In this section, you will use the *marketing* dataset from the *datarium* package in R.

The *marketing* dataset contains data on the amount of money (in thousands of dollars) that a company is willing to set aside for advertising on three different media platforms (YouTube, Facebook, and newspaper) and the correlating effect on sales. It has 200 rows and 4 columns.

To obtain the *marketing* dataset, first install the datarium package using the following code:

Note: In some environments or installations of R, the datarium package may already be included by default. It is recommended that you check the installed packages in your specific R environment before installing it. You can do this by using the installed.packages() function or checking the package list in your IDE. If the datarium package is already present, there is no need to install it again.

```
install.packages("datarium")
```

Load the *marketing* dataset into the variable md using the following code:

```
require(datarium)

md <- marketing
```

To better understand the dataset, use the function dim() to determine the dimensions of the dataset, str() to obtain information about the rows and columns of the dataset, and head() to view the first few rows of the dataset. The code to view the dataset dimensions is as follows:

```
dim(md)
```

This code produces the following output:

```
[1] 200   4
```

To display information about the rows and columns in the dataset, run the following code:

```
str(md)
```

This code produces the following output:

```
'data.frame':   200 obs. of  4 variables:

$ YouTube  : num  276.1 53.4 20.6 181.8 217 ...

$ Facebook : num  45.4 47.2 55.1 49.6 13 ...

$ newspaper: num  83 54.1 83.2 70.2 70.1 ...

$ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

Run the following code to view the first six rows of the dataset:

```
head(md)
```

This code produces the following output:

```
  YouTube Facebook newspaper sales

1  276.12    45.36     83.04 26.52

2   53.40    47.16     54.12 12.48

3   20.64    55.08     83.16 11.16

4  181.80    49.56     70.20 22.20
```

5 216.96 12.96 70.08 15.48

6 10.44 58.68 90.00 8.64

Missing values can be found in R using the function is.na(). Missing values in an R dataset are fields that contain NA. The is.na function checks each data point in a dataset and returns TRUE if it contains NA and FALSE if it does not contain NA. Tabulate these values using the table() function as follows:

table(is.na(md))

This code produces the following output:

FALSE

800

The code output above shows that none of the 800 data points in the dataset are missing values.

From the results received so far, the following statements can be made about the dataset.

- The dataset has 200 observations and 4 variables, namely, YouTube, Facebook, newspaper, and sales.
- All the values are numerical.
- The dataset does not have any missing values.

Descriptive statistics (e.g., averages) help an analyst better understand the data. The summary() function in R is used to compute summary statistics of data and models.

summary(md)

This code produces the following output:

YouTube Facebook newspaper sales

Min. : 0.84 Min. : 0.00 Min. : 0.36 Min. : 1.92

1st Qu.: 89.25 1st Qu.:11.97 1st Qu.: 15.30 1st Qu.:12.45

Median :179.70 Median :27.48 Median : 30.90 Median :15.48

Mean :176.45 Mean :27.92 Mean : 36.66 Mean :16.83

3rd Qu.:262.59 3rd Qu.:43.83 3rd Qu.: 54.12 3rd Qu.:20.88

Max. :355.68 Max. :59.52 Max. :136.80 Max. :32.40

From the code output above, note that the company spends the most on YouTube advertising (mean budget is $176,450) and the least on Facebook advertising (mean budget is $27,920).

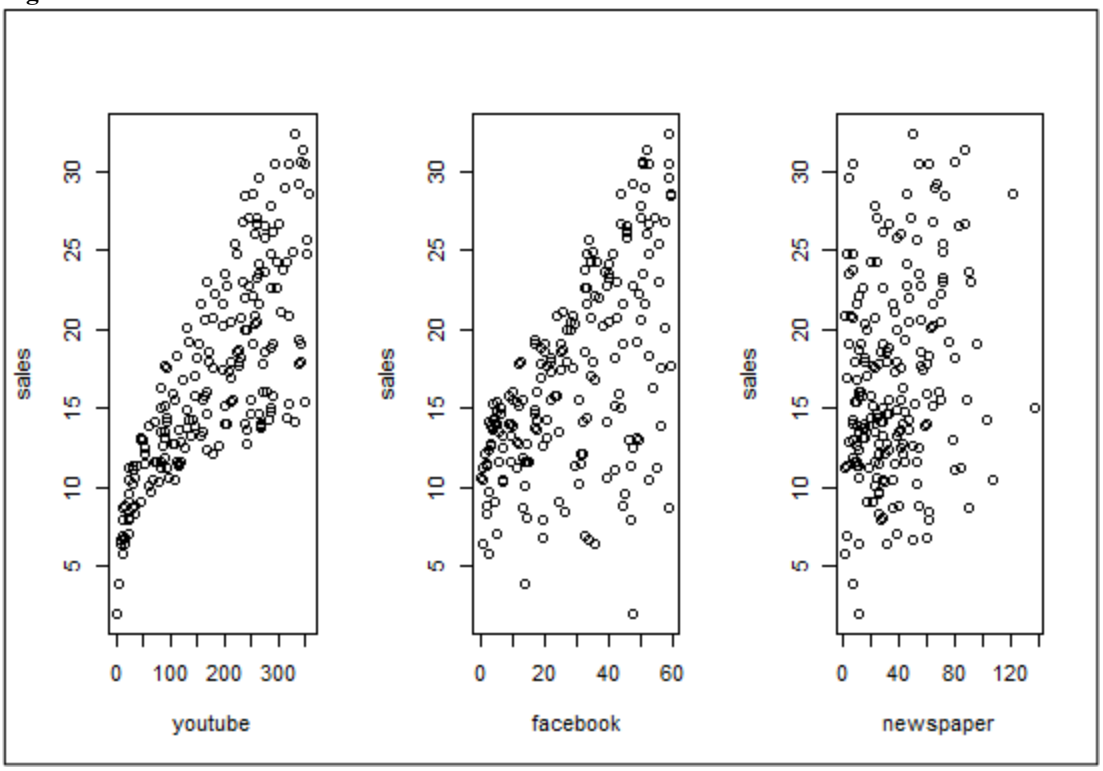
Scatter plots also help you visualize the relationship between sales and each explanatory variable (i.e., YouTube, Facebook, or newspaper). The function par(mfrow=c(nrows, ncols)) allows you to combine many plots in a single graph in R, i.e., a matrix of nrows by ncols plots. The mfrow argument specifies the dimensions of the grid, indicating the number of rows (nrows) and columns (ncols) of plots you want to arrange.

par(mfrow=c(1,3))

plot(md$youtube,md$sales,xlab = "youtube",ylab="sales")plot(md$facebook,md$sales,xlab = "facebook",ylab="sales")

plot(md$newspaper,md$sales,xlab = "newspaper",ylab="sales")

Figure 3-12



From **Figure 3-12**, the relationships between Facebook and sales, and YouTube and sales appear both positive and linear. The data is clumped in a line shape, rising from left to right. However, the relationship between YouTube and sales appears to be stronger than that between Facebook and sales because the line is clearer and the data is more tightly clustered. The relationship between newspaper and sales does not appear to be linear.

Compute the numerical value of the correlation between sales and each advertising medium can be done using the cor() function as follows:

```
cor(md)
```

This code produces the following output:

```
             YouTube   Facebook  newspaper     sales
YouTube    1.00000000 0.05480866 0.05664787 0.7822244
Facebook   0.05480866 1.00000000 0.35410375 0.5762226
newspaper 0.05664787 0.35410375 1.00000000 0.2282990
sales      0.78222442 0.57622257 0.22829903 1.0000000
```

From the last column of the R output, the correlation between sales and YouTube (0.78 to 2 decimal places) is shown to be stronger than that between sales and Facebook (0.58 to 2 decimal places). The correlation between sales and newspapers is weak (0.23 to 2 decimal places).

After the correlation analysis, further analysis of the relationship between sales and the explanatory variable YouTube is the next step because these two variables have a linear relationship and the strongest correlation.

Save the marketing data in the previously created DATA folder of your desktop as *marketing.csv* to use in later exercises of the lesson:

write.csv(md, file = 'C:/Users/INSERT-YOUR-USER-NAME-HERE/Desktop/DATA/marketing.csv', row.names = FALSE)

Correlation analysis using a real dataset in Excel

Correlation analysis using a real dataset in Excel

First, import the *marketing.csv* dataset to a new worksheet.

Create a scatter plot (or chart) to investigate the relationship between sales and YouTube using the following steps:

- Change the tab name to Data_Relationship if it is not already.
- Click on the new worksheet anywhere far away from the data.
- Click **Insert** > **Scatter** (under the charts section) to create an empty scatter chart.
- Click **Chart Design Tools** > **Select Data** to open the **Select Data Source** dialog box.
- Click **Add** in **Legend Entries (Series)** to open the **Edit Series** dialog box.
- Use the **Edit Series** dialog box to choose the appropriate data ranges for each axis. X values should correspond to YouTube values (in the range A2:A201), and Y values should correspond to sales values (in the range D2:D201); refer to **Figure 3-13**.

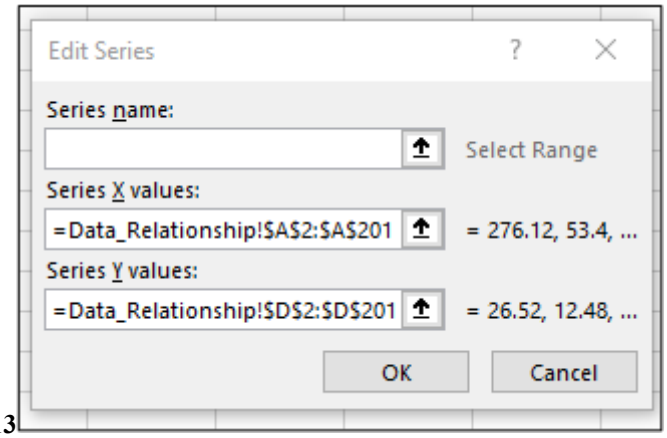


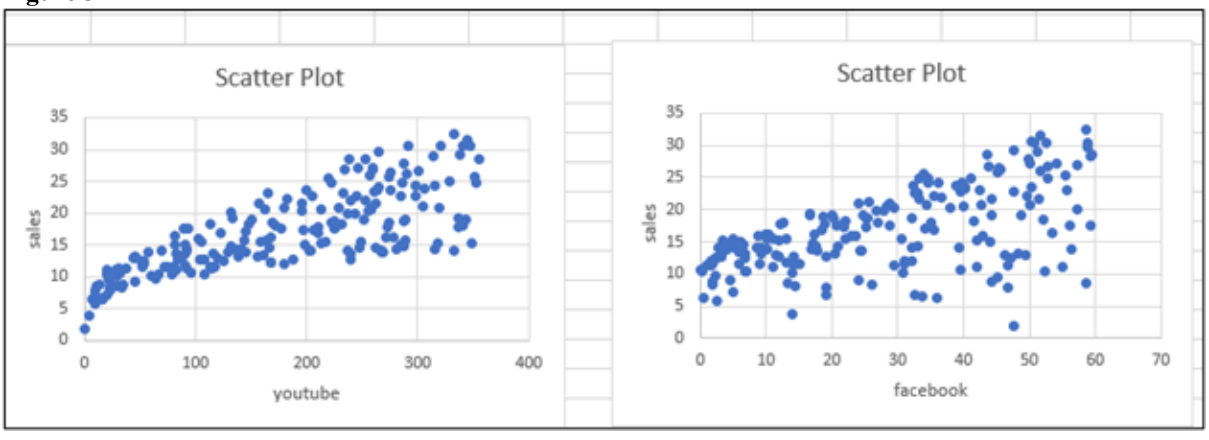
Figure 3-13

- Click **OK**. You can then add a chart title and axes titles to the graph.

Repeat the steps above to construct scatter plots of sales against Facebook and sales against newspaper.

Figure 3-14 shows scatter plots describing the relationships between YouTube and sales and Facebook and sales.

Figure 3-14

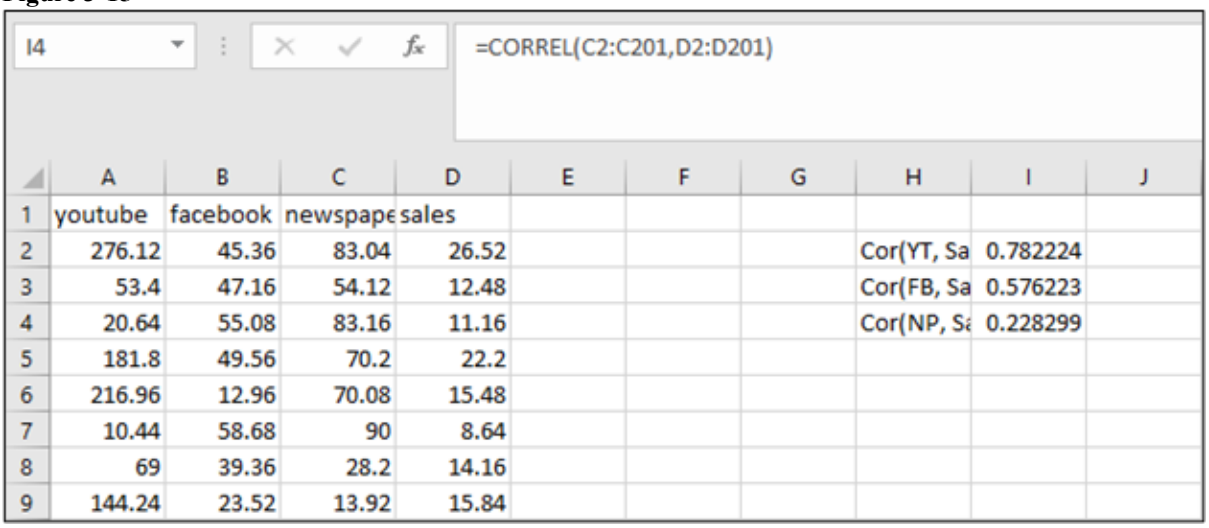


To calculate the Pearson correlation coefficient in Excel, use the function CORREL(range1, range2), where range1 and range 2 are the cell references containing the data.

Correlation formulas can be added in the cells I2, I3, and I4 of the worksheet for correlations between YouTube and sales, i.e., =CORREL(A2:A201,D2:D201), Facebook and sales, i.e., =CORREL(B2:B201,D2:D201), and newspaper and sales, i.e., =CORREL(C2:C201,D2:D201), respectively.

Figure 3-15 shows the results of these computations in Excel.

Figure 3-15



	A	B	C	D	E	F	G	H	I	J
1	youtube	facebook	newspape	sales						
2	276.12	45.36	83.04	26.52				Cor(YT, Sa	0.782224	
3	53.4	47.16	54.12	12.48				Cor(FB, Sa	0.576223	
4	20.64	55.08	83.16	11.16				Cor(NP, Sa	0.228299	
5	181.8	49.56	70.2	22.2						
6	216.96	12.96	70.08	15.48						
7	10.44	58.68	90	8.64						
8	69	39.36	28.2	14.16						
9	144.24	23.52	13.92	15.84						

Correlation and scatter charts in Excel Part 1

Correlation analysis in Excel

Correlation and scatter charts in Excel Part 2