UNIT 5 Data Analytics

Student Learning Outcomes

By the end of this chapter, students will be able to:

- Understand the role and importance of model building and their real world applications
- Build basic statistical models for real-world problems and evaluate their performance
- Understand and explain the principles of experimental design in data science
- Explain the types, uses and methods of data visualizations,
- Understand the benefits of visualizing data through descriptive statistics
- Create and interpret data visualization using data visualization software such as MS Excel, Google Sheets, Python, Tableau, and Matplotlib.

Introduction

Data analytics is the process of examining data to find useful information, patterns and trends to support decision-making.

5.1 Basic Statistical Concepts

Statistics is a branch of mathematics that helps us understand and analyze data. By using statistics, we can summarize large sets of information in a simple way, making it easier to draw conclusions. By using statistics large sets of information can be summarized in simple way making it easier to analyze and draw conclusions.

5.1.1 Measures of Central Tendency

Measures of central tendency help us identify the "center" or typical value in a dataset. There are three main measures of central tendency: mean, median, and mode. These measures give us a sense of the average or most common values of a dataset.

Mean

The mean is the average of all the numbers in a dataset. To calculate the mean, we add all the numbers together and then divide the sum by the total number of values.

Example: Imagine 5 students scored 50, 60, 70, 80, and 90 in a test. The mean score is calculated by adding all the scores and then dividing by the number of students:

$$\text{Mean} = \frac{50+40+70+80+90}{5} = 70$$

Median

The median is the middle value in a dataset when the numbers are arranged in order. If there is an odd number of values, the median is the exact middle number. If there is an even number of values, the median is the average of the two middle numbers.

Example: Using the same test scores: 50, 60, 70, 80, and 90. When we arrange these scores in ascending order (which they already are), the middle value is 70. Therefore, the median score is 70. Example with Even Number of values: If the scores were 50, 60, 70, and 80, we would take the average of the two middle scores (60 and 70):

Median = (60+70)/2= 65, so the median is 65.

The median helps us understand the middle point of the data.

Mode

The mode is the number that appears most often in a data set. There can be more than one mode if multiple numbers appear with the same highest frequency. The mode helps us identify the most frequent or common value in the data.

Example: If 5 students scored 50, 60, 70, 70, and 90, the number 70 appears twice, while all other numbers appear only once. Therefore, the mode is 70.

Example with Multiple Modes: If the scores were 50, 60, 70, 70, 60, and 90, both 60 and 70 appear twice. So, there are two modes: 60 and 70.

5.1.2 Measures of Dispersion

Measures of dispersion tell us how spread out or scattered the data is. Two common measures of dispersion are variance and standard deviation. These help us understand whether the data points are close to the average (mean) or spread far from it.

5.1.2.1 Variance

The variance shows how much the numbers in a data set differ from the mean. A higher variance means that the numbers are more spread out, while a lower variance means that the numbers are closer to the mean. To calculate variance, we use the following mathematical formula.

$$\text{Variance } (\sigma^2) = \frac{1}{N}\sum_{i=1}^{N}(x_1 - \mu)^2$$

Where: x_i represents each individual value in the data set, μ is the mean of the data set, and *N* is the total number of values in the data set.

Example: for the following two classes find out which class is more spread out by calculating their variance?

Class A: 50, 52, 55, 57, 60

Class B: 30, 45, 55, 75, 90

Steps are involved in variance calculations:

Step 1: Variance for Class A

Given Score = 50,52,55,57,60

Step 1.1: Compute the Mean (μ)

$$\mu = \frac{50+52+55+57+60}{5} = \frac{274}{5}$$

$$= 54.8$$

Step 1.2: Compute Each Squared Deviation $(x_i - \mu)^2$

x_i	$x_i - \mu$	$(x_i - \mu)^2$
50	50−54.8=−4.8	23.04
52	52−54.8=−2.8	7.84
55	55−54.8=0.2	.04
57	57−54.8=2.2	4.84
60	60−54.8=5.2	27.04

Step 1.3: Compute Variance

$$\text{Variance } (\sigma^2) = \frac{23.04+7.84+0.04+4.84+27.04}{5}$$

$$\sigma^2 = \quad 62.8/5 = 12.56$$

Step 2: Variance for Class B

Given Scores: 30,45,55,75,90

$$\mu = \frac{30+45+55+75+90}{5} = \frac{295}{5} = 59$$

x_i	$x_i - \mu$	$(x_i - \mu)^2$
30	30−59 = −29	841
45	45− 59 = −14	196
55	55 − 59 = − 4	16
75	75 − 59 = 16	256
90	90 − 59 = 31	961

Step 2.3: Compute Variance

$$\sigma^2 = \frac{841 + 196 + 16 + 256 + 961}{5}$$

$$= 2270/\ 5 = 454$$

- **Variance of Class A: 12.56**
- **Variance of Class B: 454**

This confirms that Class B has a much higher variance, meaning the scores are more spread out compared to Class A.

5.1.2.2 Standard Deviation

It is similar to variance but provides a more practical and interpretable value because it is in the same unit as the original data. The standard deviation tells us how spread out the numbers are in relation to the mean. The standard deviation is simply the square root of the variance. To calculate standard deviation, we use the following mathematical formula.

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_1 - \mu)^2}$$

Where: x represents each individual value in the data set, μ is the mean of the dataset, and N is the total number of values in the data set.

Calculating Standard Deviation:

Class A:

$$\text{Standard Deviation} = \sqrt{12.56} = 3.55$$

Class B:

$$\text{Standard Deviation} = \sqrt{456} = 21.26$$

The standard deviation for Class A is approximately 3.55, while for Class B, it is about 21.26. This means that Class A's scores are closely packed around the mean, whereas Class B's scores are more widely scattered. The standard deviation helps us easily understand how much variation exists in a dataset.

5.1.3 Introduction to Probability

Probability is the study of how likely an event is to happen. It helps us make predictions based on known information.

Example: Consider flipping a coin. There are two possible outcomes: heads or tails. Since both outcomes are equally likely, the probability of getting heads is 50% (or 1/2), and the probability of getting tails is also 50%.

We can express this mathematically as:

$$\textit{Probability} = \frac{\textit{Number of favorable outcomes}}{\textit{Total number of outcomes}}$$

In the case of the coin flip:

Probability of heads = (1 favorable outcome, 2 total outcomes) = 1/2

Probability is not just for coin flips. It is used in many areas, such as predicting the weather, making business decisions, or even predicting outcomes in sports like cricket.

For illustration, you can use the following sample data: 3, 5, 8, 8, 10, 12, 15, 15, 16, 18. But if you are interested in collecting real data from your classmates, you are welcome to do so.

Tidbits

Statistics can help us understand patterns in data, leading to better decision-making in various fields, from healthcare to marketing!

Class Activity

Instructions: You will analyze a small dataset, calculate measures of central tendency (mean, median, mode), and measures of dispersion (variance and standard deviation).

1. Collect Data: Imagine you are surveying your classmates about the number of hours they spend on homework in a week. Gather data from 10 classmates. Record the number of hours (use reasonable values , e.g., between 0 to 20 hours).
2. Calculate Measures of Central Tendency:
   - Mean: Calculate the average number of hours spent on homework.
   - Median: Determine the middle value when the hours are arranged in order.
   - Mode: Identify which number appears most frequently in your data.
3. Calculate Measures of Dispersion:
   - Variance: Use the appropriate formula to calculate variance based on your data.
   - Standard Deviation: Calculate the standard deviation using the variance obtained.
4. Reflect: Write a brief reflection (3-4 sentences) on what these calculations reveal about your classmates' study habits.

DO YOU KNOW?

Statistics is used in many everyday activities, such as predicting weather patterns or analyzing sports performance.

5.2 Data Collection and Preparation

In order to carry out any research or analysis, data collection and preparation are crucial steps. The quality and relevance of the data directly impact the results and insights drawn from the study. This section discusses various methods of data collection and how the collected data is prepared for further analysis.

5.2.1 Data Collection Methods

Data collection refers to the process of gathering relevant information for a particular purpose. Depending on the nature of the research, different methods can be used for data collection. These methods include surveys, observations, and experiments, each having its own strengths and appropriate contexts. Choosing the right method depends on the research objective and the type of data required.

5.2.1.1 Surveys

Surveys are a commonly used method for collecting large amounts of data in a structured way. They involve asking a predefined set of questions to a sample group. Surveys can be conducted using various means such as online forms, telephone calls, or face-to-face interviews.

Example: A small local grocery store in Islamabad wants to know customer preferences regarding which products they would like to see more frequently. The store creates a short survey consisting of five questions as shown below and distributes it to 50 customers over the weekend. The collected responses are then analyzed to stock products that align with customer demand, helping improve business operations.

Customer Preference Survey

1. Which product categories do you buy most often? (e.g., fruits, vegetables, dairy)
2. Are there any products you would like to see more often?
3. How often do you shop at this grocery store? (e.g., daily, weekly, monthly)
4. What influences your purchasing decisions the most? (e.g., price, quality, availability)
5. Any additional comments or suggestions?

5.2.1.2 Observations

Observation involves collecting data by watching or monitoring subjects in their natural environment. This method is useful when researchers want to gather data on behaviors or phenomena without interference.

Example: A restaurant is interested in knowing which tables are most frequently chosen by customers during lunchtime. A staff member observes the seating choices over a period of one week. Based on these observation, the restaurant arranges its seating to optimize the customer comfort and traffic flow, which helps in improving customer satisfaction and service efficiency.

5.2.1.3 Experiments

Experiments involve manipulating one or more variables to determine their effect on another variable. This method is particularly useful in scientific and engineering fields where controlled environments are necessary for accurate measurement.

Example: A school teacher wants to test whether providing students with printed notes helps improve their performance in exams. The teacher conducts an experiment with two groups of students, one receiving printed notes and the other relying solely on lectures. After one month, both groups took the same test, and the teacher compared the results to see if printed notes had a positive impact on academic performance.

5.2.2 Data Preparation

Once data has been collected, it is important to prepare it for analysis. This includes cleaning the data to remove errors or inconsistencies, organizing it in a meaningful way, and converting it into a format suitable for analysis. In cases where data is missing or incorrect, researchers may need to employ techniques such as interpolation or statistical adjustments to ensure the accuracy and reliability of the results..

Example: If survey responses contain incomplete information, missing values can be estimated based on the available data.

Proper data preparation ensures that the analysis leads to reliable and valid results.

5.2.3 Data Cleaning and Transformation

Data cleaning and transformation are important steps to prepare data for analysis. Raw data often has errors, missing values, or may be in an incorrect format. To ensure accurate results in analysis, it is important to fix these issues before moving forward.

5.2.3.1 Data Cleaning

Data cleaning means correcting or removing any problems in the data. These problems can include incorrect entries, missing values, or duplicate results. If these errors are not fixed, the results of the analysis will be misleading.

Example: Imagine a school collecting data on student scores. Some students may have entered their names incorrectly, or a few scores may be missing from the records. In this case, data cleaning would involve correcting any wrong names and including in the missing grades to complete the dataset. Table 5.1 and 5.2, illustrates the data cleaning process for student scores in a school. It shows examples of common issues such as incorrect names, missing scores and duplicate entries.

Original Data (with Errors)

Name	Scores	Class	Section
Ali	84	10	A
Alie	90	10	A
Sara		10	A

Table 5.1: Original Data with errors

Cleaned Data (After Data Cleaning)

Name	Grade	Class	Section
Ali	84	10	A
Ali	90	10	A
Sara	87	10	A

Table 5.2: Cleaned data after removing errors

5.2.3.2 Data Transformation

Once the data is clean, it often needs to be transformed it into a format that is easier to work with. This transformation may include converting data into different formats, creating new columns, or organizing data in a different way. These changes help make the data more suitable for analysis or modeling.

Example: After cleaning the student grade records, it may be necessary to transform this data for better analysis. For instance, instead of displaying grades for each individual student, the data might be aggregated or summarized to show class-level statistics such as average scores or grade distribution.

5.2.3.3 Handling Missing Data

Sometimes, data is incomplete or has missing values. There are different techniques to handle missing data. One option is to remove the rows with missing values if they are very few. Another option is to fill in the missing values with an average or with data from similar cases. The choice depends on the type of data and the amount of missing information.

Example: In the dataset of student grades, if Sara's grade is missing, this creates a challenge in assessing her performance. To address this issue, several strategies can be employed:

1. **Imputation:** One common method is to estimate the missing value using existing data. **Example:** The school can calculate the average score of all students in Sara's class. If the average score is 87, the school may assign this value to Sara's record temporarily. This approach allows the school to maintain a complete dataset while making a reasonable assumption about Sara's performance.
2. **Flagging:** The school can also keep track of Sara's missing score by adding a note in the dataset. This method indicates that Sara's score is not available, making analysts aware of the incomplete data. This approach ensures transparency while allowing the analysis to proceed without filling in the gap.
3. **Removal:** If the number of missing entries is small, the school might choose to exclude Sara's record from specific analyses. This decision is acceptable if it does not significantly impact the overall understanding of student performance. However, it risks losing valuable information about Sara.

5.3 Building Statistical Models

In this section, we will explore the basic building blocks of statistical models, including different types of models, how they are developed, and how to evaluate their performance. We'll also look at real-world examples to make the concepts easy to understand.

5.3.1 Introduction to Statistical Modeling

Statistical modeling is a introduced to analyse data to make sense of the real-world and to predict what will happen in the future. Think of it like this: if you want to know how much money you'll spend on groceries next month, you can look at your past expenses. By analyzing that data, you can create a model to help you estimate your future grocery expenses.

5.3.1.1 Model Development

Building a statistical model involves several steps. Let's break them down:

- **Step 1: Define the Problem**

  First, we need to understand the problem. Example: If we are trying to predict grocery expenses, we need to identify the factors that influence them (e.g., family size, location, or income).

- **Step 2: Collect Data**
  Next, we gather data related to the problem. In our example, we will collect data on past spending habits, number of family members, and any other factors that may affect grocery costs.
- **Step 3: Choose an Algorithm**
  Based on the problem and the data, we choose an appropriate algorithm. Algorithms are methods that help us build models. Some popular algorithms are linear regression and logistic regression, which we will later discuss in this section.
- **Step 4: Train the Model**
  The model is then trained using the collected data. This means the model learns from the data to make predictions.
- **Step 5: Evaluate the Model**
  Finally, we test the model to see how well it works by using new or unseen data. This step is very important to ensure the model makes accurate predictions.

5.3.1.2 Linear Regression

Linear regression is a widely used statistical model that helps understand the relationship between two variables. It is often used to predict one variable based on another. Let's go through a practical example to explain how it works.

Example: Imagine you run a small fruit stall in your town, and you want to predict how much money you will make each day based on the number of customers who visit your stall. The number of customers is the independent variable (the cause), and the money you earn is the dependent variable (the effect). We will use linear regression to understand this relationship and help you forecast future earnings.

- **Step 1: Collecting Data**
  To build a linear regression model, we need historical data. Let's assume you've recorded the number of customers and your daily earnings for the past 5 days:

Number of Customers	Daily Earnings (Rs.)
10	500
15	700
20	900
25	1, 100
30	1, 300

Table 5.3: Customer's data

Here, the number of customers is our independent variable (*X*), and the daily earnings are the dependent variable (*Y*).

- **Step 2: Understanding the Linear Regression Formula**
  The formula for simple linear regression is:

$$Y = \beta_0 + \beta_1 \ x + \epsilon$$

Where:

- Y is the dependent variable (in our case, daily earnings),
- X is the independent variable (the number of customers),
- β_0 is the intercept, which is the value of Y when $X = 0$,
- β_1 is the slope of the line, which shows how much Y changes with each unit change in X,
- ε is the error term, which accounts for the difference between the predicted and actual values.

- **Step 3: Building the Linear Regression Model**

  When building a linear regression model, our goal is to find the best line that explains how two things are related in this case, the number of customers and daily earnings. Here's how we get the values for the slope (40) and intercept (300):

  Understanding the Slope ($\beta_1 = 40$)

  The slope shows how much extra money we make for every new customer. Let's use our data to figure it out:

If you notice, for every 5 extra customers, earnings go up by 200 rupees. So, for each new customer:

$\beta_1 = 200/5 = 40$

This means every new customer adds 40 rupees to our earnings.

Understanding the Intercept (β_0):

The intercept (β_0) represents the earnings when no customers visit. To find this value, we look at where the line crosses the vertical axis when the number of customers is zero. In simpler terms, it tells us what the base earnings are, even if no one shows up.

Now, to find the intercept, we need to consider how much we earn when there are no customers.

We can use the equation:

Earnings = $\beta_0 + \beta_1 \times$ Customers

If we take any data point, say when there are 10 customers, the earnings are 500 rupees. Substituting these values into the equation:

$500 = \beta_0 + (40 \times 10)$

$500 = \beta_0 + 400$

Solving this gives:

$\beta_0 = 500 - 400 = 100$

This means that, based on the data, if no customers show up, you'd still expect to make 100 rupees, maybe from regular customers or other fixed earnings. So, the intercept value of 100 rupees represents the minimum amount you'd make on a day with zero customers.

- **Final Equation:**

Earnings = 100 + 40 x Customers

This equation means:

– You'll always earn 100 rupees, even if no one comes.
– Each new customer adds 40 Rs to your total.

- **Step 4: Interpreting the Model**

  Once the model is built, we can use it to predict future earnings. For example, if you expect 22 customers tomorrow, the predicted earnings would be:

  Earnings = 100 + (40 x 22) = 100 + 880 = 980 Rs

  This means that with 22 customers, you can expect to earn around 980 rupees.

- **Step 5: Testing the Model**

  After building the model, it's important to test it using new data. Let's say on the 6th day, 28 customers visited your stall, and you earned 1,250 Rs. Using the model, we can predict the earnings for 28 customers:

  Predicted Earnings = 100 + (40 x 28) = 100 + 1, 120 = 1,220 Rs.

  However, you actually earned 1,250 Rs. The difference between the predicted and actual earnings is called the error:

  Error = 1,220 - 1,250 = -30 Rs.

  While the prediction was close, it is not perfect, showing that real-world data often has some variation.

Tidbits

To improve your statistical model, consider these suggestions:

1. Use more data points for better accuracy.
2. Include relevant factors such as like family size or special events that may affect spending's.
3. Regularly update your model with new data to keep it relevant.
4. Test your predictions against actual spending to refine your approach.

5.3.1.3 Logistic Regression

Logistic regression is a powerful tool used when we want to predict an outcome that can be categorized as "yes" or "no".

Example: Let's say we want to determine whether a student will pass or fail an exam based on the number of hours they study. Instead of predicting a specific score, logistic regression helps us find the probability of passing.

Understanding Logistic Regression

Logistic regression is different from linear regression because it does not predict exact numbers. Instead, it provides a probability value between 0 and 1. This means it tells us how likely something is to happen.

5.3.1.4 Clustering Techniques

Clustering is a way of grouping similar things together based on their characteristics. Imagine you have a group of students in your class, and you want to divide them into groups based on their performance in different subjects like math, science, and English. Clustering helps us do that by creating groups of students who perform similarly.

Example: Clustering of Students by Performance

Let's say we have data for five students, showing their scores in math and English:

Student	Math Score	English Score
Basim	85	70
Umer	90	65
Anie	50	80
Tallat	40	85
Maliha	60	60

Table 5.4: Data for Clustering Techniques

We can use clustering to group these students based on their performance in these two subjects.

K-means Clustering

K-means clustering is one of the simplest and most popular techniques to group data. In K-means, we need to decide how many groups (clusters) we want. For this example, let's say we want to divide the students into two clusters: one for students who are strong in math and one for students who are strong in English. The algorithm will group students with similar performance in math and English together by calculating the distance between their scores and finding patterns. It will assign students like Basim and Umer (who are good at math) to one group and students like Anie and Tallat (who are good at English) to another group.

5.3.2 Evaluating and Interpreting Models

Once a model is built, it is important to check how well it performs and to understand the results it provides. This is called model evaluation.

5.3.2.1 Performance Metrics

Performance metrics help us measure how well a model is doing. Some common metrics include:

- **Error Metrics**

  Error metrics measure how much the model's predictions differ from the actual values. In our grocery example, if the model predicts a monthly grocery bill of

8,000 rupees but the actual bill is 10,000 rupees, the difference is the error.

- **Accuracy Metrics**

  Accuracy metrics tell us how many of the model's predictions were correct. For example, if a model predicts whether a student will pass or fail an exam, accuracy measures the percentage of correct predications made by the model.

5.3.2.2 Interpreting Outputs

Interpreting a model's output means understanding what the results reveals.

Drawing Conclusions from Insights

For example, if our linear regression model shows that number of hours studied strongly affects exam scores, we can conclude that students should study more hours to improve their scores.

5.3.2.3 Ethical Considerations

When building models, it is important to consider the ethical implications, such as fairness and privacy.

Fairness and Bias

A model should be fair and unbiased. For example, if a model is used to decide who gets a loan, it should not unfairly favor one group of people over another.

Data Privacy

When using personal data to build models, it is important to respect privacy. For example, if a company is using customer data to build models, they should ensure the data is secure and not shared without permission.

Tidbits

Always visualize your data before building models and test your results on new data for better accuracy.

5.4 Introduction to Data Visualization

Data visualization is the process of representing data in a visual format, such as graphs or charts. It helps us to quickly identify patterns, trends, and insights from the data.

5.4.1 Types of Visualizations

Data visualization is a powerful way to understand complex information. Different types of visualizations serve various purposes, making it easier to interpret and analyze data. Below are some common types of visualizations explained in detail:

5.4.1.1 Bar Charts

Bar charts are ideal for comparing different categories. Each bar represents a category, and the height (or length) of the bar indicates the value associated with that category.

Example: Imagine you want to compare the sales figures for different products in a store. A bar chart can visually represent this data.

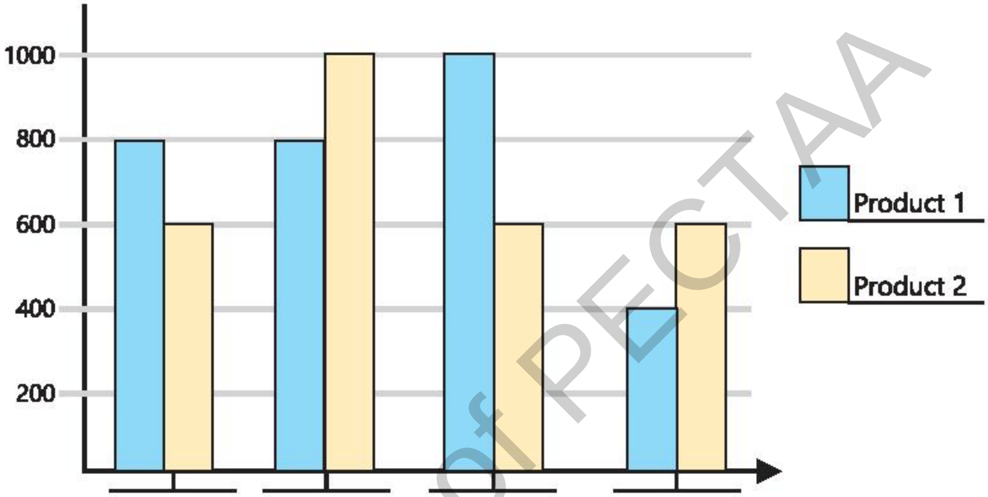

Figure 5.1: A bar chart showing sales figures of different products

5.4.1.2 Line Graphs

Line graphs are used to show trends over time. They plot data points and connect them with a line, making it easy to observe changes.

Example: If you track the temperature over a week, a line graph will show how the temperature rises and falls each day.

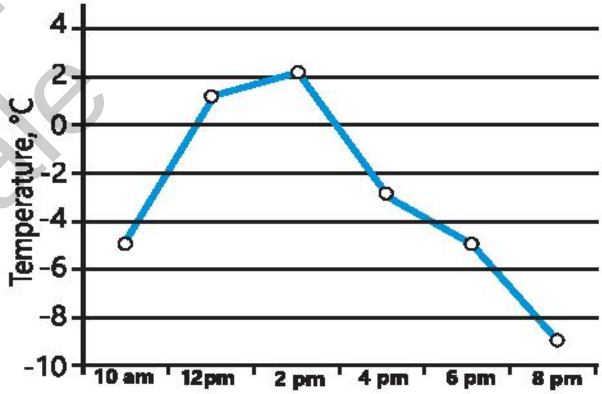

Figure 5.2: A line graph showing variation of temperature over time

5.4.1.3 Histograms

Histograms are used to show the distribution of a dataset. They group data into bins or intervals, allowing you to see how frequently values occur within those ranges.

Example: If you want to analyze how students performed in math exam, a histogram can show the distribution of scores.

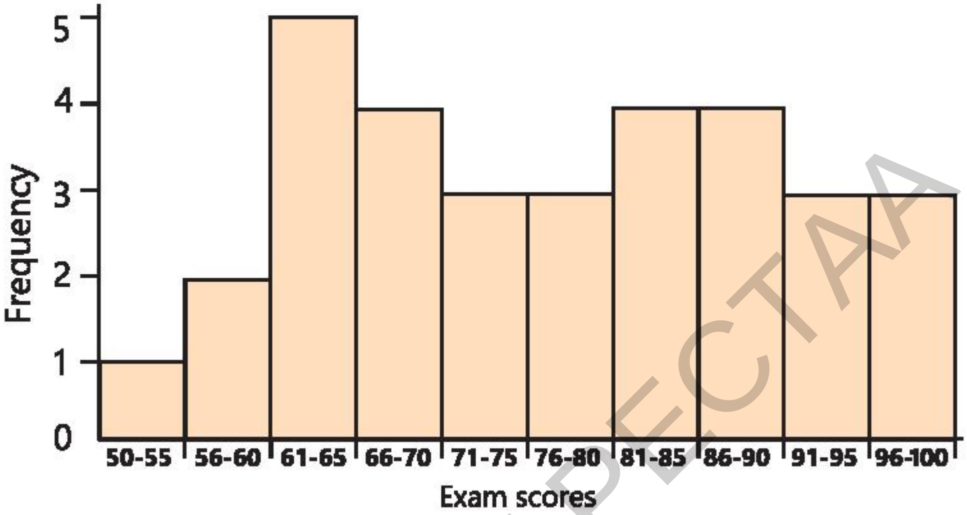

Figure 5.3: Example of a Histogram showing the distribution of exam scores

5.4.1.4 Scatterplots

Scatterplots are used to display the relationships between two variables. Each point on the graph represents an observation, and the position indicates values for both variables.

Example: A scatterplot can be used to explore the relationship between the number of hours studied and exam scores. (see Figure 5.4)

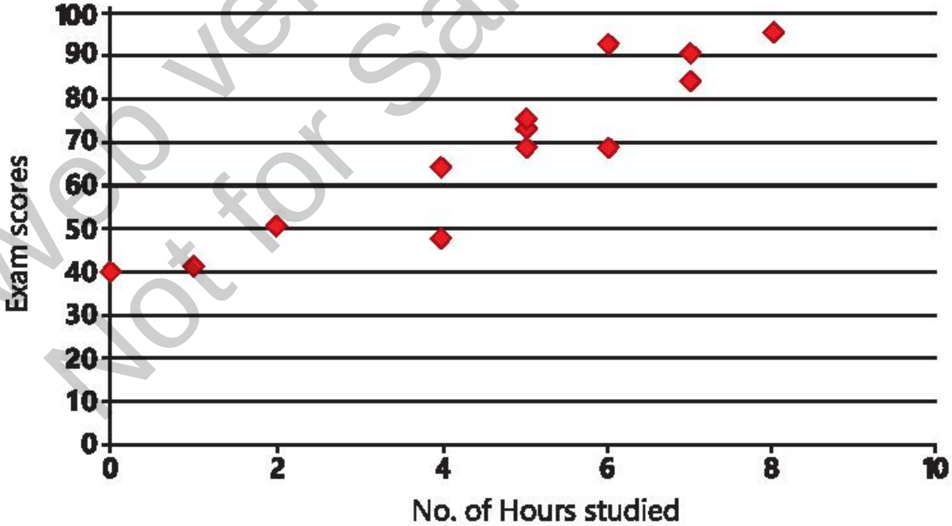

Figure 5.4: Scatterplot showing the relationship between hours studied and exam scores

5.4.1.5 Boxplots

Boxplots, or whisker plots, summarize data distribution by displaying the median, quartiles, and potential outliers. They provide a visual summary of data variability and spread.

Example: A boxplot can be used to compare the exam scores of different classes to see which class performed better overall.

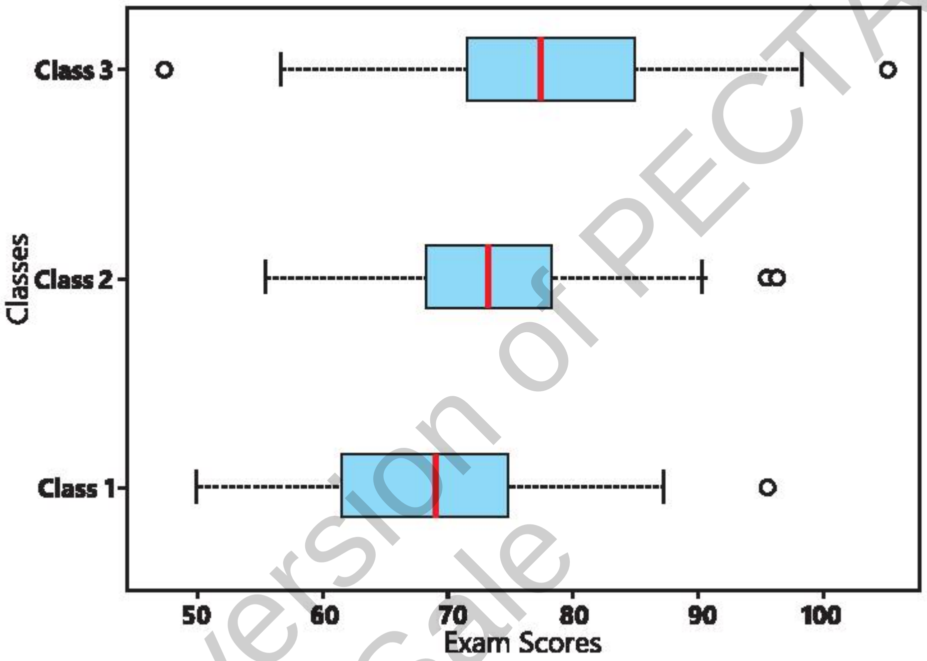

Figure 5.5: A Boxplot, showing class scores performance of three classes

5.5 Tools for Data Visualization

As we discussed in the above section visualization data helps us make sense of large amounts of information by turning numbers into easy-to-understand charts and graphs. In this section, we will discuss tools that can be used to create these visualizations and guide you through how to create and interpret them step by step.

There are many tools available for creating data visualizations, but some of the easiest to use are ones you may already be familiar with, such as such as Microsoft Excel and Google Sheets. These tools are widely accessible and provide straightforward methods for creating to create charts, graphs, and other visual representations of data.

5.5.1 Using Excel and Google Sheets for Visualization

Excel and Google Sheets: These tools allow you to easily enter your data and then generate a variety of visualizations such as bar charts, line graphs, and scatterplots.

Example: Let's say you run a small business in your local area, and you want to track how many products you sell each month. You can enter the data for each month in Excel or Google Sheets, and with just a few clicks, you can create a bar chart to see which month had the most sales.

5.2.2 Creating and Interpreting Visualizations

Step-by-Step Guide: Here's a simple guide to creating a visualization in Excel or Google Sheets.

1. **Enter Your Data:** Start by entering your data into the spreadsheet. Example: In one column, you could have the months (January, February, etc.), and in another column, the sales figures for each month.
2. **Select the Data:** Highlight the data you want to visualize by clicking and dragging your mouse over the cells.
3. **Choose a Chart Type:** Click on the "Insert" tab and select the type of chart you want to create (bar chart, line graph, etc.).
4. **Customize the Chart:** You can add labels to your chart to make it clearer, such as labeling the x-axis with the months and the y-axis with the sales figures. This makes the chart easier to interpret.
5. **Understanding Statistical Representations:**

   When you create a visualization, it's important to understand what the chart is telling you.

EXERCISE

Multiple Choice Questions

1. An example of a basic statistical model:

   a) Linear Regression b) Neural Networks

   c) Decision Trees d) Support Vector Machines

2. The activity involved in experimental design in data science:

   a) Creating visualizations

   b) Collecting and analyzing data systematically

   c) Writing code for machine learning

   d) Building databases

3. A commonly used tool for creating data visualizations:

   a) MS Excel b) Python (Matplotlib)

   c) Tableau d) All of the above

4. The meaning of the slope in a linear regression model:

   a) The intercept of the model

   b) The change in the dependent variable for a unit change in the independent variable

   c) The error term

   d) The mean of the data

5. An example of a real-world application of statistical models:

   a) Predicting house prices

   b) Creating social media posts

   c) Designing websites

   d) Writing essays

6. Option not considered a benefit of data visualization:

   a) Identifying trends and patterns

   b) Communicating insights effectively

   c) Making data more complex

   d) Summarizing large datasets

7. A primary goal of K-Means Clustering:

   a) To classify data into predefined categories

   b) To group data into clusters based on similarity

   c) To predict continuous outcomes

   d) To reduce the dimensionality of data

8. The meaning of "K" in K-Means Clustering:

a) Number of features in the dataset

b) Number of clusters to be formed

c) Number of iterations required for convergence

d) Number of data points in the dataset

Short Questions

1. What is the importance of building statistical models in real-world applications?
2. Name one basic statistical model used for predicting outcomes and explain its purpose.
3. List two types of data visualizations and describe when you would use each.
4. How does visualizing data help in understanding descriptive statistics?

Long Questions

1. Explain the role and importance of statistical models in solving real-world problems.
2. Describe the steps involved in building a basic statistical model (e.g., linear regression). Include details on data collection, model training, and evaluation.
3. Discuss the types of data visualizations and their uses.
4. Explain data collection methods.
5. Discuss the concept of measure of tendency with example.