Chapter 13: Descriptive Statistics – Thống kê mô tả

Exercise 1: PimaIndiansDiabetes

- Cài package "mlbench"
- Load dataset PimaIndiansDiabetes
- In 10 dòng đầu của dataset
- In thông tin của dataset
- In thống kê chung của dataset
- Cho biết số dòng, cột của dataset
- Với diabetes, hãy cho biết có bao nhiêu mẫu "pos"/ bao nhiêu mẫu "neg". Nhận xét.
- Tính phương sai của tất cả các thuộc tính trong dữ liệu
- Tính standard deviation của tất cả các thuộc tính trong dữ liệu
- Cài package "e1071"
- Tính skewness của tất cả các thuộc tính trong dữ liệu
- Tính kurtosis của tất cả các thuộc tính trong dữ liệu
- Thực hiện correlation matrix để xem xét mối tương quan giữa các thuộc tính trong dữ liệu. Trực quan hóa kết quả.
- Vẽ boxplot của các thuộc tính kiểm tra dữ liệu có outlier hay không?
- Cho biết giá trị ở phân vị thứ 0.05, 0.15, 0.25, 0.50, 0.75, 0.90 của dữ liệu. Biểu diễn phân vị và giá trị tương ứng trên biểu đồ.

In [1]:
```
install.packages("mlbench")
```

```
Installing package into 'C:/Users/ktphuong/Documents/R/win-library/3.3'
(as 'lib' is unspecified)

package 'mlbench' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ktphuong\AppData\Local\Temp\RtmpSg00cI\downloaded_packages
```

In [2]:
```
# load the library
library(mlbench)
# load the dataset
data(PimaIndiansDiabetes)
# display first 10 rows of data
head(PimaIndiansDiabetes, n=10)
```

pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
6	148	72	35	0	33.6	0.627	50	pos
1	85	66	29	0	26.6	0.351	31	neg
8	183	64	0	0	23.3	0.672	32	pos
1	89	66	23	94	28.1	0.167	21	neg
0	137	40	35	168	43.1	2.288	33	pos
5	116	74	0	0	25.6	0.201	30	neg
3	78	50	32	88	31.0	0.248	26	pos
10	115	0	0	0	35.3	0.134	29	neg
2	197	70	45	543	30.5	0.158	53	pos
8	125	96	0	0	0.0	0.232	54	pos

In [3]:
```
str(PimaIndiansDiabetes)
```

```
'data.frame':  768 obs. of  9 variables:
 $ pregnant: num  6 1 8 1 0 5 3 10 2 8 ...
 $ glucose : num  148 85 183 89 137 116 78 115 197 125 ...
 $ pressure: num  72 66 64 66 40 74 50 0 70 96 ...
 $ triceps : num  35 29 0 23 35 0 32 0 45 0 ...
 $ insulin : num  0 0 0 94 168 0 88 0 543 0 ...
 $ mass    : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ pedigree: num  0.627 0.351 0.672 0.167 2.288 ...
 $ age     : num  50 31 32 21 33 30 26 29 53 54 ...
 $ diabetes: Factor w/ 2 levels "neg","pos": 2 1 2 1 2 1 2 1 2 2 ...
```

In [4]:
```
summary(PimaIndiansDiabetes)
```

```
   pregnant         glucose          pressure         triceps     
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
    insulin           mass           pedigree           age         diabetes  
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   neg:500  
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   pos:268  
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00            
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24            
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00            
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00            
```

In [5]:
```
dim(PimaIndiansDiabetes)
```

```
768  9
```

In [6]:
```
# distribution of class variable
y <- PimaIndiansDiabetes$diabetes
cbind(freq=table(y), percentage=prop.table(table(y))*100)
```

	freq	percentage
neg	500	65.10417
pos	268	34.89583

In [7]:
```
# calculate variance for all attributes
sapply(PimaIndiansDiabetes[,1:8], var)
```

pregnant	11.3540563206215
glucose	1022.24831425196
pressure	374.647271227184
triceps	254.473245328118
insulin	13281.1800779552
mass	62.1599839573827
pedigree	0.109778637873139
age	138.303045890374

In [8]:
```
# calculate standard deviation for all attributes
sapply(PimaIndiansDiabetes[,1:8], sd)
```

pregnant	3.36957806269887
glucose	31.9726181951362
pressure	19.3558071706448
triceps	15.9522175677276
insulin	115.244002351338
mass	7.88416032037545
pedigree	0.331328595012775
age	11.7602315406787

In [9]:
```
remove.packages("e1071")
```

```
Removing package from 'C:/Users/ktphuong/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
```

In [10]:
```
install.packages("e1071")
```

```
Installing package into 'C:/Users/ktphuong/Documents/R/win-library/3.3'
(as 'lib' is unspecified)


package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ktphuong\AppData\Local\Temp\RtmpSg00cI\downloaded_packages
```

In [11]:
```
# calculate skewness for each variable
library(e1071)
skew <- apply(PimaIndiansDiabetes[,1:8], 2, skewness)
# display skewness
skew
```

pregnant	0.898154872604808
glucose	0.173075366346189
pressure	-1.83641264105356
triceps	0.10894563103235
insulin	2.2633825833595
mass	-0.427307333732463
pedigree	1.91241792381955
age	1.12518804431459

In [12]:
```
kur <- apply(PimaIndiansDiabetes[,1:8], 2, kurtosis)
# display kurtosis
kur
```

pregnant	0.142183957122032
glucose	0.619369451373168
pressure	5.11750995409889
triceps	-0.530936920676741
insulin	7.13313491538188
mass	3.2449626779631
pedigree	5.52853885711983
age	0.621726907802014

In [13]:
```
# calculate a correlation matrix for numeric variables
correlations <- cor(PimaIndiansDiabetes[,1:8])
# display the correlation matrix
correlations
```

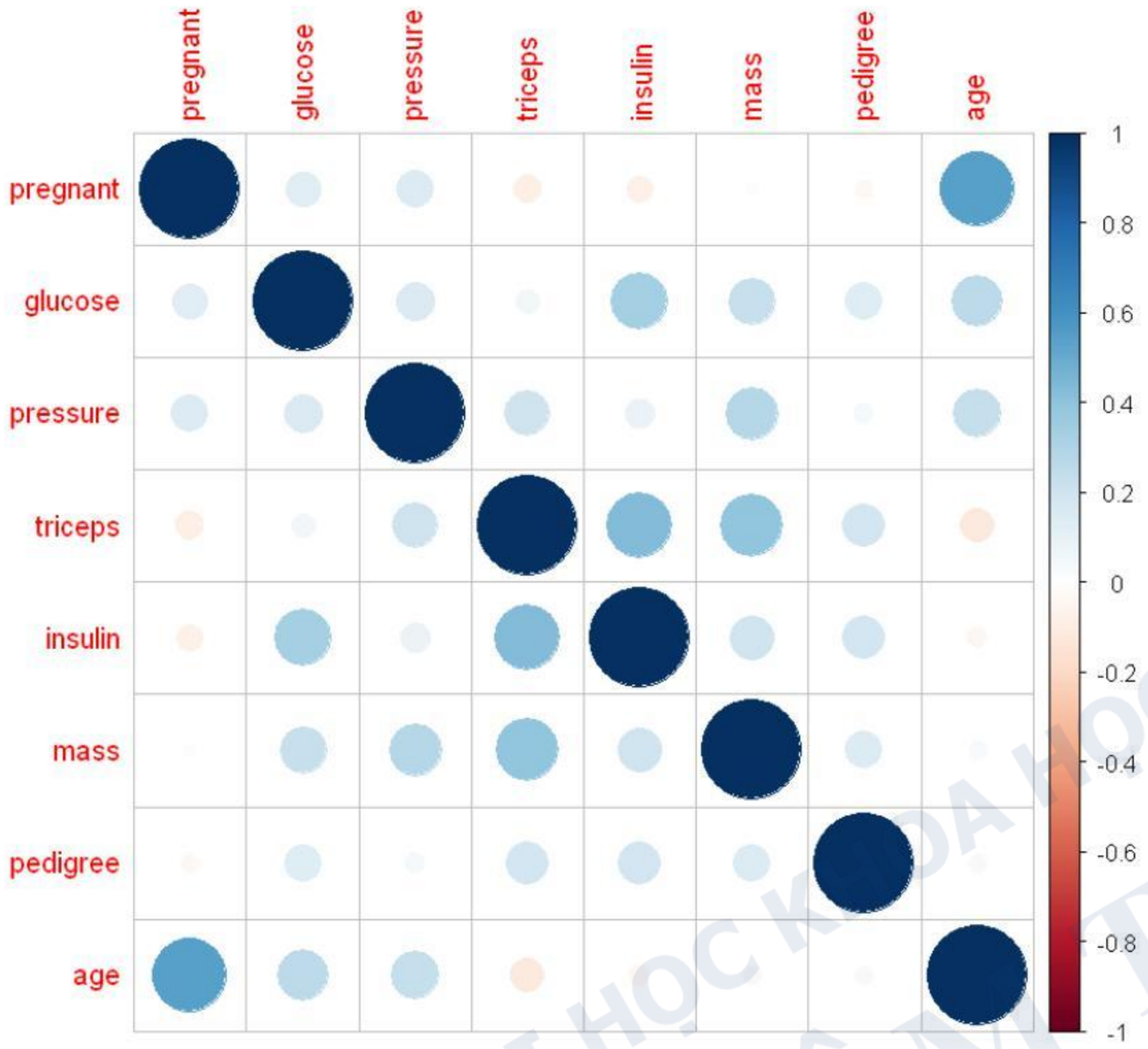
	pregnant	glucose	pressure	triceps	insulin	mass	pedigree
pregnant	1.00000000	0.12945867	0.14128198	-0.08167177	-0.07353461	0.01768309	-0.03352267
glucose	0.12945867	1.00000000	0.15258959	0.05732789	0.33135711	0.22107107	0.13733730
pressure	0.14128198	0.15258959	1.00000000	0.20737054	0.08893338	0.28180529	0.04126495
triceps	-0.08167177	0.05732789	0.20737054	1.00000000	0.43678257	0.39257320	0.18392757
insulin	-0.07353461	0.33135711	0.08893338	0.43678257	1.00000000	0.19785906	0.18507093
mass	0.01768309	0.22107107	0.28180529	0.39257320	0.19785906	1.00000000	0.14064695
pedigree	-0.03352267	0.13733730	0.04126495	0.18392757	0.18507093	0.14064695	1.00000000
age	0.54434123	0.26351432	0.23952795	-0.11397026	-0.04216295	0.03624187	0.03356131

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

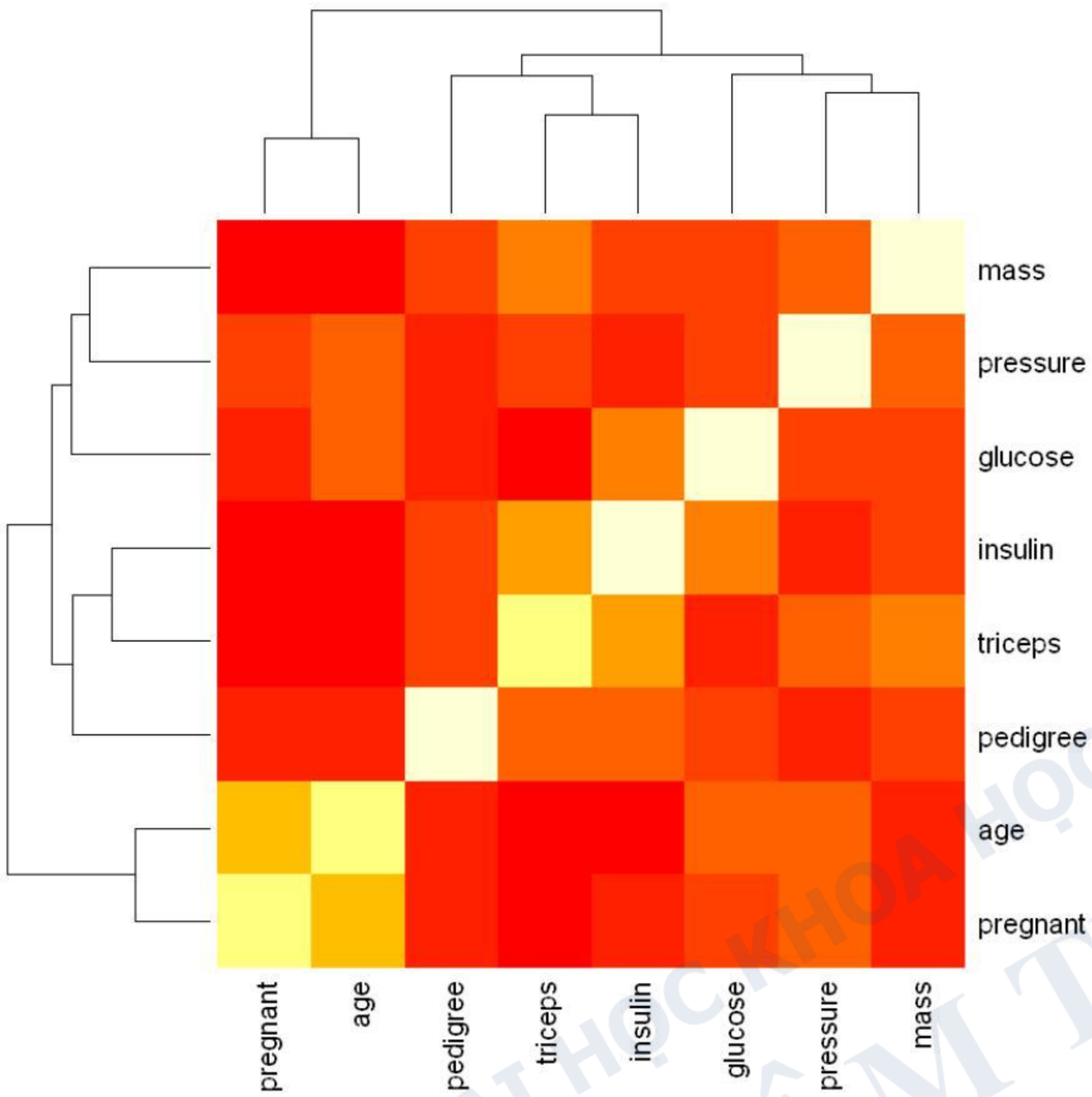
```
Installing package into 'C:/Users/ktphuong/Documents/R/win-library/3.3'
(as 'lib' is unspecified)


package 'corrplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ktphuong\AppData\Local\Temp\RtmpSg00cI\downloaded_packages
```

In [15]:
```
# link: http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-cor
```

In [16]:
```
library(corrplot)
corrplot(correlations)
```



In [17]:
```
heatmap(x = correlations)
```



In [18]:
```
ages = quantile(PimaIndiansDiabetes$age, c(0.05, 0.15, 0.25, 0.50, 0.75, 0.90))
ages
```

5%	21
15%	22
25%	24
50%	29
75%	41
90%	51

In [19]:
```
# Plot the line.
plot(c(0.05, 0.15, 0.25, 0.50, 0.75, 0.90), ages, type = "o")
```