
UNIT 2

DIAGNOSTIC DATA ANALYSIS

Structure

2.1 Introduction

2.2 Expected Learning Outcomes

2.3 Diagnostics Analysis

2.3.1 Data Mining

2.3.2 Correlation Analysis

2.4 Summary

2.5 Answers

2.6 References/Further Readings

2.1 INTRODUCTION

In our earlier unit-1 we learned about the Descriptive Data Analysis, here in this unit we are extending our discussion for Diagnostic Data Analysis, by the end of this unit we will understand that Descriptive data analysis and diagnostic data analysis are two closely related but conceptually different stages of data analysis, each serving a distinct purpose in understanding data.

Descriptive data analysis focuses on *what has happened* in the data. Its primary objective is to summarize, organize, and present raw data in a meaningful and understandable way. In this type of analysis, data is condensed using measures such as mean, median, mode, range, variance, standard deviation, percentages, and frequency distributions. Visual tools like tables, bar charts, pie charts, histograms, and line graphs are commonly used to highlight patterns, trends, and overall behavior of the dataset. For example, in a business context, descriptive analysis may show total sales for each month, average customer spending, or the distribution of product defects. However, descriptive analysis does not explain *why* these patterns occur; it only provides a clear snapshot of the current or past state of the data.

In contrast, **diagnostic data analysis** goes a step further by addressing the question of *why something happened*. It builds upon descriptive results and attempts to identify the underlying causes, relationships, and drivers behind observed patterns or anomalies. Diagnostic analysis uses more advanced techniques such as data mining, correlation analysis, drill-down analysis, segmentation, and comparative analysis. For instance, if descriptive analysis shows a sudden drop in sales in a particular month, diagnostic analysis investigates possible reasons—such as changes in customer behavior, pricing, competition, supply issues, or marketing performance. By examining relationships between variables (for example, advertising spend and sales, or temperature and product demand), diagnostic analysis provides insights into cause-and-effect or strong associations.

In practice, diagnostic analysis cannot exist without descriptive analysis, because one must first understand *what* is happening before exploring *why* it is happening. Together, they form a logical progression in data analysis: descriptive analysis provides clarity and structure to data, while diagnostic analysis transforms that clarity into actionable insights by uncovering causes and relationships.

From a learning perspective, the key difference lies in intent and depth. Descriptive analysis answers “*What is happening or what has happened?*” and is mainly concerned with summarization and presentation. Diagnostic analysis answers “*Why did it happen?*” and emphasizes reasoning, explanation, and interpretation. While descriptive analysis relies mostly on basic statistical measures and visualization, diagnostic analysis heavily depends on data mining algorithms (such as classification, clustering, and association rule mining) and correlation techniques (such as Pearson, Spearman, or Kendall correlation) to uncover hidden patterns and relationships.

It can be understood that, at the core of diagnostic analysis lie **data mining techniques** and **correlation analysis**, which together enable analysts to explore large datasets, detect hidden patterns, and identify meaningful relationships among variables.

Let’s extend our discussion for Diagnostic Data Analysis at large, in the subsequent section.

2.2 EXPECTED LEARNING OUTCOMES

After completing this unit, you should be able to:

- Understand Diagnostic Analysis
- Understand Data Mining
- Understanding Correlation Analysis

2.3 DIAGNOSTICS ANALYSIS

Diagnostic analysis is a crucial stage in data analytics that goes beyond simply describing what has happened and focuses on understanding **why it happened**. While descriptive analysis summarizes historical data using measures such as mean, median, charts, and tables, diagnostic analysis investigates the underlying causes, patterns, and relationships hidden within the data. It plays a key role in transforming raw data into meaningful insights that support informed decision-making.

In the context of **Data Mining**, diagnostic analysis involves systematically exploring large datasets to discover significant patterns, trends, anomalies, and associations. Data mining techniques such as classification, clustering, association rule mining, and anomaly detection help analysts drill down into data to identify factors that influence outcomes. These techniques enable organizations to diagnose performance issues, customer behavior changes, operational inefficiencies, and risk factors by uncovering relationships that are not immediately visible

through simple summaries.

Correlation Analysis is a fundamental statistical tool used in diagnostic analysis to measure the strength and direction of relationships between variables. By analyzing correlation coefficients, analysts can determine whether variables move together positively, negatively, or not at all. Correlation analysis helps answer diagnostic questions such as whether an increase in one variable is associated with a change in another, and to what extent. Although correlation does not imply causation, it provides strong evidence for identifying potential drivers that require further investigation.

Together, Data Mining and Correlation Analysis form the backbone of diagnostic analytics. Data mining techniques help extract and structure relevant information from complex datasets, while correlation analysis quantifies relationships among variables to support logical reasoning. In diagnostic analysis, these tools are widely applied across domains such as business intelligence, healthcare, finance, marketing, and operations to identify root causes of problems, explain performance variations, and support strategic decisions. The interrelation between Descriptive analysis, Diagnostic Analytics, Data Mining and Correlation Analysis will be more clear from the examples given below:

Example-1: To understand how data mining and correlation analysis relate to diagnostic analysis, consider a practical example from a business context. Suppose an e-commerce company observes a sudden drop in monthly sales. A descriptive analysis may report figures such as total sales, average order value, or number of customers, confirming that sales have indeed declined. However, it does not explain the reason behind this decline. At this stage, diagnostic analysis is applied to investigate the underlying causes.

Using data mining techniques, the company can explore large volumes of transactional and customer data to discover hidden patterns. For example, clustering algorithms may group customers based on purchasing behavior and reveal that a specific customer segment has sharply reduced its purchases. Classification algorithms may identify that customers exposed to delayed deliveries or negative product reviews are more likely to stop buying. Association rule mining might show that customers who experienced payment failures are also likely to abandon future purchases. These insights help diagnose *why* sales dropped rather than merely stating *that* they dropped.

At the same time, correlation analysis plays a crucial role in diagnostic analysis by quantifying the strength and direction of relationships between variables. For instance, the company may calculate the correlation between delivery time and repeat purchases and find a strong negative correlation, indicating that longer delivery times are associated with fewer repeat orders. Similarly, a high negative correlation between product return rates and customer satisfaction scores may suggest that frequent returns contribute to declining loyalty. While correlation does not imply causation, it provides strong evidence pointing toward potential causes that warrant deeper investigation.

Example-2 : In another example from healthcare, a hospital may observe an increase in patient readmission rates. Diagnostic analysis supported by data mining can identify

patterns such as specific treatments, age groups, or hospital departments associated with higher readmissions. Correlation analysis may reveal that readmission rates are strongly correlated with shorter initial hospital stays or lack of follow-up care. Together, these techniques help healthcare administrators diagnose operational or clinical factors contributing to the problem.

Thus, data mining contributes to diagnostic analysis by automatically discovering patterns, trends, anomalies, and groupings in large and complex datasets, while correlation analysis supports diagnostic reasoning by measuring relationships among variables and highlighting influential factors. In diagnostic analysis, these techniques are often used together: data mining identifies what patterns exist, and correlation analysis helps explain how variables are related.

From the above examples it can be concluded that the diagnostic data analysis relies heavily on data mining and correlation analysis to move from observation to explanation. By uncovering hidden patterns and examining relationships among variables, these techniques enable analysts to answer critical “why” questions, making diagnostic analysis an essential step for informed decision-making and problem solving.

Overall, diagnostic analysis bridges the gap between “**what happened**” and “**why it happened.**” By leveraging data mining methods and correlation analysis, analysts gain deeper insights into data behavior, enabling organizations to move from observation to explanation and lay the foundation for *predictive and prescriptive analytics to be discussed in subsequent units i.e. Unit 3 and Unit 4.*

So, we learned that, at the core of diagnostic analysis lie **data mining techniques** and **correlation analysis**, which together enable analysts to explore large datasets, detect hidden patterns, and identify meaningful relationships among variables.

In our subsequent section we are going to discuss the role of Data Mining in performing Diagnostic data analysis.

☛ Check Your Progress 1

Q1. Differentiate between Descriptive Data Analysis and Diagnostic Data Analysis.

.....

.....

Q2. What is Diagnostic Data Analysis? Explain its importance in decision-making.

.....

.....

Q3. Explain the role of Data Mining and Correlation Analysis in Diagnostic Analysis with examples.

.....

.....

2.3.1 DATA MINING

Data mining provides the computational and algorithmic foundation for diagnostic analysis. It involves extracting useful patterns, relationships, and knowledge from large and complex datasets that are otherwise difficult to analyse manually. We learned from the examples cited above in section 2.3 that, while performing diagnostic analysis, data mining helps to:

- **Identify root causes** behind outcomes such as declining sales, system failures, or customer churn
- **Discover hidden patterns** that are not immediately visible through simple summaries
- **Segment the data** into meaningful groups to compare behavior across categories
- **Detect anomalies and exceptions** that may explain unusual events

Several data mining algorithms are particularly useful in performing diagnostic analysis a few are listed below:

1. **Classification algorithms** (Decision Trees, Naïve Bayes, k-Nearest Neighbours):
Used to determine factors responsible for categorical outcomes, such as loan default or disease diagnosis.
2. **Clustering algorithms** (K-Means, Hierarchical Clustering, DBSCAN):
Help identify groups with similar characteristics to compare behaviors and diagnose differences.
3. **Association Rule Mining** (Apriori, FP-Growth):
Discovers relationships between variables, such as product combinations influencing sales or symptoms linked to diseases.
4. **Anomaly / Outlier Detection:**
Identifies unusual patterns that may signal errors, fraud, or rare events influencing results.
5. **Regression Analysis:**
Helps quantify the impact of independent variables on a dependent variable, supporting cause-effect investigation.

Note: *In this section we are going to discuss the above-mentioned algorithms in brief, as there will be detailed discussion on these algorithms in other courses of this programme.*

Now in the remaining part of this section we are going to discuss the various data mining algorithms, which are particularly useful in performing diagnostic analysis, and the same are mentioned above.

- 1) **Classification algorithms** are an important group of data mining techniques used in diagnostic data analysis to identify and explain the factors responsible for **categorical outcomes**. These algorithms assign data instances to predefined classes based on input

features and are widely used to diagnose outcomes such as *loan default vs. non-default*, *disease present vs. absent*, or *spam vs. non-spam*.

- **Decision Trees** classify data by repeatedly splitting it into smaller subsets based on the most influential attributes. Each internal node represents a decision condition, and each leaf node represents a final class. In diagnostic analysis, decision trees are especially useful because they clearly show which factors lead to a particular outcome. For example, in loan default analysis, a decision tree may reveal that low income combined with poor credit history and high loan amount leads to a higher probability of default.
- **Naïve Bayes** is a probabilistic classification algorithm based on Bayes' theorem and the assumption of independence among predictors. Despite this simplifying assumption, it performs well in many real-world problems. In diagnostic contexts, Naïve Bayes helps estimate the probability of a categorical outcome given certain conditions. For instance, in disease diagnosis, it can compute the likelihood of a disease based on symptoms such as fever, blood pressure, and test results.
- **k-Nearest Neighbours (k-NN)** is a distance-based classification algorithm that assigns a class to a data point based on the majority class among its k closest neighbors. It is particularly useful when class boundaries are complex and non-linear. In diagnostic analysis, k-NN can help identify outcomes by comparing new cases with historically similar cases, such as predicting customer churn by analyzing behavior patterns of similar past customers.

Overall, classification algorithms play a vital role in diagnostic analysis by **explaining why a specific categorical outcome occurs** and identifying the key factors influencing that outcome.

2) **Clustering algorithms** are unsupervised data mining techniques used in diagnostic data analysis to identify **groups of data points with similar characteristics** without predefined class labels. These algorithms help analysts compare behaviors across different groups and diagnose why certain patterns or outcomes differ within a dataset.

- **K-Means clustering** partitions data into a predefined number of clusters by minimizing the distance between data points and their cluster centroids. In diagnostic analysis, K-Means is useful for segmenting entities such as customers, products, or regions to identify groups with similar behavior. For example, customer clusters based on spending patterns can help diagnose why certain segments respond differently to marketing strategies.
- **Hierarchical clustering** builds a tree-like structure (dendrogram) that shows how data points are progressively grouped based on similarity. It allows analysts to examine clustering at different levels of granularity. In diagnostic analysis,

hierarchical clustering is helpful for understanding nested relationships, such as grouping patients based on medical profiles to diagnose variations in treatment outcomes.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** groups data points based on density rather than distance. It is particularly effective in identifying irregularly shaped clusters and detecting noise or outliers. In diagnostic analysis, DBSCAN helps identify unusual behaviors or rare patterns, such as detecting abnormal transaction clusters that may explain fraud or system anomalies.

Overall, clustering algorithms support diagnostic analysis by revealing **hidden group structures**, enabling meaningful comparisons, and helping explain differences in behavior across segments.

3) **Association Rule Mining** is a data mining technique used in diagnostic data analysis to uncover **hidden relationships and co-occurrence patterns** among variables in large datasets. It helps explain why certain events or outcomes occur together by identifying rules of the form *if-then*, such as *if a customer buys item A, then they are likely to buy item B*. This makes association rule mining especially useful in understanding interactions between factors that influence outcomes.

- The **Apriori algorithm** identifies frequent item sets by repeatedly scanning the dataset and using the principle that all subsets of a frequent itemset must also be frequent. In diagnostic analysis, Apriori helps determine which combinations of variables are strongly associated. For example, in retail analytics, Apriori can reveal that customers who purchase smartphones and earphones are also likely to purchase protective cases, helping diagnose sales patterns and cross-selling opportunities.
- The **FP-Growth (Frequent Pattern Growth)** algorithm improves efficiency by avoiding repeated database scans. It uses a compact tree structure called an FP-tree to extract frequent item sets more quickly. In diagnostic contexts, FP-Growth is particularly effective for large datasets, such as medical records, where it can identify combinations of symptoms or treatments that commonly occur together, aiding in disease diagnosis and treatment evaluation.

Overall, association rule mining plays a key role in diagnostic analysis by revealing **interdependencies between variables**, enabling analysts to understand how combinations of factors contribute to observed outcomes.

4) **Anomaly or Outlier Detection** is an important data mining technique used in diagnostic data analysis to identify **unusual, rare, or unexpected patterns** in data that deviate significantly from normal behavior. Such anomalies often indicate errors, fraud, system failures, or exceptional events that strongly influence overall results. By detecting these irregularities, analysts can diagnose underlying issues that may not be visible through average-based or summary statistics.

Several algorithms are commonly used for anomaly detection. **Statistical methods**, such as Z-score and Interquartile Range (IQR), identify outliers by measuring how far a data point lies from the central tendency of the dataset. These methods are simple and effective when data follows a known distribution. **Distance-based techniques**, including k-Nearest Neighbors (k-NN), detect anomalies by identifying points that are far away from their nearest neighbors, making them useful when abnormal behavior differs significantly from the majority of observations.

Density-based algorithms, such as DBSCAN and Local Outlier Factor (LOF), identify anomalies based on regions of low data density. These methods are particularly effective in complex datasets where normal behavior forms dense clusters and anomalies appear in sparse regions. **Isolation Forest**, a tree-based algorithm, isolates anomalies by randomly partitioning the data; since anomalies are easier to separate, they require fewer splits, making this method efficient for large datasets.

In diagnostic analysis, anomaly detection helps explain unusual outcomes, such as sudden spikes in transactions indicating fraud, abnormal sensor readings pointing to equipment failure, or rare medical cases affecting health outcomes. By identifying and analyzing these outliers, organizations can better understand exceptional events and take corrective or preventive actions.

- 5) **Regression Analysis** is a fundamental analytical technique used in diagnostic data analysis to **quantify the relationship between a dependent variable and one or more independent variables**. It helps investigate cause-effect relationships by estimating how changes in explanatory variables influence an outcome. By providing numerical coefficients and statistical significance measures, regression analysis enables analysts to identify key drivers behind observed trends and variations.

Several regression algorithms are commonly used in diagnostic analysis. **Simple Linear Regression** examines the effect of a single independent variable on a dependent variable, such as assessing how advertising spend influences sales. **Multiple Linear Regression** extends this approach to multiple predictors, allowing analysts to evaluate the combined impact of factors like price, income, and promotion on demand. **Logistic Regression** is used when the dependent variable is categorical, such as predicting loan default or disease occurrence, by estimating probabilities rather than continuous values.

Advanced regression techniques include **Ridge Regression** and **Lasso Regression**, which address multicollinearity and improve model stability by adding regularization terms. **Polynomial Regression** captures non-linear relationships between variables, while **Elastic Net** combines features of Ridge and Lasso for better variable selection. In diagnostic analysis, these models help isolate influential factors, measure their relative importance, and support evidence-based explanations of outcomes.

Overall, regression analysis plays a critical role in diagnostic analytics by translating relationships into quantifiable effects, thereby strengthening cause–effect investigation and decision-making.

Now it is the time to understand the difference between Correlation analysis and regression analysis, as they are the two closely related statistical techniques, but they serve different purposes and are used in different analytical contexts.

Regression analysis, on the other hand, focuses on examining the **cause–effect relationship** between variables. It clearly distinguishes between a dependent variable (outcome) and one or more independent variables (predictors). Regression analysis not only identifies relationships but also quantifies the impact of independent variables on the dependent variable by estimating regression coefficients. These coefficients indicate how much the dependent variable is expected to change for a unit change in an independent variable. Unlike correlation, regression analysis supports prediction and forecasting through regression equations and models. The relationship in regression is directional, moving from independent variables to the dependent variable, and the variables are not interchangeable.

Correlation analysis is primarily concerned with measuring the **degree and direction of association** between two variables. It answers the question of whether variables are related and how strongly they move together. The relationship measured through correlation is symmetrical, meaning there is no distinction between dependent and independent variables—both variables are treated equally. The result of correlation analysis is a correlation coefficient, usually ranging between -1 and $+1$, where positive values indicate a direct relationship, negative values indicate an inverse relationship, and values close to zero indicate little or no linear association. Correlation analysis does not explain cause-and-effect relationships and cannot be used for prediction; instead, it serves as an exploratory tool to identify relationships that may warrant further investigation, especially in diagnostic data analysis.

In practical applications, correlation analysis is often used as a preliminary step to detect associations between variables, while regression analysis is applied afterward to model and explain these relationships in detail. Together, they complement each other in diagnostic data analysis by first identifying related variables and then quantifying and validating their influence on outcomes.

In the subsequent section we are going to extend our discussion for Diagnostic Analysis to the level of correlation analysis.

☛ Check Your Progress 2

Q1. What is the role of Data Mining in Diagnostic Data Analysis?

.....

.....

Q2. Explain any three data mining algorithms used in diagnostic analysis with examples

.....
.....
Q3. Differentiate between Correlation Analysis and Regression Analysis.
.....
.....

2.3.2 CORRELATION ANALYSIS

Correlation analysis is a core statistical technique in **diagnostic data analysis**, whose primary objective is to explain **why certain outcomes occur** by examining relationships among variables. While descriptive analysis tells us *what has happened*, diagnostic analysis seeks to uncover *why it happened*, and correlation analysis provides one of the most powerful tools for this investigation.

You have already learned various statistical measures such as measures of central tendency, measures of dispersion, moments, skewness, and kurtosis, all of which analyze variables individually. However, in many practical situations, it is necessary to examine two variables simultaneously in order to understand the relationship between them.

This section introduces the concept of correlation, which focuses on studying the linear relationship between two or more variables. For the sake of basic understanding, one can say that *“when two variables are related in such a way that change in the value of one variable affects the value of another variable, then variables are said to be correlated or there is correlation between these two variables.”*

You will learn how to compute the correlation coefficient under different conditions and understand its properties. Therefore, before beginning this section, it is recommended that you revise the concepts of arithmetic mean and variance, as they will help you grasp the fundamentals of correlation more effectively.

Further, in diagnostic data analysis, the basic understanding of correlation is essential, as it enables analysts to move beyond simple averages and summaries and investigate the interdependencies that exist within the data.

At the most fundamental level, correlation analysis is used to measure the degree and direction of association between two variables. It helps determine whether two variables move together, whether a change in one variable is accompanied by a change in the other, and whether the relationship between them is positive, negative, or nonexistent. The strength of this relationship is represented numerically by a correlation coefficient, which generally lies between -1 and $+1$. A value of $+1$ indicates a perfect positive relationship (i.e. *Positive Correlation*), -1 indicates a perfect negative relationship (i.e. *Negative Correlation*), and 0 signifies the absence of a linear relationship.

It can be understood that according to the direction of change in variables there are two **types of correlation:**

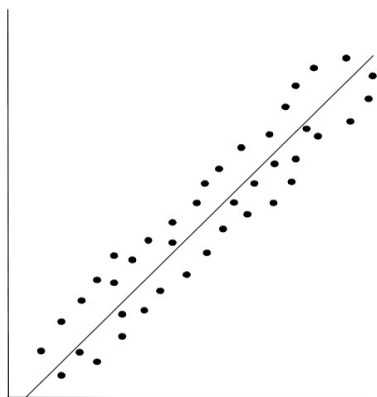
1. **Positive Correlation:** The Correlation between two variables is said to be positive if the values of the variables deviate in the same direction i.e. if the values of one variable increase (or decrease) then the values of other variables also increase (or decrease).
2. **Negative Correlation:** The Correlation between two variables is said to be negative if the values of variables deviate in opposite directions i.e. if the values of one variable increase (or decrease) then the values of other variables decrease (or increase).

So far as the graphical interpretation is concerned, the scatter diagram is quite useful to understand the type of relationship between the variables. It can be understood that the Scatter diagram is a statistical tool for determining the potentiality of correlation between dependent variable and independent variable. The scatter diagram does not talk about the exact relationship between two variables but it indicates whether they are correlated or not.

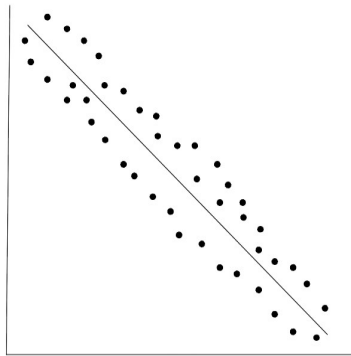
Let (X_i, Y_i) where $i = 1, 2, \dots, n$ be the bivariate distribution. If the values of the dependent variable Y_i are plotted against corresponding values of the independent variable X_i in the XY plane, such a diagram of dots is called a scatter diagram or dot diagram. *It is to be noted that scatter diagrams are not suitable for large numbers of observations.*

In order to interpret from the scatter diagram, one needs to analyse the pattern followed by the dots in the scatter diagram, and interpretations can be made on the basis of the following:

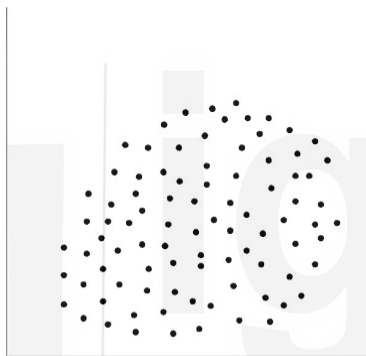
1. If dots are in the shape of a line and the line rises from left bottom to the right top, then correlation is said to be perfectly positive, as shown below.



2. If dots in the scatter diagram are in the shape of a line and line moves from left top to right bottom, then correlation is perfectly negative, as shown below.



3. If dots of scatter diagram do not show any trend there is no correlation between the variables, as shown below.



Coefficient of Correlation: Scatter diagram tells us whether variables are correlated or not. But it does not indicate the extent of which they are correlated. Coefficient of correlation gives the exact idea of the extent of which they are correlated. Coefficient of correlation measures the intensity or degree of linear relationship between two variables.

There are some Common Correlation Measures Used in Diagnostic Analysis, as diagnostic analysis progresses, different correlation measures are chosen based on data type and complexity, a few are listed below:

- **Pearson Correlation:** Used to measure linear relationships between continuous variables. It is widely applied in business, finance, and engineering diagnostics.
- **Spearman Rank Correlation:** Used when relationships are monotonic but not necessarily linear, or when data is ordinal. It is valuable in diagnostic analysis when assumptions of normality are violated.
- **Kendall's Tau:** Focuses on rank concordance and is more robust for small datasets or data with ties.

At an intermediate level, analysts learn to select the appropriate correlation measure based on data distribution, scale, and analytical Objectives.

Now we will discuss these Common Correlation Measures Used in Diagnostic Analysis, let's begin with Pearson Correlation.

Pearson Correlation:

Pearson correlation is one of the most widely used statistical techniques in data science for measuring the strength and direction of a linear relationship between two continuous variables. In diagnostic data analysis, it plays an important role in identifying variables that are closely associated with an outcome and may act as potential drivers behind observed patterns. Pearson correlation focuses on how two variables vary together and helps determine whether an increase or decrease in one variable is accompanied by a corresponding change in another variable, as well as the nature and intensity of this relationship.

The result of Pearson correlation is expressed through the Pearson correlation coefficient, usually denoted by r . The value of r lies between -1 and $+1$. A value of $+1$ indicates a perfect positive linear relationship, meaning that both variables move in the same direction in exact proportion. A value of -1 indicates a perfect negative linear relationship, where one variable increases while the other decreases in exact proportion. A value of 0 suggests that there is no linear relationship between the variables. In diagnostic analysis, values of r closer to ± 1 highlight strong associations that may require deeper investigation.

The Pearson correlation coefficient is calculated using the formula

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

where x and y represent individual observations of the two variables, and $\bar{x}$ and $\bar{y}$ represent their respective means. The numerator of this formula measures the extent to which the two variables vary together, while the denominator standardizes this covariance by accounting for the variability of each variable individually. This standardization ensures that the coefficient is unit-free and confined within the range of -1 to $+1$.

Alternate formula (in terms of Variance and covariance of the data): If **X** and **Y** are two random variables, then the **correlation coefficient** between **X** and **Y** is denoted by **r** and is defined as:

$$r = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X) V(Y)}}$$

Here, **Corr(X, Y)** indicates the correlation coefficient between the two variables **X** and **Y**.

Where **Cov(X, Y)** is the covariance between **X** and **Y**, defined as:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

And **V(X)**, the variance of **X**, is defined as:

$$V(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

And **V(Y)**, the variance of **Y**, is defined as:

$$V(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

To illustrate the concept, consider a dataset where advertising expenditure and sales revenue are observed together.

Example: Suppose a data scientist wants to diagnose whether **advertising spend** is related to **sales**.

Advertising Spend (X)	Sales (Y)
10	40
20	50
30	60
40	70
50	80

Step 1: Calculate the means

$$\bar{x} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30 \quad \bar{y} = \frac{40 + 50 + 60 + 70 + 80}{5} = 60$$

Step 2: Create deviation table

X	Y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(\bar{x})^2$	$(\bar{y})^2$
10	40	-20	-20	400	400	400
20	50	-10	-10	100	100	100
30	60	0	0	0	0	0
40	70	10	10	100	100	100
50	80	20	20	400	400	400

$$\sum(x - \bar{x})(y - \bar{y}) = 1000 \sum(x - \bar{x})^2 = 1000 \sum(y - \bar{y})^2 = 1000$$

Step 3: Apply Pearson correlation formula

$$r = \frac{1000}{\sqrt{1000 \times 1000}} = \frac{1000}{1000} = 1$$

As the computed value of r is 1, it indicates a strong positive linear relationship, suggesting that higher advertising spend is strongly associated with higher sales.

In diagnostic data analysis, such a result helps analysts identify advertising as a potential driver of sales performance, although it does not by itself confirm causation.

Example: Given Data (5 days), A data analyst wants to check whether Digital Ad Spend (X) is linearly related to Daily Sales (Y) for a small online store.

Day	Ad Spend X (₹ '000)	Sales Y (₹ '000)
1	2	20
2	4	22
3	6	25
4	8	29
5	10	35

Solution: Using the correlation formula: $r = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}$

following Tasks are to be performed to calculate the coefficient of correlation (r), using the above formula $\bar{x}$ $\bar{y}$

1. Compute $\bar{x}$ and $\bar{y}$
2. Compute $\text{Cov}(X, Y)$
3. Compute $V(X)$ and $V(Y)$
4. Find r and interpret it for diagnostic analysis (does spend relate to sales?)

Step 1: Means

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6 \quad ; \quad \bar{y} = \frac{20 + 22 + 25 + 29 + 35}{5} = \frac{131}{5} = 26.2$$

Step 2: Build calculation table

X	Y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(\bar{x})^2$	$(\bar{y})^2$
2	20	-4	-6.2	24.8	16	38.44
4	22	-2	-4.2	8.4	4	17.64
6	25	0	-1.2	0	0	1.44
8	29	2	2.8	5.6	4	7.84
10	35	4	8.8	35.2	16	77.44

Now sum:

$$\begin{aligned} \sum(x - \bar{x})(y - \bar{y}) &= 24.8 + 8.4 + 0 + 5.6 + 35.2 = 74 \\ \sum(x - \bar{x})^2 &= 16 + 4 + 0 + 4 + 16 = 40 \\ \sum(y - \bar{y})^2 &= 38.44 + 17.64 + 1.44 + 7.84 + 77.44 = 142.8 \end{aligned}$$

Step 3: Covariance and variances (using 1/n as in your formula)

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{n} \sum(x - \bar{x})(y - \bar{y}) = \frac{74}{5} = 14.8 \\ V(X) &= \frac{1}{n} \sum(x - \bar{x})^2 = \frac{40}{5} = 8 \\ V(Y) &= \frac{1}{n} \sum(y - \bar{y})^2 = \frac{142.8}{5} = 28.56 \end{aligned}$$

Step 4: Correlation

$$r = \frac{14.8}{\sqrt{8 \times 28.56}} = \frac{14.8}{\sqrt{228.48}} = \frac{14.8}{15.115} \approx 0.98$$

Since r is close to +1, there is a very strong positive linear relationship between Ad Spend and Sales. This supports a diagnostic insight that higher ad spend is strongly associated with higher sales

Pearson correlation is especially advantageous in diagnostic analysis because it helps analysts move beyond simple averages and summaries to examine interdependencies between variables. It is often used as a preliminary diagnostic tool to filter relevant variables before applying more advanced techniques such as regression analysis or causal modelling. By highlighting which variables are strongly associated, *Pearson correlation supports hypothesis generation and guides deeper analytical investigation.*

We can say that the *Pearson correlation is highly relevant in diagnostic analysis* because of the following reasons:

- *Identifies variables that move together*
- *Helps narrow down potential explanatory factors*
- *Supports hypothesis generation*
- *Acts as a foundation for regression and predictive modelling*

However, for an expert-level understanding, it is important to recognize the *limitations of Pearson correlation*. It captures only linear relationships, is sensitive to outliers, and does not establish cause-and-effect relationships. Therefore, in diagnostic data analysis, Pearson correlation should be interpreted carefully and used in combination with domain knowledge, regression analysis, and other data mining techniques.

Limitations of Pearson correlation (Diagnostic Perspective), For expert diagnostic use, it is important to note that:

- Pearson correlation captures only linear relationships
- It is sensitive to outliers
- It does not imply causation

Correlation vs. Causation in Diagnostic Analysis: At an expert level, understanding the limitations of correlation it is important to understand that the Correlation analysis:

- Identifies associations, not causality
- Cannot, by itself, prove that one variable cause another
- Must be interpreted in conjunction with domain knowledge, regression analysis, experiments, or causal models

It is to be noted that in diagnostic data analysis, correlation is used as **evidence**, not as proof. Strong correlations suggest where causes may exist, but further analytical or experimental validation is required.

Therefore, Pearson correlation should be used as a diagnostic indicator, not as final proof.

It is learned from the above discussion that the Pearson correlation is a foundational tool for diagnostic data analysis in data science. It provides a clear and quantitative measure of linear association between variables, helping analysts identify influential factors, understand data behavior, and lay the groundwork for more advanced analytical methods.

Spearman Rank Correlation:

Spearman rank correlation is a non-parametric statistical measure used to assess the strength and direction of association between two variables based on their ranks rather than their actual values. It is especially useful in diagnostic data analysis when the data does not satisfy the assumptions required for Pearson correlation, such as normality or linearity, or when the data is ordinal in nature.

Spearman rank correlation examines whether **two variables move together in a consistent (monotonic) manner**, meaning that as one variable increases, the other either consistently increases or consistently decreases. Unlike Pearson correlation, Spearman correlation does not require the relationship to be linear; it only needs to be monotonic.

In diagnostic data analysis, Spearman correlation is valuable when:

- Data contains **outliers**
- Data is **ordinal or ranked**
- The relationship is **non-linear but monotonic**
- Exact numerical values are less reliable than relative ordering

The Spearman rank correlation coefficient is denoted by ρ (**rho**).

The value of Spearman's rank correlation coefficient lies between **-1 and +1**:

- $\rho = +1 \rightarrow$ Perfect positive monotonic relationship
- $\rho = -1 \rightarrow$ Perfect negative monotonic relationship
- $\rho = 0 \rightarrow$ No monotonic relationship

In diagnostic analysis, higher absolute values of ρ indicate stronger associations that may point to important explanatory variables.

When there are **no tied ranks**, the Spearman rank correlation coefficient is calculated using the formula:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Explanation of Symbols:

- d = difference between the ranks of corresponding observations
- n = number of observations
- $\sum d^2$ = sum of squares of rank differences

This formula measures how closely the rankings of two variables agree.

Note: When tied ranks are present, an adjusted formula using Pearson correlation on ranks is applied, it is out of the scope and we are not going to discuss this case here.

Example: Given Data:

Employee	Experience (Years)	Performance Score
A	2	60
B	4	65
C	6	70
D	8	85
E	10	90

A data scientist wants to diagnose whether employee experience is related to job performance ratings. Where the performance is ranked subjectively.

Solution: Since performance is ranked subjectively, Spearman rank correlation is appropriate.

Step 1: Assign Ranks

Employee	Experience	Rank X	Performance	Rank Y
A	2	1	60	1
B	4	2	65	2
C	6	3	70	3
D	8	4	85	4
E	10	5	90	5

Step 2: Compute Rank Differences

Employee	Rank X	Rank Y	d = X – Y	d ²
A	1	1	0	0
B	2	2	0	0
C	3	3	0	0
D	4	4	0	0
E	5	5	0	0

$$\sum d^2 = 0$$

Step 3: Apply the Formula

$$\rho = 1 - \frac{6 \times 0}{5(5^2 - 1)} = 1$$

A Spearman correlation coefficient of $\rho = 1$ indicates a **perfect positive monotonic relationship** between experience and performance. In diagnostic data analysis, this suggests that higher experience consistently corresponds to higher performance ratings, making experience a strong diagnostic factor.

Example: Given Data

Employee	Training Hours (X)	Performance Score (Y)
A	12	80
B	20	82
C	16	85
D	25	90
E	10	70
F	18	78

A data scientist wants to examine whether there is a relationship between hours of online training completed (X) and performance ranking (Y) of employees.

Solution: Since performance is ranked and the relationship may not be perfectly linear, Spearman rank correlation is appropriate.

Employee	Training Hours (X)	Performance Score (Y)
A	12	80
B	20	82
C	16	85
D	25	90
E	10	70
F	18	78

Step 1: Assign Ranks :

Rank X		Rank X
X		Rank X
10		1
12		2
16		3
18		4
20		5
25		6

Rank Y

Y	Rank Y
70	1
78	2
80	3
82	4
85	5
90	6

Step 2: Compute d and d²

Employee	Rank X	Rank Y	d	d ²
A	2	3	-1	1
B	5	4	1	1
C	3	5	-2	4
D	6	6	0	0
E	1	1	0	0
F	4	2	2	4

$$\sum d^2 = 1 + 1 + 4 + 0 + 0 + 4 = 10$$

Step 3: Apply Spearman Rank Correlation Formula

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 10}{6(36 - 1)} = 1 - \frac{60}{210} = 1 - 0.286 = 0.714$$

A Spearman correlation of 0.714 indicates a strong positive monotonic relationship, but not a perfect one. This suggests that higher training hours generally lead to better performance, but other factors (such as experience, motivation, or role complexity) also influence performance. This insight is valuable for diagnostic analysis, as it highlights training as an important but not exclusive driver of performance.

Kendall's Tau:

Kendall's Tau correlation is a **non-parametric measure of association** used to determine the **strength and direction of relationship between two variables based on the relative ordering (ranks) of data rather than their actual values**. It is especially useful in **diagnostic data analysis** when the dataset is small, contains tied ranks, or when the relationship between variables is ordinal and monotonic rather than strictly linear.

Kendall's Tau focuses on the **consistency of ordering between two variables**. Instead of measuring how far values deviate from the mean (as in Pearson correlation), Kendall's Tau evaluates whether the **relative rankings of observations agree or disagree** across two variables.

In diagnostic data analysis, this is particularly important when:

- Data is **ordinal or ranked**
- There are **many tied values**
- The sample size is **small**
- Robustness and interpretability are preferred over sensitivity to outliers

Kendall's Tau is denoted by **τ (tau)**.

For any pair of observations (y_i) and (y_j):

- The pair is **concordant** if:
 - $x_i > x_j$ and $y_i > y_j$, or
 - $x_i < x_j$ and $y_i < y_j$
- The pair is **discordant** if:
 - $x_i > x_j$ and $y_i < y_j$, or
 - $x_i < x_j$ and $y_i > y_j$

Kendall's Tau is based on comparing the **number of concordant and discordant pairs** in the dataset.

The basic formula for Kendall's Tau (τ) is:

$$\tau = \frac{C - D}{\frac{1}{2}n(n - 1)}$$

Where:

- C = number of concordant pairs
- D = number of discordant pairs
- n = number of observations
- $\frac{1}{2}n(n - 1)$ = total number of possible pairs

This formula measures how much agreement exists between the rankings of two variables.

The value of Kendall's Tau lies between **-1 and +1**:

- $\tau = +1$ → Perfect agreement in rankings
- $\tau = -1$ → Perfect disagreement in rankings
- $\tau = 0$ → No association in rankings

In diagnostic data analysis, higher absolute values of τ indicate stronger ordinal association and suggest meaningful diagnostic relationships.

Example: Given Data

Project	Complexity Rank (X)	Completion Time Rank (Y)
A	1	2
B	2	1
C	3	4
D	4	3

A data scientist wants to diagnose whether **project complexity ranking** is related to **time-to-completion ranking** for completed projects, Using Kendall's Tau correlation.

Solution:

Step 1: List all possible pairs

For $n = 4$, total pairs:

$$\frac{4(4 - 1)}{2} = 6$$

Pairs:

- (A, B)
- (A, C)
- (A, D)
- (B, C)
- (B, D)
- (C, D)

Step 2: Determine Concordant and Discordant Pairs

Pair	X Order	Y Order	Result
A, B	1 < 2	2 > 1	Discordant
A, C	1 < 3	2 < 4	Concordant
A, D	1 < 4	2 < 3	Concordant
B, C	2 < 3	1 < 4	Concordant
B, D	2 < 4	1 < 3	Concordant
C, D	3 < 4	4 > 3	Discordant

- Concordant pairs (C) = 4
- Discordant pairs (D) = 2

Step 3: Apply Kendall's Tau Formula

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)} = \frac{4 - 2}{6} = \frac{2}{6} = 0.33$$

A Kendall's Tau value of **0.33** indicates a **moderate positive ordinal association** between project complexity and completion time. In diagnostic data analysis, this suggests that **higher complexity projects tend to take longer**, but the relationship is not perfectly consistent, indicating the influence of other factors such as team size, experience, or tools used.

Important Note: In this example, the data consists of **ranked variables**—project complexity rank and completion time rank. Since the information is ordinal rather than continuous, Karl Pearson's coefficient of correlation is not appropriate. Pearson correlation requires numerical data measured on an interval or ratio scale and assumes a linear relationship and approximate normality. In this case, the actual numerical distances between ranks are not meaningful; only the **order of observations** matters. Therefore, Pearson correlation would give misleading results.

While **Spearman rank correlation** is suitable for ranked data, it measures association based on **differences in ranks** and treats all rank differences equally. Spearman's method is more sensitive to larger rank gaps and assumes that the monotonic relationship between variables is relatively smooth. In small datasets, such as the one used in the example, Spearman correlation can exaggerate the strength of association because a single change in rank can significantly affect the coefficient. This makes it less reliable for fine-grained diagnostic interpretation when the number of observations is limited.

Kendall's Tau, on the other hand, is especially well suited for this example because it is based on the **comparison of concordant and discordant pairs**. Instead of focusing on the magnitude of rank differences, Kendall's Tau evaluates how consistently the ordering of one variable agrees with the ordering of the other. This pairwise comparison approach provides a more **robust and intuitive measure of association**, particularly for small samples and diagnostic analysis.

From a diagnostic perspective, Kendall's Tau is more meaningful because it directly answers the question: *What proportion of pairs are ordered consistently across the two variables?* In the given example, this helps assess how reliably higher project complexity is associated with longer completion time, without being overly influenced by individual rank shifts. Moreover, Kendall's Tau handles tied ranks and ordinal inconsistencies more effectively, making it a preferred choice when data quality or scale is limited.

Overall, Kendall’s Tau is more relevant than Spearman rank correlation and Karl Pearson’s coefficient in this example because the data is ordinal, the sample size is small, and the diagnostic goal is to assess **consistency of ordering rather than strength of linear or monotonic relationships**.

It is important to note that Kendall’s Tau is particularly valuable because it is robust to the presence of outliers, handles tied ranks more effectively than Spearman’s rank correlation, performs well even with small datasets, and offers a probability-based interpretation of agreement between variables. When compared with Spearman correlation, Kendall’s Tau is generally more conservative and easier to interpret, while Spearman is more sensitive to large differences in ranks. Additionally, Kendall’s Tau tends to perform better in situations involving small samples and tied observations, as it directly measures the degree of pairwise agreement between the rankings of two variables.

Overall, we learned that Kendall’s Tau correlation is a powerful and robust tool for **diagnostic data analysis**, particularly when dealing with ranked, ordinal, or non-linear data. By focusing on concordant and discordant pairs, it provides a clear and intuitive measure of association that supports root-cause investigation and informed analytical reasoning.

Check Your Progress 3

Q1. What is Correlation Analysis? Explain the meaning and range of the correlation coefficient. Also, Differentiate between Positive and Negative Correlation with examples.

.....
.....

Q2. Compare Pearson, Spearman, and Kendall correlation methods.

.....
.....

2.5 SUMMARY

Diagnostic Data Analysis is an advanced stage of data analysis that focuses on answering the question **“why did something happen?”**. While descriptive analysis explains what has occurred, diagnostic analysis goes deeper to identify the underlying causes, patterns, and relationships within the data. A key component of this process is **Data Mining**, which provides the computational tools and algorithms required to extract meaningful patterns from large and complex datasets. Techniques such as classification, clustering, association rule mining, anomaly detection, and regression analysis help analysts identify root causes, segment data into meaningful groups, detect unusual patterns, and quantify the impact of various factors on outcomes. These methods enable organizations to move from observation to explanation, making data-driven decision-making more effective.

Another fundamental component of diagnostic analysis is **Correlation Analysis**, which examines the strength and direction of relationships between variables. It helps determine whether changes in one variable are associated with changes in another, using measures such as Pearson, Spearman, and Kendall correlation coefficients. The correlation coefficient ranges from -1 to $+1$, indicating negative, positive, or no relationship. While correlation does not imply causation, it plays a crucial role in identifying potential drivers and guiding further investigation. Together, data mining and correlation analysis form the backbone of diagnostic analytics, enabling analysts to uncover hidden relationships, explain observed patterns, and generate actionable insights for solving real-world problems and improving decision-making.

2.5 SOLUTIONS/ANSWERS

☛ Check Your Progress 1

- Q1. Differentiate between Descriptive Data Analysis and Diagnostic Data Analysis.
Solution: Refer to section 2.1 & 2.3
- Q2. What is Diagnostic Data Analysis? Explain its importance in decision-making.
Solution: Refer to section 2.1 & 2.3
- Q3. Explain the role of Data Mining and Correlation Analysis in Diagnostic Analysis with examples.
Solution: Refer to section 2.1 & 2.3

☛ Check Your Progress 2

- Q1. What is the role of Data Mining in Diagnostic Data Analysis?
Solution: Refer to section 2.3.1
- Q2. Explain any three data mining algorithms used in diagnostic analysis with examples
Solution: Refer to section 2.3.1
- Q3. Differentiate between Correlation Analysis and Regression Analysis.
Solution: Refer to section 2.3.1

☛ Check Your Progress 3

- Q1. What is Correlation Analysis? Explain the meaning and range of the correlation coefficient. Also, Differentiate between Positive and Negative Correlation with examples.
Solution: Refer to section 2.3.2
- Q2. Compare Pearson, Spearman, and Kendall correlation methods.
Solution: Refer to section 2.3.2

2.6 FURTHER READINGS

1. *Handbook of Statistical Analysis and Data Mining Applications* by Robert Nisbet, John Elder, and Gary Miner, published by Elsevier (2nd Edition, 2017, ISBN: 978-0124166455).
2. *Data Mining and Analysis: Fundamental Concepts and Algorithms* by Mohammed J. Zaki and Wagner Meira Jr., published by Cambridge University Press (1st Edition, 2014, ISBN: 978-0521766333).
3. *Core Concepts in Data Analysis: Summarization, Correlation and Visualization* by Boris Mirkin, published by Springer (1st Edition, 2011, ISBN: 978-0857292872).
4. *Applied Predictive Modeling*, Max Kuhn and Kjell Johnson, 1st Edition, Springer, 2013.
5. *An Introduction to Statistical Learning with Applications in R*, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 1st Edition, Springer, 2013.
6. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Trevor Hastie, Robert Tibshirani and Jerome Friedman, 2nd Edition, Springer, 2009.
7. *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, published by Springer (2nd Edition, 2021, ISBN: 978-1071614174).
8. *Time Series Analysis: Forecasting and Control* by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung, published by Wiley (5th Edition, 2015, ISBN: 978-1118675021)