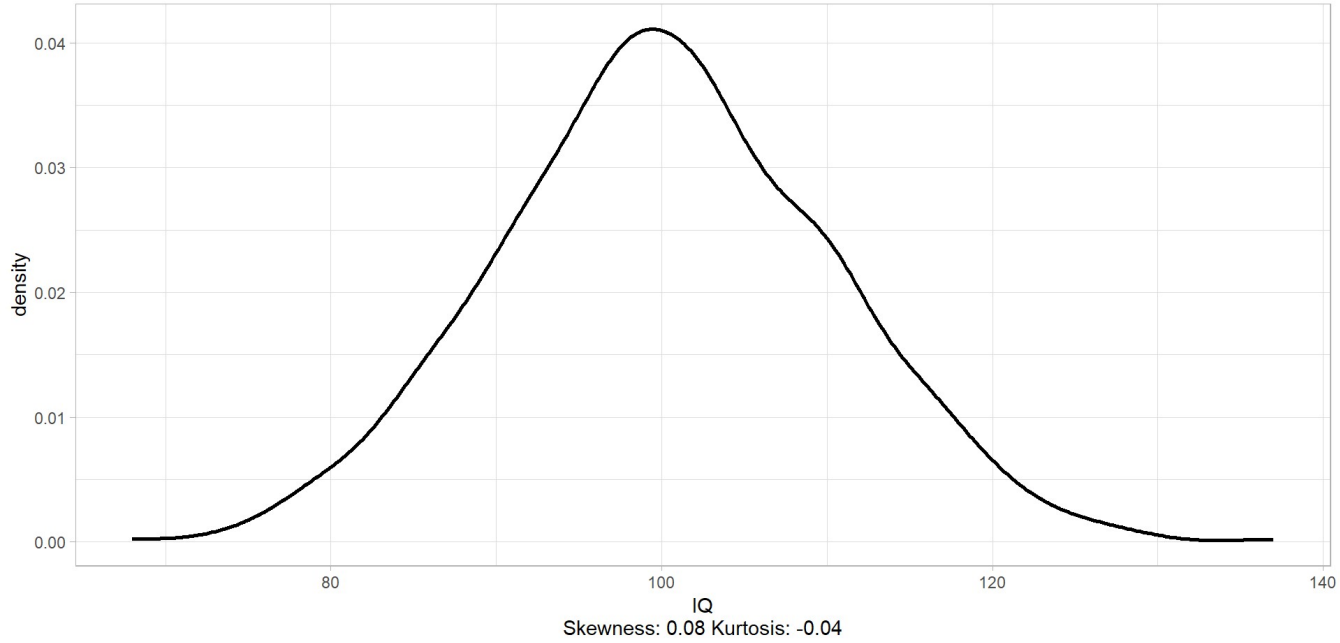
Density plot for all Numerical variables

```
ExpNumViz(data,gp=NULL,type=1,nlim=NULL,fname=NULL,col=NULL,Page=c(2,2),sample=sn)
```

```
## $`0`
```





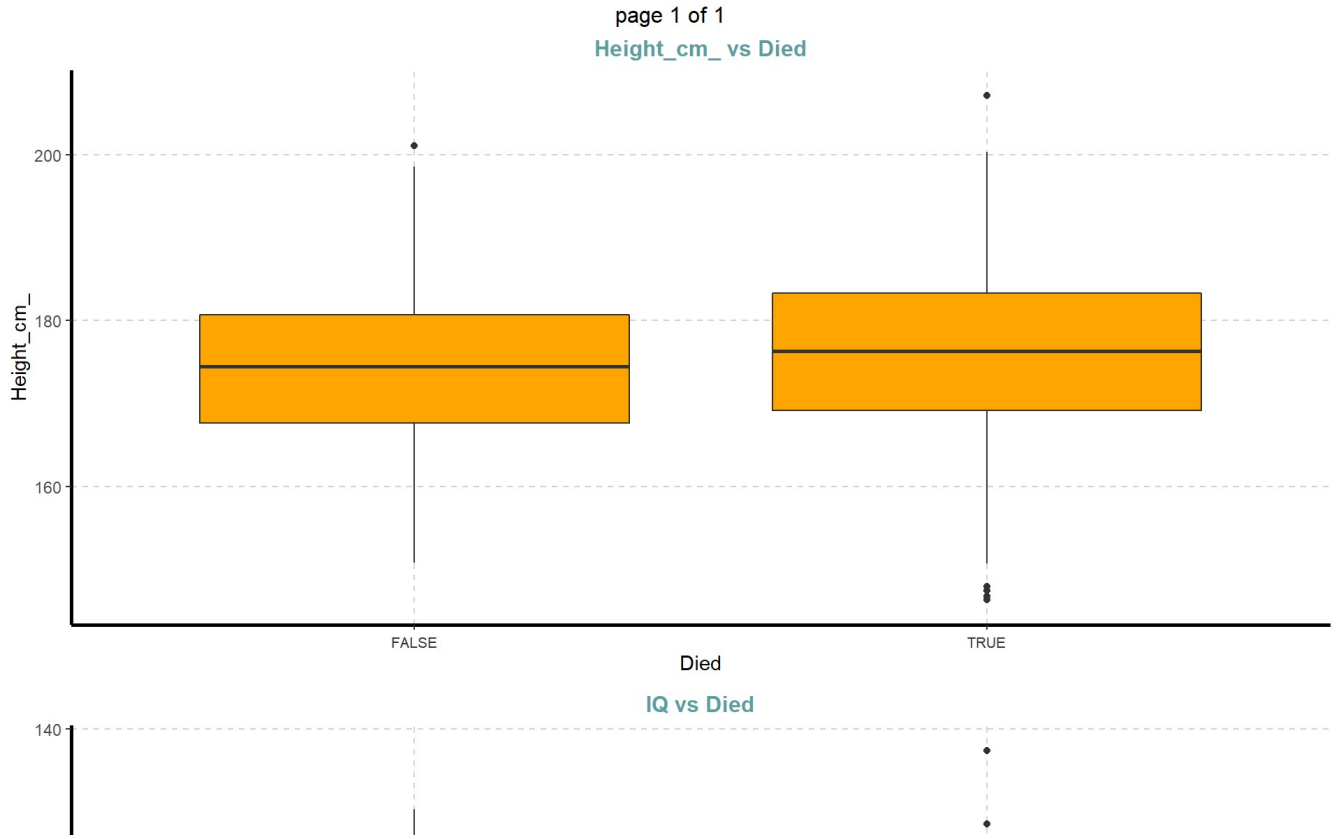


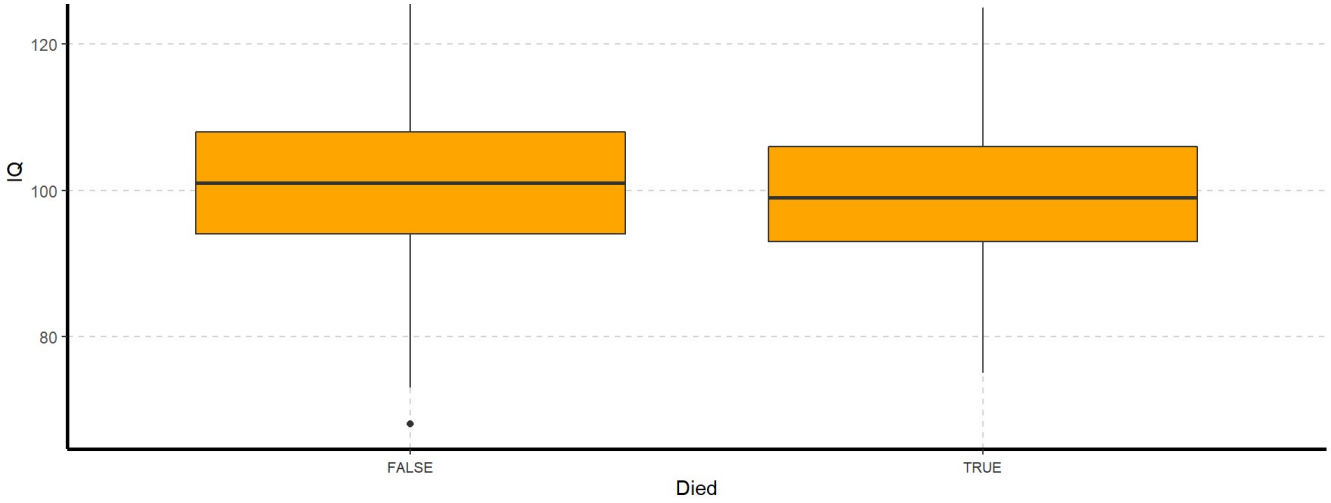
Box plots for all numeric features vs categorical dependent variable - Bivariate comparision only with categories

Boxplot for all the numeric attributes by each category of **Died**

```
ExpNumViz(data,gp=Target,type=2,nlim=NULL,fname=NULL,col=NULL,Page=c(2,2),sample=sn)
```

```
## $`0`
```





4. Summary of categorical variables

Summary of categorical variable

Cross tabulation with target variable

- Custom tables between all categorical independent variables and traget variable **Died**

```
ExpCTable(data,Target=Target,margin=1,clim=10,nlim=NULL,round=2,bin=NULL,per=F)
```

VARIABLE <chr>	CATEGORY <chr>	Died:FALSE <dbl>	Died:TRUE <dbl>	TOTAL <dbl>
Race	Asian	14	14	28
Race	Bi-Racial	5	13	18
Race	Black	56	58	114
Race	Hispanic	75	71	146
Race	NA	52	55	107
Race	Native	2	5	7
Race	Other	0	1	1
Race	White	261	318	579
Race	TOTAL	465	535	1000
Sex	Female	204	275	479

1-10 of 12 rows Previous **1** 2 Next

Information Value

```
ExpCatStat(data,Target=Target,Label=label,result = "IV",clim=10,nlim=5,Pclass=Rc)
```

Variable <fctr>	Target <fctr>	Class <chr>	Out_1 <int>	Out_0 <int>	TOTAL <int>	Per_1 <dbl>	Per_0 <dbl>	Odds <dbl>	WOE <dbl>
Race	Died	Asian	14	14	28	0.030	0.026	1.154	0.143
Race	Died	Bi-Racial	5	13	18	0.011	0.024	0.458	-0.781
Race	Died	Black	56	58	114	0.120	0.108	1.111	0.105

Variable <fctr>	Target <fctr>	Class <chr>	Out_1 <int>	Out_0 <int>	TOTAL <int>	Per_1 <dbl>	Per_0 <dbl>	Odds <dbl>	WOE <dbl>
Race	Died	Hispanic	75	71	146	0.161	0.133	1.211	0.191
Race	Died	NA	52	55	107	0.112	0.103	1.087	0.083
Race	Died	Native	2	5	7	0.004	0.009	0.444	-0.812
Race	Died	Other	0	1	1	0.000	0.002	0.000	0.000
Race	Died	White	261	318	579	0.561	0.594	0.944	-0.058
Sex	Died	Female	204	275	479	0.439	0.514	0.854	-0.158
Sex	Died	Male	261	260	521	0.561	0.486	1.154	0.143

1-10 of 10 rows | 1-10 of 13 columns

Statistical test

```
ExpCatStat(data,Target=Target,Label=label,result = "Stat",clim=10,nlim=5,Pclass=Rc)
```

Variable <fctr>	Target <fctr>	Uni... <fctr>	Chi-squared <fctr>	p-value <fctr>	df <fctr>	IV Value <fctr>	Cramers V <fctr>	Degree of Association <fctr>
Race	Died	8	NaN	NaN	7	0.024	NaN	Very Weak
Sex	Died	2	5.356	0.021	1	0.023	0.07	Very Weak

2 rows | 1-9 of 10 columns

5. Distributions of categorical variables

Graphical representation of all categorical variables

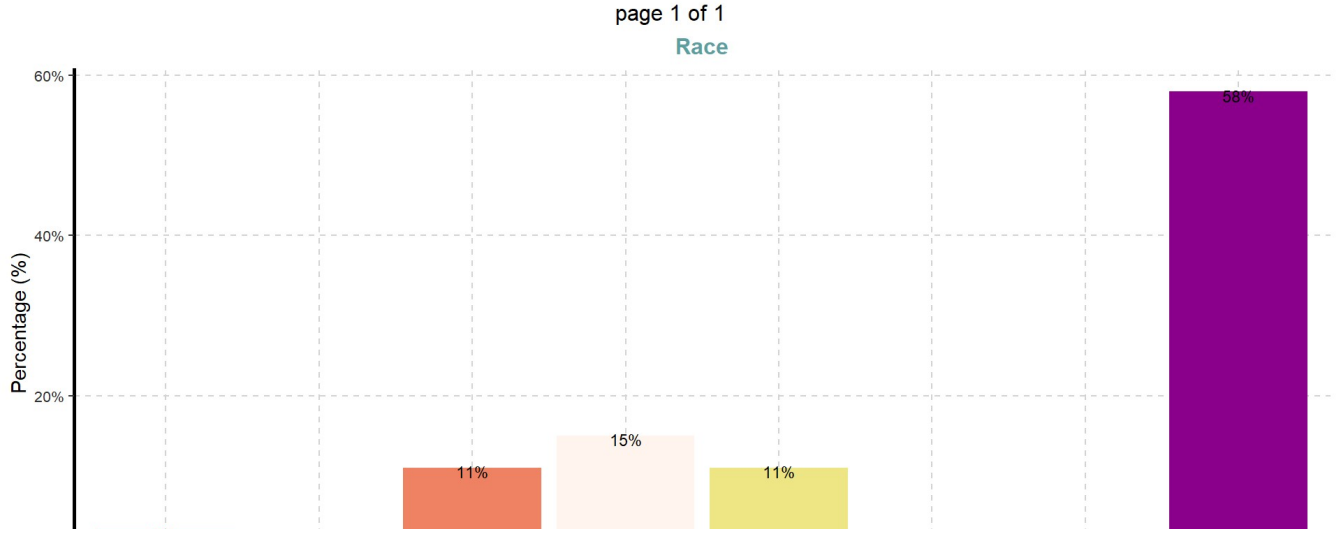
- Bar plot (Univariate)
- Stacked Bar plot (Bivariate)

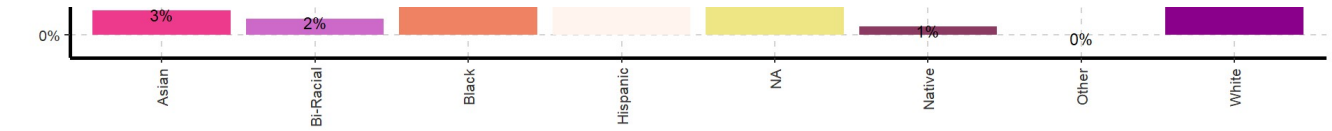
Bar plots for all categorical variables

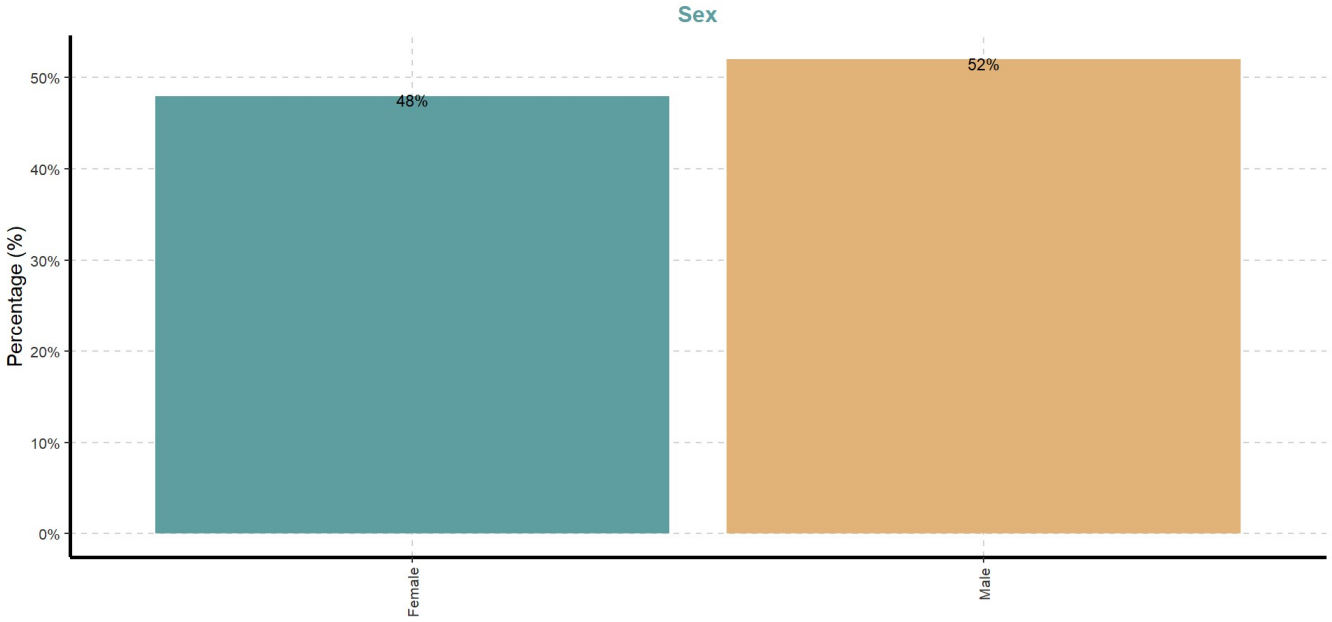
- Bar plot with vertical or horizontal bars for all categorical variables

```
ExpCatViz(data,gp=NULL,fname=NULL,clim=10,margin=2,Page = c(2,2),sample=sc)
```

```
## $`0`
```



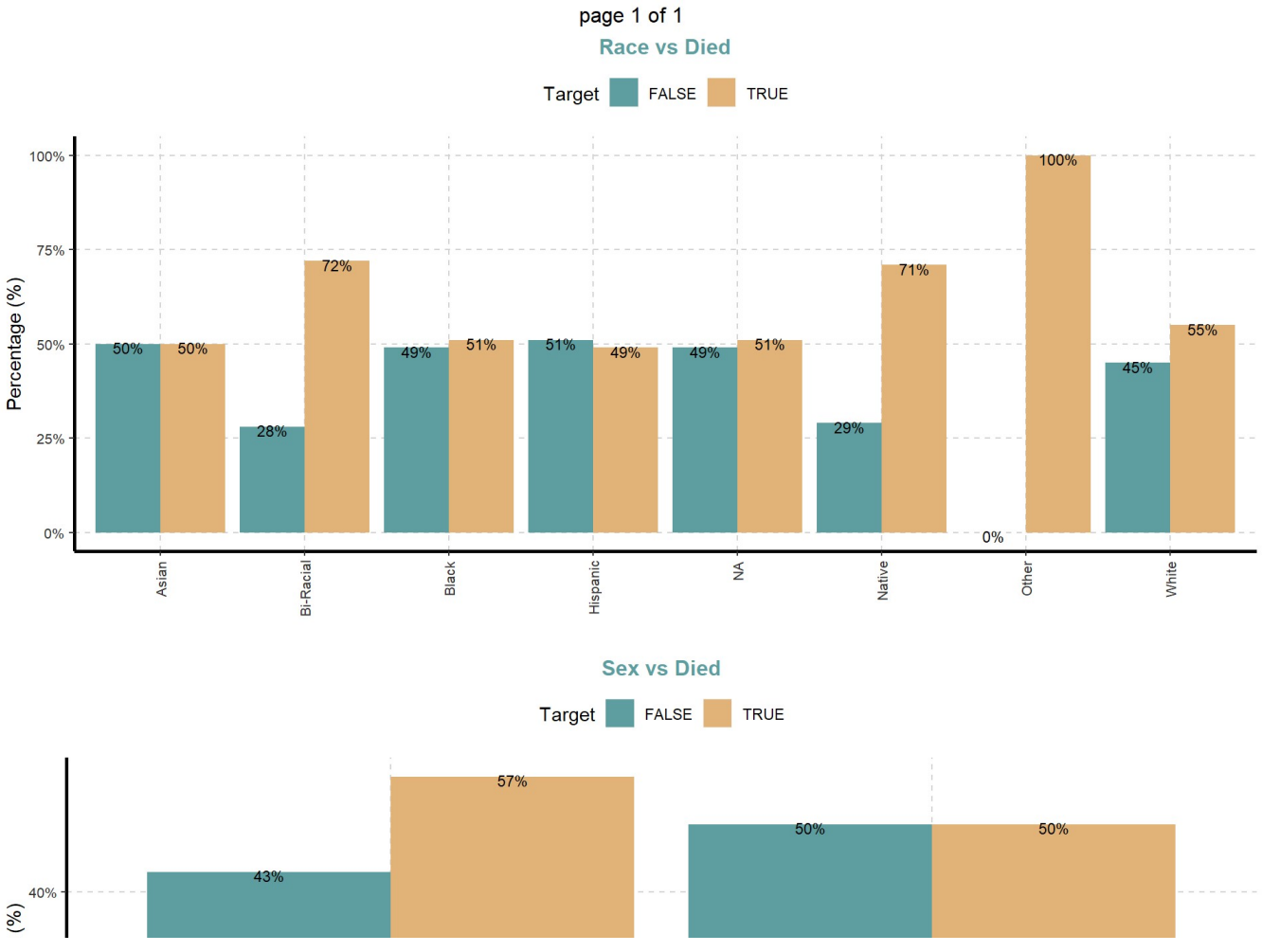


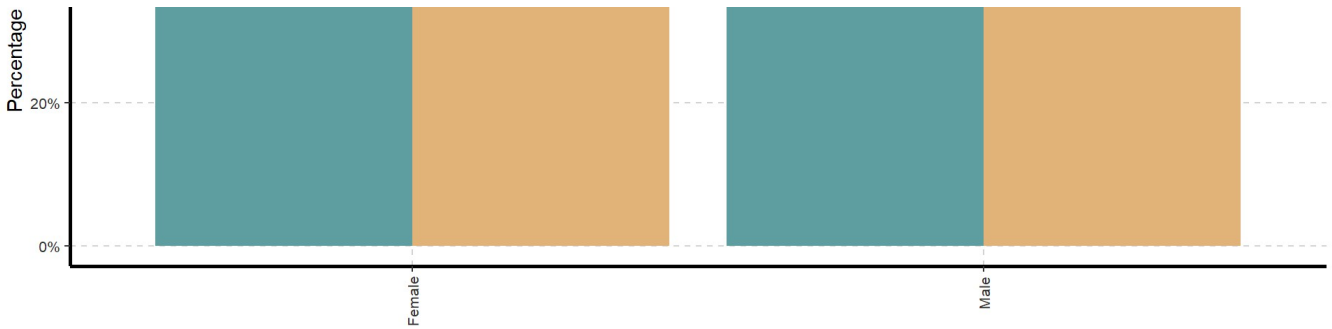


- Stacked bar plot with vertical or horizontal bars for all categorical variables

```
ExpCatViz(data,gp=Target,fname=NULL,clim=10,margin=2,Page = c(2,2),sample=sc)
```

```
## $`0`
```




Percentage
20%
0%
Female
Male