Bivariate and Multivariate Graphical Data Analysis

1. Bivariate analysis

Covariance Code Example 1.1 Covariance is a statistical representation of the degree to which two variables vary together. Basically, covariance is a number that reflects the degree to which two variable vary together. If the greater values of one variable correspond with the greater values of the other variable, or for the smaller values, then the variables show similar behavior, the covariance is a positive. If the greater values of one variable correspond to the smaller values of the other, the variables tend to show opposite behavior, the covariance is negative. If one variable is greater and paired equally often with both greater and lesser values on the other, the covariance will be near to zero.

Example

```r
# Question: Find the covariance of eruption duration and waiting time in the data set faithful

# Printing out the the first 6 rows of the dataset
head(faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```r
# Assigning the eruptions column to the variable eruptions
eruptions <- faithful$eruptions

# Assigning the waiting column to the variable waiting
waiting<- faithful$waiting

# Using the cov() function to determine the covariance
cov(eruptions, waiting)
```

```
## [1] 13.97781
```

The covariance of eruption duration and waiting time is about 13.98. It indicates a positive linear relationship between the two variables.

CHALLENGE

```
# Question: Find out the covariance of Bwt and Hwt in the cats dataset

# Finding out the covariance

# Assigning the Bwt column to the variable Bwt
#Bwt <- cats$Bwt

# Assigning Hwt column to the variable Hwt
#Hwt <- cats$Hwt

# Determining the covaraince
#cov(Bwt, Hwt)
```

Correlation Coefficient Code Example 1.2 The correlation coefficient of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related. If the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

Example

```
# Question: Find the correlation coefficient of eruption duration and waiting time in the faithful data

# Assigning the eruptions column to the variable eruptions
eruptions <- faithful$eruptions

# Assigning the waiting column to the variable waiting
waiting<- faithful$waiting

# Using the cor() function to determine the covariance
cor(eruptions, waiting)
```

```
## [1] 0.9008112
```

The correlation coefficient of eruption duration and waiting time is 0.90081. Because it is close to 1, we can conclude that *the variables are positively linearly related.*

```
## Challenge
# Question: Find out the covariance of Bwt and Hwt in the cats data set below:

# Previewing the cats dataset by first importing the Mass library
# then displaying the first 6 records of this database
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.5
```

```
head(cats)
```

```
##   Sex Bwt Hwt
## 1   F 2.0 7.0
## 2   F 2.0 7.4
## 3   F 2.0 9.5
## 4   F 2.1 7.2
## 5   F 2.1 7.3
## 6   F 2.1 7.6
```

```
cor(cats$Bwt, cats$Hwt)
```

```
## [1] 0.8041274
```

CHALLENGE

```
# Question: Create a correlation matrix in R using the corr() function

# Hint: http://bit.ly/RDocumentationCorrMatrix

url = "http://bit.ly/HousingDatainR"

housing <- read.csv(url)

# Previewing our data
head(housing)
```

```
##
## 1
## 2
## 3
## 4
## 5
## 6   <meta name=description content=Kaggle is the world&#x2019;s largest data science community with
```

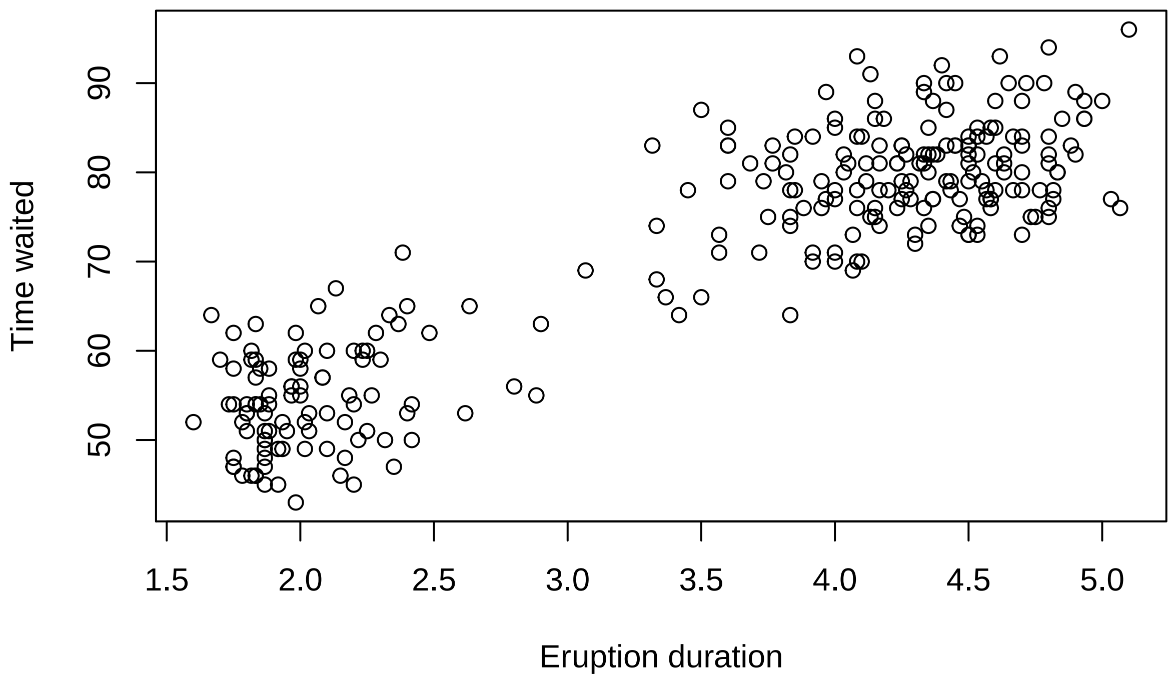
- This data can not be found

2. Graphical Techniques

Scatterplot Code Example 2.1 A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis. Scatter plots are used when you want to show the relationship between two variables. They are sometimes called correlation plots because they show how two variables are correlated.

```
# Question: Create a scatter plot of the eruption durations and waiting intervals from the faithful dat

# Assigning the eruptions column to the variable eruptions
eruptions <- faithful$eruptions
```

```
# Assigning the waiting column to the variable waiting
waiting <- faithful$waiting

# Creating the scatter plot using eruptions and waiting
plot(eruptions, waiting, xlab="Eruption duration", ylab="Time waited")
```



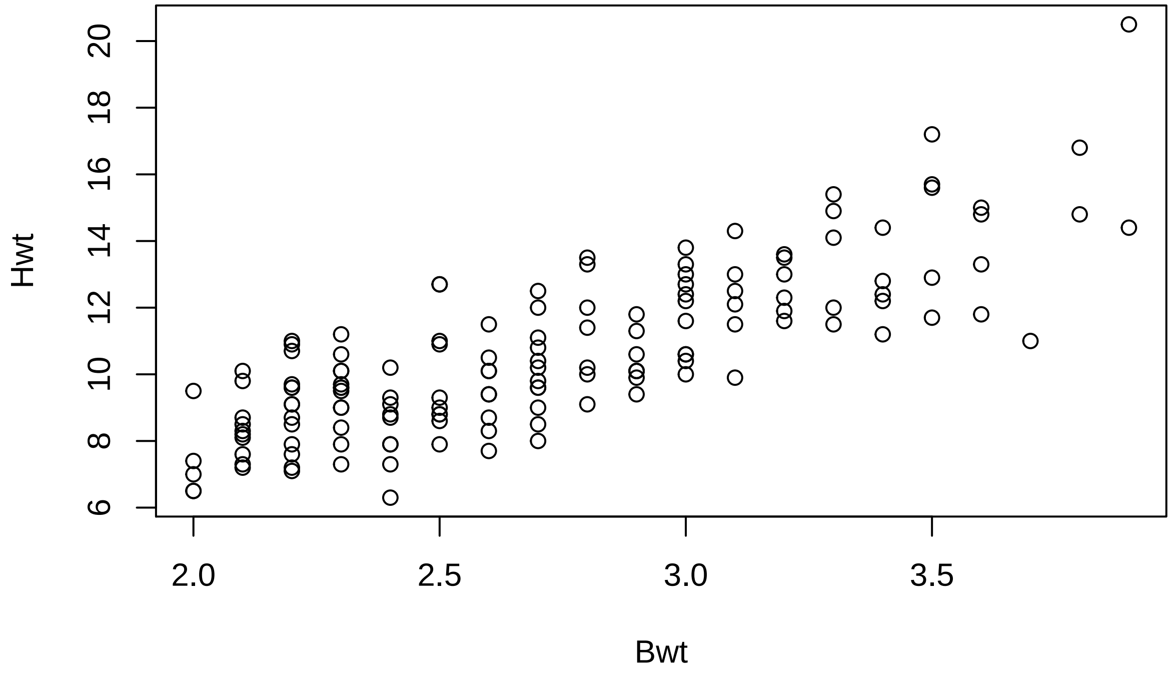
The scatter plot above reveals a positive linear relationship between eruptions and waiting.

CHALLENGE

```
# Question: Using the cats dataset, create a scatter plot of the Bwt and Hwt variables.
# Does it reveal any relationship between these variables?

head(cats)
```

```
##   Sex Bwt Hwt
## 1   F 2.0 7.0
## 2   F 2.0 7.4
## 3   F 2.0 9.5
## 4   F 2.1 7.2
## 5   F 2.1 7.3
## 6   F 2.1 7.6
```

```r
# Assigning the Bwt column to the variable Bwt
Bwt <- cats$Bwt

# Assigning Hwt column to the variable Hwt
Hwt <- cats$Hwt

plot(Bwt, Hwt, xlab="Bwt", ylab="Hwt")
```


Hwt
Bwt