Data, Data Sources and Visualization

Data

- Data refers to factual information, often in the form of numbers, words, measurements, or observations, that can be used as a basis for reasoning, discussion, or calculation

Data and its types

What to know about Data?

Gather and understand	Gather and understand the data
Identify	Identify the type of each data file
Identify	Identify the type of attributes in each data file

Gather relevant data

Data

For example: Restaurant business
consumer_survey.xls
dining_preferences.xls
Feedback.txt
RD_Insert_Script.txt
restaurant_cuisine.accdb
restaurant_cuisine.xls
restaurant_parking,xls
reviews.xml
user_rating.xls

Source: Infosys Springboard

Classify the data

Structured data

Data in DBs, Spreadsheets.....

Semi-structured data

Emails, xml file, html file........

Unstructured data

Images, Videos, text files....

eg: Categories Restaurant data

consumer_survey.xls
dining_preferences.xls
RD_Insert_Script.txt
restaurant_cuisine.accdb
restaurant_cuisine.xls
restaurant_parking,xls
user_rating.xls
reviews.xml
Feedback.txt

Structured data

- consumer_survey.xls
- dining_preferences.xls
- RD_Insert_Script.txt
- restaurant_cuisine.accdb
- restaurant_cuisine.xls
- restaurant_parking,xls
- user_rating.xls

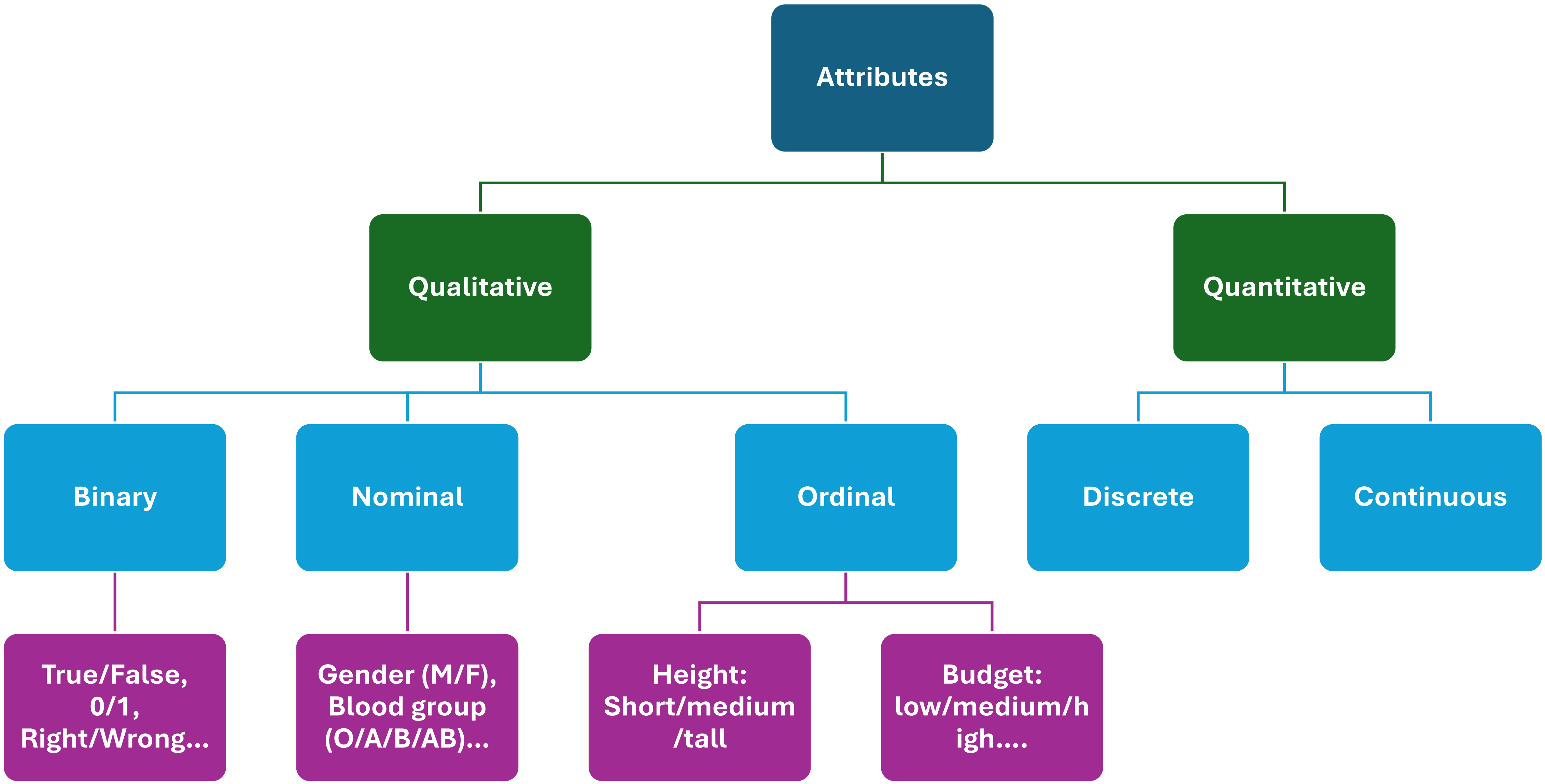
Semi-structured data

- reviews.xml

Unstructured data

- Feedback.txt

Identifying types of attributes in structured data



Classifying attributes of a structured dataset Consumer Survey.xls file

Consumer Survey.xls

Gender	Male/Female; Qualitative (Binary)
Smoker	True/False; Qualitative (Binary)
marital_status	Single/Married/Widow; Qualitative (Nominal)
Income	low/medium/high; Qualitative (Ordinal)
FRVPM-Freq of restaurant visits per month	Quantitative (Discrete)
AERPM-Avg expense in Restaurants per month	Quantitative (Continuous)

Types of Datasets

Numerical Datasets – Contain numerical values used for statistical analysis.
Example: Sales figures, stock market data.

Categorical Datasets – Contain classified data with labels.
Example: Customer segmentation (New, Returning, VIP).

Time-Series Datasets – Data collected over time intervals.
Example: Weather reports, sensor readings.

Spatial (Geospatial) Datasets – Data related to locations and geography.
Example: GPS coordinates, satellite images.

Text Datasets – Unstructured text data used in NLP.
Example: Tweets, chatbot conversations.

Image & Video Datasets – Used in AI and computer vision.
Example: Facial recognition databases, medical scans.

Transactional Datasets – Generated from business transactions.
Example: E-commerce purchase history, banking transactions.

Data Quality & Issues

High-quality data is critical for accurate analysis and decision-making. Poor data quality leads to incorrect insights and business risks.

Key Dimensions of Data Quality

Accuracy – Data should be correct and free from errors.

Completeness – No missing values or gaps in data.

Consistency – Uniform format and values across different sources.

Timeliness – Data should be up to date and relevant.

Validity – Data should conform to predefined formats and rules.

Uniqueness – No duplicate or redundant records.

Common Data Issues

Missing Data – Incomplete records causing bias in analysis.

Duplicate Data – Multiple records for the same entity.

Inconsistencies – Data mismatch across different systems.

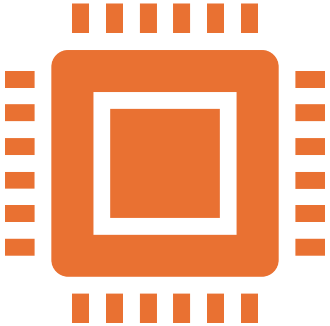
Incorrect Data – Human or system errors in data entry.

Data Drift – Changes in data patterns over time affecting model accuracy.

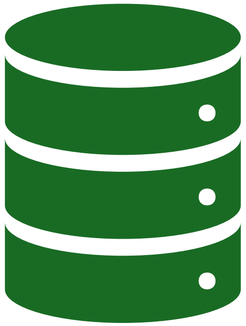
Bias in Data – Unrepresentative data leading to skewed results.

Data Modeling

Data Modeling



The process of analyzing and defining all the different data types your business collects and produces, as well as the relationships between those bits of data.



It helps visualize data, understand its structure, and design databases or data warehouses to store and manage data efficiently

The benefits of data modeling

- Improved Data Understanding
- Enhanced Data Management
- Better Data Analysis

Steps in Data Modeling

- Understanding Business Requirements
- Conceptual Modeling
- Logical Modeling
- Physical Modeling
- Implementation and Maintenance

Conceptual Modeling

Identifying Entities and Attributes:

Identify the key entities (e.g., customers, products, orders) and their attributes (e.g., customer ID, product name, order date).

Defining Relationships:

Determine the relationships between entities (e.g., one-to-many, many-to-many).

Creating a Conceptual Diagram:

Represent the entities, attributes, and relationships in a conceptual diagram, which is a high-level representation of the data model.

Logical Modeling

Refining the Model:

Refine the conceptual model by adding more details, such as data types, constraints, and indexes.

Choosing a Data Model:

Select the appropriate data model for the specific use case, such as relational, dimensional, or graph models.

Creating a Logical Diagram:

Represent the refined data model in a logical diagram, which is a more detailed representation than the conceptual diagram

Physical Modeling

Designing the Database:

Design the physical database structure, including tables, columns, and relationships.

Choosing a Database System:

Select the appropriate database system for the specific use case, such as SQL databases or NoSQL databases.

Optimizing Performance:

Optimize the database structure for performance, including indexing and partitioning.

Implementation and Maintenance

Implementing

Implementing the Model: Implement the data model in the chosen database system.

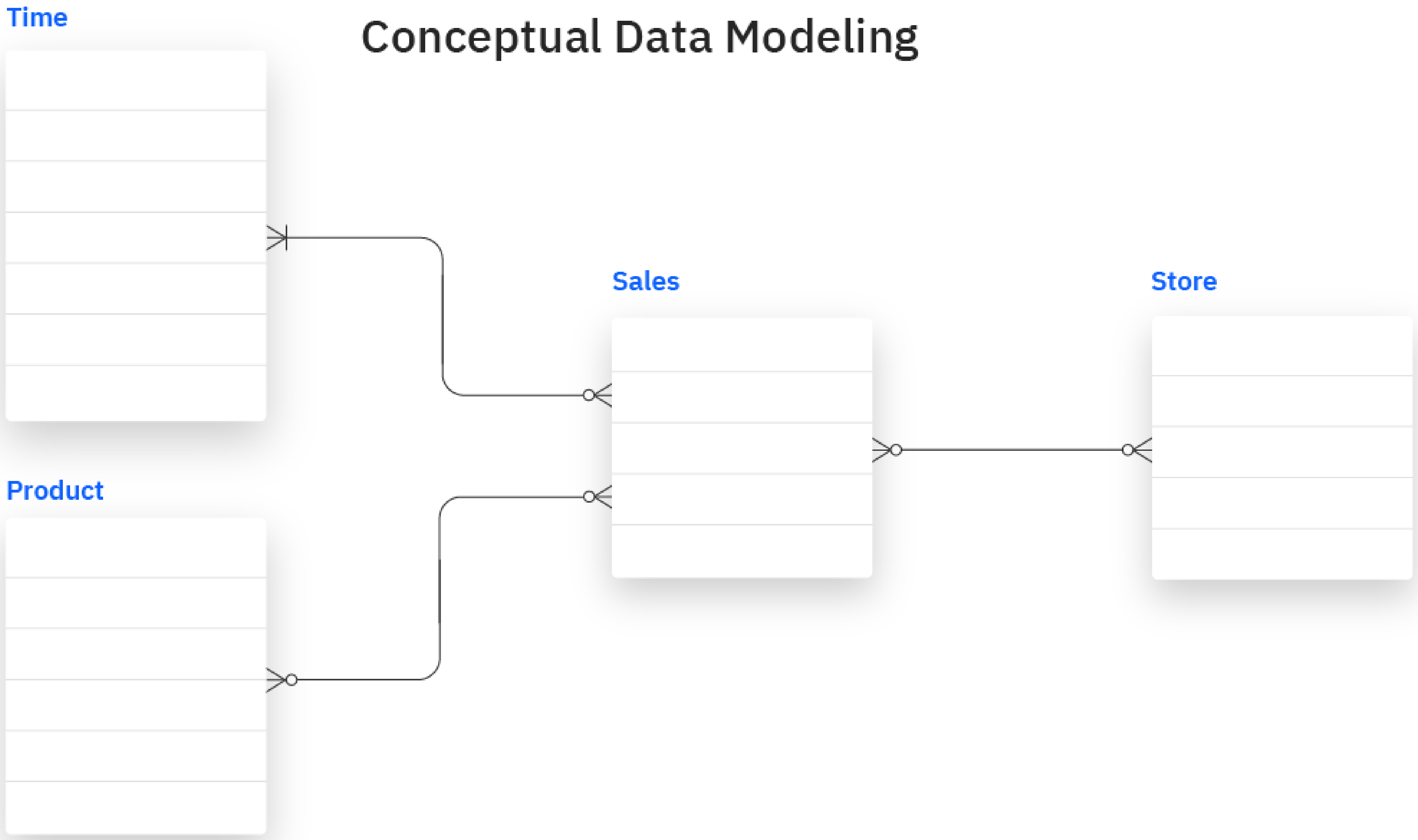
Testing

Testing and Validation: Test the data model to ensure that it meets the requirements and that the data can be accessed and manipulated correctly.

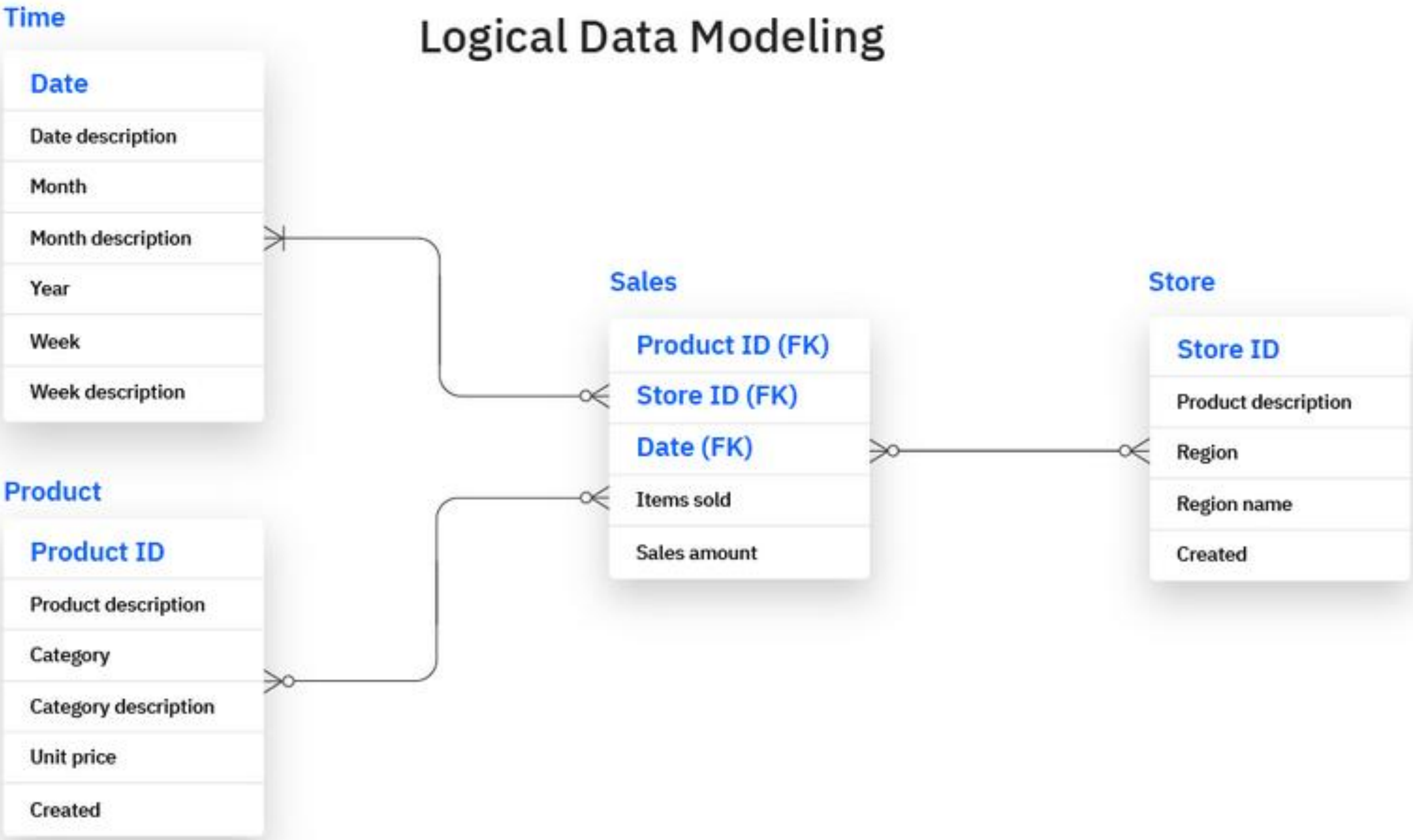
Maintaining

Maintaining the Model: Maintain the data model as the business requirements change, ensuring that the model remains accurate and relevant.

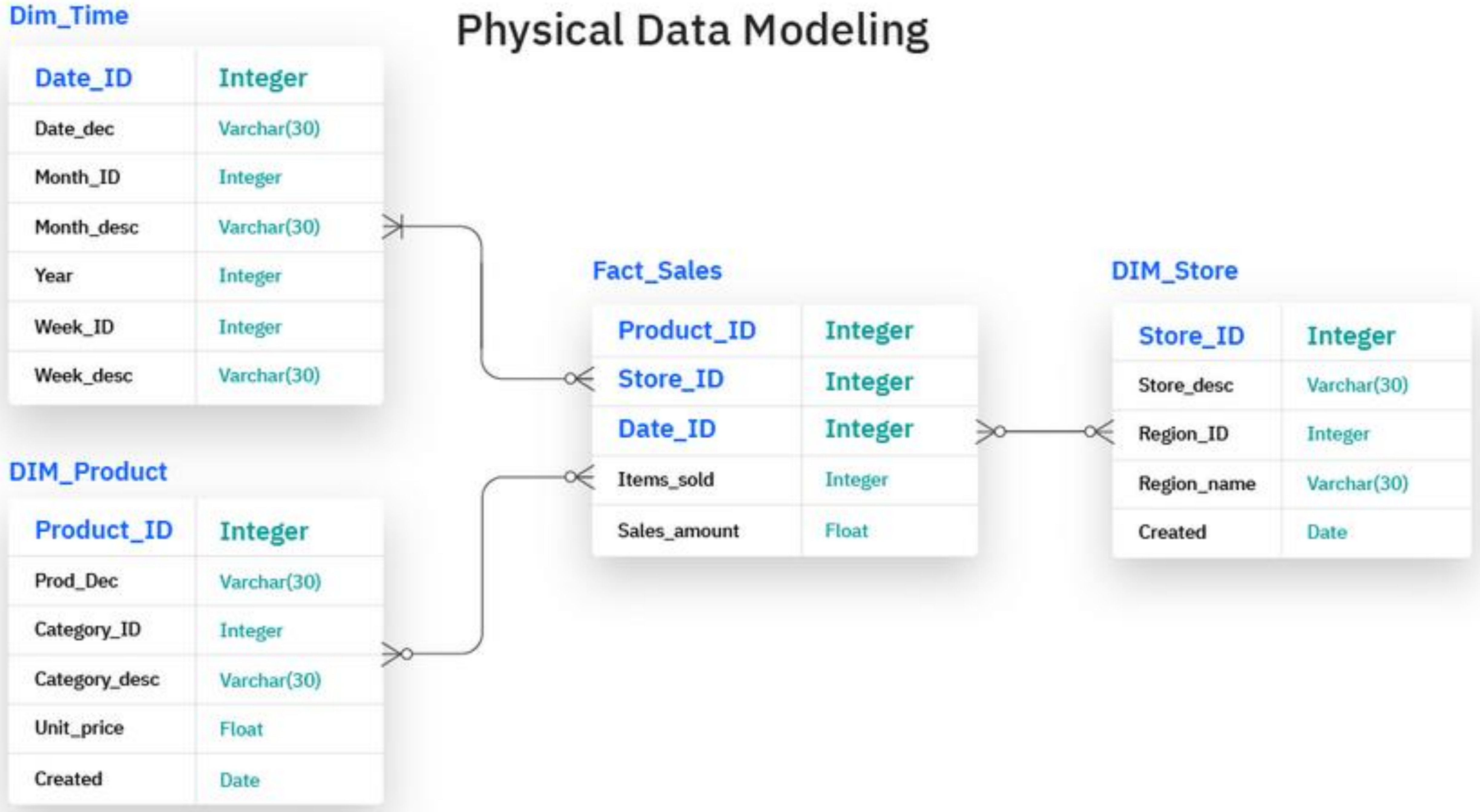
Conceptual Data Modeling



Logical Data Modeling



Physical Data Modeling



Practice

1. Data model for an online bookstore

Example Entities

- Customer
- Book
- Order
- Payment

2. Design a Data Model for a Healthcare Management System

Example entities

- Patient,
- Doctor,
- Appointment,
- Prescription

Data Analysis

- Data analysis is an aspect of data science and data analytics that is all about analyzing data for different kinds of purposes.
- The data analysis process involves inspecting, cleaning, transforming and modeling data to draw useful insights from it.

Types of Data Analysis

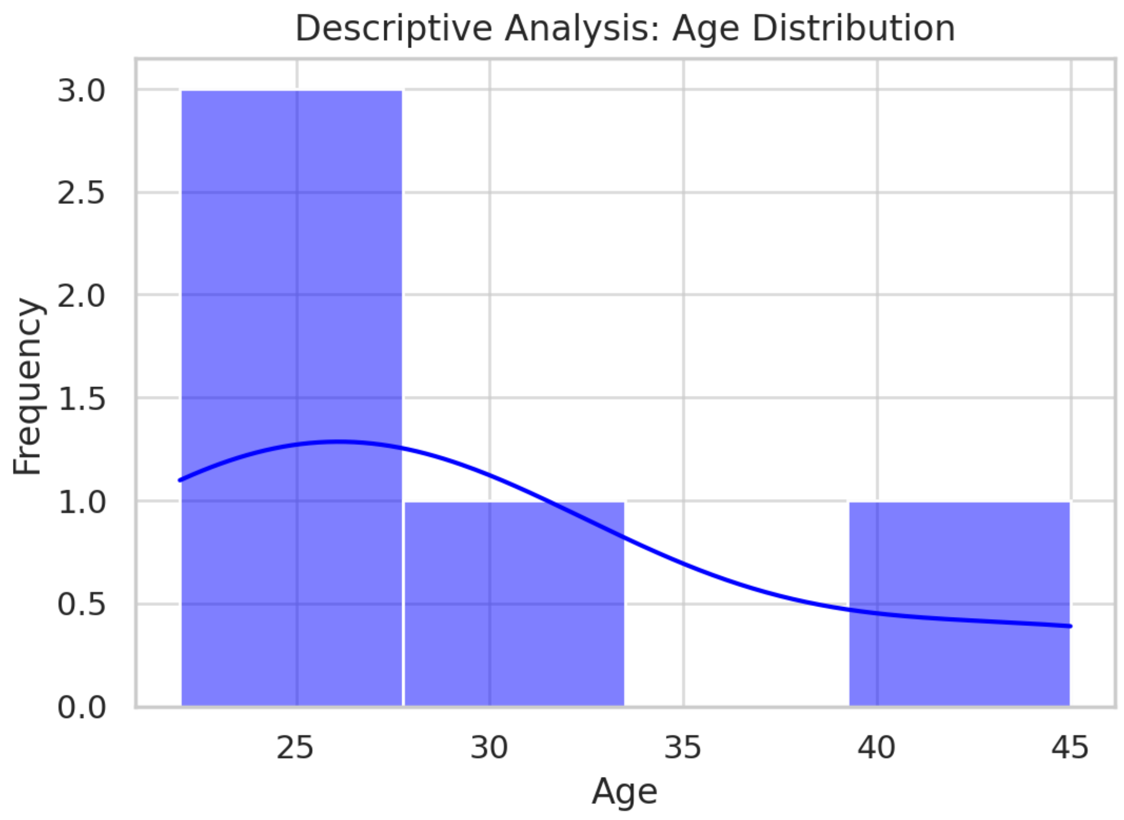
1. Descriptive analysis
2. Diagnostic analysis
3. Exploratory analysis
4. Inferential analysis
5. Predictive analysis
6. Causal analysis
7. Mechanistic analysis
8. Prescriptive analysis

Survived	Pclass	Sex	Age	Fare	Embarked
1	1	Female	25	71.28	C
0	3	Male	30	7.93	S
1	2	Female	22	13.00	S
0	3	Male	45	8.05	S
1	1	Female	27	52.00	C

Subset of Titanic dataset

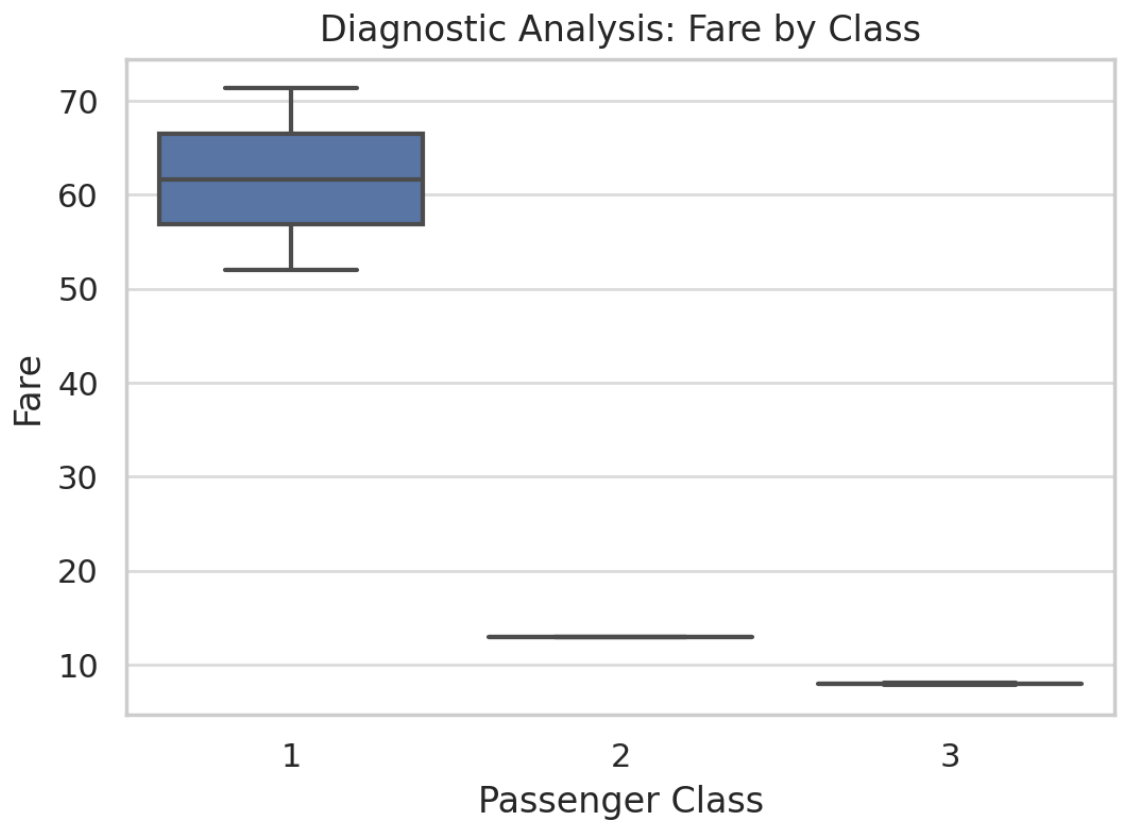
Descriptive analysis

- Descriptive analysis is the first step in analysis where you summarize and describe the data you have using descriptive statistics, and the result is a simple presentation of your data.
- These plots visualize the distribution of a dataset, providing insights into measures of central tendency (mean, median) and variability (standard deviation, IQR).
- Example: A histogram showing the age distribution of customers in a retail store.



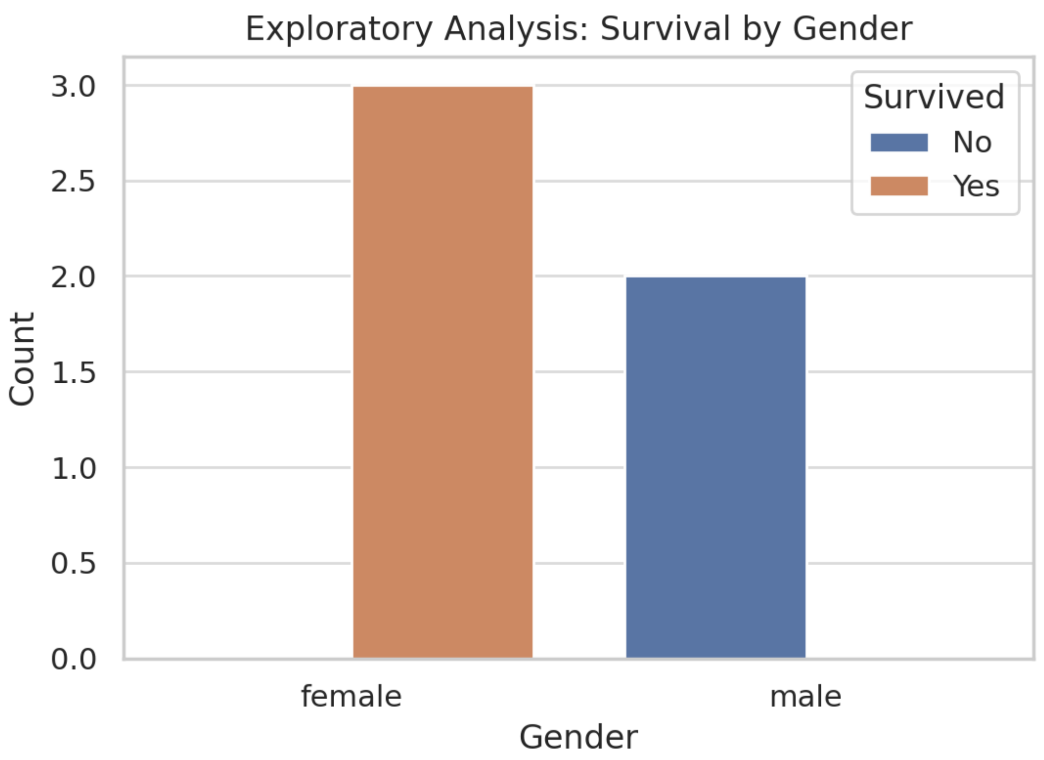
Diagnostic Analysis

- Diagnostic analysis seeks to answer the question “Why did this happen?” by taking a more in-depth look at data to uncover subtle patterns.
- Diagnostic analysis identifies relationships and patterns that explain the causes of observed outcomes.
- Example: A scatter plot to diagnose the relationship between marketing spend and sales revenue.



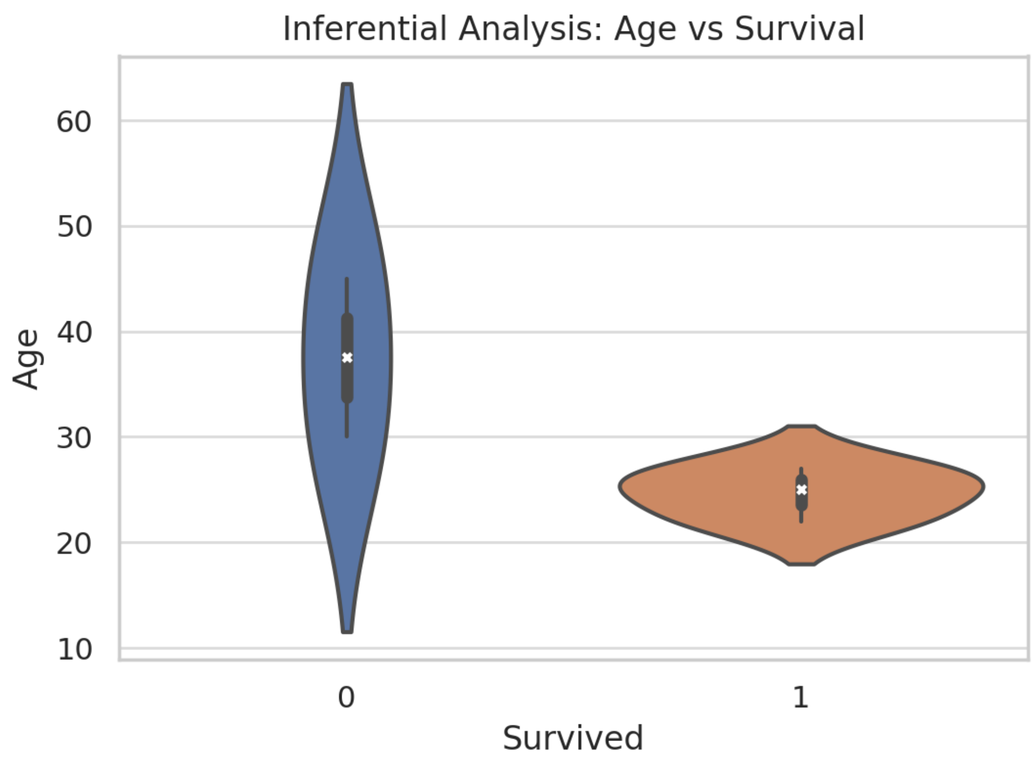
Exploratory Analysis

- These plots help visualize relationships, patterns, and anomalies across multiple variables in a dataset.
- Example: A pair plot to explore correlations between various financial indicators like revenue, profit, and expenses.



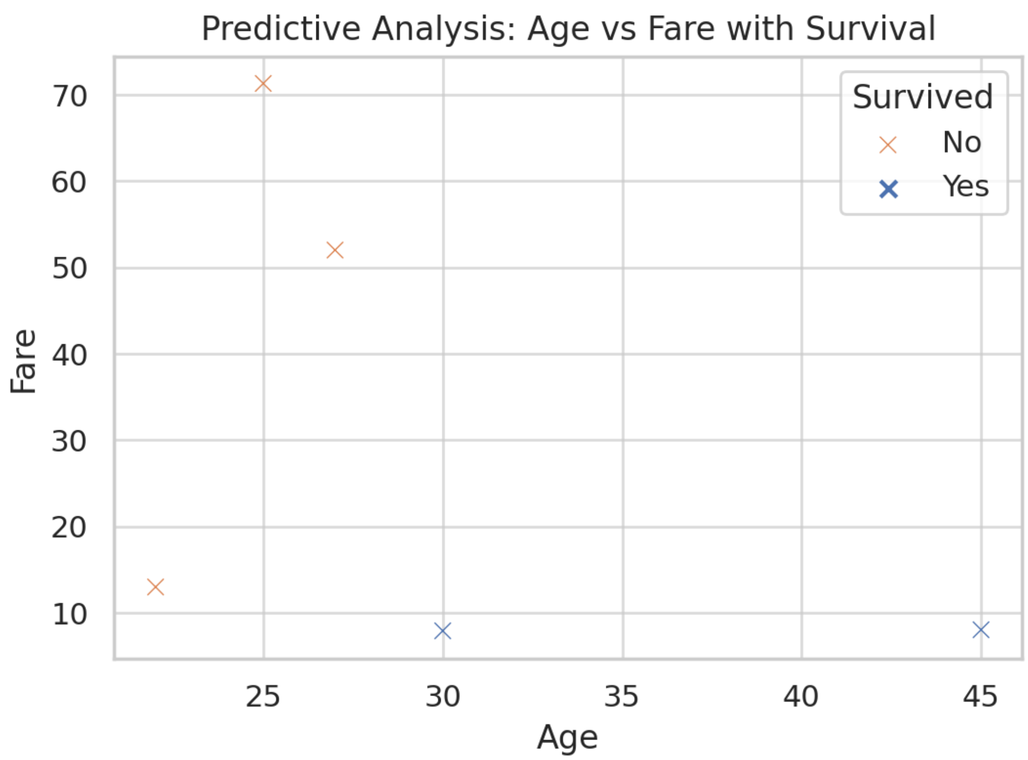
Inferential Analysis

- Inferential analysis draws conclusions about a population based on sample data. Confidence interval plots help visualize the uncertainty in estimates.
- **Example:** A confidence interval plot showing the average customer satisfaction score for different store locations.



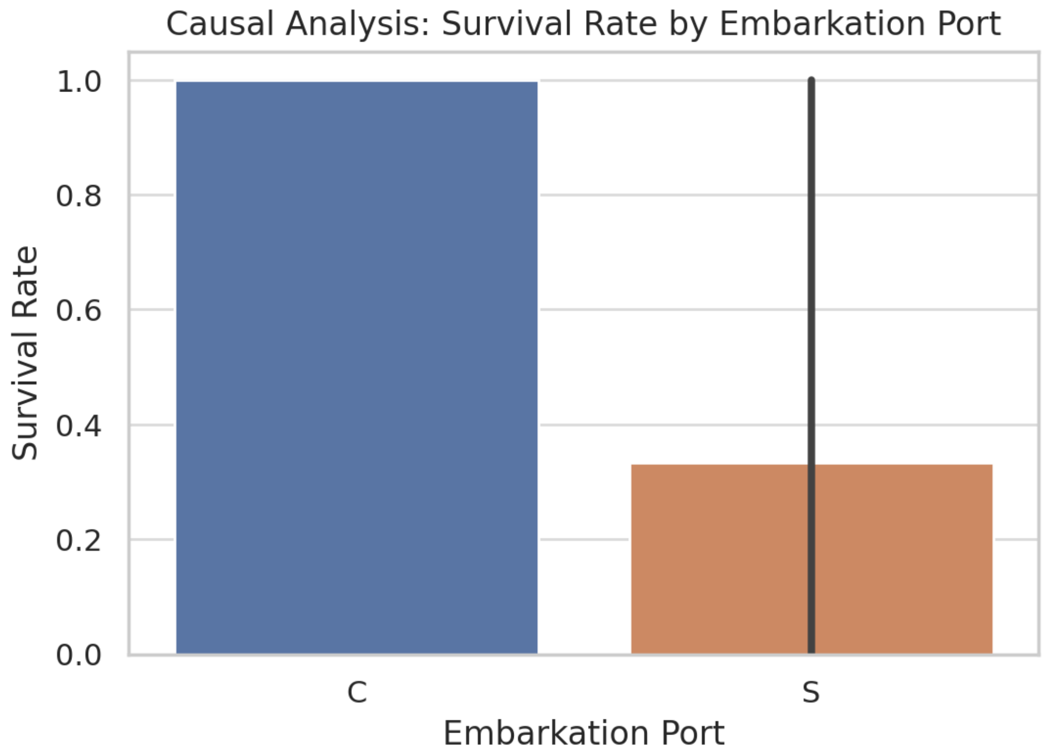
Predictive Analysis

- Predictive analysis uses historical data to make forecasts. Regression plots visualize the model's predictions against actual data points.
- Example: A line plot showing predicted sales for the next quarter using time series forecasting.



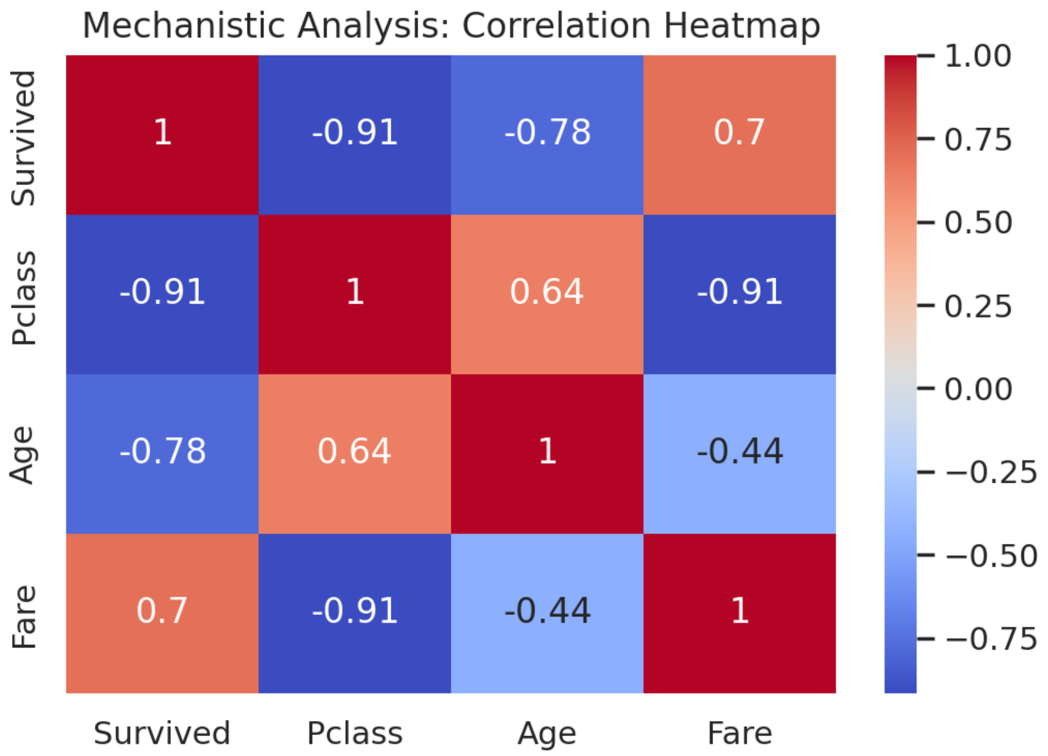
Causal Analysis

- Causal analysis determines cause-and-effect relationships between variables.
- Example: A causal impact plot to analyze the effect of a marketing campaign on website traffic.



Mechanistic Analysis

- Mechanistic analysis models complex systems to understand how different components interact.
- Example: A simulation plot showing the effect of varying production rates on inventory levels.



Prescriptive Analysis

- Prescriptive analysis provides recommendations based on predictive insights. Decision trees illustrate different scenarios and their potential outcomes.
- Example: A decision tree plot to suggest the best loan approval strategy based on customer data.

Association Analysis

- Association Analysis is a data mining technique used to discover relationships or patterns between items in large datasets. It is widely used in market basket analysis, recommendation systems, fraud detection, and web usage mining.
- To find frequent item-sets and association rules that describe how items are related within a dataset.
- **Association Rules**
- Association rules are statements in the form of:
- If $X \Rightarrow Y$, Which means, **if item X appears, item Y is also likely to appear**.

Association Analysis

- Example:
- **{Bread, Butter}** → **{Milk}** (People who buy bread and butter often buy milk)
- **{Laptop}** → **{Mouse}** (People who buy a laptop are likely to buy a mouse)

Association Analysis

Example:

•**{Bread, Butter}** → **{Milk}** (People who buy bread and butter often buy milk)

•**{Laptop}** → **{Mouse}** (People who buy a laptop are likely to buy a mouse)

Key metrics used to evaluate association rules:

1. Support
Measures how frequently an itemset appears in the dataset.

$$Support(X) = \frac{\text{Frequency of X in dataset}}{\text{Total transactions}}$$

UPES
UNIVERSITY OF TOMORROW

Association Analysis

Example: If **Milk** appears in **30 out of 100 transactions**, then:

$$Support(Milk) = \frac{30}{100} = 30\%$$

2. Confidence

Measures how often **Y** appears when **X** is present.

$$Confidence(X \Rightarrow Y) = \frac{\text{Support}(\text{X} \cup \text{Y})}{\text{Support}(\text{X})}$$

Example: If **Bread** appears in **50 transactions**, and in **40** of them, **Milk** is also bought:

$$Confidence(Bread \Rightarrow Milk) = \frac{40}{50} = 80\%$$

Association Analysis

3. Lift

Measures how much **stronger** the association is compared to a random occurrence.

$$Lift(X \Rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)}$$

If Lift > 1: X and Y are positively correlated (buying one increases the likelihood of buying the other).

If Lift < 1: X and Y are negatively correlated (buying one reduces the likelihood of buying the other).

Problems

A retailer wants to analyze buying patterns based on **500 transactions** in a week:

- {Laptop} appears in 100 transactions.
- {Laptop, Mouse} together appear in 60 transactions.
- {Mouse} appears in 150 transactions.

Questions:

1. What is the **confidence** of the rule **{Laptop} → {Mouse}**?
2. What is the **confidence** of the rule **{Mouse} → {Laptop}**?

UPES
UNIVERSITY OF TOMORROW

Problems: Supermarket Transactions

Transaction Dataset

Transaction ID	Items Purchased
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Bread, Butter, Eggs
T5	Bread, Butter, Eggs

Step 1: Compute Support

- Support(Milk)
- Support(Bread)
- Support(Butter)
- Support({Milk, Bread})
- Support({Bread, Butter})

Step 2: Compute Confidence

- Confidence(Milk → Bread)
- Confidence(Bread → Butter)

Step 3: Compute Lift

- Lift(Milk → Bread)
- Lift(Bread → Butter)

UPES
UNIVERSITY OF TOMORROW

Problems

Transaction Data

Transaction ID	Items Purchased
T1	Apple, Banana, Milk
T2	Apple, Banana
T3	Apple, Banana, Milk
T4	Banana, Milk, Bread
T5	Apple, Bread
T6	Banana, Bread
T7	Apple, Banana, Bread

Lift(Apple → Banana) 0.87

Lift(Banana → Bread) 0.60

UPES
UNIVERSITY OF TOMORROW

Applications of Market Basket analysis

Retail:

- Optimize product placement (**e.g., placing Milk near Bread**).
- Identify frequently bought-together items for promotions.

E-commerce & Recommendations:

- Suggest items frequently bought together (**Amazon's "Customers who bought this also bought..."**).
- Improve personalized recommendations.

Healthcare:

- Analyze patient symptoms and medications that are frequently prescribed together.

Finance:

- Detect fraud by identifying unusual spending patterns.

Data Pipelines

A **data pipeline** is a set of processes that automate the movement, transformation, and processing of data from source to destination.

It ensures data is collected, cleaned, enriched, and stored efficiently for analysis or machine learning.

Key Components of a Data Pipeline

Data Ingestion – Collecting data from different sources (databases, APIs, logs, IoT devices, etc.).

Data Processing (ETL/ELT) – Transforming raw data into a structured format.

- **ETL (Extract, Transform, Load)**: Data is transformed before loading into the data warehouse.
- **ELT (Extract, Load, Transform)**: Data is loaded first and transformed within the data warehouse.

Data Storage – Storing data in a data warehouse, data lake, or a database.

Data Analysis & Consumption – Querying, reporting, or using data for machine learning.

Orchestration & Monitoring – Managing dependencies, scheduling tasks, and ensuring system reliability.

UPES
UNIVERSITY OF TOMORROW

Common Data Pipeline Patterns

Batch Processing Pipeline
• Processes data in chunks at scheduled intervals.
• Suitable for large-scale ETL workloads.
• **Example:** Nightly aggregation of customer transactions for financial reporting.

Technology Stack:
- Apache Spark, Apache Hadoop, Airflow, AWS Glue

Lambda Architecture (Hybrid Batch + Stream)
• Combines batch and real-time processing.
• **Example:** A weather app that uses real-time sensor data for short-term forecasts and batch data for long-term trends.
Layers:
• **Batch Layer:** Stores historical data.
• **Speed Layer:** Processes real-time data.
• **Serving Layer:** Merges both for a unified view.
Technology Stack:
- Apache Kafka, Apache Spark, HDFS, NoSQL Databases

Streaming Data Pipeline
• Processes data in real-time or near real-time.
• Suitable for applications like fraud detection, live analytics, and IoT.
• **Example:** Monitoring website clicks or detecting fraudulent credit card transactions.
Technology Stack:
- Apache Kafka, Apache Flink, Spark Streaming, AWS Kinesis

Data Lake + Data Warehouse Hybrid
• Stores **raw data** in a **data lake** (e.g., AWS S3, Azure Data Lake).
• Transforms and moves structured data into a **data warehouse** (e.g., Snowflake, Redshift).
• **Example:** An e-commerce company storing all transactions in a data lake but using a warehouse for analytics.
Technology Stack:
- AWS S3, Azure Data Lake, Snowflake, BigQuery

UPES
UNIVERSITY OF TOMORROW

Best Practices for Data Pipelines

✓ **Use a Scalable Architecture** – Design for growing data volume.

✓ **Ensure Data Quality** – Use validation and anomaly detection.

✓ **Automate Orchestration** – Schedule and monitor pipelines with Apache Airflow.

✓ **Optimize Performance** – Use caching, indexing, and parallel processing.

✓ **Implement Security & Governance** – Encrypt data, use access controls, and comply with GDPR.