Applications using 

Analysis of numerical variables

Suppose we have loaded the package *ggplot2* that contains the dataset *diamonds*

To view the dataset *diamonds*
```
View(diamonds)
```

To define the variable *carat* from the dataset *diamonds*
```
carat=diamonds$carat
```

To compute the mean of the variable *carat*
```
m=mean(carat)
```

To compute the median of the variable *carat*
```
Me=median(carat)
```

To compute the minimum value of the variable *carat*
```
min(carat)
```

To compute the maximum value of the variable *carat*
```
max(carat)
```

To compute the first quartile of the variable *carat*
```
Q1=quantile(carat,0.25)
```

To compute the third quartile of the variable *carat*
```
Q3=quantile(carat,0.75)
```

To compute the p-quantile of the variable *carat*
```
Qp=quantile(carat,p)
```

To compute the frequency of the value 0.21 for the variable *carat*
```
length(carat[carat==0.21])
```

To compute minimum value, maximum value, mean, median, first and third quartile of the variable *carat*
```
summary(carat)
```

To draw the histogram of the variable *carat*
```
hist(carat)
```

To draw the histogram of the variable *carat* with 50 bins
```
hist(carat,50)
```

To draw the histogram of the variable *carat* with densities

```
hist(carat,freq=F)
```

To draw the histogram (with densities and 50 bins) and the density plot (in blue) of the va riable *carat*

```
hist(carat,50,freq=F)
lines(density(carat),col="blue")
```

To define the variable *price* from the dataset *diamonds*

```
price=diamonds$price
```

To draw the scatterplot between variables *carat* and *price*

```
plot(carat,price)
```

To draw a smoothed scatterplot between variables *carat* and *price*

```
smoothScatter(carat,price)
```

To compute the correlation between variables *carat* and *price*

```
cor(carat,price)
```

To draw the scatterplot matrix of the variables *carat*, *price,* and *x* of the dataset *diamonds*

```
X=diamonds$x
pairs(cbind(carat,price,x))
```

Analysis of categorical variables

To define the categorical variables *cut* and *color*

```
cut=diamonds$cut
color=diamonds$color
```

To display the categories of the categorical variable *cut*

```
unique(cut)
```

To display the absolute frequencies of the categorical variable *cut* and identify the mode

```
table(cut)
```

Draw a bar plot of the categorical variable *cut*

```
plot(cut)
```

Display the percentages of the categorical variable *cut*

```
100*table(cut)/length(cut)
```

or

```
100*prop.table(table(cut))
```

Draw the pie chart of the variable *cut*

```
pie(table(cut))
```

Draw a bar plot of the categorical variables *cut* and *color*

```
plot(cut,color)
```

Data handling

To check the presence of missing values in a dataset
```
anyNA(diamonds)
```

If the output is FALSE, we have no missing value.
If the output is TRUE, we have one or more missing values, and we can identify them
```
colSums(is.na(diamonds))
```

If you want to delete the rows with NA (if any), we define a new dataset
```
diamonds2=na.omit(diamonds)
```

To check the presence of possible outliers for the variable *carat*
```
boxplot(x)
```

If there are possible outliers they can be listed using
```
outliersx=boxplot(x)$out
```

To check the number of categories of the variable *color*
```
unique(color)
```

Regression analysis

Suppose we have imported the dataset *DatasetMarketing* reporting for 170 companies the following variables:

Nationality
Youtube advertising expenses
Facebook advertising expenses
Newspaper advertising expenses
Sales

The aim is the estimate of the best model to predict the sales of a company

First, check for the presence of missing values

```
anyNA(DatasetMarketing)
```

Given the answer (TRUE), we check how many missing values are present.

```
colSums(is.na(DatasetMarketing))
```

Given that there is just one missing value, we can omit the row containing it and define a new dataset

```
DatasetMarketing2=na.omit(DatasetMarketing)
```

To define the variables

```
nationality=DatasetMarketing2$nationality
youtube=DatasetMarketing2$youtube
facebook=DatasetMarketing2$facebook
news=DatasetMarketing2$newspaper
sales=DatasetMarketing2$sales
```

Then, we check for possible outliers and inaccuracies.

Next step is the estimate of the complete regression model

```
mod0=lm(sales ~ nationality+youtube+facebook+news)
```

To estimate the best model and display the output

```
modB=step(mod0)
summary(modB)
```

To make a prediction with the following categories/values:

Nationality=US

Youtube expenses = 300

Facebook expenses = 100

Newspaper expenses = 100

```
newdata=data.frame(nationality="US",youtube=300,facebook=100,news=100)
predict(modB,newdata)
```