Chapter 4: Data Analysis

By Ramesh Tamang

4.1 Data analytics: Descriptive, diagnostic, predictive and prescriptive analytics

What is Data Analysis?

- **Data Analysis** is the process of systematically examining and modeling data to extract useful information, identify patterns, draw conclusions and support decision-making.
- **Types of data analytics**
  1. Descriptive
  2. Diagnostic
  3. Predictive
  4. Prescriptive

Example Use Case

- Scenario: E-commerce platform analyzing customer behavior.
  1. Descriptive: Analyze last quarter's sales data.
  2. Diagnostic: Identify why sales dipped in a specific month.
  3. Predictive: Forecast next quarter's sales.
  4. Prescriptive: Recommend marketing strategies to boost sales.

1. Descriptive Analytics

- Focuses on summarizing historical/past data to understand **what has happened.**
- It is the simplest form of analytics and serves as the foundation for other types of analytics (diagnostic, predictive, prescriptive).
- Tools: Python (pandas, matplotlib, seaborn).
- Example: Analyzing sales data to determine the total revenue over a quarter."

Techniques in Descriptive Analytics

Summary Statistics

- Measures of Central Tendency: Mean, median, and mode.
- Measures of Dispersion: Range, variance, standard deviation.
- Measures of Distribution: Skewness, kurtosis.
- Percentiles: Understanding data distribution at specific thresholds (e.g., 25th, 50th, 75th percentiles).

Data Aggregation

- Summing or averaging values across groups (e.g., sales by region).
- Compute sums, averages, counts, and maximum/minimum values.
- Example: Customer data might be segmented by demographics, region, or product category to uncover insights.

Techniques in Descriptive Analytics (contd.)

Cross-Tabulation

- Examining relationships between multiple categorical variables in a tabular format.

Time-Series Analysis

- Understanding trends over time or
- Observing patterns over time (e.g., seasonality in sales).

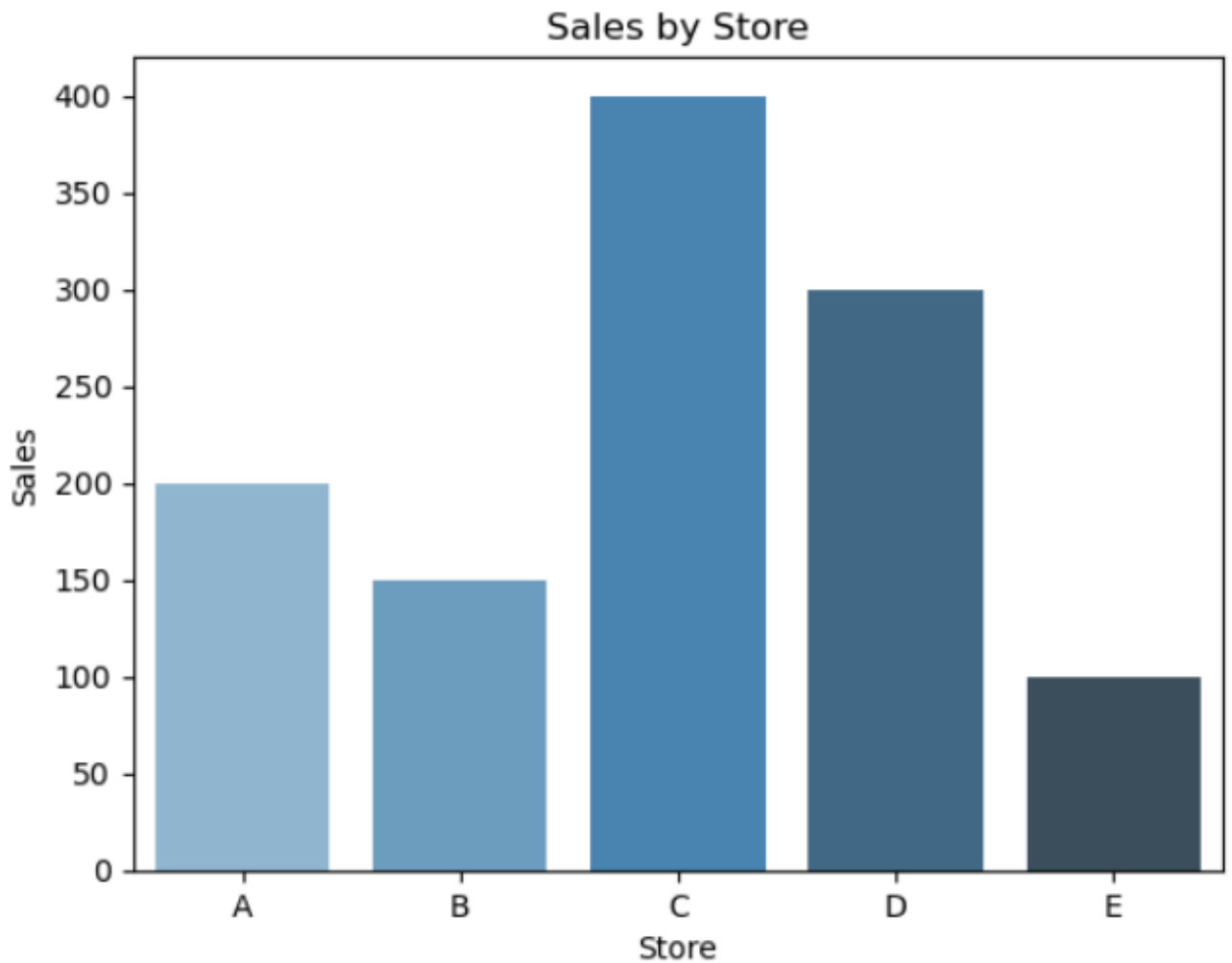
Data Visualization

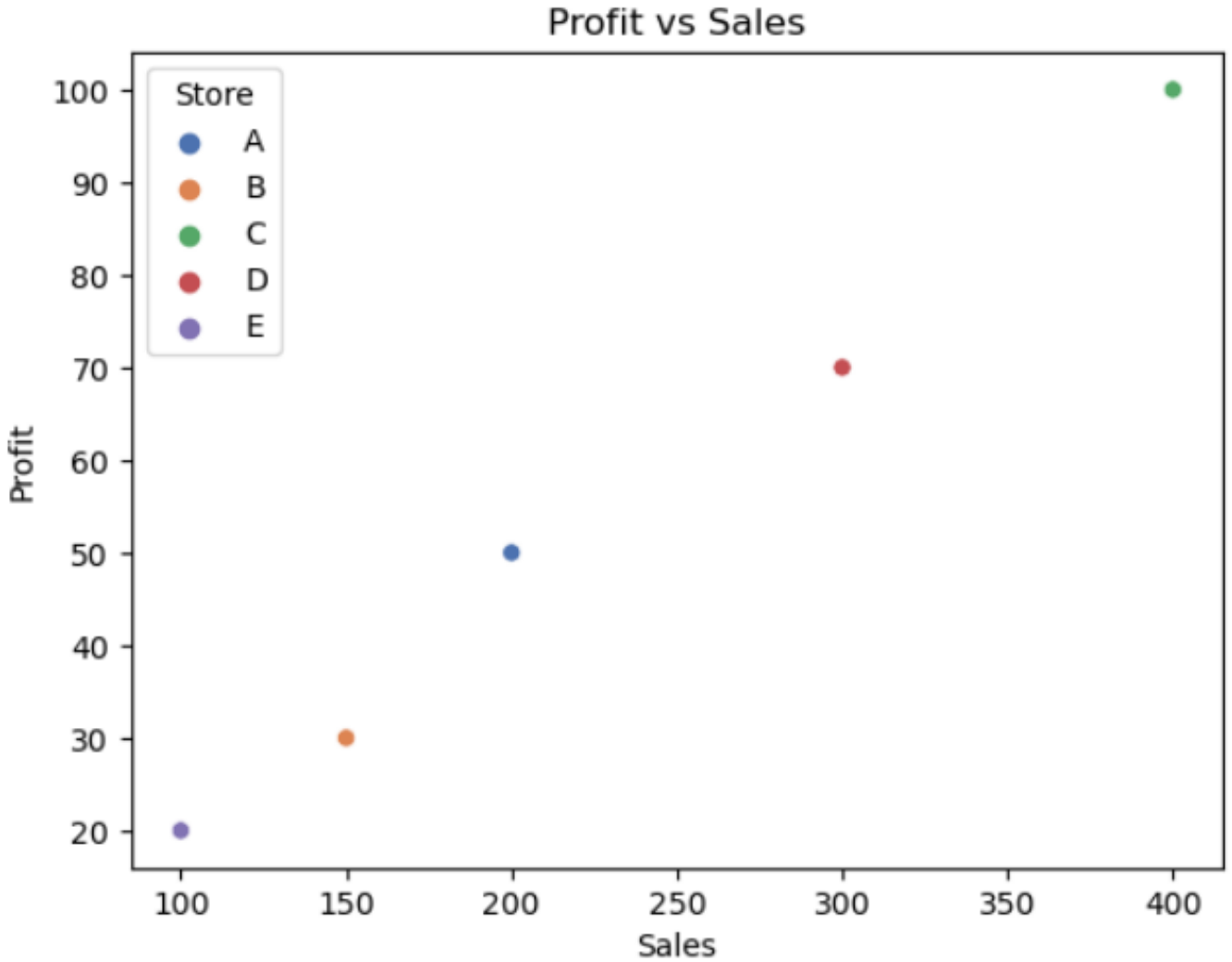
- Bar charts
- Line graphs
- Pie charts
- Heatmaps

```python
data = {
    "Store": ["A", "B", "C", "D", "E"],
    "Sales": [200, 150, 400, 300, 100],
    "Profit": [50, 30, 100, 70, 20]
}
```

```
Summary Statistics:
             Sales      Profit
count     5.000000    5.000000
mean    230.000000   54.000000
std     120.415946   32.093613
min     100.000000   20.000000
25%     150.000000   30.000000
50%     200.000000   50.000000
75%     300.000000   70.000000
max     400.000000  100.000000

Total Sales: 1150
Average Sales: 230.0
Top Performing Store: C
```




Profit vs Sales
Store
A
B
C
D
E
Profit
100
90
80
70
60
50
40
30
20
100
150
200
250
300
350
400
Sales

```
# Sample data
data = pd.DataFrame({
    "Gender": ["Male", "Female", "Female", "Male", "Male", "Female"],
    "Preference": ["Product A", "Product B", "Product A", "Product B", "Product A", "Product B"]
})
```

Preference	Product A	Product B
Gender		
Female	1	2
Male	2	1

2. Diagnostic Analytics

- Diagnostic analytics examines why certain events happened in the past by analyzing data patterns and relationships.
- Explores causal relationships to explain outcomes.
- Diagnostic Analytics answers the question: “Why did this happen?”
- Key Objectives
  - Determine causal relationships between variables.
  - Identify patterns or anomalies (unexpected behaviors) in historical data.
  - Provide insights for decision-making by answering "Why?" questions.

Steps in Diagnostic Analytics:

1. Understand the Problem:
    - Clearly define the question to address, e.g., "Why did sales decline in Q3?"
2. Collect and Clean Data:
    - Gather relevant datasets.
    - Perform preprocessing (e.g., handle missing values).
3. Analyze Relationships:
    - Apply statistical and visualization techniques to identify dependencies between variables.
4. Identify Root Causes:
    - Use methods like correlation analysis, anomaly detection, and root cause analysis.
5. Communicate Insights:
    - Present findings clearly with appropriate visualizations.

Techniques in Diagnostic Analytics

1. Correlation Analysis:
   - Measures the strength and direction of relationships between two variables.
   - Example: Correlation between marketing spend and sales.
2. Root Cause Analysis (RCA):
   - Identifies the primary cause of a problem.
   - Example: Using **Pareto Analysis** or charts to identify major contributors to customer complaints.
   - Example: To analyze why sales have dropped using the **5 Whys Analysis**.
3. Anomaly Detection:
   - Identifies unusual data points.
   - Example: Detecting sudden spikes in website traffic or detecting fraudulent transactions in financial data.

Techniques in Diagnostic Analytics (contd.)

4. Drill-Down Analysis:
   - Break down data into more detailed levels.
   - Example: Breaking down sales data by region and product.
5. Hypothesis Testing
   - Involves statistical tests to confirm or reject assumptions about data relationships.
   - Common Tests:
     - t-test: Comparing means of two groups.
     - Chi-square test: Testing relationships between categorical variables.
6. Time-Series Analysis:
   - Examines data trends over time to identify patterns or irregularities.
   - Example: Analyzing seasonal fluctuations in sales.

Techniques in Diagnostic Analytics (contd.)

7. Comparative Analysis:
    - Compares different datasets or time periods to identify key differences.
    - Example: Comparing Q1 and Q2 sales performance.
8. Data Mining
    - Identifying patterns, trends, and relationships within large datasets.
    - Common Techniques:
        - Clustering
        - Decision Trees

Tools for Diagnostic Analytics

- Key Python Libraries
  - pandas: Data manipulation.
  - numpy: Numerical computations.
  - scipy: Hypothesis testing.
  - statsmodels: Statistical modeling.
  - scikit-learn: Machine learning and data mining.

Applications

Business:

- Identifying reasons for decreased revenue.
- Diagnosing high employee turnover rates.

Healthcare:

- Exploring causes of increased patient readmissions.

Education:

- Investigating poor student performance in specific subjects.

Example: 5 Whys Analysis

1. Why did sales drop?
   - Because fewer customers are making purchases.
2. Why are fewer customers making purchases?
   - Because customers are not showing interest in the products.
3. Why are customers not showing interest in the products?
   - Because the product offerings have become outdated or less appealing compared to competitors.
4. Why have the product offerings become outdated or less appealing?
   - Because the product development team has not introduced new products or features in a while.
5. Why has the product development team not introduced new products or features?
   - Because there is a lack of investment in market research and innovation.

Examples

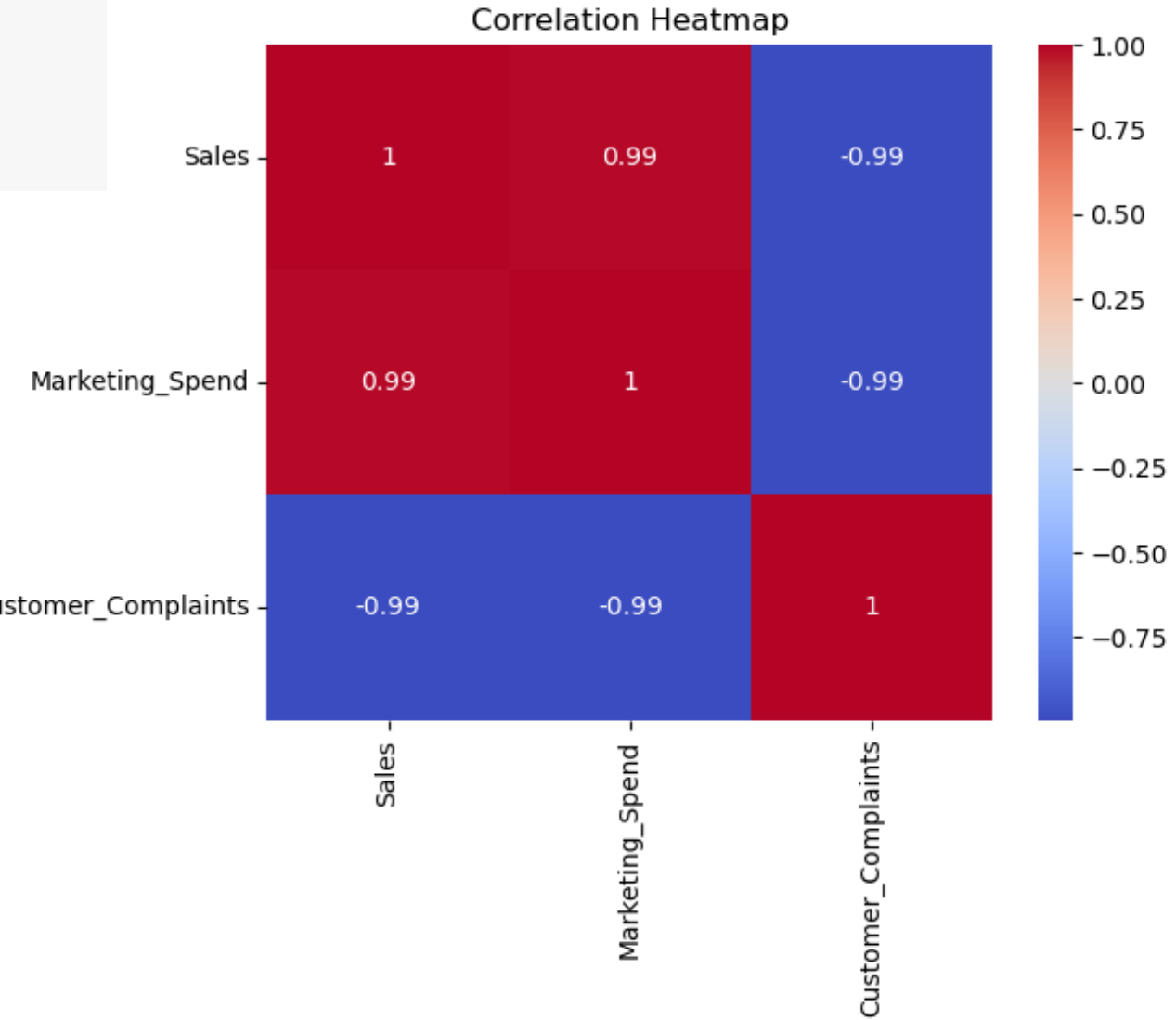
Scenario: An e-commerce platform experiences a drop in sales in Q3.

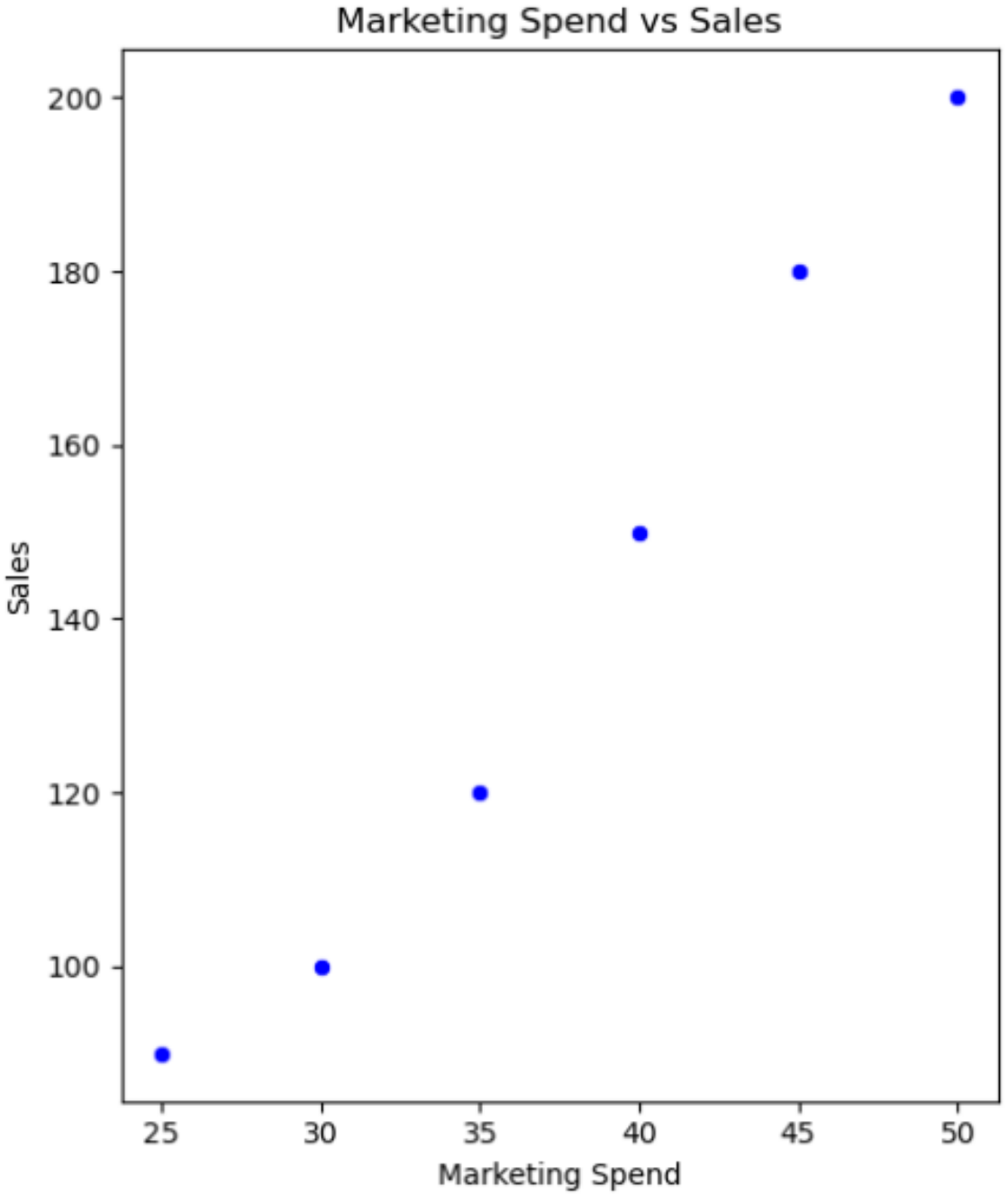
Steps:

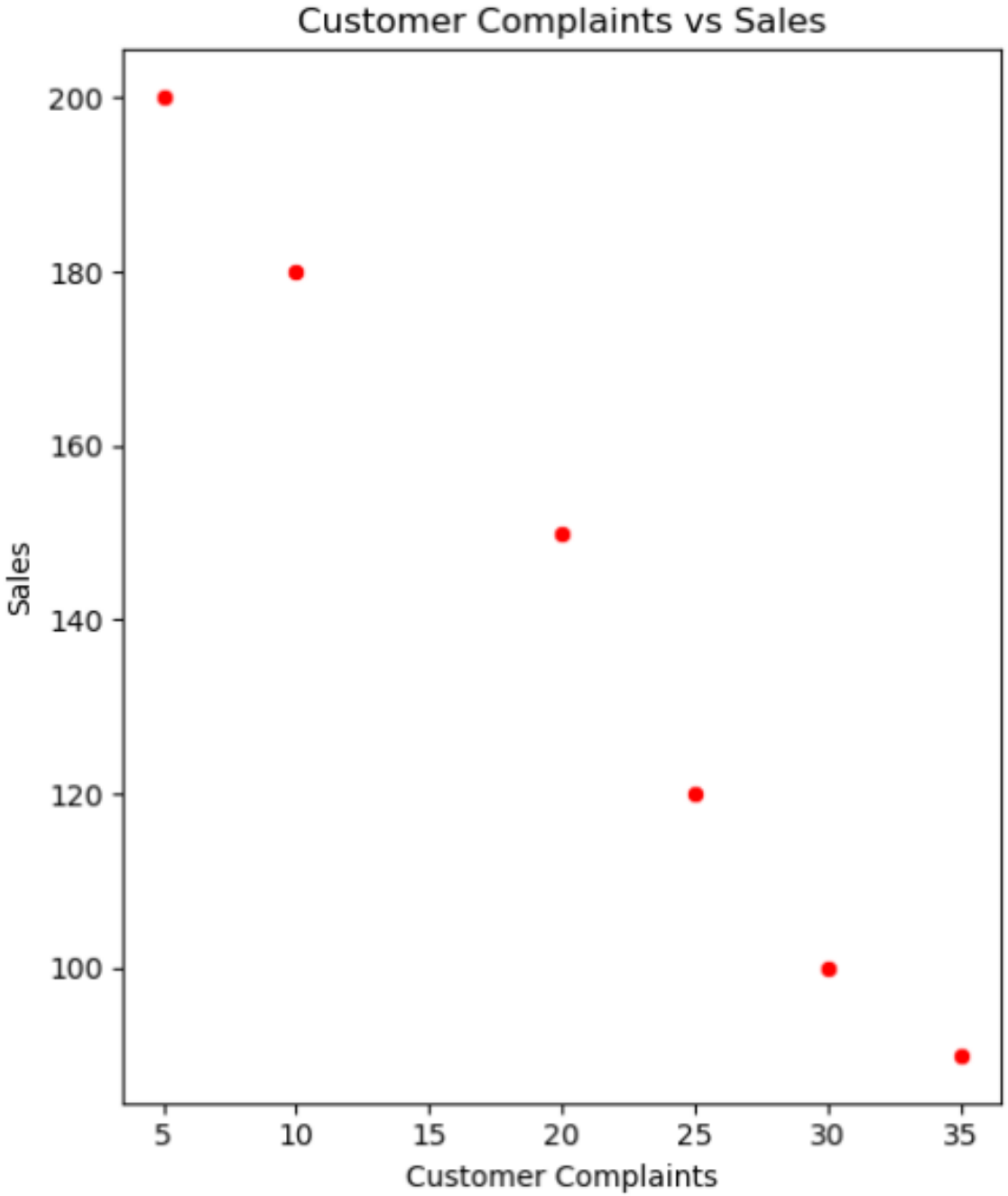
1. Define the Problem:
   - "Why did sales drop in Q3?"
2. Collect Data:
   - Historical sales data, customer feedback, marketing spend.
3. Analyze Relationships:
   - Use correlation and drill-down analysis to identify possible factors.
4. Identify Root Causes:
   - Late deliveries and reduced marketing spend are primary causes.

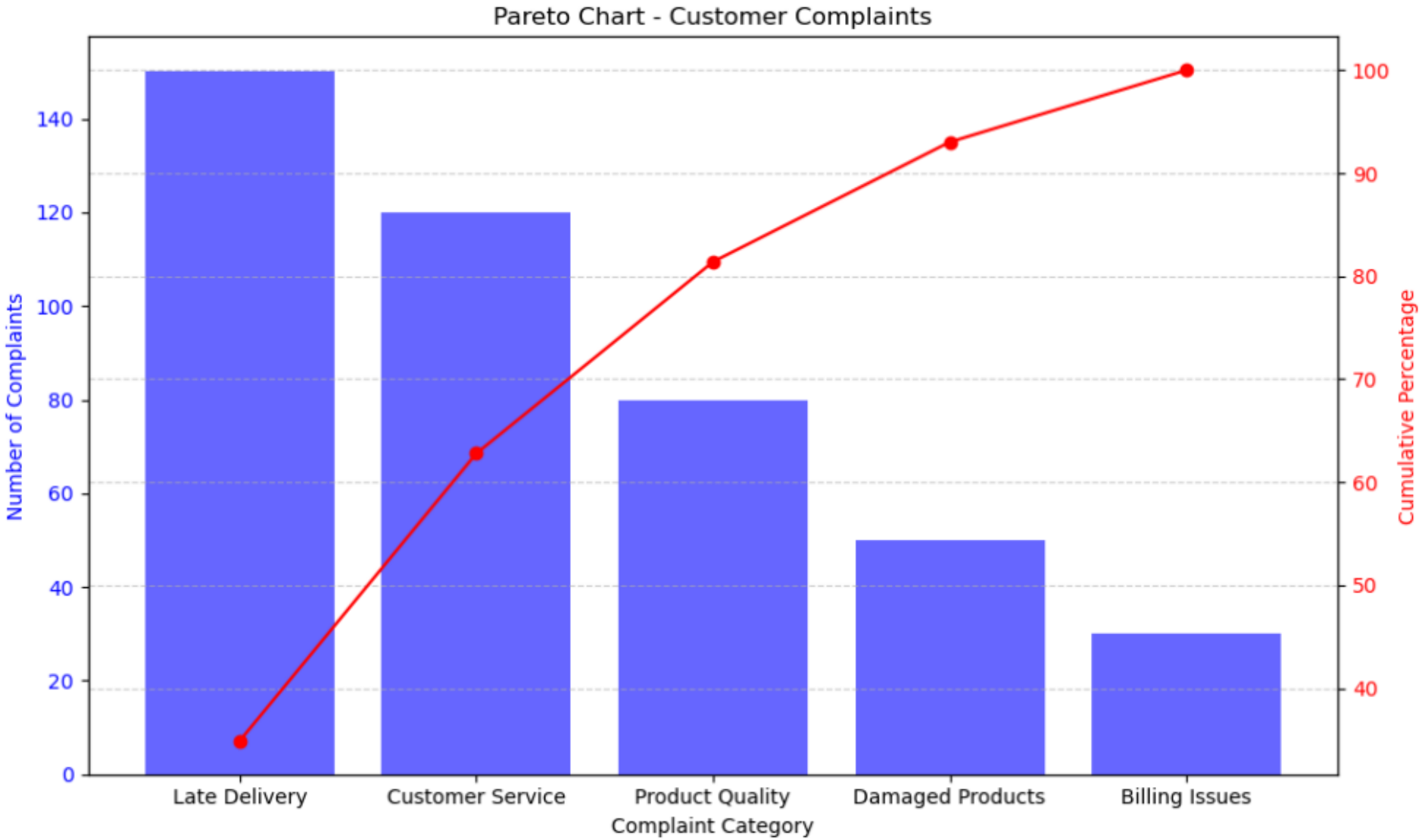
```python
data = {
    'Month': ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun'],
    'Sales': [200, 180, 150, 120, 100, 90],
    'Marketing_Spend': [50, 45, 40, 35, 30, 25],
    'Customer_Complaints': [5, 10, 20, 25, 30, 35]
}
```

```
Correlation Matrix:
                        Sales  Marketing_Spend  Customer_Complaints
Sales                1.000000         0.990038            -0.994534
Marketing_Spend      0.990038         1.000000            -0.992161
Customer_Complaints -0.994534        -0.992161             1.000000
```




Marketing Spend vs Sales
Sales
200
180
160
140
120
100
25
30
35
40
45
50
Marketing Spend


Customer Complaints vs Sales
Sales
200
180
160
140
120
100
5
10
15
20
25
30
35
Customer Complaints


Pareto Chart - Customer Complaints
Number of Complaints
Cumulative Percentage
0
20
40
60
80
100
120
140
40
50
60
70
80
90
100
Late Delivery
Customer Service
Product Quality
Damaged Products
Billing Issues
Complaint Category

3. Predictive Analytics

- Uses historical/past data to forecast future outcomes.
- Predictive Analytics involves using statistical techniques, data mining, and machine learning to make predictions about future outcomes based on historical data.
- **Key Objectives**
  - To predict trends, behaviors, and events.
  - To aid in proactive decision-making.

Applications of Predictive Analytics

- Customer Churn Prediction:
  - Predicting the likelihood of customers discontinuing a service.
  - Example: Telecom companies use predictive analytics to identify at-risk customers.
- Demand Forecasting:
  - Estimating future product demand in supply chains.
- Healthcare:
  - Predicting disease outbreaks or patient readmissions.

Techniques

- Regression Analysis
  - Purpose: Predict a continuous outcome (e.g., sales, temperature).
  - Techniques: Linear Regression, Polynomial regression, etc.
- Classification
  - Purpose: Categorize data into predefined groups (e.g., fraud detection).
  - Techniques: Logistic regression, support vector machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), Decision Trees, Random Forest, etc.
- Time-Series Analysis
  - Purpose: Predict future trends or values based on historical time-ordered data.
  - Techniques: ARIMA (Auto-Regressive Integrated Moving Average), Exponential Smoothing, LSTM (Long Short-Term Memory)

Techniques (contd.)

- Clustering
  - Purpose: Group similar data points for pattern recognition.
  - Techniques: K-Means Clustering, Hierarchical Clustering
- Ensemble Methods
  - Purpose: Combine multiple models for better accuracy and robustness.
  - Techniques: Bagging, Boosting
- Neural Networks
  - Purpose: Model complex patterns and relationships in data.
  - Techniques: Feedforward Neural Networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs):

Tools

- Python Libraries:
  - scikit-learn for regression, classification, and clustering.
  - statsmodels for statistical modeling.
  - xgboost for boosting techniques.
  - TensorFlow and PyTorch for neural networks.

Benefits of Predictive Analytics

- **Improved Decision-Making**: Provides actionable insights for strategic decisions.
  - Example: Retailers can forecast demand and plan inventory accordingly.
- **Cost Savings:** Preventative measures can save resources and optimizes
  - Example: Manufacturing companies predict machine failures and schedule maintenance to avoid costly breakdowns.
- **Better Resource Allocation**
  - Example: Call centers predict peak hours to schedule agents effectively.

Benefits of Predictive Analytics (contd.)

- **Risk Mitigation**: Identifies and minimizes risks proactively.
  - Example: Banks use predictive models to detect fraudulent transactions and block them in real time.
- **Competitive Advantage**: Enables businesses to stay ahead by anticipating market trends.
  - Example: E-commerce companies analyze purchasing patterns to launch trending products faster than competitors.

4. Prescriptive Analytics:

- Recommends actions based on predictions or
- To recommend specific actions based on predictive models and analytics.
- It goes beyond forecasting outcomes to suggest strategies that achieve desired goals.

Key Features

- Action-Oriented:
  - Focuses on actionable recommendations rather than just insights.
- Optimization:
  - Employs mathematical models to determine the best course of action.
- Real-Time Decision-Making:
  - Uses live data and feedback loops for timely interventions.

Techniques in Prescriptive Analytics

- Optimization:
  - Mathematical methods for finding the best solutions under given constraints.
  - Examples: Linear programming
- Simulation:
  - Models real-world processes to evaluate different strategies.
  - Examples: Monte Carlo simulation.
- Machine Learning:
  - Recommends actions based on patterns in historical data.
  - Examples: Reinforcement learning.
- Decision Analysis:
  - Uses decision trees or influence diagrams to evaluate actions.
  - Examples: Risk-reward analysis.

Applications of Prescriptive Analytics

- Logistics and Supply Chain Optimization
  - Determines the best routes, schedules, and inventory levels to reduce costs and improve efficiency.
  - Example: Optimizing shipping routes to minimize transportation costs and delivery times.
- Healthcare Treatment Planning
  - Recommends personalized treatment strategies to improve patient outcomes.
  - Example: Allocating hospital resources, such as staff and beds, based on real-time patient needs.

Applications of Prescriptive Analytics

- Energy Management
  - Identifies optimal strategies for balancing energy supply and demand, ensuring efficient resource utilization.
  - Example: Power companies optimize energy distribution to prevent outages and minimize costs.
- Marketing Campaign Design
  - Suggests effective marketing channels and strategies to maximize customer engagement and conversions.
  - Example: Designing personalized discounts and promotions based on customer preferences.
- Risk Management in Finance
  - Provides strategies to mitigate financial risks and optimize investment portfolios.
  - Example: Dynamically reallocating assets in response to market conditions.

Benefits of Prescriptive Analytics

- Proactive and Actionable Insights
  - Offers specific recommendations to address challenges or capitalize on opportunities.
  - Example: Retailers adjust pricing strategies in real-time to align with competitor actions and market demand.
- Operational Optimization
  - Improves resource allocation and efficiency by suggesting the best possible actions.
  - Example: Airlines optimize crew schedules to reduce delays and enhance operational performance.
- Scenario Planning and What-If Analysis
  - Allows organizations to evaluate multiple scenarios and implement the most effective strategy.
  - Example: Urban planners test different traffic management policies to improve city traffic flow.

- Real-Time Decision-Making
  - Enables timely interventions by providing actionable recommendations based on live data.
  - Example: Stock trading systems execute immediate buy/sell decisions using live market trends.
- Improved ROI and Cost Efficiency
  - Maximizes returns by prescribing actions that optimize costs and benefits.
  - Example: Manufacturing facilities reduce production waste while maintaining high-quality standards.

Benefits of Prescriptive Analytics

- Proactive Decision-Making:
  - Suggests actions to achieve objectives before issues arise.
- Improved Efficiency:
  - Optimizes resource utilization and reduces waste.
- Enhanced Customer Satisfaction:
  - Provides personalized experiences through tailored recommendations.
- Risk Reduction:
  - Identifies optimal solutions that minimize potential risks.
- Competitive Advantage:
  - Helps businesses adapt quickly to market changes.

Aspect	Predictive Analytics	Prescriptive Analytics
Focus	Forecasts future trends or outcomes.	Recommends actionable strategies based on forecasts.
Actionability	Provides insights but no direct actions.	Offers specific courses of action to achieve objectives.
Example in Retail	Predicts customer demand for the next month.	Suggests optimal inventory levels and sourcing strategies.
Example in Healthcare	Identifies patients at risk of readmission.	Recommends treatment plans to reduce readmission rates.

4.2 Exploratory data analysis using descriptive statistics

Exploratory Data Analysis (EDA)

- Exploratory Data Analysis (EDA) is a process of analyzing and summarizing datasets using statistical and visualization techniques to uncover patterns, relationships, anomalies, and insights.
- Descriptive statistics provide the foundation for EDA by summarizing and describing the characteristics of data.

Steps in EDA Using Descriptive Statistics

1. Data Overview
    - Inspect dataset size, shape, and data types.
    - Check for missing values or inconsistencies.

2. Summary Statistics: Provides key measures of central tendency, dispersion, and shape.

Key Metrics:

- Central Tendency:
    - Mean: Average of the data.
    - Median: Middle value of sorted data.
    - Mode: Most frequent value.

Steps in EDA Using Descriptive Statistics (contd.)

- Dispersion:
  - Range: Difference between the maximum and minimum values.
  - Variance: Measure of data spread.
  - Standard Deviation: Square root of variance, indicating variability.
- Shape:
  - Skewness: Measures asymmetry of the data distribution.
  - Kurtosis: Indicates the peakedness or flatness of the distribution.

3. Distribution Analysis
   - Understands how data points are distributed across the range.
   - Tools: Histograms, box plots.

Steps in EDA Using Descriptive Statistics (contd.)

4. Correlation and Covariance
   - Correlation: Measures the strength and direction of the linear relationship between two variables ($-1 \leq r \leq 1$).
   - Covariance: Indicates how two variables change together.
   - Tools: Correlation heatmaps and scatter plots.
5. Outlier Detection
   - Detects abnormal or extreme values that may impact analysis.
   - Tool: Z-Score: Measures how far a value is from the mean in terms of standard deviations.

4.3 Data Visualization

4.3 Data Visualization

- Data Visualization is the graphical representation of data using charts, graphs, and other visual elements.
- It helps interpret complex datasets easily, revealing patterns, trends, and insights at a glance.

Purpose of Data Visualization

- Simplify Data Interpretation:
  - Transforms raw data into meaningful visuals for better understanding.
- Highlight Patterns and Trends:
  - Reveals relationships and trends that might not be apparent in tabular data.
- Support Decision-Making:
  - Provides actionable insights to guide business strategies.
- Enhance Communication:
  - Enables effective storytelling by conveying insights visually.

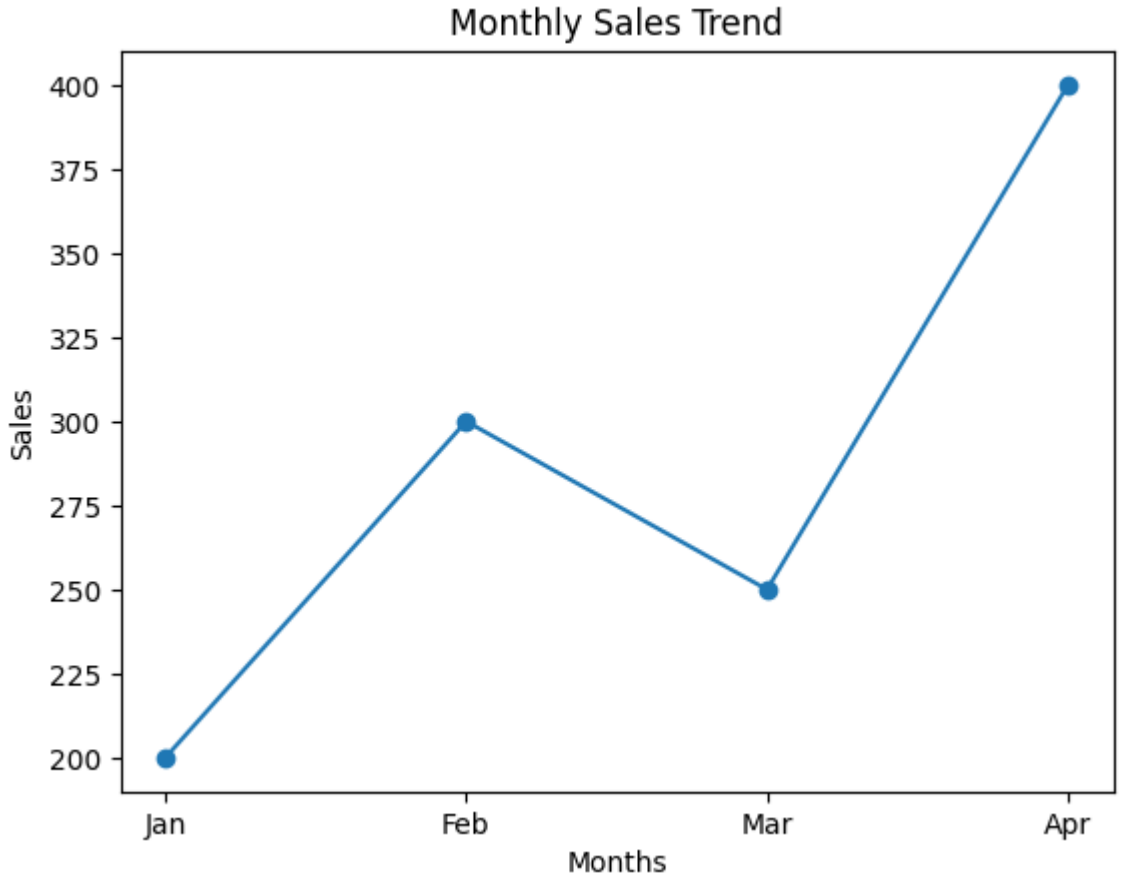
Common Visualization Techniques

- Line Charts: Show trends over time.
- Bar Charts: Compare categorical data.
- Pie Charts: Shows proportions or percentages of a whole.
- Heatmaps: Visualize correlations.
- Box Plots: Summarize data distribution and detect outliers.
- Scatter Plots: Shows relationships between two variables.
- Histograms: Displays the frequency distribution of a dataset.

Line Charts: Show trends over time

Example: Stock price movements or monthly sales trends.

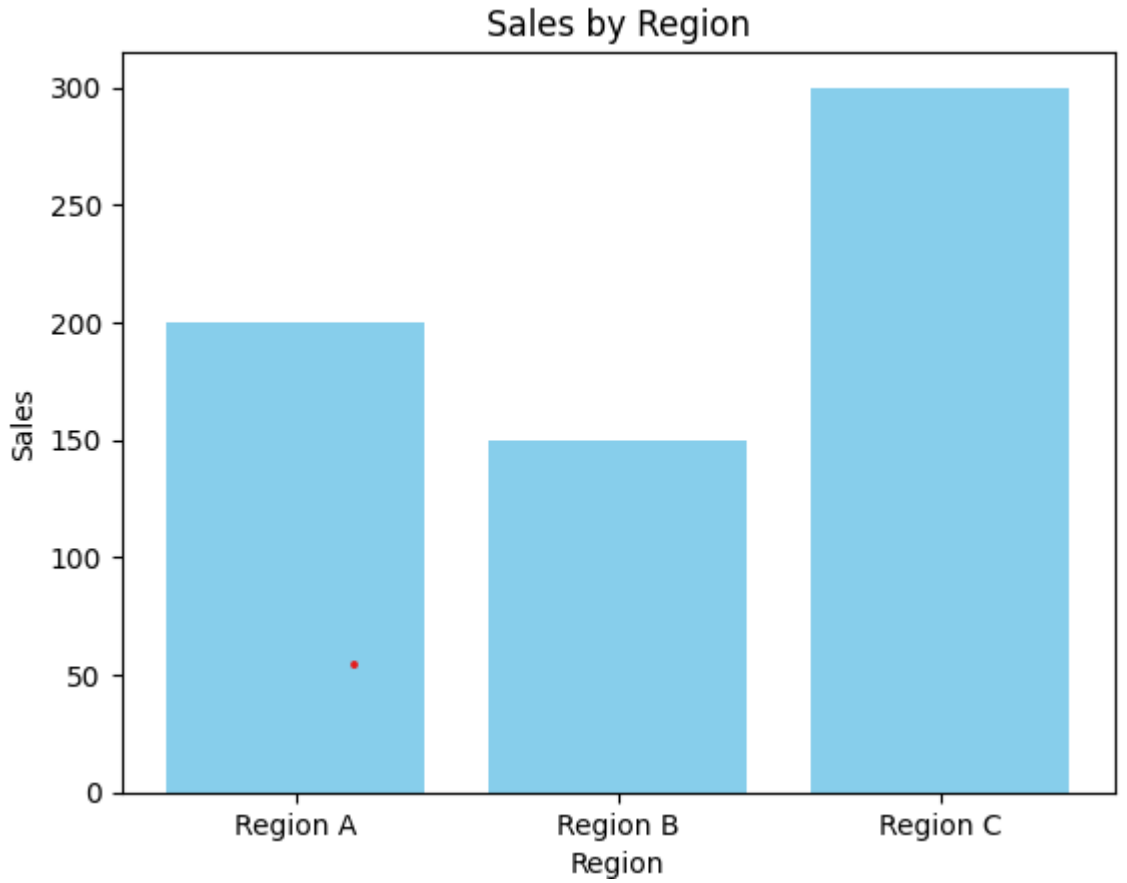
months = ['Jan', 'Feb', 'Mar', 'Apr’]
sales = [200, 300, 250, 400]



Bar Charts: Compare categorical data

Example: Comparing sales figures across different regions.

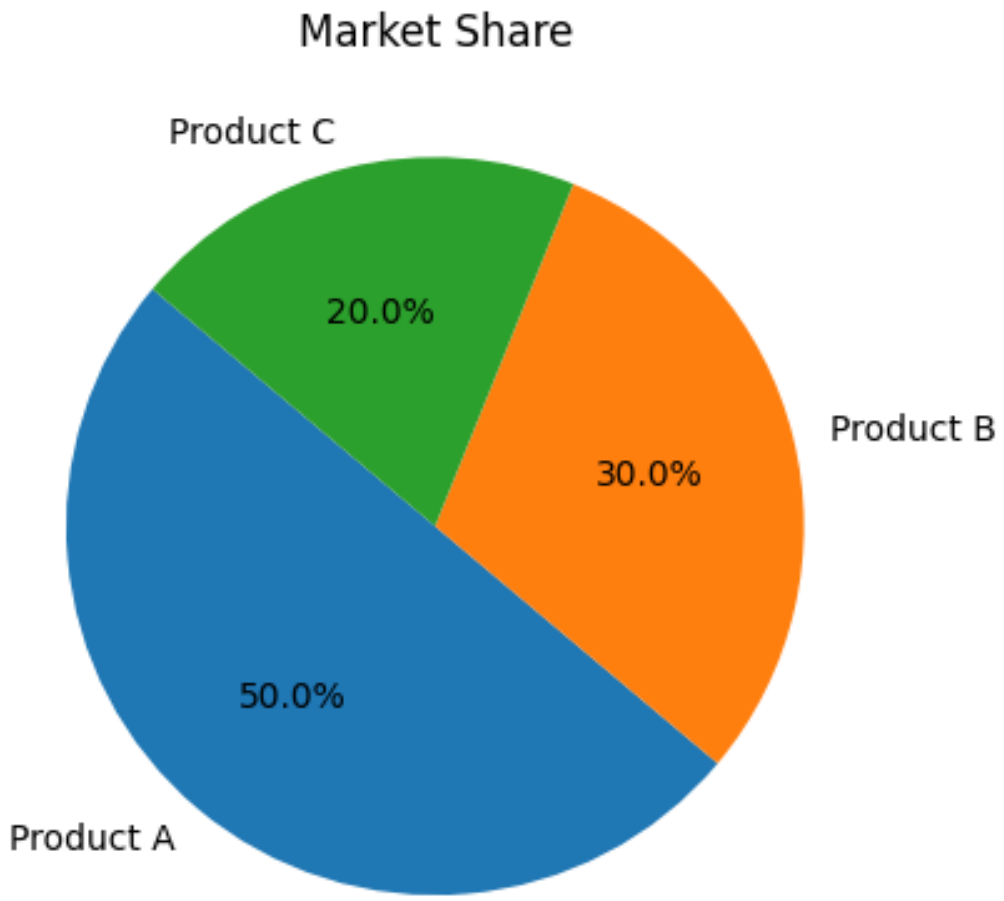
categories = ['Region A', 'Region B', 'Region C’]
values = [200, 150, 300]



Pie Charts: Shows proportions or percentages of a whole

Example: Market share of products.

labels = ['Product A', 'Product B', 'Product C']
sizes = [50, 30, 20]

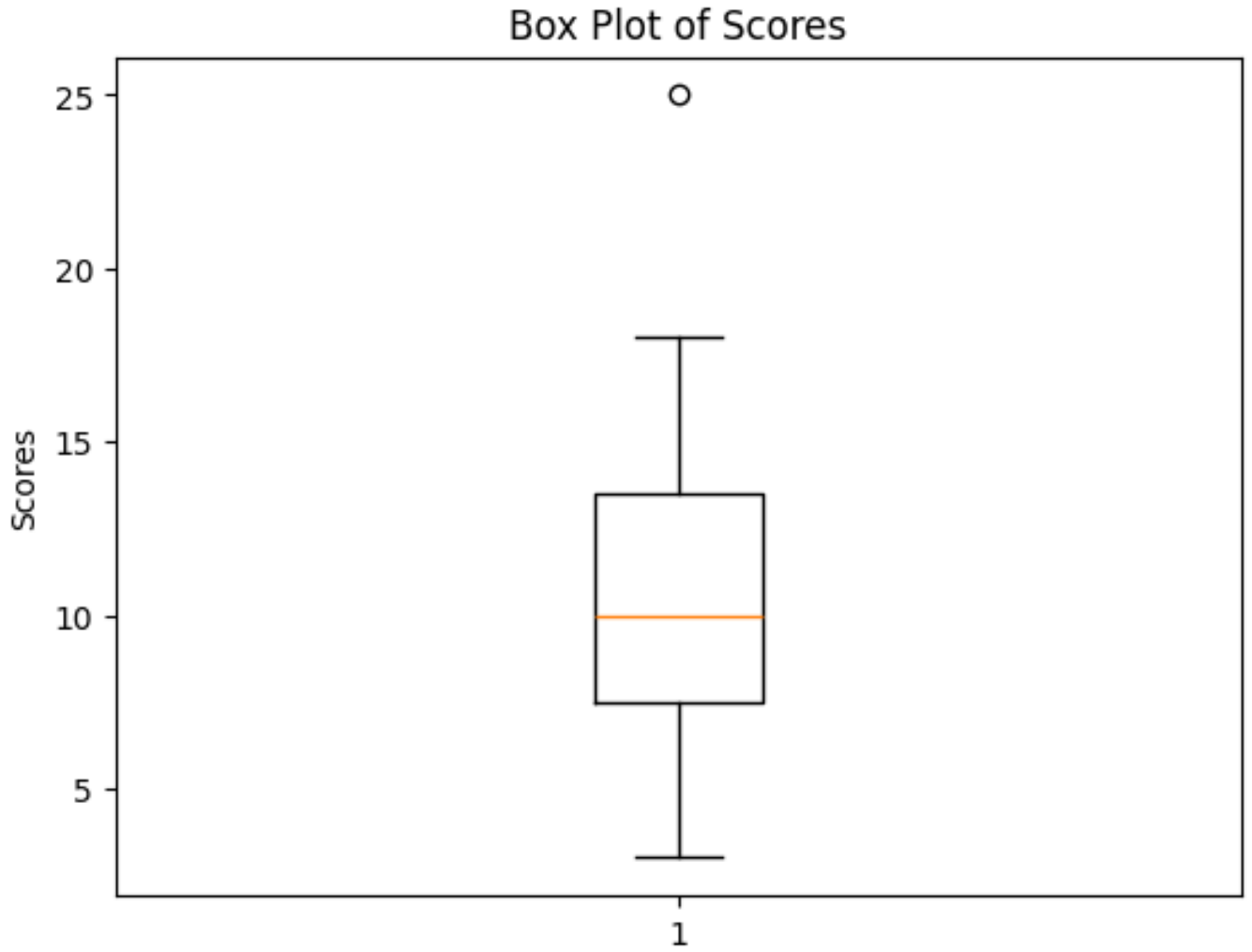


Heatmaps: Visualize correlations.

Box Plots: Summarize data distribution and detect outliers

Example: Distribution of exam scores across a class.

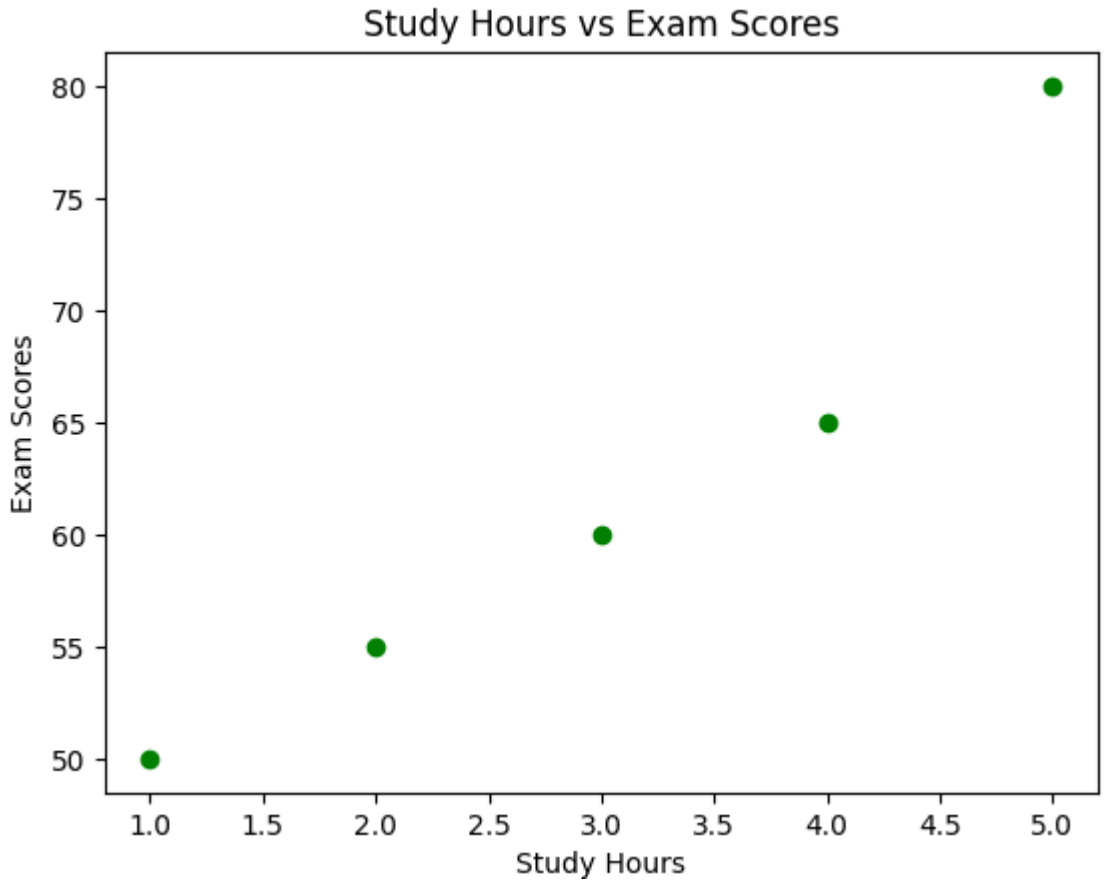
data = [3, 3, 7, 8, 8, 10, 11, 12, 15, 18, 25]



Scatter Plots: Shows relationships between two variables

- **Example**: Correlation between study hours and exam scores.

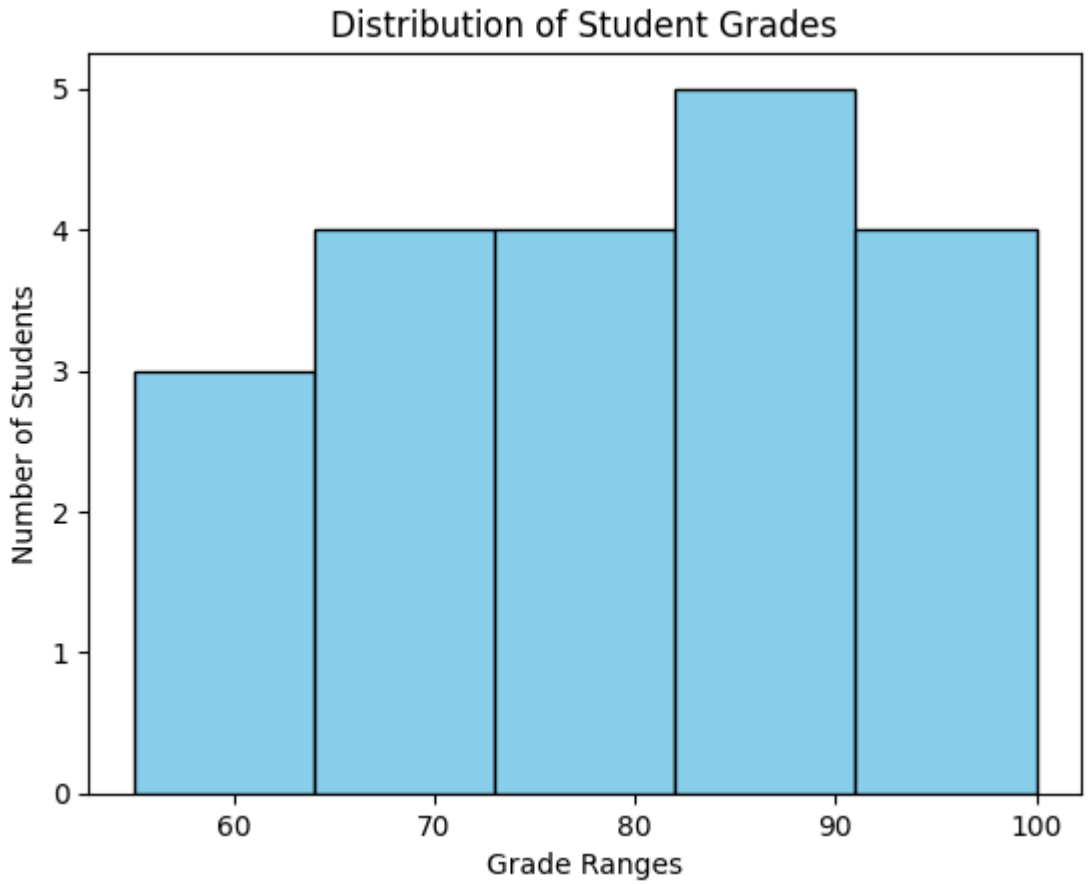
```
hours = [1, 2, 3, 4, 5]
scores = [50, 55, 60, 65, 80]
```



Histograms: Displays the frequency distribution of a dataset

Example: Distribution of student grades in a class.

grades = [55, 62, 77, 85, 90, 65, 72, 88, 92, 76, 84, 95, 67, 78, 82, 94, 60, 70, 80, 100]



4.4 Principles of Effective Data Visualization

- Clarity
  - Avoid unnecessary clutter, use clear labels and legends and ensure visuals are easy to interpret.
- Focus
  - Highlight the most critical aspects of the data
- Data Integrity
  - Ensure the data is represented accurately and without bias.
  - Represent data proportions and relationships truthfully.
- Aesthetics
  - Use appropriate colors, fonts, and layouts.
- Consistency
  - Use uniform styles, scales, labels, and colors.

4.6 Feature Engineering

- Definition: The process of creating new features or modifying existing ones to improve model performance.

Process

1. Understanding the Problem:
   - Before creating features, understand the problem you're trying to solve.
     - Defining the objective: What are you trying to predict or classify?
     - Identifying the target variable: The variable you're trying to predict.
     - Exploring the data: Understanding the data types, distributions, and any initial patterns.
2. Extracting Features:
   - This involves selecting the relevant variables or columns from your dataset.
     - Domain knowledge: Using your understanding of the domain to pick relevant features.
     - Exploratory Data Analysis (EDA): Visualizing and summarizing the data to find important features.

3. Transforming Features:
   - Normalization: Scaling numerical data to a standard range, usually [0, 1].
   - Standardization: Adjusting data to have a mean of 0 and a standard deviation of 1.
   - Encoding Categorical Variables: Converting categorical data into numerical values using techniques like One-Hot Encoding.
   - Creating New Features: Combining or deriving new features from existing ones. For example, creating a 'total_spent' feature from 'quantity' and 'price'.
     - Polynomial Features: Create interaction terms (e.g., product of two features or higher-order terms) to capture non-linear relationships.
     - Date and Time Features: Break down dates into day, month, year, weekday, hour, etc. Time-related features can have seasonality or cyclicity that is important to capture.
     - Binning: Convert continuous variables into discrete bins or ranges (e.g., age groups like 0-20, 21-40, etc.) to reduce noise or capture certain patterns.
4. Handling Missing Data

5. Feature Selection:
   - Statistical Tests: Using methods like correlation coefficients, to see how each feature relates to the target variable.
   - Model-Based Selection: Using algorithms like Lasso Regression, Random Forests, etc., which can rank features by importance during the model training process.
   - Dimensionality Reduction Techniques: Principal component analysis (PCA), Liner Factor Model which help reduce the number of features while preserving the variance in the data.
6. Feature Scaling
   - Scaling features is important, especially for algorithms (NNs, logistic regression) sensitive to the scale of data:
   - Min-Max Scaling, Standard Scaling

4.6 Feature Engineering

- Definition: The process of creating new features or modifying existing ones to improve model performance.

Techniques

- Feature Selection: Eliminate irrelevant or redundant variables.
  - Example: Correlation-based feature removal.
- Feature Extraction: Reduce dimensions using PCA.
- Feature Transformation: Scale or normalize features.
  - Example: Min-max scaling.
- Encoding Categorical Variables:
  - One-hot encoding, ordinal encoding.
- Creating interaction terms (e.g., multiplying two variables).