Chapter 13: Descriptive Statistics – Thống kê mô tả

Ex3: Students

- Cho dữ liệu students.xls
- Đọc dữ liệu, kiểm tra và loại bỏ các dòng/các cột chứa dữ liệu NA
- Tạo dữ liệu mới từ cột thứ 4 đến hết. In thông tin của dataset
- In thống kê chung của dataset
- Vẽ pie chart thống kê sinh viên theo Major, theo Gender, theo Student.Status
- Tính phương sai của tất cả các thuộc tính số dữ liệu
- Tính standard deviation của tất cả các thuộc tính số trong dữ liệu
- Tính skewness của tất cả các thuộc tính số trong dữ liệu
- Tính kurtosis của tất cả các thuộc tính số trong dữ liệu
- vẽ histogram cho các cột Age, SAT, Average.score..grade., Height..in.

Gợi ý

In [1]:
```
# Load the library into R workspace.
library("xlsx")
```

```
Loading required package: rJava

Loading required package: xlsxjars
```

In [2]:
```
df <- read.xlsx("students.xls", sheetName ="Full")
print("Content of file:")
head(df)
```

```
[1] "Content of file:"
```

ID	LastÂ.Name	First.Name	City	State	Gender	Student.Status	Major	Country	Age
1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30
2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19
3	DOE01	JOE01	Elmira	New York	Male	Graduate	Math	US	26
4	DOE02	JOE02	Lackawana	New York	Male	Graduate	Econ	US	33
5	DOE03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37
6	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25

In [3]:
```
df <- df[colSums(!is.na(df)) > 0]
```

In [4]:
```
df <- na.omit(df)
```

In [5]:
```
df <- df[, 4:14]
head(df)
```

City	State	Gender	Student.Status	Major	Country	Age	SAT	Average.score..grade.	H
Los Angeles	California	Female	Graduate	Politics	US	30	2263	67.00000	
Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63.00000	
Elmira	New York	Male	Graduate	Math	US	26	2221	78.11328	
Lackawana	New York	Male	Graduate	Econ	US	33	1716	77.80859	
Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65.00000	
Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69.00000	

In [6]:
```
str(df)
```

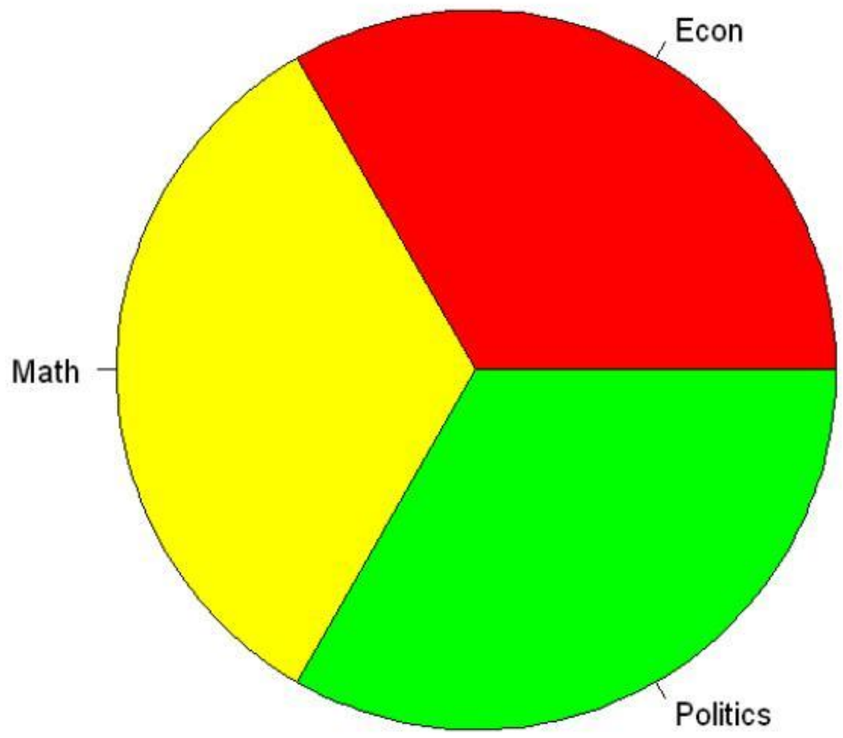
```
'data.frame':   30 obs. of  11 variables:
 $ City                          : Factor w/ 29 levels "Acme","Amsterdam",..:
17 25 9 14 7 27 6 15 20 22 ...
 $ State                         : Factor w/ 26 levels "Argentina","Arizon
a",..: 4 2 15 15 17 8 16 9 5 15 ...
 $ Gender                        : Factor w/ 2 levels "Female","Male": 1 1 2 2
2 2 2 1 1 1 ...
 $ Student.Status                : Factor w/ 2 levels "Graduate","Undergraduat
e": 1 2 1 1 1 1 1 2 2 1 ...
 $ Major                         : Factor w/ 3 levels "Econ","Math",..: 3 2 2
1 1 1 3 3 2 2 ...
 $ Country                       : Factor w/ 11 levels "Argentina","Bulgari
a",..: 10 10 10 10 10 6 10 10 3 10 ...
 $ Age                           : num  30 19 26 33 37 25 39 21 18 33 ...
 $ SAT                           : num  2263 2006 2221 1716 1701 ...
 $ Average.score..grade.         : num  67 63 78.1 77.8 65 ...
 $ Height..in.                   : num  61 64 73 68 71 67 70 62 62 66 ...
 $ Newspaper.readership..times.wk.: num  5 7 6 3 6 5 5 5 6 5 ...
```

In [7]:
```
summary(df)
```

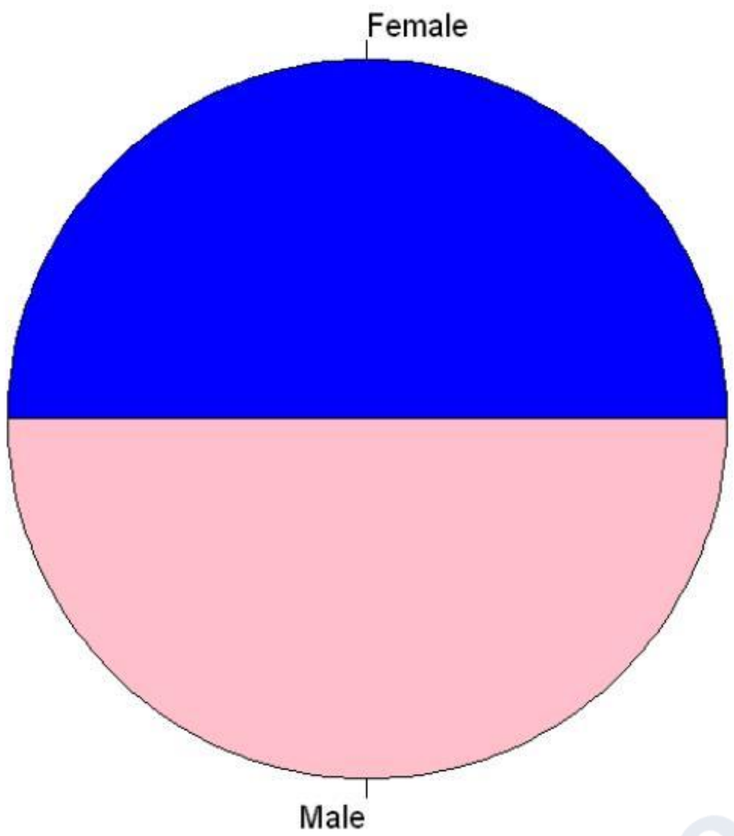
```
          City            State       Gender          Student.Status
New York    : 2   New York  : 5   Female:15   Graduate     :15
Acme        : 1   Argentina : 1   Male  :15   Undergraduate:15
Amsterdam   : 1   Arizona   : 1
Beijing     : 1   Bulgaria  : 1
Buenos Aires: 1   California: 1
Caracas     : 1   Canada    : 1
(Other)     :23   (Other)   :20
     Major          Country         Age             SAT
Econ    :10   US         :20   Min.   :18.0   Min.   :1338
Math    :10   Argentina: 1   1st Qu.:19.0   1st Qu.:1658
Politics:10   Bulgaria : 1   Median :23.0   Median :1817
              Canada   : 1   Mean   :25.2   Mean   :1849
              China    : 1   3rd Qu.:30.0   3rd Qu.:2032
              Holland  : 1   Max.   :39.0   Max.   :2309
              (Other)  : 5
 Average.score..grade.  Height..in.    Newspaper.readership..times.wk.
Min.   :63.00         Min.   :59.00   Min.   :3.000
1st Qu.:72.00         1st Qu.:63.00   1st Qu.:4.000
Median :79.75         Median :66.50   Median :5.000
Mean   :80.40         Mean   :66.43   Mean   :4.867
3rd Qu.:88.00         3rd Qu.:70.75   3rd Qu.:6.000
Max.   :95.88         Max.   :75.00   Max.   :7.000
```

In [8]:
```
major <- table(df$Major)
colors1 <- c("red", "yellow", "green")

gender <- table(df$Gender)
colors2 <- c("blue", "pink")

status <- table(df$Student.Status)
colors3 <- c("violet", "cyan")
```

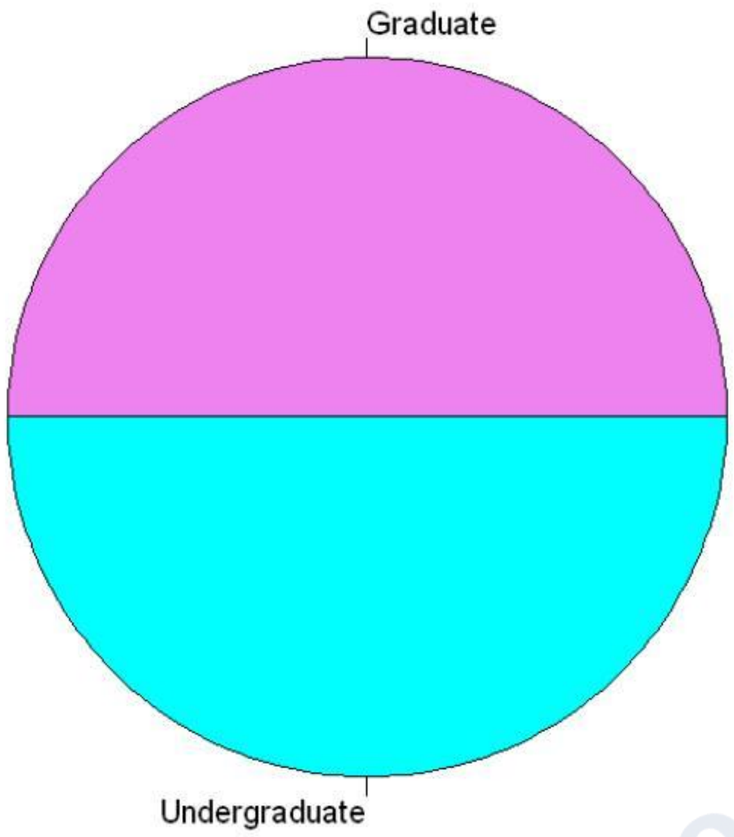
In [9]:
```
pie(major, col=colors1, main="Number of Students per Major")
```



In [10]:
```
pie(gender, col=colors2, main="Male vs Female")
```



In [11]:
```
pie(status, col=colors3, main="Graduate vs Undergraduate")
```



In [12]:
```
# calculate variance for all attributes
sapply(df[7:11], var)
```

Age	47.2
SAT	75686.7137931035
Average.score..grade.	102.130104261266
Height..in.	21.7022988505747
Newspaper.readershi…	1.6367816091954

In [13]:
```
# calculate sd for all attributes
sapply(df[7:11], sd)
```

Age	6.87022561492707
SAT	275.112184014273
Average.score..grade.	10.1059440064383
Height..in.	4.65857261943771
Newspaper.readershi…	1.27936765989898

In [14]:
```
#install.packages("e1071")
```

In [15]:
```
# calculate skewness for each variable
library("e1071")
skew <- apply(df[,7:11], 2, skewness)
# display skewness
skew
```

Age	0.502709664447605
SAT	0.14044712797574
Average.score..grade.	-0.101374236291567
Height..in.	0.155085443150385
Newspaper.readershi…	-0.0468347397810847

In [16]:
```
kur <- apply(df[,7:11], 2, kurtosis)
kur
```

Age	-1.20242842095183
SAT	-1.04281692418247
Average.score..grade.	-1.15697430829857
Height..in.	-1.21584757091849
Newspaper.readershi…	-1.14165472583428

In [17]:
```
par(mfrow=c(2,2))

# Create the histogram.
hist(df$Age, main = "Ages",
     xlab = "Ages",
     xlim = c(min(df$Age)-5, max(df$Age)+5),
     col = "orange",
     breaks = 10,
     border = "blue",
     # so lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$Age))

hist(df$SAT, main = "SAT",
     xlab = "SAT",
     xlim = c(min(df$SAT)-10, max(df$SAT)+10),
     col = "orange",
     breaks = 10,
     border = "blue",
     # so lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$SAT))

hist(df$Average.score..grade., main = "Avg Score Grade",
     xlab = "Avg score",
     xlim = c(min(df$Average.score..grade.)-10, max(df$Average.score..grade.)+10
     col = "orange",
     breaks = 10,
     border = "blue",
     # so lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$Average.score..grade.))

hist(df$Height..in., main = "Height in",
     xlab = "Height in",
     xlim = c(min(df$Average.score..grade.)-10, max(df$Average.score..grade.)+10
     col = "orange",
     breaks = 10,
     border = "blue",
     # so lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$Height..in.))
```