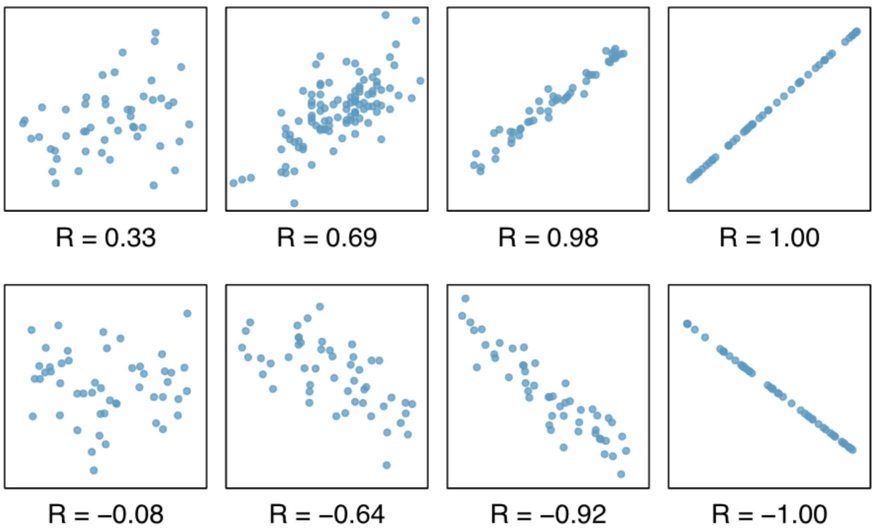
Tutorial 1



Example 1:

Below is a table displaying the number of employees (x1) and the profits per employee (x2) for 16 publishing firms. Employees are recorded in 1000s of employees and profits per employee are recorded in $1000s.

publishing firms	Profits ($1000s)	Employees (1000s)	publishing firms	Profits ($1000s)	Employees (1000s)
1	33.5	9.4	9	9.8	10.7
2	31.4	6.3	10	9.1	9.9
3	25	10.7	11	8.5	26.1
4	23.1	7.4	12	8.3	70.5
5	14.2	17.1	13	4.8	14.8
6	11.7	21.2	14	3.2	21.3
7	10.8	36.8	15	2.7	14.6
8	10.5	28.5	16	-9.5	26

Commented [BM1]: ***sample covariance:***
sample variance: measure of spread.
Sd: This measure of variation uses the same units as the observations.
sample correlation coefficient: This measure of the linear association between two variables does not depend on the units of measurement.

Commented [BM2]: https://courses.lumenlearning.com/odessa-introstats1-1/chapter/line-fitting-residuals-and-correlation/

The correlation coefficient is a dimensionless measure of association.

Commented [BM3]: Book: Applied multivariate statistical analysis, 6th Edition
Example 1.3 Page34 pdf

Commented [BM4]: يوجد أدناه جدول يعرض عدد الموظفين (س) والأرباح لكل موظف (ص) لـ 16 شركة نشر/ طباعه. يتم تسجيل الموظفين بآلاف ويتم تسجيل الأرباح لكل موظف في 1000 دولار.

https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/v/impact-of-removing-outliers-on-regression-lines

https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/v/interpreting-computer-regression-data

A. Compute the sample means $\bar{x}_1$ and $\bar{x}_2$, the sample variances $\boldsymbol{s}_{11}$ and $\boldsymbol{s}_{22}$**.** Compute the sample covariance $\boldsymbol{s}_{12}$ and the sample correlation coefficient $\boldsymbol{r}_{12}$. Interpret these quantities.
B. Display the sample mean array $\bar{x}$, the sample variance-covariance array S_n, and the sample correlation array **R** .
C. Construct a scatter plot of the data and comment on the appearance of the diagrams (using R).
D. Infer the sign of the sample covariance $\boldsymbol{s}_{12}$ from the scatter plot.

Solution:

A.

$\bar{x}_1 = 12.3188$; $\bar{x}_2 = 20.7063$; $s_{11} = 123.6683$; $s_{22} = 251.4340$

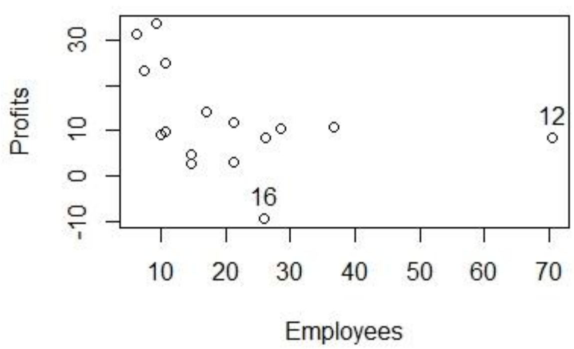
$\sqrt{s_{11}} = 11.1206$; $\sqrt{s_{22}} = 15.8567$

$s_{12} = s_{21} = -67.2615$

Pearson's correlation coefficient $r_{12} = r_{21} = \frac{S_{12}}{\sqrt{S_{11}}\sqrt{S_{22}}} = -0.381$ week negative correlation.

B. $\bar{x} = \begin{bmatrix} 12.3188 \\ 20.7063 \end{bmatrix}$, $S_n = \begin{bmatrix} 123.6683 & -67.2615 \\ -67.2615 & 251.4340 \end{bmatrix}$, $R = \begin{bmatrix} 1 & -0.3814 \\ -0.3814 & 1 \end{bmatrix}$

C.



We note two "unusual" observations. The twelfth publishing firm is the largest firm in terms of number of employees, but is "typical" in terms of profits per employee. The sixteenth publishing firm has a "typical" number of employees, but comparatively small (negative) profits per employee.

The sample correlation coefficient computed from the values of x1 and x2 is

$$r_{12}^2 = \begin{cases} -0.39 & for\ all\ 16\ firms \\ -0.55 & for\ all\ firms\ except\ specimen\ 12^{th} \\ -0.39 & for\ all\ firms\ except\ specimen\ 16^{th} \\ -0.50 & for\ all\ firms\ except\ specimen\ 16^{th}\ and\ 12^{th} \end{cases}$$

D. The sign of the sample covariance s_{12} **is negative**

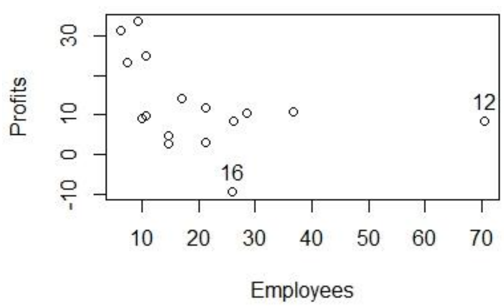
R code

```
##Example 1
rm(list=ls())
data1 <- read.table("C:/Desktop/stat438/Example 1-Profits per employee_TXT.tx
t",header = TRUE,row.names = 1)

# Mean
mean(data1$Profits); mean(data1$Employees)

[1] 12.31875

[1] 20.70625

apply(data1,2,mean) # 2:columns, 1:rows

  Profits Employees 
 12.31875  20.70625 

#var & sd
apply(data1,2,var)

  Profits Employees 
 123.6683  251.4340 

apply(data1,2,sd)

  Profits Employees 
 11.12062  15.85667 

# Correlation matrix
cor(data1)

            Profits Employees
Profits    1.000000 -0.381439
Employees -0.381439  1.000000

round(cor(data1),digits=2) #rounded to 2 decimals

          Profits Employees
Profits      1.00     -0.38
Employees   -0.38      1.00
```

```
# Covariance matrix
cov(data1)

             Profits Employees
Profits   123.66829 -67.26146
Employees -67.26146 251.43396

var(data1)

             Profits Employees
Profits   123.66829 -67.26146
Employees -67.26146 251.43396

## NOTE
summary(data1)

# If you need more descriptive statistics, use stat.desc() from the package {
pastecs}:
#install.packages("pastecs")
library(pastecs)

stat.desc(data1)
```

plot

```
# Scatterplot
plot(data1$Employees, data1$Profits, xlab="Employees", ylab="Profits")
identify(data1$Employees, data1$Profits) # Hit Esc key once you have selected
the point.
```




```
data1[12,]

   Profits Employees
12     8.3      70.5

data1[16,]

   Profits Employees
16    -9.5        26
```

```r
#Calculate correlation with removal of outliers.
round(cor(data1[-12,]),digits=3)
```

```
          Profits Employees
Profits     1.000    -0.553
Employees  -0.553     1.000
```

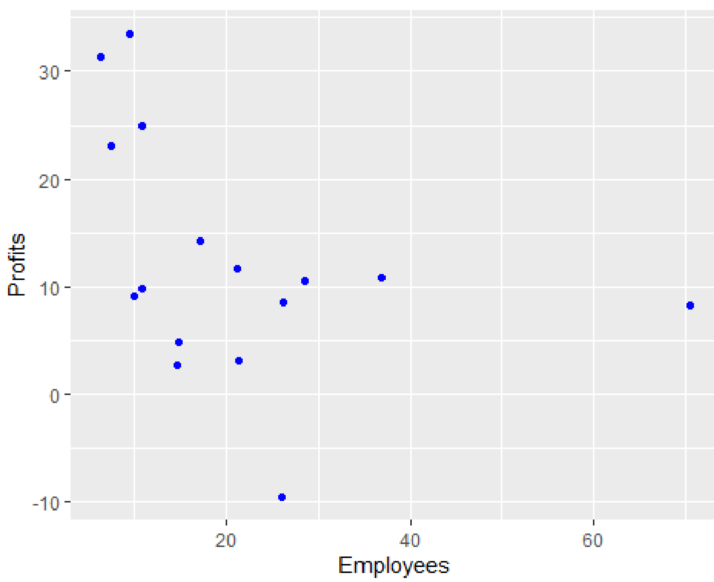
```r
round(cor(data1[-16,]),digits=3)
```

```
          Profits Employees
Profits     1.000    -0.394
Employees  -0.394     1.000
```

```r
round(cor(data1[c(-12,-16),]),digits=3)
```

```
          Profits Employees
Profits     1.000    -0.504
Employees  -0.504     1.000
```

```r
#other method for Scatterplot in {ggplot2} packages
library(ggplot2)
ggplot(data1) +
  aes(x = Employees, y = Profits) +
  geom_point(colour = "blue")
```



Example 2:

The following diagrams are display seven pairs of measurements (x_1, x_2) :

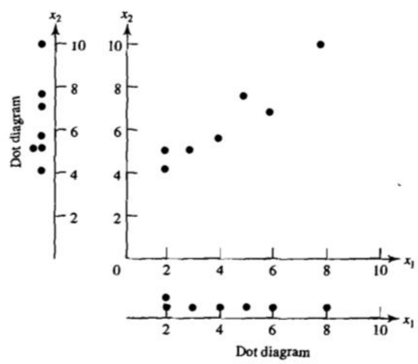


Figure 1.1 A scatter plot and marginal dot diagrams.

A. Infer the sign of the sample covariance $\boldsymbol{s_{12}}$ from the scatter plot.
B. Compute the sample means $\bar{x}_1$and $\bar{x}_2$, the sample variances $\boldsymbol{s_{11}\, and\, s_{22}.}$ Compute the sample covariance $\boldsymbol{s_{12}}$ and the sample correlation coefficient $\boldsymbol{r_{12}}$. Interpret these quantities.
C. Display the sample mean array $\bar{x}$, the sample variance-covariance array S_n, and the sample correlation array **R** .

<u>**Solution :**</u>

A. The sign of the sample covariance $\boldsymbol{s_{12}}$ **is positive.**

B. From scatter plot, we can construct table as

x_1	2	2	3	4	5	6	8
x_2	4	5	5	5.5	7.5	7	10

The information contained in the dot diagrams can be used to calculate the sample means, sample variances. And the information contained in the scatter plot can be used to calculate covariance.

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad ,k = 1,2,..,p$$

$$\bar{x}_1 = 4.29 \qquad , \qquad \bar{x}_2 = 6.29$$

The $sample\ covariance$: $S_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad i = 1,2,,..,p \quad ,k = 1,2,\dots,p$

$$S_1^2 = S_{11} = \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2 = \frac{1}{7}[(2-4.29)^2 + \cdots + (8-4.29)^2] = 4.20$$

$$S_2^2 = S_{22} = \frac{1}{7}[(4-6.29)^2 + \cdots + (10-6.29)^2] = 3.56$$

$$S_{12} = \frac{1}{7}[(2-4.29)(4-6.29)+..+(8-4.29)(10-6.29)] = 3.70$$

$$sample\ correlation\ coefficient: r_{12} = r_{21} = \frac{S_{12}}{\sqrt{S_{11}}\sqrt{S_{22}}} = \frac{3.70}{\sqrt{4.20}\sqrt{3.56}} = 0.96$$

C. $\bar{x} = \begin{bmatrix} 4.29 \\ 6.29 \end{bmatrix}$, $S_n = \begin{bmatrix} 4.20 & 3.70 \\ 3.70 & 3.56 \end{bmatrix}$, $R = \begin{bmatrix} 1 & 0.96 \\ 0.96 & 1 \end{bmatrix}$

Example 3: (Multiple scatter plot for paper strength measurement)

Commented [BM5]: Example 1.5 page 36 pdf.

BOX PLOT in R:
https://stackoverflow.com/questions/11134348/how-can-i-identify-the-labels-of-outliers-in-a-r-boxplot

https://statsandr.com/blog/outliers-detection-in-r/

Paper is manufactured in continuous sheets several feet wide. Because of the orientation of fibers within the paper, it has a different strength when measured in the direction produced by the machine than when measured across, or at right angles to, the machine direction. The measured values includes

$x_1 = density\ (grams/cublic\ centimeter)$

$x_2 = strength\ (pounds)\ in\ the\ machine\ direction$

$x_3 = strength\ (pounds)\ in\ the\ cross\ direction$

A. Construct a Multiple scatter plot of the data and comment on the appearance of the diagrams. (using R program)
B. Display the sample mean array $\bar{x}$, the sample variance-covariance array S_n, and the sample correlation array **R** (using R)

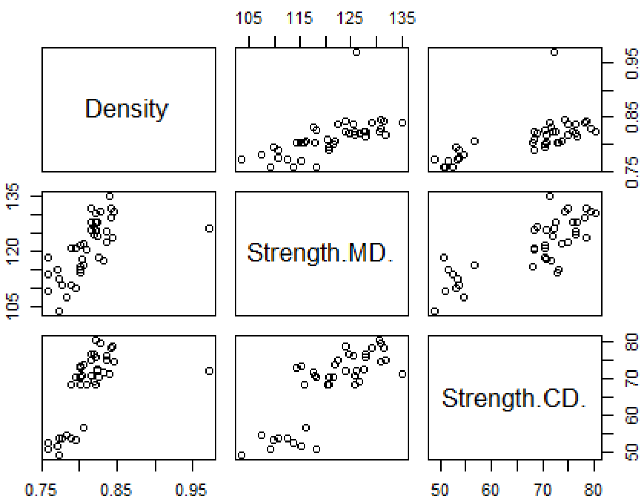
Solution:

A. In scatter plot matrix, there is one unusual observation: the density of specimen 25. Some of the scatter plots have patterns suggesting that there are two separate clumps of observations.

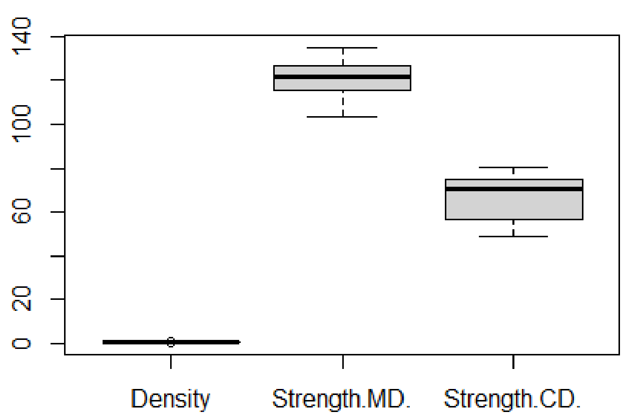
B. $\bar{x} = \begin{bmatrix} 0.8119 \\ 120.9534 \\ 67.7232 \end{bmatrix}$, $S_n = \begin{bmatrix} 0.001 & 0.168 & 0.225 \\ & 59.321 & 60.993 \\ & & 95.857 \end{bmatrix}$, $R = \begin{bmatrix} 1 & 0.615 & 0.647 \\ & 1 & 0.809 \\ & & 1 \end{bmatrix}$

R code

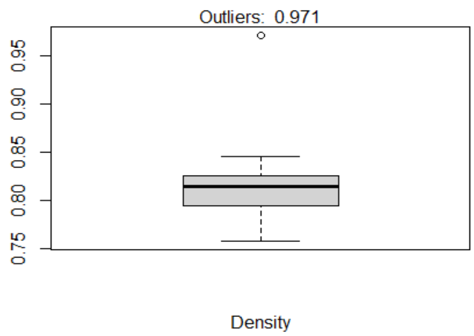
```
##Example 3

data3 <- read.table("C:/Desktop/stat438/Example 3- Paper-Quality Measurements
_TXT.txt",header = TRUE, row.names = 1)

#the first column ID to become the vector of row names of the data frame, wit
h row.names = 1.

apply(data3,2,mean) # 2:columns, 1:rows

     Density Strength.MD. Strength.CD.
   0.8118537  120.9534146   67.7231707

# Covariance matrix
round(var(data3), digits=4)

             Density Strength.MD. Strength.CD.
Density       0.0013       0.1684       0.2252
Strength.MD.  0.1684      59.3211      60.9925
Strength.CD.  0.2252      60.9925      95.8567

# Correlation matrix
round(cor(data3), digits =4)

             Density Strength.MD. Strength.CD.
Density        1.000       0.6150       0.6470
Strength.MD.   0.615       1.0000       0.8088
Strength.CD.   0.647       0.8088       1.0000

## plot
# multiple scatterplots
pairs(data3)
```




```
#boxplot
boxplot(data3)
```



```
out=boxplot(data3$Density,xlab="Density")$out

mtext(paste("Outliers: ", paste(out, collapse = ", "))) #Write Text into the Margins of a Plot
```




```
# Extract the outliers from the original data frame
data3[data3$Density %in% out,]

   Density Strength.MD. Strength.CD.
25   0.971        126.1         72.1
```

Example 4: (Looking for lower-dimensional structure)

A zoologist obtained measurement on n = 25 lizard known scientifically as *Cophosaurus texanus*. The weight, or mass, is given in grams while the snout-vent length (SVL) and hind limb span (HLS) are given in millimetres. The data are displayed in Table 1.3.

Table 1.3 Lizard Size Data

Lizard	Mass	SVL	HLS	Lizard	Mass	SVL	HLS
1	5.526	59.0	113.5	14	10.067	73.0	136.5
2	10.401	75.0	142.0	15	10.091	73.0	135.5
3	9.213	69.0	124.0	16	10.888	77.0	139.0
4	8.953	67.5	125.0	17	7.610	61.5	118.0
5	7.063	62.0	129.5	18	7.733	66.5	133.5
6	6.610	62.0	123.0	19	12.015	79.5	150.0
7	11.273	74.0	140.0	20	10.049	74.0	137.0
8	2.447	47.0	97.0	21	5.149	59.5	116.0
9	15.493	86.5	162.0	22	9.158	68.0	123.0
10	9.004	69.0	126.5	23	12.132	75.0	141.0
11	8.199	70.5	136.0	24	6.978	66.5	117.0
12	6.601	64.5	116.0	25	6.890	63.0	117.0
13	7.622	67.5	135.0				

Source: Data courtesy of Kevin E. Bonine.

Commented [BM6]:

Descriptive statistics in R:
https://statsandr.com/blog/descriptive-statistics-in-r/

Correlation coefficient and correlation test in R:
https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/

3D-scater plot in R :
http://www.sthda.com/english/wiki/scatterplot3d-3d-graphics-r-software-and-data-visualization

A. Construct a 3D-scatter plot of the data and comment on the appearance of the diagrams. (using R program)

Solution:

From three-dimensional scatter plot, it is clearly most of the variation is scatter about a one-dimensional straight line. Knowing the position on a line along the major axes of the cloud of points would be almost as good as knowing the three measurements Mass, SVL, and HLS.

However, this kind of analysis can be misleading if one variable has a much larger variance than the others. Consequently, we first calculate the standardized values

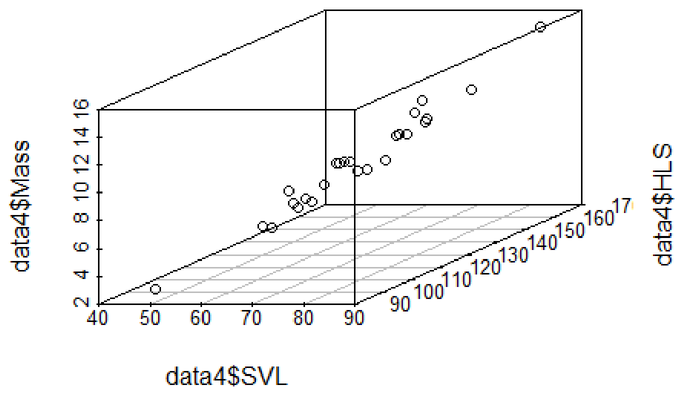
$$Z_{jk} = \frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}}$$

so the variables contribute equally to the variation in the scatter plot.

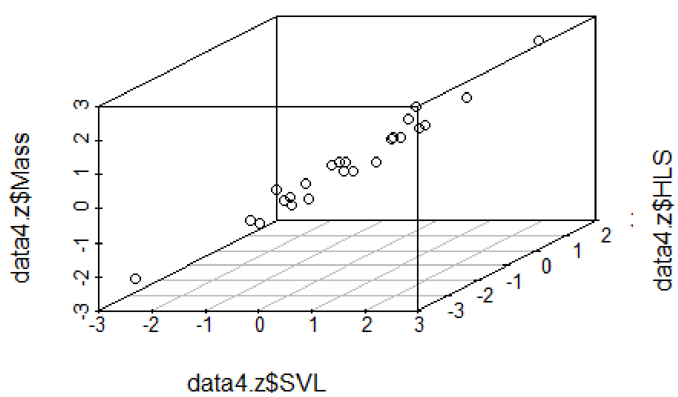
From three-dimensional scatter plot for the standardized variable, most of the variation can be explained by a **single variable** determined by line through the cloud of point.

R code

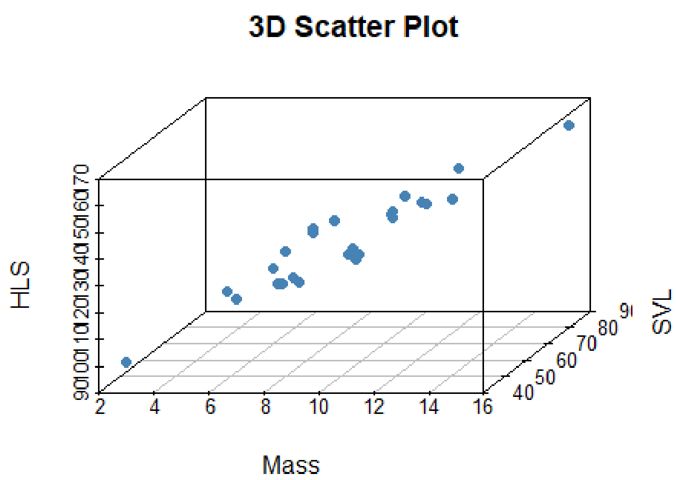
```
## Example 4
rm(list=ls())
data4 <- read.table("C:/Desktop/stat438/Example 4- Lizard Size Data_TXT.txt"
                   ,header = TRUE,row.names = 1)


# Standardize Data
data4.z <- as.data.frame(scale(data4[,-4], center = TRUE, scale = TRUE))

# 3D-scatterplot
# install.packages("scatterplot3d")
library("scatterplot3d")
scatterplot3d(data4$SVL,data4$HLS,data4$Mass)
```




```r
scatterplot3d(data4.z$SVL,data4.z$HLS,data4.z$Mass)
```




```r
# Change the angle of point view.Also change title,color,shape
scatterplot3d(data4[,-4], angle = 55,
              color="steelblue",pch = 16,
              main="3D Scatter Plot",
              xlab = "Mass",
              ylab = "SVL",
              zlab = "HLS")
```



Example 5: (Looking for group structure in three dimensions)

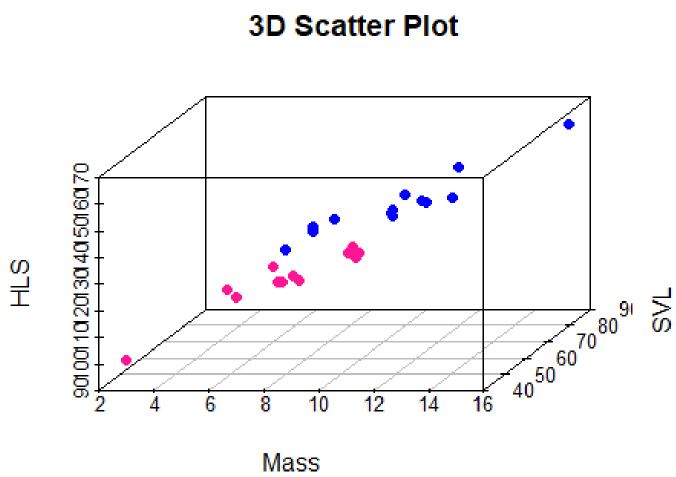
Commented [BM7]: Example 1.7 page 40 pdf

Referring to Example 4, it is interesting to see if male and female lizard occupy different parts of three-dimensional space containing the size data.

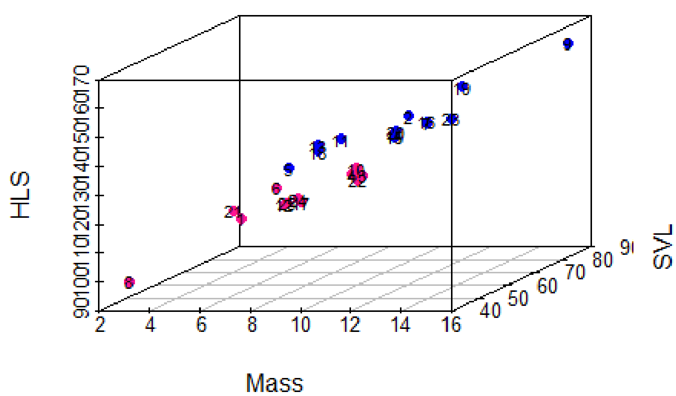
Solution:

From 3-D scatter plot of male and female lizards, we note that the males are typically larger than females.

```
## Example 5
# Change point colors by groups
colors <- c("deeppink", "blue")
colors <- colors[as.numeric(as.factor(data4$Gender))]
scatterplot3d(data4[,-4], angle = 55,
              color=colors,pch = 16,
              main="3D Scatter Plot",
              xlab = "Mass",
              ylab = "SVL",
              zlab = "HLS")
```




```r
#Add point labels
s3d<-scatterplot3d(data4[,-4],color=colors,pch = 16)
text(s3d$xyz.convert(data4[,-4]), labels = rownames(data4),cex= 0.7, col = "b
lack")
```



Example 1.9: (Rotated plots in three dimensions) page 44 pdf .

Example 1.9 (Rotated plots in three dimensions) Four different measurements of lumber stiffness are given in Table 4.3, page 186. In Example 4.14, specimen (board) 16 and possibly specimen (board) 9 are identified as unusual observations. Figures 1.12(a), (b), and (c) contain perspectives of the stiffness data in the x_1, x_2, x_3 space. These views were obtained by continually rotating and turning the three-dimensional coordinate axes. Spinning the coordinate axes allows one to get a better

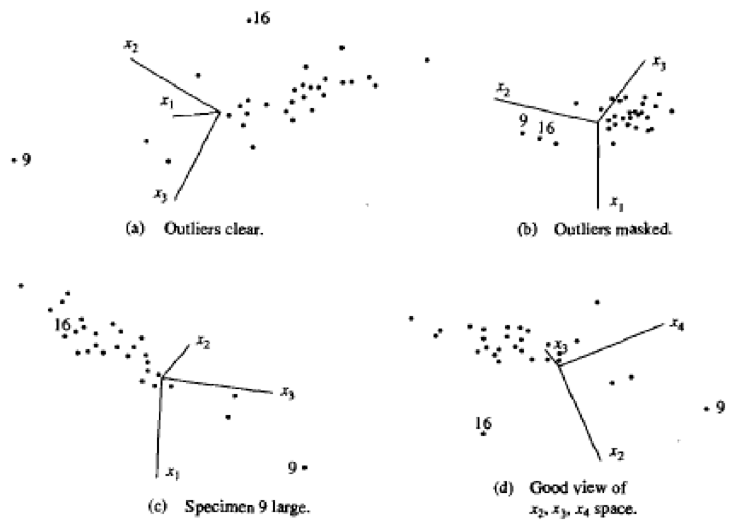


Figure 1.12 Three-dimensional perspectives for the lumber stiffness data.

understanding of the three-dimensional aspects of the data. Figure 1.12(d) gives one picture of the stiffness data in x_2, x_3, x_4 space. Notice that Figures 1.12(a) and (d) visually confirm specimens 9 and 16 as outliers. Specimen 9 is very large in all three coordinates. A counterclockwiselike rotation of the axes in Figure 1.12(a) produces Figure 1.12(b), and the two unusual observations are masked in this view. A further spinning of the x_2, x_3 axes gives Figure 1.12(c); one of the outliers (16) is now hidden.

Additional insights can sometimes be gleaned from visual inspection of the slowly spinning data. It is this dynamic aspect that statisticians are just beginning to understand and exploit. ■

HW1

Exercise 1.22: Oxygen-consumption data

Researchers interested in assessing pulmonary function in nonpathological populations asked subjects to run on a treadmill until exhaustion. Samples of air were collected at definite intervals and the gas contents analyzed. The results on 4 measures of oxygen consumption for 25 males and 25 females are given in **Table 6.12** on page 348. The variables were

x_1: resting volume O_2 (L/min)

x_2: resting volume O_2 (mL/kg/min)

x_3: $maximum\ volume\ O_2\ (L/min)$

x_4: $maximum\ volume\ O_2\ (mL/kg/min)$

Using appropriate computer software,

A. View the entire data set in three dimensions employing various combinations of three variables to represent the coordinate axes. Begin with the x_1, x_2, x_3 space.
B. Check this data set for outliers.
C. Compute $\bar{x}$, S_n and **R** arrays**.** Interpret the pairwise correlations.
D. Would the correlation in Part C change if you change the unit of measurement? Explain.

(Use R program)

Commented [BM8]:

A >> 2 Mark
B >> 1 Mark
C >> 1 Mark
D>> 1 Mark

Commented [BM9]: اضفته من سؤال
1.27 (c)
Answer:
The correlation coefficient is a **unitless** measure of association. The correlation in (c) would not change .

in Book : page 29
The sample correlation coefficient is a standardized version of the sample co
variance,

Exercise 1.2:
A morning newspaper lists the following used-car prices for a foreign compact *with* age x_1 measured in years and selling price x_2 measured in thousands of dollars:

x_1	1	2	3	3	4	5	6	8	9	11
x_2	18.95	19	17.95	15.54	14	12.95	8.94	7.49	6	3.99

A. Construct a scatter plot of the data and comment on the appearance of the diagrams (Manually and R).
B. Infer the sign of the sample covariance s_{12} from the scatter plot.
C. Compute the sample means $\bar{x}_1$ and $\bar{x}_2$, the sample variances s_{11} and s_{22}**.** Compute the sample covariance s_{12} and the sample correlation coefficient r_{12}. Interpret these quantities.
D. Display the sample mean array $\bar{x}$, the sample variance-covariance array S_n, and the sample correlation array **R**.

Commented [BM10]:
A >> 2 Mark
B >> 1 Mark
C >> 1 Mark
D>> 1 Mark