8 Correlation Analysis

Success is not the key to Happiness; Happiness is the key to success.
If you love what you are doing you will be successful.

—Albert Schweitzer

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Understand the concept of correlation and its role in analytics.
- Learn to calculate correlation between two continuous random variables.
- Understand the difference between correlation and causation.
- Understand correlation between a continuous variable and a discrete variable.
- Learn to calculate correlation between two binary variables.

Correlation

Correlation is a statistical measure of an associative relationship between two random variables. It is not necessarily a causal relationship. Correlation is an important concept in analytics, as it helps identify variables that may be used in model building and is also useful for identifying issues such as multi-collinearity that can destabilise regression-based models. Correlation is also useful for finding proxy variables in analytics model building.

8.1 | Introduction to Correlation

One of the more challenging tasks in analytics, especially in predictive analytics, is identifying the variables or features that may be associated with the response or outcome variable of interest. Organizations collect data on several variables, whose numbers can sometimes run into thousands (including derived variables such as ratios and interactions). For example, mobile service providers collect data regarding variables such as call duration, number of calls, phone numbers to which calls have been made, number of calls received, characteristics of the device that was used to make the call, location (and the mobile tower that the phone was attached to), time between calls, last recharge (in case of pre-paid

mobile services), recharge amount, service plan (in case of post-paid connection), number of messages sent, number of messages received, apps downloaded, time spent on surfing the internet, and so on. The number of variables collected, and new variables generated (through feature engineering), may exceed several thousands. A few of these variables are government regulatory requirements that mobile service providers are expected to collect and store. The idea behind the collection of all these variables is to find answers to questions such as

1. Which customer is likely to churn?
2. How can the revenue generated from a customer be increased?
3. What is Customer Lifetime Value (CLV)?
4. What is the best service plan for a customer?
5. What recommendations regarding service plan can be made to a customer?

Finding answers to these questions involves building predictive/prescriptive analytics models. Model building involves identifying relevant variables from thousands that are available to build the model (in analytics terminology, this is called *feature selection*). Taking all the variables simultaneously to create a model can result in problems such as multi-collinearity, which can destabilise the model and is also time consuming since most predictive analytics models involve matrix operations such as matrix inverse calculation. So, the knowledge of how different variables are related to one another is important in building analytical models. **Correlation** is a measure of the strength and direction of the relationship that exists between two random variables; in other words, it is a measure of association between two variables. It is measured using the correlation coefficient. We will be discussing different types of correlation coefficients (depending on the scale of measurement of the variables involved) in this chapter.

Correlation is only an associative relationship and not necessarily a causal one. Thus, the user should know that two variables may have a high correlation coefficient value, without the existence of any direct dependence between these variables.

Mozart Effect

Rauscher *et al.* (1993) made the following claim in their paper published in Nature:

> "We performed an experiment in which students were each given three sets of standard IQ spatial tasks. Each task was preceded by 10 minutes of (1) Listening to Mozart's Sonata for two piano's in D major, K488, (2) Listening to a relaxation tape and (3) Silence. Performance was improved for those tasks immediately following the first condition compared to the other two."

Many researchers questioned this claim, but many others strongly believed in it – having interpreted it incorrectly. But media went overboard as usual and started projecting the article using the headline, "Mozart makes you smart". In 1998, the Governor of Georgia, Zell Miller, proposed to spend USD 105,000 to distribute music CDs to newborn babies each year (Sack, 1998). Miller argued that listening to classical music at a young age improves intelligence. The major problem with this is that Miller assumed a causal relationship between classical music and intelligence.

Causal relationships imply that the occurrence of one event causes the other to occur (effect). Correlation between two events does not imply causation.

8.2 | Pearson Correlation Coefficient

Pearson Product Moment Correlation (in short Pearson Correlation) is used for measuring the strength and direction of the linear relationship between two continuous random variables, X and Y. For example, consider two variables – average call duration (variable Y) and age (variable X). We may like to know whether average call duration is related to the age of the caller. It is possible that there may not be any relationship between age and the average call duration. A simple approach for checking the existence of an association relationship is to draw a scatter plot. In Table 8.1, we have age of customer and average call duration (measured in seconds) from sample data; the corresponding scatter plot is shown in Figure 8.1.

Table 8.1 | Data on age and average call duration (in seconds)

Age	14	15	18	19	20	24	25	27	29	30
Call duration	540	544	567	548	550	520	512	516	511	511
Age	33	36	38	39	40	41	42	43	45	48
Call duration	490	487	472	460	455	463	440	422	411	397

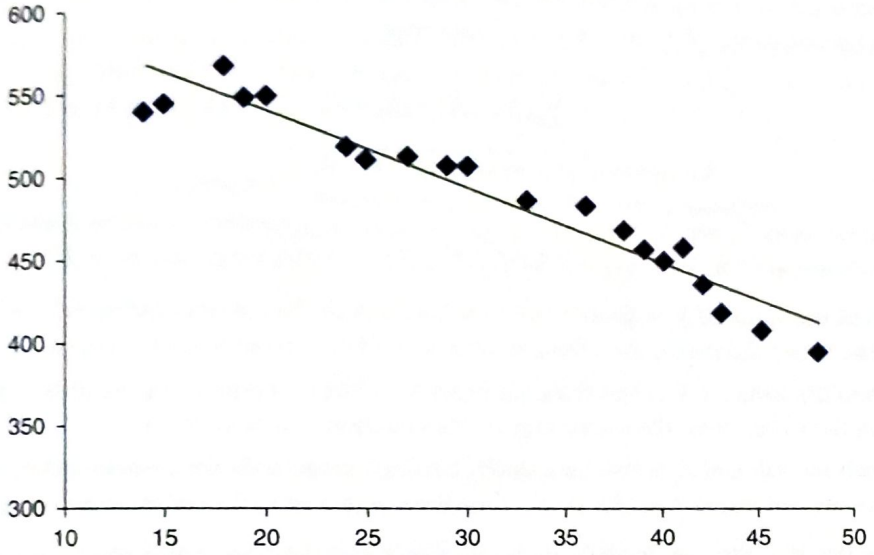

Figure 8.1 | Association relationship between age and average call duration.

In Figure 8.1, we can see that the average call duration (Y) decreases as the age of the customer (X) increases. We can calculate the strength of the linear association relationship using a numerical measure called correlation coefficient. In the next section, we will discuss mathematical equations for calculating Pearson Product Moment Correlation Coefficient.

8.2.1 Calculation of Pearson Product Moment Correlation Coefficient

Pearson Product Moment Correlation is used when we are interested in finding a linear relationship between two continuous random variables (that is, the variable should be either a ratio or an interval scale). When we try to measure how the change in one variable (say variable Y) is related to changes in another variable (say variable X), one issue that we need to consider is the measurement scale and unit of measurement of the two variables. In the example discussed in Table 8.1, the variable age is measured in years and call duration in seconds. The ranges of the two variables of interest can be different; we thus need to standardize the variables for measurement of correlation. The mathematical theory of correlation was developed by Francis Galton and Karl Pearson (Pearson, 1920).

Let X_i be different values of the variable X and Y_i be different values of Y, such that (X_i, Y_i) form dyads. Then the standardized values of X_i and Y_i are given by

$$Z_X = \left(\frac{X_i - \bar{X}}{\sigma_X} \right) \tag{8.1}$$

$$Z_Y = \left(\frac{Y_i - \bar{Y}}{\sigma_Y} \right) \tag{8.2}$$

where $\bar{X}$ and $\bar{Y}$ are mean values of random variables X and Y; σ_X and σ_Y are the corresponding standard deviations. The Pearson's Correlation coefficient is given by

$$r = \frac{\sum_{i=1}^{n} Z_X Z_Y}{n} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \times (Y_i - \bar{Y})}{n\sigma_X \sigma_Y} \tag{8.3}$$

where n is the number of cases in the sample. The formula in Equation (8.4) is also frequently used to account for the degrees of freedom and is recommended when the standard deviation is calculated from a sample. For large samples, the correlation coefficients calculated using Equations (8.3) and (8.4) will converge.

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \times (Y_i - \bar{Y})}{(n-1)S_X S_Y} \tag{8.4}$$

In Equation (8.4), S_X and S_Y are the standard deviations of random variables X and Y calculated from the sample. We can note the following properties from Equation (8.3):

1. When the value of X_i is greater than the mean with the corresponding value of Y_i also greater than the mean, then the numerator in the equation will be positive.
2. When the value of X_i is less than the mean with the corresponding value of Y_i also less than the mean, then the numerator in the equation will be positive.
3. When the value of X_i is less (or greater than) the mean with the corresponding value of Y_i greater (or lesser than) the mean, then the numerator in the equation will be negative.

It is possible that we may have combinations of the three cases listed above in a dataset. Thus, the numerator in Equation (8.3) is likely to be positive, negative or zero. The value of Pearson's Correlation coefficient lies between −1 and +1. Equation (8.3) is mathematically equivalent to Equations (8.5), (8.6) and (8.7):

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \tag{8.5}$$

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \times \sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}} \tag{8.6}$$

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{8.7}$$

In Equation (8.7), Cov(X, Y) is the covariance between random variables X and Y, and is given by

$$\text{Cov}(X,Y) = E((X_i - \bar{X})(Y_i - \bar{Y})) \tag{8.8}$$

8.2.2 Properties of Pearson Correlation Coefficient

1. The value of correlation coefficient always lies between −1 and +1. High absolute value of r, $|r|$, indicates a strong association relationship between the two variables.
2. Positive value of r indicates positive correlation (as the value of X increases, the value of Y also increases) and negative value of r indicates negative correlation (as the value of X increases, the value of Y decreases).
3. The sign of the correlation coefficient is same as the sign of covariance between the two random variables.

4. Assume that the value of **Pearson Correlation coefficient** between X and Y is r. Let Z_1 and Z_2 be linear combinations of X and Y ($Z_1 = A + BX$ and $Z_2 = C + DY$). Then, the correlation coefficient between Z_1 and Z_2 will be r when the signs of B and D are the same (both are positive or negative) and $-r$ when the signs of B and D are opposite.
5. Mathematically, the square of the correlation coefficient is equal to the coefficient of determination (R^2) of the linear regression model, that is $r^2 = R^2$.
6. The Pearson Correlation coefficient may be zero even when there is a strong non-linear relationship between variables X and Y (Reed, 1917). Thus, low correlation coefficient value cannot be taken as evidence of no relationship.

Pearson Correlation coefficient is a measure of linear relationship. Pearson Correlation may not capture the existence of non-linear relationships.

Example 8.1

The average share prices of two companies over the past 12 months are shown in Table 8.2. Calculate the Pearson Correlation coefficient.

Table 8.2 | Share prices (monthly average) of two companies over the last 12 months

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

The average values are: $\bar{X} = 292.9717$ and $\bar{Y} = 229.8292$.

Solution

The following equation is used to calculate the correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

The calculations are shown in Table 8.3.

Table 8.3 | Calculation of correlation coefficient

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
274.58	219.50	−18.39	−10.33	189.97	338.25	106.6917
287.96	242.92	−5.01	13.09	−65.61	25.12	171.3699
290.35	245.90	−2.62	16.07	−42.13	6.87	258.2717
320.07	256.80	27.10	26.97	730.86	734.32	727.4259
317.40	240.60	24.43	10.77	263.11	596.74	116.0109

(Continued)

Table 8.3 | *(Continued)*

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
319.53	245.23	26.56	15.40	409.02	705.35	237.1857
301.52	232.09	8.55	2.26	19.33	73.07	5.111367
271.75	222.65	−21.22	−7.18	152.35	450.36	51.54043
323.65	231.74	30.68	1.91	58.62	941.16	3.651284
259.80	214.43	−33.17	−15.40	510.82	1100.36	237.1343
263.02	201.86	−29.95	−27.97	837.72	897.10	782.2743
286.03	204.23	−6.94	−25.60	177.70	48.19	655.3173
Sum				3241.77	5916.89	3351.98

From Table 8.3, we have

$$\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y}) = 3241.77$$

$$\sum_{i=1}^{12}(X_i - \bar{X})^2 = 5916.89$$

$$\sum_{i=1}^{12}(Y_i - \bar{Y})^2 = 3351.98$$

$$\text{Correlation coefficient } r = \frac{3241.77}{\sqrt{5916.89} \times \sqrt{3351.98}} = 0.7279$$

In Microsoft Excel, CORREL(array 1, array 2) will give the Pearson Product Moment correlation value.

8.2.3 Explainable and Spurious Correlation

We classify correlation into 1. Explainable correlation and 2. Spurious correlation. If the correlation between two variables can be explained using logical reasoning, we call it explainable correlation; otherwise, it will be labelled as a spurious correlation. One of the major problems with correlation is the possibility of spurious correlation between two random variables, which in many cases is caused due to some other latent variable (hidden variable) that influences both variables. The following are a few examples of spurious correlation (some of them can be explained through hidden variable) between two random variables:

1. ***Crime rate versus ice cream sale:*** *It has been reported that the sales of ice cream and crime rates are positively correlated (Levitt and Dubner, 2009). Obviously, ice cream is not driving crime rates. In this case, the hidden variable is the temperature – ice cream sales increase during summer, a season that sees an increase in crime as a result of people going on vacation, leaving their homes to become easy targets.*
2. ***Doctors and deaths:*** *The number of doctors is positively correlated with the number of deaths in villages, that is, as the number of doctors increases, the number of deaths also increase. We can be sure that doctors are not causing an increase in deaths (Young, 2001). In this case, the need for doctors is more in villages with poor public health.*
3. ***Divorce rate in Maine and per capita consumption of margarine:*** *The divorce rate in Maine was highly correlated with per capita consumption of margarine (based on data between 1999 and 2009. The correlation coefficient value was 0.9926 (Source: tylervigen.com[1]). This is completely spurious since we are not able to find any hidden variable that is causing this high correlation.*

[1] http://www.tylervigen.com/spurious-correlations

4. **Skirt Length (Hemline) Theory:** This theory, first proposed by the economist George Taylor in 1926, claimed that the average length of women's skirts and the economy are related (Kim and Kim 2012). According to this theory, skirts get shorter when the economy is doing well and get longer when the economy is doing badly. The reason for longer skirts being in fashion during times of economic downturn, according to Taylor, was that women could no longer afford stockings (Fitzgerald, 2012).
5. **Proximity to Freeways Causes Autism:** Volk *et al.* (2011) claimed that residential proximity to freeways caused autism. The study also claimed that living near other major roads at birth was not associated with autism.

8.2.4 Hypothesis Test for Correlation Coefficient

For any two sets of data, the Pearson Correlation coefficient calculated using Equation (8.7) is likely to be non-zero. Many thumb rules exist to classify the correlation value as no correlation, low correlation, medium correlation and high correlation (Monroe and Stuit, 1933). For example, a correlation coefficient value of less than 0.2 can be considered as signifying negligible correlation, and a value above 0.7 as high correlation (Monroe and Stuit, 1933). We would like to know what the minimum value of the Pearson Correlation coefficient should be before it can be considered statistically significant. Let ρ be the population correlation coefficient. The null and alternative hypotheses are given by

H_0: $\rho = 0$ (there is no correlation between two random variables)

H_A: $\rho \neq 0$ (there is a correlation between two random variables)

The sampling distribution of correlation coefficient r follows an approximate t-distribution with $(n-2)$ degrees of freedom (df), where n is the number of records in the sample used for calculating the correlation coefficient. Two degrees of freedom are lost since we estimate two mean values from the data. The mean of the sampling distribution is ρ and the corresponding standard deviation is (Ezekiel, 1941)

$$\sqrt{\frac{1-r^2}{n-2}} \tag{8.9}$$

The t-statistic for null hypothesis is given by

$$t_{\alpha/2,n-2} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} \tag{8.10}$$

When the null hypothesis is $\rho = 0$, the test statistic in Equation (8.10) becomes

$$t_{\alpha/2,n-2} = r\sqrt{\frac{n-2}{1-r^2}} \tag{8.11}$$

Example 8.2

For Example 8.1, conduct the following two hypothesis tests at $\alpha = 0.05$:

(a) The correlation between the share prices of two companies is zero.

(b) The correlation between the share prices of two companies is at least 0.5.

Solution

(a) The null and alternative hypotheses are:

H_0: $\rho = 0$

H_A: $\rho \neq 0$

The corresponding t-statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.7279\sqrt{\frac{12-2}{1-0.7279^2}} = 3.3569$$

Note that this is a two-tailed test and the critical t-value at $\alpha = 0.05$ and $df = 10$ is 2.2281 [which can be obtained using the Excel function TINV(0.05, 10)]. Since the calculated t-statistic is higher than the critical t-value, we reject the null hypothesis and conclude that there is a statistically significant correlation between the share prices of the two companies. The corresponding p-value is 0.0072 (In Excel T.DIST.2T(3.3569, 10) = 0.0072).

(b) The null and alternative hypotheses are given by

$$H_0: \rho \leq 0.5$$

$$H_A: \rho > 0.5$$

The corresponding t-statistic is

$$t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7279-0.5}{0.2168} = 1.05$$

This is a right-tailed test and the corresponding t-critical value is 1.8124 [the corresponding Excel function is T.INV(0.1, 10)]. The calculated t-value is less than the critical value of t, and thus we retain the null hypothesis and conclude that the correlation between share prices of the two companies is less than or equal to 0.5. The corresponding p-value is 0.1592 [T.DIST.RT(1.05, 10) = 0.1592].

(a) Less than 0.85 (b) More than 0.85
(c) 0.85 (d) −0.85

5. The covariance between two random variables is 0.5; correlation between these two random variables will be
(a) At least 0. 5
(b) At most 0.5
(c) 0.25
(d) Cannot predict without the standard deviation values.

Exercises

1. Professor Bell at Bellandur University, Bangalore believes that the Cumulative Grade Point Average (CGPA) of students is negatively correlated with usage (measured in average minutes per day) of smartphones. Table 8.10 shows the CGPA and smartphone usage in minutes per day of 40 students.
(a) Calculate the Pearson Correlation coefficient between CGPA and mobile phone usage of students.
(b) Conduct a hypothesis test at $\alpha = 0.01$ to check whether CGPA and mobile phone usage are negatively correlated.
(c) Professor Bell believes the correlation is less than −0.4. Conduct a hypothesis test at $\alpha = 0.1$ to check whether the claim is correct.

Table 8.10 | Data of CGPA and mobile phone usage (average minutes per day)

CGPA	2.65	2.25	1.86	1.47	2.10	1.94	2.71	1.83	2.65	2.04
Phone usage	75	89	65	136	95	103	74	109	75	98
CGPA	2.54	2.16	2.28	2.47	2.18	2.57	1.97	2.87	2.10	3.28
Phone usage	60	93	88	81	92	78	102	70	95	89
CGPA	2.78	2.44	1.87	2.50	2.24	2.01	2.17	2.20	2.05	1.63
Phone usage	72	82	107	80	89	100	92	91	98	123
CGPA	2.28	2.63	2.86	2.24	2.44	2.69	2.22	3.07	1.77	3.03
Phone usage	88	76	70	89	82	74	90	65	113	66

2. Mr Chellappa is the founder of Oho Productions, which produces movies in various Indian languages. Mr Chellappa believes the length of the movie (measured in minutes) is related to its box office collections. Table 8.11 shows length of the movie (in minutes) and box office collections (in millions of rupees). Use an appropriate hypothesis test to check whether there is any correlation between the length of the movie and box office collections (in millions of rupees) at a significance level of 0.05.

Table 8.11 | Data on length of the movie and box office collections (in millions of rupees)

Length of the movie	121	79	170	160	77	147	115	76	110	141
Box office collection	1,078	415	441	1,192	258	1185	139	427	309	411
Length of the movie	100	82	82	114	110	163	92	172	142	136
Box office collection	506	441	595	1,728	1507	518	1463	1356	1014	422
Length of the movie	143	108	154	140	177	97	106	163	142	115
Box office collection	508	1262	1783	1,281	1253	1178	1103	454	301	296

3. Table 8.12 provides ranking of Indian states based on corruption and Table 8.13 provides ranking based on literacy rate. Calculate the Spearman rank correlation between corruption rank and literacy rank.

Table 8.12 | Rank based on corruption (1 implies high corruption)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	1	2	3	4	5	6	7	8
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	9	10	11	12	13	14	15	16

Table 8.13 | Rank based on literacy rate (1 implies high literacy)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	16	12	10	11	7	15	4	9
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	2	5	13	1	8	14	6	3

Conduct a hypothesis test to check whether corruption and literacy rate are negatively correlated at $\alpha = 0.05$.

4. Harrison Seth, Dean of a Business School, believes the outgoing salary of his MBA students may be correlated with their undergraduate specialization. Harrison believes students with engineering specialization in undergrad received higher salaries compared to those with other degrees. Table 8.14 shows the outgoing salary (in millions of rupees) of MBA graduates and their discipline in undergraduate (1 = engineering and 0 = non-engineering). Calculate the correlation between salary and engineering discipline.

Table 8.14 | Salary (in millions of rupees) and undergraduate degree
(1 = engineering and 0 = non-engineering)

Degree	0	1	0	1	0	0	1	0	0	1
Salary	3.3	2.22	1.82	2.55	1.84	2.53	2.87	2.39	2.32	2.79
Degree	1	1	0	1	0	0	1	1	0	0
Salary	2.22	2.31	2.05	2.04	1.7	2.28	2.56	3.13	2.26	2.56
Degree	0	0	0	0	1	0	0	0	1	1
Salary	2.03	1.45	1.62	0.92	2.31	2.37	1.59	2.56	3.13	3

5. Telepower is a telephone service provider which collects data on customer churn and the number of mobile handsets used by each customer. Table 8.15 shows the data in which Y denotes churn ($Y = 1$ implies churn and $Y = 0$ implies no churn), and variable X denotes the number of handsets used by the customer where $X = 0$ implies the customer uses single handset and $X = 1$ implies the customer uses more than one handset for making phone calls. Calculate the Phi-coefficient for the data shown in Table 8.15.

Table 8.15 | Number of handsets (X) and customer churn (Y)

X	1	1	0	0	0	1	1	1	1	1
Y	1	1	1	1	0	0	1	0	1	1
X	0	1	1	1	1	0	0	1	1	1
Y	0	1	0	1	1	0	0	1	1	1
X	1	1	1	0	1	0	1	0	1	1
Y	0	1	1	0	1	0	0	1	1	1
X	1	1	1	1	0	1	1	0	1	1
Y	0	1	0	1	1	1	1	0	0	1
X	0	0	1	0	1	0	1	1	0	1
Y	0	0	1	1	1	0	0	1	1	1

14.4 Clustering Algorithms

The distance measures discussed in Section 14.2 can be used to group data into useful and differentiable groups. Clustering algorithms group data into a finite number of mutually exclusive subsets. Assume that S is the set of all observations. Non-overlapping clustering algorithms attempt to create subsets (C_j) of S, such that $S = \bigcup_{j=1}^{n} C_j$, and for any two subsets C_i and C_j, $C_i \cap C_j = \phi$ (null set) for $i \neq j$. There are many clustering algorithms that use different logic and distance/similarity measures. The objective of the clustering techniques is to create groups such that the variation within the group is minimized and

the variation between the groups is maximized. In this section, we will be discussing the two most frequently used clustering methods, namely, *K*-means clustering and Hierarchical clustering. The following steps are followed in clustering algorithms:

1. Variable selection.
2. Deciding the distance/similarity measure for measuring distance/dissimilarity between the observations.
3. Deciding the number of clusters.
4. Validation of the clusters.

We will discuss these steps in detail next.

14.4.1 Variable Selection

Ketchen and Shook (1996) suggest inductive, deductive, and cognitive approaches for variable selection. Inductive is basically an exploratory approach and starts with as many variables as possible. On the other hand, in deductive variable selection, suitability of the variable and theoretical basis influence the selection of variables. Under cognitive variable selection, expert opinion plays a major role (Ketchen and Shook, 1996).

14.4.2 Deciding Distance/Similarity Measures

Choosing the right distance/similarity measure plays an important role in developing clusters. For example, Euclidean distance is valid only for variables under interval and ratio scales. However, for qualitative variables, Euclidean distance is not valid. Similarity measures such as Jaccard coefficient should be used for binary variables. Cosine similarity can be used for both quantitative and qualitative variables. If the data has both qualitative and quantitative variables, then one may have to use measures such as Gower's distance.

14.4.3 Number of Clusters

Several approaches are available for deciding the number of clusters such as *CH* index [Eq. (14.13)], Hartigan statistic [Eq. (14.14)], Silhouette statistic [Eq. (14.15)], and elbow method in which the ideal number of clusters is given by the position of elbow in an *L*-shaped curve.

14.4.4 Cluster Validation

The clusters created should be validated for consistency using different algorithms to ensure that the clusters represent the structures that exist in the population. Halkidi *et al.* (2001) suggest the following measures to validate the clusters:

1. **Compactness:** Closeness of each member of a cluster which can be measured through variance.
2. **Separation:** Distance between different clusters.

14.5 | *K*-Means Clustering

K-means clustering is one of the frequently used clustering algorithms. It is a non-hierarchical clustering method in which the number of clusters (K) is decided *a priori*. The observations in the sample are assigned to one of the clusters (say $C_1, C_2, \ldots, C_K$). The following steps are used in *K*-means clustering algorithm:

1. Choose K observations from the data that are likely to be in different clusters. There are many ways of choosing these initial K values; the easiest approach is to choose observations that are farthest (in one of the parameters of the data).
2. The K observations chosen in step 1 are the centroids of those clusters.

3. For the remaining observations, find the nearest cluster using appropriate distance/similarity measure. Add the new observation (say observation j) to the nearest cluster. Adjust the centroid value after adding a new observation to the cluster.
4. Repeat step 3 till all observations are assigned to a cluster.
5. Repeat the process of identifying the cluster for each observation by finding the minimum distance between the observation and centroid of various clusters. This step is carried out since the centroid keeps moving with the addition of observations.

Note that centroids keep moving when new observations are added; also observations may move to different clusters. The following criteria can be used as stopping criteria to stop K-means clustering:

1. The maximum number of iterations is achieved (usually chosen as 10)
2. No change is observed in cluster membership.
3. Negligible change in centroid of the clusters.

An important aspect of K-means clustering is choosing the appropriate value of K. Initially the value of K is a guess; however, it can be decided based on several measures such as $CH(K)$ index, Silhouette coefficient and elbow method.

K-means clustering is used to group 149 Bollywood movies (Data file: Bollywood Data Clustering.xls). The variables used along with descriptive statistics are given in Table 14.9.

Table 14.9 | Bollywood movie data for clustering

Variable	Minimum	Maximum	Mean	Standard Deviation
Box-office collection	0.01 (in crores)	735 (in crore)	55.67	94.49
Profit	−56.80	650.00	26.24	79.09
Earnings ratio (Ratio of box-office collection over budget)	0.01	9.17	1.77	1.84
Budget	1.8 (crores)	150 (crores)	29.43 (crores)	28.25 (crores)
YouTube views	4354	23171067	3337919.91	3504406.99
YouTube likes	1	101275	7877.54	12748.04
YouTube dislikes	1	11888	1207.82	1852.69

K-mean clustering output for $K = 3$ using SPSS is shown in Tables 14.10–14.13.

Table 14.10 | Final cluster centres

	Cluster		
	1	2	3
Box_Office_Collection	306.10	72.89	32.42
Profit	215.801666	34.2598	10.9492
Earning_Ratio	3.40	2.08	1.53
Budget	90.3	38.6	21.5
Youtube_Views	16399358	5506403	1542508
Youtube_Likes	52311	12857	2871
Youtube_Dislikes	7169	2068	448

Table 14.10 shows the mean values of variables in each cluster (in which the value of K is chosen as 3). Cluster centers in Table 14.10 help us to identify characteristics of various clusters. For example, the budget (and profit) of movies for cluster 1 is much higher than clusters 2 and 3. The number of Youtube likes is much higher for cluster 1 compared to clusters 2 and 3.

Table 14.11 | Distances between final cluster centres

Cluster	1	2	3
1		10893027.904	14856933.416
2	10893027.904		3963907.285
3	14856933.416	3963907.285	

Table 14.11 shows the Euclidean distance between centroids of the three clusters. A large value of distance between cluster centres indicate better separation between clusters. However, note that the scales of different variables are different and hence ideally the data should be normalized before performing cluster analysis. ANOVA for the variables used in clustering is shown in Table 14.12. Higher value of *F* indicates higher level of contribution of that variable in clustering. Variable that are not significant (say significance or p-value greater than 0.05) imply that the average values of those variable in different clusters are not significantly different. We can observe in Table 14.12, that all the variables are statistically significant.

Table 14.12 | ANOVA for variables used in clustering[a]

	Cluster		Error			
	Mean Square	*df*	Mean Square	*df*	*F*	Sig.
Box_Office_Collection	221299.22	2	6020.03	146	36.76	0.000
Profit	120704.64	2	4687.10	146	25.75	0.000
Earning_Ratio	12.98	2	3.24	146	4.00	0.020
Budget	16121.67	2	588.17	146	27.41	0.000
Youtube_Views	775557272679141	2	1825027216893.38	146	424.95	0.000
Youtube_Likes	7709212606.51	2	59133256.50	146	130.37	0.000
Youtube_Dislikes	151589730.46	2	1402919.62	146	108.05	0.000

[a]The *F*-tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The number of observations in three different clusters is shown in Table 14.13.

Table 14. 13 | Number of cases in each cluster

Cluster	1	6
	2	45
	3	98
Valid		149
Missing		0

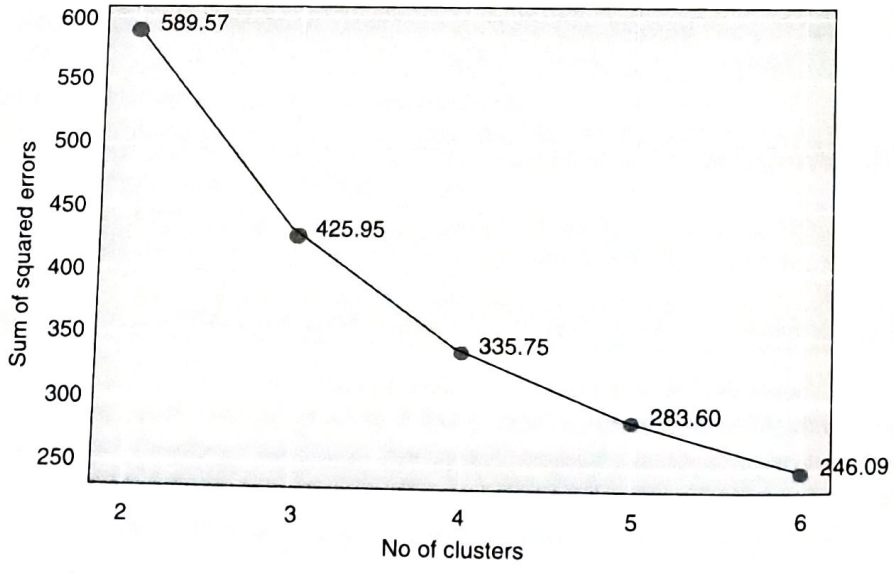

Figure 14.3 | Elbow curve for the Bollywood data.

The elbow curve for the clusters developed using the Bollywood data is shown in Figure 14.3 (obtained using R programming language). Based on the elbow curve we can conclude that the optimal number of clusters in this case is 3 (bend seems to appear when the number of clusters is 3). Elbow curve is a plot of decrease in of sum of squared errors (SSE) versus change in the number of clusters (as the value of K increases). A very small decrease in SSE will indicate insignificant improvement in the model. This is captured by the elbow of the curve.

14.6 Hierarchical Clustering

Hierarchical clustering is a clustering algorithm which uses the following steps to develop clusters:

1. Start with each data point in a single cluster.
2. Find the data points with shortest distance (using an appropriate distance measure) and merge them to form a cluster.
3. Repeat step 2 until all data points are merged to form a single cluster.

The above procedure is called **agglomerative hierarchical cluster**. Agglomerative hierarchical clustering is explained by using the data in Table 14.14. There are 8 data points ($D_1, D_2, \ldots, D_8$) in Table 14.14 and distances between each of these observations are given.

Table 14.14 | Data points with distances

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
D_1	0	0.45	0.36	0.71	1.00	0.27	0.38	0.21
D_2	0.45	0	0.19	0.36	0.54	0.30	0.91	0.76
D_3	0.36	0.19	0	0.87	0.54	0.72	0.28	0.64
D_4	0.71	0.36	0.87	0	0.34	0.51	0.43	0.72
D_5	1.00	0.54	0.54	0.34	0	0.65	0.57	0.41
D_6	0.27	0.30	0.72	0.51	0.65	0	0.33	0.68
D_7	0.38	0.91	0.28	0.43	0.57	0.33	0	0.44
D_8	0.21	0.76	0.64	0.72	0.41	0.68	0.44	0

In Table 14.14, the closest data points are D_2 and D_3. So, D_2 and D_3 will be merged to form the first cluster, say C_1. Table 14.14 can be modified to denote this cluster as shown in Table 14.15. The distances between cluster C_1 and other data points are calculated based on the maximum distance between the data points in the cluster and other data points which are not part of cluster C_1 (use of average distance is also practiced while calculating distance between an observation and a cluster). For example, the distance between cluster C_1 and data point D_1 is 0.45 since the distance between D_1 and D_2 is 0.45 and the distance between D_1 and D_3 is 0.36. The process is repeated till all data points become part of one single cluster. The agglomerative hierarchical clustering can be represented using a tree-like structure called dendrogram (Figure 14.2).

Table 14.15 | Data points with distances after first cluster

	D_1	$C_1 = \{D_2, D_3\}$	D_4	D_5	D_6	D_7	D_8
D_1	0	0.45	0.71	1.00	0.27	0.38	0.21
$C_1 = \{D_2, D_3\}$	0.45	0	0.87	0.54	0.72	0.91	0.76
D_4	0.71	0.87	0	0.34	0.51	0.43	0.72
D_5	1.00	0.54	0.34	0	0.65	0.57	0.41
D_6	0.27	0.72	0.51	0.65	0	0.33	0.68
D_7	0.38	0.91	0.43	0.57	0.33	0	0.44
D_8	0.21	0.76	0.72	0.41	0.68	0.44	0

Table 14.16 | Data points with distances after second cluster

	$C_1 = \{D_2, D_3\}$	D_4	D_5	D_6	D_7	$C_2 = \{D_1, D_8\}$
$C_2 = \{D_1, D_8\}$	0.76	0.71	1.00	0.68	0.44	0
$C_1 = \{D_2, D_3\}$	0	0.87	0.54	0.72	0.91	0.76
D_4	0.87	0	0.34	0.51	0.43	0.72
D_5	0.54	0.34	0	0.65	0.57	1.00
D_6	0.72	0.51	0.65	0	0.33	0.68
D_7	0.91	0.43	0.57	0.33	0	0.44

The next cluster will be D_1 and D_8. Table 14.16 shows the table after cluster $C_2 = \{D_1, D_8\}$. The process is repeated till all observations become part of one cluster. For the Bollywood data, the SPSS output for hierarchical clustering is shown in Tables 14.17 and 14.18. For demonstration, only the first 8 observations are used.

Table 14.17 | Squared Euclidean distance between cases

	Squared Euclidean Distance							
Case	1	2	3	4	5	6	7	8
1	0	2.51E + 13	5.04E + 13	6.73E + 13	7.41E + 13	7.43E + 13	8.05E + 13	8.3E + 13
2	2.51E + 13	0	4.37E + 12	1.02E + 13	1.29E + 13	1.3E + 13	1.57E + 13	1.68E + 13
3	5.04E + 13	4.37E + 12	0	1.22E + 12	2.27E + 12	2.32E + 12	3.52E + 12	4.05E + 12
4	6.73E + 13	1.02E + 13	1.22E + 12	0	1.61E + 11	1.73E + 11	5.92E + 11	8.2E + 11
5	7.41E + 13	1.29E + 13	2.27E + 12	1.61E + 11	0	2.53E + 08	1.36E + 11	2.55E + 11
6	7.43E + 13	1.3E + 13	2.32E + 12	1.73E + 11	2.53E + 08	0	1.25E + 11	2.4E + 11
7	8.05E + 13	1.57E + 13	3.52E + 12	5.92E + 11	1.36E + 11	1.25E + 11	0	1.87E + 10
8	8.3E + 13	1.68E + 13	4.05E + 12	8.2E + 11	2.55E + 11	2.4E + 11	1.87E + 10	0

Table 14.17 provides squared Euclidean distance between the 8 cases considered for clustering. The schedule of the clustering is provided in Table 14.18. Initially all 8 observations are in individual clusters. In the first stage (first row of Table 14.18), cases 5 and 6 are merged to form a cluster since they have the minimum distance (Table 14.17) among all cases. Columns 5 and 6 report when the cluster has appeared in the immediate past merging stage. For example, in stage 3, cases 4 and 5 are merged, but case 5 appears in stage 1 earlier. This is reflected in column 6. In stage 4, cases 4 and 7 are merged, and case 4 has already appeared in stage 3 (column 5) and case 7 has appeared in stage 2 (column 6). The last column indicates in which stage the cases merged in the current stage appear again. In Table 14.18, the coefficients denote the distance between the clusters.

Table 14.18 | Agglomeration schedule

	Cluster Combined			Stage Cluster First Appears		Next
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Stage
1	5	6	252516641.660	0	0	3
2	7	8	18725562050.840	0	0	4
3	4	5	166865087578.522	0	1	4
4	4	7	361200948423.055	3	2	5
5	3	4	2674635820607.503	0	4	6
6	2	3	12184893092907.690	0	5	7
7	1	2	64974081024938.125	0	6	0

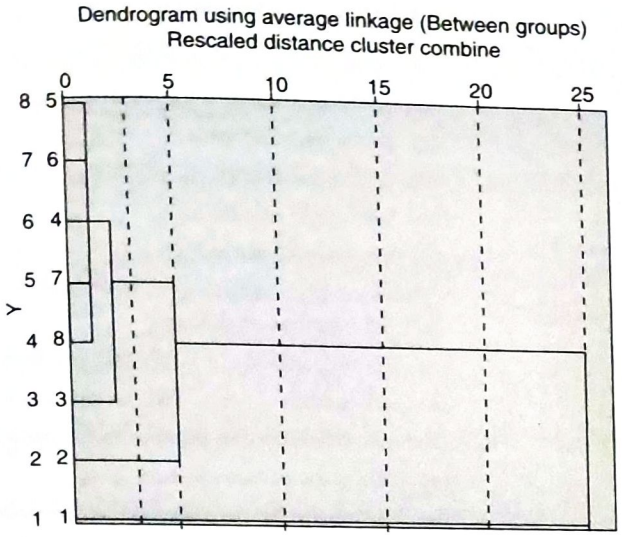

Figure 14.4 | Dendrogram for movie clustering.

The corresponding dendrogram is shown in Figure 14.4. Dendrogram is a pictorial representation of merging of various cases as the Euclidean distance is increased. The distance is rescaled to a scale between 0 and 25 in Figure 14.4. By drawing a vertical line at different values of re-scaled distance, one can identify the clusters. For example, assume that the vertical line is at a scaled distance of 20 (as shown in a dotted line in Figure 14.4). Then there will be 2 clusters $C_1 = \{1\}$ and $C_2 = \{2, 3, 4, 5, 6, 7, 8\}$. If the vertical line is around the rescaled distance of 4, then the number of clusters will be 3 and they are $C_1 = \{1\}$, $C_2 = \{2\}$, $C_3 = \{3, 4, 5, 6, 7, 8\}$.

There are large number of clustering algorithms reported in literature which differ mostly by distance and similarity measures used and logic used for deriving clusters. It is important that the algorithm is able to identify the underlying structure of the clusters for any meaningful use. Once the clusters have been identified, the decision maker has to derive strategies for different clusters to maximize the value generated from each cluster.

Summary

1. Clustering is an unsupervised learning algorithms that divides the dataset into mutually exclusive and exhaustive subsets (in non-overlapping clusters) that are homogeneous within the group and heterogeneous between the groups.
2. Clustering is one of the frequently used techniques and practitioners first cluster the data and develop predictive models for each cluster for better management.
3. Several distance measures such as Euclidian distance, Manhattan distance are used in clustering algorithms. Similarity measures such as Jaccard coefficient and Gower's similarity are used depending on the data type.
4. *K*-means clustering and Hierarchical clustering are two popular techniques used for clustering.
5. One of the decisions to be taken during clustering is to decide on the number of clusters. Usually this is carried out using elbow curve. The cluster number at which the elbow (bend) occurs in the elbow curve is the optimal number of clusters.

Multiple Choice Questions

1. Which of the following techniques is an unsupervised learning algorithm?
   (a) Logistic regression
   (b) Multiple linear regression
   (c) Clustering
   (d) Classification trees
2. Euclidean distance can be used to calculate the distance between variables only when
   (a) The variables are measured in either ratio or interval scale
   (b) The variables are measured in ratio scale
   (c) The variables are measured using ordinal scale
   (d) The variables are nominal scale variables

dataset

Exercises

1. The movie ratings given by 4 customers (C_1, C_2, C_3, and C_4) on five movies (A, B, C, D and E) are given in Table 14.19.

Table 14.19 | Movie ratings by customers

Movies → Customer↓	A	B	C	D	E
C_1	4	1	3	2	4
C_2	3	2	4	4	2
C_3	3	3	4	4	4
C_4	3	3	2	3	3

Use cosine similarity to find who among customers C_1, C_2, and C_3, is the closest to customer C_4.

2. An online store sells products under 8 categories labelled: A, B, …, H. The past purchase details of 7 customers are given in Table 14.20.

Table 14.20 | Purchase history of products

Product → Customer↓	A	B	C	D	E	F	G	H
C_1	1	0	0	1	1	0	1	1
C_2	1	0	1	1	1	1	0	0
C_3	0	1	1	0	0	0	1	1
C_4	1	0	0	1	1	0	0	0
C_5	1	1	1	0	0	0	0	1
C_6	0	0	1	1	0	0	1	0
C_7	1	1	0	0	0	1	1	1

where

$$a_{ij} = \begin{cases} 1, & \text{Customer } i \text{ purchased product } j \\ 0 & \text{otherwise} \end{cases}$$

Use Jaccard coefficient to find the customer who is closest to customer C_1.

3. Customer feedbacks on 5 training programs (on a 5-point scale) by 6 customers are provided in Table 14.21.

Table 14.21 | Feedback on training programs

	M_1	M_2	M_3	M_4	M_5
C_1	2	4	2	4	3
C_2	4	3	2	4	5
C_3	1	2	3	2	4
C_4	4	4	2	4	3
C_5	2	1	2	2	3
C_6	2	1	1	4	4

(a) Use cosine similarity to identify the customer who is closest to customer 1.

(b) Calculate correlation between different customers. Which customer has the highest correlation with customer 1?

(c) What is your conclusion based answers to questions (a) and (b)?

Table 14.22 | Amount spent by customers on apparel and beauty and healthcare products (in thousands of rupees)

Customer	Apparel	Beauty and Healthcare	Customer	Apparel	Beauty and Healthcare
1	21.1	0.7	11	5.2	16.2
2	15.23	5.5	12	14.2	2.9
3	5.22	18.6	13	4.4	19.4
4	31.1	1.8	14	4.25	15.5
5	6.12	21.5	15	22.3	0.9
6	14.5	8.2	16	7.9	18.8
7	8.5	16.2	17	13.4	4.2
8	26.5	2.2	18	30.6	1.9
9	4.34	17.7	19	14.4	6.28
10	13.75	7.3	20	6.25	9.98

4. An online grocery store has captured amount spent per annum (in Indian rupees) by 20 customers on apparel and beauty and healthcare products. The data is shown in Table 14.22.
   - **(a)** Use K-means algorithm to find ideal number of clusters and cluster characteristics.
   - **(b)** Calculate the cluster centres of the clusters identified in (a).
   - **(c)** Calculate the distances between the clusters identified in (a).
   - **(d)** Use Hierarchical clustering to find the appropriate clusters for the data in Table 14.22.
5. The dataset usedcars.xls has details of 1008 used cars along with the following variables: 1. Brand, 2. Car model, 3. Resale price, 4. Mileage, 5. Seat capacity, 6. Vehicle type, 7. Fuel type, 8. Transmission, 9. Parking sensor, 10. Airbag, 11. Cruise Control, 12. Keyless entry, 13. Alloy wheel, 14. ABS, 15. Climate control, 16. Rear AC vent and 17. Power Steering
   - **(a)** For the dataset given, which distance measure is more appropriate?
   - **(b)** Use the distance measure identified in (a) and cluster the data. Identify the cluster characteristics.
   - **(c)** Use only the numerical variables (resale price, mileage, and capacity) in the dataset and build clusters. Compare clusters developed in (b) and (c). Which clusters are better? Justify your answer.
6. Table 14.23 shows the Euclidean distance between 8 records. Records R_1, R_2 and R_3 are grouped under cluster 1, records R_4, R_5 and R_6 are grouped under cluster 2 and records R_7, R_8 and R_9 are grouped under cluster 3.

Table 14.23 | Distance between records

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
R_1	0	2	3	4	5	6	7	8	9
R_2		0	4	5	6	7	8	2	8
R_3			0	5	6	7	8	3	7
R_4				0	4	6	8	4	6
R_4					0	3	4	5	5
R_6						0	3	4	3
R_7							0	4	2
R_8								0	1
R_9									0

   - **(a)** Calculate the Silhouette distance for record 9 (R_9), comment whether R_9 is in the right cluster.
   - **(b)** For business reasons, it was decided that cluster 3 can have only 2 records. If we have to remove one record from cluster 3, which one should be removed and to which cluster the removed record should be added?